Generation and analysis of a telomere-to-telomere reference assembly and pangenome of lettuce (*Lactuca spp*.) with reference to repertoires of disease resistance genes

By

SAGAYAMARY SAGAYARADJ

DISSERTATION

Submitted in partial satisfaction of the requirement for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics & Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Richard Michelmore, Chair

Paul Gepts

Julin N Maloof

Committee in Charge

2023

Acknowledgements

First and foremost, I would like to express my deep gratitude to my advisor, Professor Richard Michelmore, for his invaluable advice and unwavering support. Without his guidance and persistent help, this dissertation would not have been possible. I would like to thank Professor Paul Gepts and Professor Julin Maloof for their treasured support as members of my dissertation committee.

My gratitude extends to all the members of the Michelmore Lab for providing a great atmosphere and for all their help over the years. I owe a great deal of gratitude, especially to Dr. Alexander Kozik, Dr. Dean Lavelle, and Dr. Kyle Fletcher who provided me with great insights and valuable feedback on various aspects of my project.

Special thanks to my colleagues in BASF for supporting me throughout this Ph.D. journey. I want to express my gratitude to Dr. Andreas Sewing, Dr. Manuel Rosas, Dr. Peter Visser, Dr. Joris Benschop, and Dr. Uwe Thissen for making this Ph.D. possible.

On a personal note, I want to give all the credit in the world to my two beautiful daughters, Olivia and Melania, who have been my support all through graduate school. Mere words are not enough to express how much they both mean to me. My special thanks to my wonderful siblings, Jaralin, Margaret, Amal, and Alphonsa for always being there for me.

This dissertation is dedicated to my Mom, who is my rock, for always being on my side, encouraging me to keep going and to never give-up!!!

ABSTRACT

Pangenome analysis becomes increasingly necessary as multiple genomes are sequenced from the same species. Lettuce (*Lactuca sativa* L.) is a commercially important crop with an annual farm-gate value of more than $3.1 billion in the United States. Whole genome re-sequencing efforts are underway to identify variations among different lettuce cultivars and wild germplasm. This dissertation reports on the generation and annotation of a new high-quality, telomere-to-telomere v11 reference genome assembly of *L. sativa* cv. Salinas based on Pacific BioSciences High-Fidelity reads, as a foundation for pangenome analyses. Chromosome-scale, high-quality assemblies were also generated for four domesticated genotypes of *L. sativa* (cv. La Brillante, cv. Ninja, PI251246, VIAE) and two wild accessions of *L. serriola* (US96UC23, Armenian 999). Several contemporary, publicly available, graph-based pangenome tools were evaluated for their ability to explore the large genome (~2.7 Gb) and high repeat content of *Lactuca* spp. Based on these assemblies, a pangenome of ~3 Gb encoding a total of 212,497 genes was generated. These genes were classified into 36,959 orthologous gene families, of which 23,751 were core families and 9,864 were dispensable families. Structural variants were assessed relative to the reference genome. Results from this pangenome analysis will allow the mapping of introgressed segments and a better understanding of structural and functional differences specific to a genotype. This dissertation provides a workflow for expanded pangenome analyses as more genome assemblies of *Lactuca* spp. become available in the near future. The pangenome resources will provide a foundation for syntenic inferences across multiple genotypes and species in the lettuce genepool and facilitate map-based cloning of agriculturally important genes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Overview of Lettuce

Lettuce (*Lactuca sativa* L.) is a commercially important fresh leaf crop and one of the most widely consumed vegetables in the world. Lettuce is mainly domesticated in temperate and subtropical climates. It is one of the most valuable vegetable crops in the U.S., with an annual production of more than 8 billion pounds and a farm gate value of more than $3.4 billion (Agricultural Statistics Service, 2020). Lettuce is amenable to classical and molecular genetic analyses. The generation time is usually three to five months depending on the genotype and environment, allowing for multiple lettuce generations each year. Lettuce can be routinely and efficiently transformed using *Agrobacterium tumefaciens* and is amenable to a variety of biotechnological approaches, including genome editing (Michelmore et al., 1987) .

Lettuce is a member of the Compositae (Asteraceae) family, which contains a large number of flowering plants in terms of number of species and diversity of habitats colonized. Popular members of the Compositae family include endive, chicory, artichoke, sunflower, and safflower. In total, more than 27 million hectares of Compositae species are planted worldwide, of which lettuce, sunflower, and artichoke are the genetically best characterized (Reyes-Chin-Wo et al., 2017). The Compositae is thought to have originated in the mid-Eocene (45–49 Myr) and expanded greatly during the Oligocene (28–36 Myr). It encompasses 1,620 recognized genera and at least 23,600 species, constituting approximately 10% of all angiosperms (Lindqvist, 1960). Over 200 species have been domesticated for a wide variety of uses. The genus *Lactuca* consists of about 100 species,

three of which, *L. serriola* (prickly lettuce), *L. saligna* (willowleaf lettuce), and *L. virosa* (bitter lettuce), are wild species sexually compatible to varying degrees with *L. sativa* (de Vries, 1997).

Domesticated lettuce displays enormous morphological diversity. There are five major types of lettuce based on their morphological characteristics: crisphead (iceberg), loose-leaf, romaine, butterhead, and stem (Lebeda et al., 2013). These basic types of lettuce often form the basis for grouping lettuce as is commonly seen in supermarkets. Each of these types consists of numerous cultivars, each distinguished by morphology, cultural adaptations, and resistance to diseases. Despite the variation in their morphologies, different lettuce types share common domestication traits, such as broad leaves, absence of spines on leaves and stems, and loss of shattering seeds (de Vries, 1997). However, the molecular basis for domestication and divergence between the various horticultural forms of lettuce remains little studied. Understanding the genetic and genomic landscape underlying these lettuce types is of considerable importance to lettuce breeding.

Domestication of wild lettuce species has led to the loss of prickles from leaves and stems, less latex and tissue bitterness, loss of seed shattering, reduced suckering, slow bolting, and increased seed size. Human selection and breeding efforts have also resulted in changes in size, shape, color, texture, and taste of leaves and plants, heading habits, resistance to diseases and insects, yield, and adaptation to different geographic areas and environments (Mou, 2011). Several wild forms of lettuce are suitable for animal food or for oil from the seeds. These landrace lettuce cultivars still exist in Egypt today and those suitable for oil production have large seeds with a high oil content.

1.1.1 Domestication of lettuce

L. sativa was likely domesticated from one or more weedy relatives in Egypt, the Mediterranean, the Middle East, or southwest Asia (Lindqvist, 1960). Domesticated lettuce was first documented on the walls of Egyptian tombs in approximately 2,500 BC, suggesting that lettuce has been domesticated for at least 4,500 years (de Vries, 1997). Research suggested that 4,000 years ago, Egyptians started to cultivate wild lettuce (*L. serriola*) in Africa, and this species is thought to be the ancestor of modern lettuce cultivars (Harlan, 1986). However, in 1960, Lindqvist proposed that *L. serriola* and another unknown species may have been involved in the domestication of domesticated lettuce (Lindqvist, 1960). In 1991, Kesseli et al., (1991) suggested a polyphyletic origin of *L. sativa* using restriction fragment length polymorphism (RFLP) loci. A recent study by Wei et al., (2021) revealed that the Middle East, including Transcaucasia, Iran, and Asia Minor, was a major domestication center, where wheat, barley, oat, chickpea, and lentil have been discovered in the archaeological records (Figure 1.1). Among the six geographic groups identified here, the Caucasus was likely the center of lettuce domestication considering the highest nucleotide diversity and the smallest genetic differentiation from domesticated lettuce. The findings also showed gene flow from Southern European populations to domesticated lettuce, which agrees with an early cultivation history in Greece and Italy. These findings were further supported by the phylogenetic results of genomic regions associated with domestication traits in domesticated lettuce. The genomic region genetically determining seed shattering shares a close relationship with the Caucasian *L. serriola* population, indicating that lettuce was domesticated in this region. In contrast, leaf morphology was determined by a 600 kb region on Chromosome 3 shared among most of the domesticated and the Southern

European lettuce accessions, indicating that this trait was introgressed from wild species during crop improvement in Southern Europe.

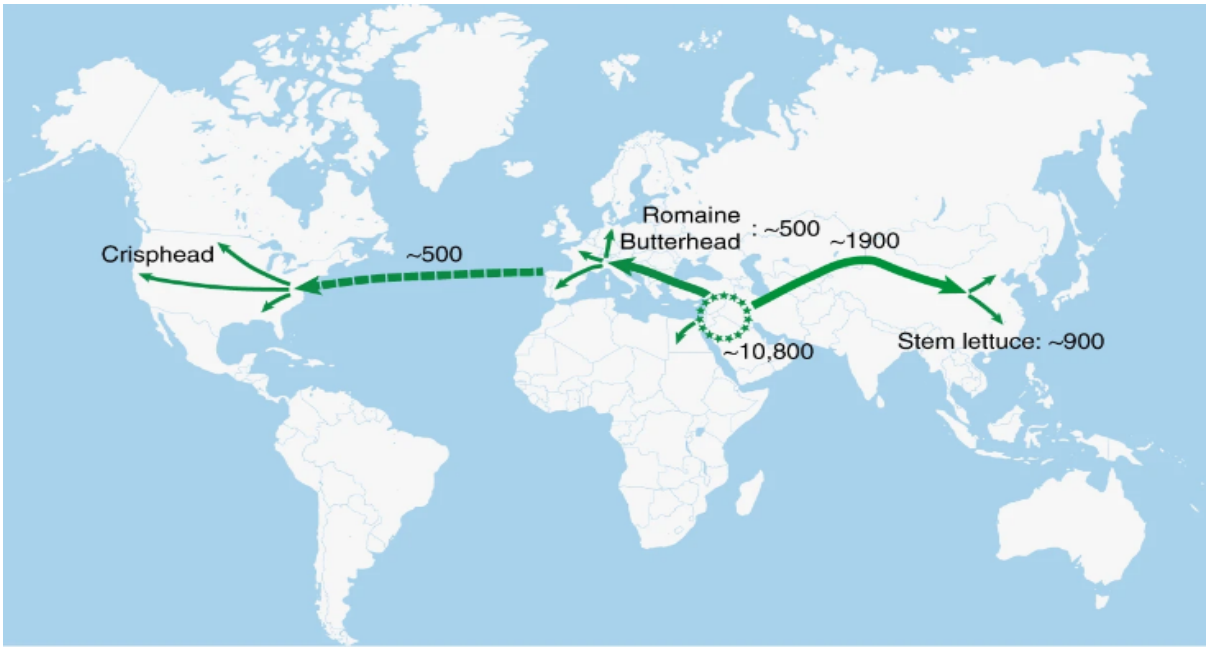

Figure 1.1*. Map showing the origin and dispersal of domesticated lettuce* (from Zhang et al., 2017).

Based on genomic information, the ancestor of domesticated lettuce is now concluded to have been domesticated from a common ancestor of *L. serriola* that underwent a single domestication event (Wei et al., 2021). Mutations in *L. serriola* led to the appearance of favorable traits, particularly forms without spines on stems and leaves and plants with large seeds. Individuals with desired traits were then selected and further modified to fit human needs. Several morphological types of lettuce have evolved since the initial domestication event and subsequent improvement (diversification), which contributed to the different leaf-based lettuce cultivars (butterhead, crisphead, loose-leaf, romaine) and those used for stems and oil.

1.1.2 Phylogenetic relationships of *Lactuca* spp.

High throughput DNA sequencing makes it possible to analyze germplasm to explore the genetic resources and domestication history of *Lactuca* species. The Centre for Genetic Resources, the Netherlands (CGN), maintains extensive collections of agricultural and horticultural crops, including one of the largest collections of *Lactuca* spp. worldwide. Recently, Wei et al., (2021.) used 445 *Lactuca* accessions from the CGN collection representing the major lettuce types and wild relatives to conduct a phylogenetic analysis. The completely inter-fertile taxa, domesticated *L. sativa* and the wild species, *L. aculeata, L. altaica, L. dregeana,* and *L. serriola* form the primary gene pool (GP1). A single species, *L. saligna* comprises the secondary gene pool (GP2). Several species, including *L. virosa*, are in the tertiary gene pool (GP3). This study also showed that *L. serriola* from the Caucasus represents promising genetic resources for breeding programs as populations from this area showed the highest nucleotide diversity.

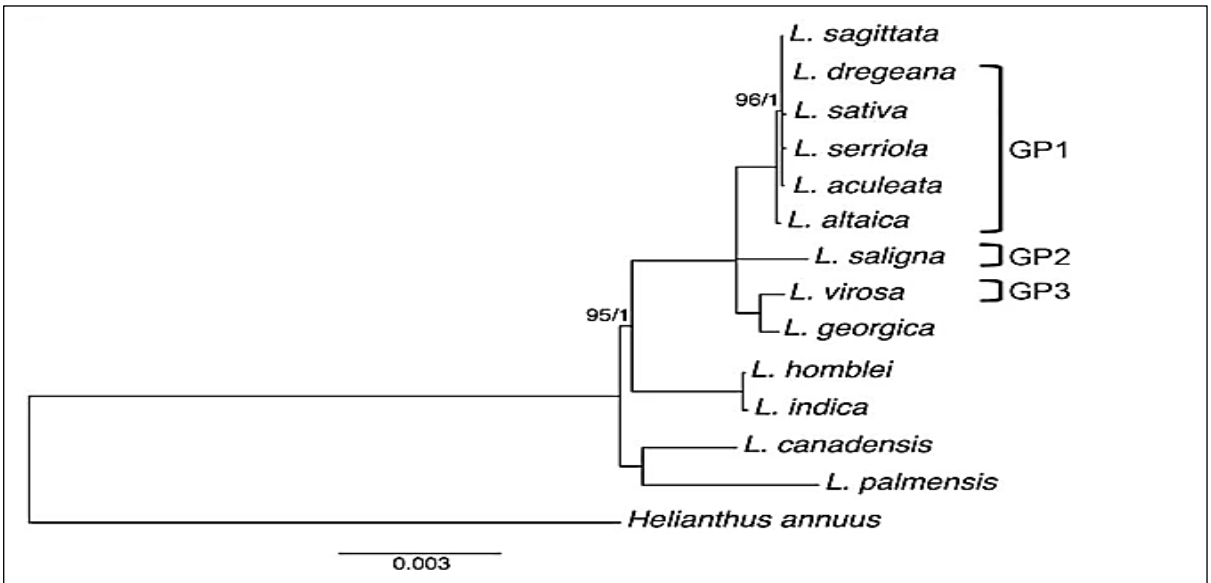

Figure 1.2. *Phylogenetic tree showing 12 Lactuca species and the outgroup H. annuus (from Wei et al., 2021).*

Lettuce gene pools can provide rich genetic resources for improving lettuce growth, with respect to resistance to abiotic and biotic stressors. All lettuce cultivars and sexually compatible *Lactuca* spp. are self-fertilizing diploids with 2n = 2x = 18 chromosomes. Crosses between *L. sativa* and *L. serriola* are fully fertile, while crosses between *L. serriola* and *L. saligna*, and between *L. sativa* and *L. saligna* are partly fertile (Jeuken et al., 2001; Thompson et al., 1941; Zohary, 1991). Crosses between *L. sativa* and *L. virosa* require embryo rescue to be successful (D'andrea et al., 2008). *L. serriola* from GP1 possess interesting alleles for acquiring water and fertilizer in soil, increasing germination, and improving seed longevity (Argyris et al., 2005; Johnson et al., 2000; Schwember & Bradford, 2010). *L. aculeata* from GP1, *L. saligna* from GP2, *L. virosa* from GP3, and *L. tatarica*, *L. biennis*, *L. canadensis, L. homblei, L. indica*, and *L. perennis* all showed high resistance to downy mildew (Jeuken et al., 2008). These species may provide rich genetic resources for domesticated lettuce. In addition, *L. orientalis* could be a potential resource to improve growth, development, and resistance to diseases (Wei et al., 2016).

1.2 Genetic and Genomic Resources for Lettuce

With the advent of DNA-based markers in the 1980s, the widespread use of markers in molecular breeding began. To create genetic maps for crop improvement, a number of DNA marker technologies have been developed, including RFLP, random amplified polymorphic (RAPD), simple sequence repeats or microsatellites (SSR), sequence characterized amplified region (SCAR), Amplified Fragment Length Polymorphism (AFLP), and single nucleotide polymorphism (SNP). These marker technologies have been used to explore the relationships between lettuce cultivars and wild relatives. Several lettuce genetic

maps based on RFLP, RAPD, AFLP, SSR, and EST markers have been published to study the genetics of various traits (Table 1.1).

Table 1.1. *Populations of lettuce and wild relatives analyzed genetically using molecular markers.*

Population name	**Population type**	**Population size**	**Marker type**	**Trait evaluated**	**Reference source**
L. sativa cv. Calmar x *L. sativa* cv. Kordaat	Intraspecific	350 / $F_{2:3}$	RFLP	downy mildew resistance	Landry et al., (1987)
L. sativa cv. Salinas x *L. sativa* cv. Green Lakes	Intraspecific	1429 / $F_{2:3}$	RFLP	resistance to corky root	Brown & Michelmore (1988)
L. sativa cv. Calmar x *L. sativa* cv. Kordaat	Intraspecific	$F_{2:3}$	RFLP and RAPD	downy mildew resistance	Kesseli et al. (1994)
L. sativa cv Salinas x *L. serriola*	Interspecific	100 / $F_{2:3}$	AFLP	root architecture	Johnson et al., (2000)
L. sativa cv. Olof x *L. saligna*	Interspecific	180 / $F_{2:3}$	AFLP	downy mildew resistance	Jeuken et al., (2001)
L. sativa cv Salinas x *L. serriola* UC96US23	Interspecific	103 / F_8 RILs	RFLP	thermo-tolerance	Argyris et al., (2005)

L. sativa cv Salinas x *L. serriola*	Interspecific	113 / F_9 RILs	SNPs	shelf life, leaf area, leaf thickness, leaf dry and fresh weight, epidermal cell area, epidermal cell number	Zhang et al., (2007)
L. sativa cv Salinas x *L. serriola*	Interspecific	89 / F_8 RILs	SNPs	seed longevity	Schwember et al., (2010)
L. sativa cv Salinas x *L. serriola*	Interspecific	89 / F_8 RILs	RFLP	seed priming	Schwember & Bradford (2010)
L. sativa cv Salinas x *L. serriola*	Interspecific	114 / F_8 RILs	SNPs	domestication traits (germination time, rosette leaf length, plant height, number of stem leaves)	Hartman et al., (2012)
L. sativa cv Salinas x *L. serriola*	Interspecific	114 / F_8 RILs	SNPs	germination rate, biomass, days to first flower, seed output	Hartman et al., (2012)
L. serriola and *L. sativa* cv Dynamite	Interspecific	558 / F_2 RILs	SSR, SNPs	drought, salinity and nutrient deficiency	Uwimana et al., (2012)
L. sativa cv Salinas x *L. serriola*	Interspecific	114 / F_8 RILs	AFLP, SNPs / 1,513	fitness related	Hartman et al., (2013)
L. sativa cv Saladin x L.	Intraspecific	254 / F_5 RILs	AFLP, SSR / 424	postharvest discoloration	Atkinson et al., (2013)

sativa cv Iceberg					
L. sativa cv Salinas x *L. serriola* US96UC23	Interspecific	213 / $F_{7:8}$ RILs	SNPs / 13,943	ultra-dense genetic map	Truco et al., (2013)
L. sativa cv Grand Rapids x *L. sativa cv* Iceberg	Interspecific	90 / F_6 RILs	SNPs	downy mildew	Lebeda et al., (2013); Simko et al., (2013); van Treuren et al., (2011)

Lettuce genomics accelerated with the availability of the *L. sativa* cv. Salinas draft genome (Reyes-Chin-Wo et al., 2017). A draft genome of *L. sativa* cv. Salinas has been assembled that covers 2.3 Gb of the total estimated 2.7 Gb lettuce genome (Reyes-Chin-Wo et al., 2017). This version 8 genome assembly was built mostly with Illumina (short-read) and medium-coverage Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing data. Genetic linkage was used to assign scaffolds to chromosomal linkage groups. *In vitro* proximity ligation was used to generate large super-scaffolds for each chromosome based on long-range contact frequencies between scaffolds. The draft lettuce genome was assembled into 168,554 contigs comprising 2.3 Gb with a contig N50 of 200 kb (Reyes-Chin-Wo et al., 2017). The genome assembly was predicted to have 36,136 protein coding genes. The genome assembly of lettuce was one of the more complete for any plant species reported at the time, particularly for genomes larger than 2 Gb with a high repeat content. The lettuce genome assembly revealed a family-specific whole genome triplication event and provides a reference genome for the Compositae family. The assembly also showed that 26% of the genome in the triplicated regions contains 30% of all genes that are

enriched for regulatory sequences and depleted for genes involved in defense. The v8 genome assembly has been adopted as the reference genome by NCBI and reannotated (https://www.ncbi.nlm.nih.gov/assembly/GCF_002870075.3/).

GenBank (https://www.ncbi.nlm.nih.gov102/genbank/), CGNB (https://www.cng-b.org/), and the Lettuce Genome Resource (https://lgr.genomecenter.ucdavis.edu/) are three places where lettuce-related omics data are concurrently aggregated and made available. Our lab maintains a number of lettuce databases that the general public may access at http://michelmorelab.ucdavis.edu. Access to data produced as part of the Next-Generation Lettuce Breeding: Genes to Growers (G2G) and CLGRP-funded research is available through the G2G website (http://scri.ucdavis.edu/). The ultra-dense map is accessible as genetic chromosomal pseudomolecules using our Gbrowse genome viewer, which is available at http://gviewer.gc.ucdavis.edu/cgi-bin/gbrowse/lettucePublic/. These databases are continuously being updated to make it easier for disease-centric, breeder-oriented viewpoints to acquire marker information for breeding purposes. Recently the lettuce breeding group (Beijing Academy of Agriculture and Forestry Sciences (BAAFS)) completed the genome assembly of stem lettuce, which is now available at the Lettuce Genome Database (lettucegdb.com). Similarly, Wageningen University is actively developing the genomic assemblies of *L. saligna* and *L. virosa,* as well as an extensive expression atlas of lettuce transcriptomic data.

1.3 Advancement in Genome Sequencing Technologies

1.3.1 Long-read sequencing technologies

A new era in genomics began with the introduction of single-molecule, third-generation sequencing technologies, primarily represented by Pacific Biosciences (PacBio)

and Oxford Nanopore Technologies (ONT). Numerous investigations are swiftly incorporating these technologies, adding to the body of scientific information gathered over the previous decades based on short-read sequencing techniques. The assembly of several highly contiguous crop genomes is made possible by recent developments in long-read technology (Koren & Phillippy, 2015). The typical primary read lengths produced by PacBio SMRT and Oxford Nanopore sequencing are over 60 kb. These read lengths are longer than the majority of simple repetitions in many genomes, making it possible to achieve highly contiguous genome assemblies. Due to read length restrictions and the high error rate associated with long-read sequencing, highly repetitive regions of the genome, such as centromeres, telomeres, and nucleolar organizing regions (NORs), are still mostly poorly assembled.

Even though PacBio and ONT sequencing have solved the read length barrier and ONT can produce extremely long reads (the longest being > 4 Mb), the inherent 5 to 15% per base error rate causes incorrect or incomplete assemblies, even when they are highly contiguous. SMRT and ONT sequencing have demonstrated their value for genome assembly; however, they require significant computationally time-consuming error correction. To overcome this challenge, improvements are continually being made to long-read sequencing technologies to increase the accuracy of base calls from the raw reads. Recently, PacBio Circular Consensus Sequencing (CCS) HiFi sequencing has been developed, which generates highly accurate reads around in the 15–20 kb range (99.9% accuracy). This can provide highly accurate as well as contiguous complex genome assemblies (Chin et al., 2016; Wenger et al., 2019).

1.3.2 Long-range scaffolding technologies

In parallel to the development of long-read DNA sequencing, several technologies have been developed for scaffolding contigs to provide chromosome-level genome assemblies. One is optical mapping, which can build ordered maps of up to several hundred kb-long DNA molecules., High-throughput fingerprinting systems, such as the Saphyr system (Bionano Genomics, bionanogenomics.com), have only been widely applied in recent years, even though it was created in 1993 (Schwartz et al., 1993). Using fluorescently labeled enzymes, Saphyr system can determine the physical distances between sequence-specific sites along DNA molecules. Individual optical maps can be merged into consensus maps to construct major contigs or discover significant and intricate structural variations (Nagarajan et al., 2008). Unlike assemblies of sequencing reads, which are often challenged by repeated sequences, optical maps reveal tandem arrays of repeats; however, they are prone to breaking at sections where two sites are closely positioned on opposing strands. Therefore, sequencing data and optical maps can be coupled to enhance assembly accuracy. Several researchers have utilized the powerful combination of optical maps and long-read assembly contigs to scaffold assemblies of plant genomes (Jiao et al., 2017; Schnable et al., 2009).

Chromosome-scale assembly may also be assisted by chromosome conformation capture sequencing (Hi-C) (Miiro et al., 2009). Hi-C was initially developed to examine the three-dimensional architecture of chromosomes by ligating and sequencing spatially proximal DNA using paired-end sequencing. Although not all Hi-C read pairs are adjacent on chromosomes, intrachromosomal regions interact more frequently than those from other chromosomes and the majority of them originate from two closely spaced regions; the contact frequency between regions reduces as the linear distance between them increases.

Consequently, the Hi-C read pairs provide mid- to long-range, and even centromere-spanning, information about the linear distance between regions, which can be used for scaffolding assemblies. Studies have shown that Hi-C read pairs can produce comparable improvements in assembly contiguity to optical consensus maps (Hosmani et al., 2019; Kronenberg et al., 2018). In addition, they can be coupled to boost the contiguity of an assembly since they help link diverse complex genomic regions.

1.4 Pangenome Studies in Plants

Sequencing and assembly of the genome of a single individual cannot capture the sequence diversity of a whole species. The decrease in sequencing costs and developments in next-generation sequencing technologies have allowed the cost-effective sequencing of multiple genomes of the same species. The sequencing and comparisons of many accessions has revealed that SNPs, minor indels, and structural variants (SVs) account for the majority of variations across the genomes. There is a growing realization that the presence of these variants renders a single reference genome incapable of providing a comprehensive inventory of the genetic diversity contained by a species. The two primary components of the pangenome are the core genome and the auxiliary genome. The core genome is the portion of the genome composed of the DNA sequences that are conserved among all accessions. The term “dispensable genome” refers to the sequences found in some but not all pangenome individuals. The dispensable genome also consists of individual-specific genetic content. Increasingly, pan genomic techniques are used in plant science research to facilitate the discovery of crop genetic heterogeneity (Zhang et al., 2021). Compared to single reference genomes, pangenomes can contain more of the variation repertoire of a given species or genus, allowing for faster and more precise definition of SVs and their effect on

phenotype. From disease resistance to plant shape and productivity, pan genomics is revolutionizing our understanding of the genetic diversity underpinning important agronomic variables (Bayer et al., 2020; Golicz, Batley, et al., 2016a; Khan et al., 2020).

1.4.1 Strategies for pangenome construction

Several approaches have been applied to building a pangenome; these approaches are broadly classified as sequence based and presence/absence (PAV) gene content based. The sequence based pangenome includes comparative *de novo* assembly, iterative mapping and assembly, and the map-to-pan/graph-based methodology (Figure. 1.3). The comparative *de novo* assembly strategy seeks to assemble the whole genomes of all accessions in order to enhance the resolution of repetitive regions and copy number variation (CNV). This method is susceptible to a number of technical restrictions, such as a high cost of data production, high computing resource needs, and artefactual inconsistencies in assembly and annotation, which can lead to false presence/absence variation (PAV) calls (Khan et al., 2020; Sherman & Salzberg, 2020a). The iterative mapping and assembly method is based on sequentially mapping reads from all individuals to the reference genome and then updating the original reference with assembled unmapped reads, resulting in a new pangenome reference (Golicz et al., 2016). Iterative mapping and assembly permits PAV calls at every gene locus without using orthologous gene clustering and is applicable for analyzing PAVs across large population-based short-read datasets, but it lacks precise positioning of accessory sequences and is likely to result in under-representation of repetitive sequences due to errors in mapping and assembly of reads representing repeats. Due to the combined advantages and individual limits of *de novo* and iterative mapping methods, pangenome investigations of a species should ideally incorporate both techniques.

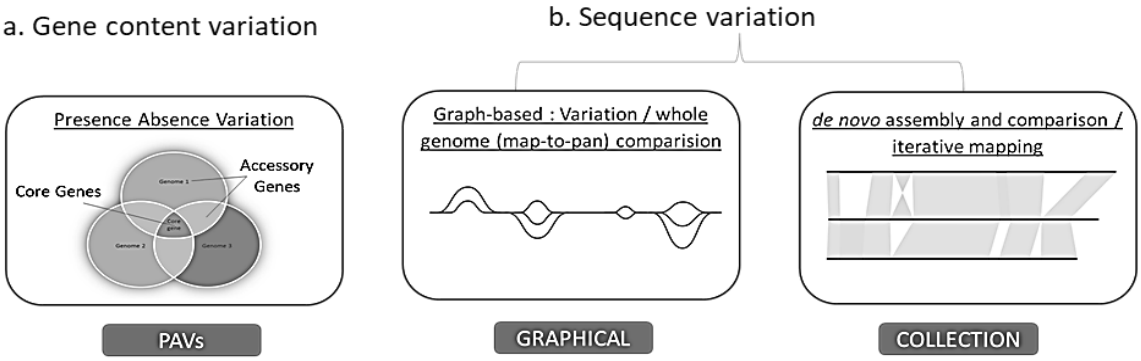

Figure 1.3. *Pan-genome approaches* (adapted from Bayer et al., 2020).

The map-to-pan/graph-based strategy is based on the generation and mapping of several high/low-quality *de novo* assemblies to an existing reference genome. The quality of the assembly is one of the most important factors influencing the pangenome analysis and is typically evaluated using parameters such as the total assembly span, N50/N90 size of scaffolds, number of scaffolds, the length of the longest scaffold, and the proportion of conserved core eukaryotic genes mapped to databases.

Gene-level analysis is frequently used to determine core and variable sequences in whole genome comparisons. These often rely on orthologous gene grouping, which can lead to errors in the assembly of highly duplicated crop genomes (Khan et al., 2020). However, the recent advent of long-read sequencing technologies, such as PacBio and Oxford Nanopore, has considerably simplified *de novo* assembly procedures (Gordon et al., 2014; Montenegro et al., 2017). The table below shows the several pangenome efforts that have been carried out in various crop species.

Table 1.2. *Examples of crop pangenome studies using different approaches (della Coletta et al., 2021).*

Pangenome approach	Pangenome Species	No. of accessions	Trait studied	Reference source
Iterative assembly	*Oryza sativa* (rice)	62	flowering time, stress tolerance, grain weight etc.	Zhao et al., (2018)
Map-to-pan	*Oryza sativa* (rice)	3,010	flowering time, disease resistance, grain length	Wang et al., (2018)
Iterative assembly	*Brassica napus* (cabbage)	53	Disease resistance	Hurgobin et al., (2018)
Iterative assembly	*Helianthus annuus* (sunflower)	493	Disease resistance	Hubner et al., (2019)
Iterative assembly	*Solanum lycopersicum* (tomato)	725	Disease resistance and fruit flavor	Gao et al., (2019)
De novo	*Sesamum indicum* (sesame)	5	Disease resistance and biosynthetic pathways	Yu et al., (2019)
Iterative assembly	*Brassica napus*	50	Disease resistance	Dolatabadian et al., (2020)
De novo	*Brassica napus* (rapeseed)	9	Seed weight and flowering time	Song et al. (2020)

Iterative assembly	*Cajanus cajan* (pigeon pea)	89	Self-fertilization and disease resistance	Zhao et al., (2020)
De novo, graph	*Glycine max* (soybean)	29	Iron uptake	Liu et al., (2020)
De novo	*H. vulgare* (barley)	20	Yield	Jayakodi et al., (2020)
De novo	*Malus domestica* (apple)	91	Fruit quality	Sun et al., (2020)
De novo	*Juglans* spp. (walnut)	6	Disease resistance	Trouern-Trend et al., (2020)
De novo	*Zea mays* (maize)	6	Biosynthesis pathway	Haberer et al., (2020)
De novo	*Triticum aestivum* (bread wheat)	10	Stress resistance, grain quality, disease resistance and yield	Walkowiak et al., (2020)
Iterative assembly	*Sorghum bicolor* (sorghum)	177	Drought resistance	Ruperao et al., (2021)
Iterative assembly	*Brassica oleracea*	243	Disease resistance and stress resistance	Bayer et al., (2021)
Iterative assembly	*Gossypium hirsutum* (cotton)	1,581	Fiber development, flowering time and yield	Li et al., (2021)

1.5 Disease Resistance in Plants

Many disease resistance genes have been characterized as simple Mendelian traits (Matvienko et al., 2013). There is extensive genetic information on monogenic resistance genes, and numerous resistance genes with qualitative phenotypes have been identified and introduced into domesticated genotypes. In contrast, some resistances have quantitative phenotypes and may be polygenically determined. Molecular methods, such as genome-wide analyses and the availability of cloned resistance genes, offer possibilities for rapid characterization and exploitation of wild germplasm and have the potential to allow for more durable resistance (Christopoulou et al., 2015).

Plants have developed a two-layer immune system against microbial pathogens and pests (Jones et al., 2006). In the first layer of defense, transmembrane pattern recognition receptors (PRRs), often with extracellular leucine rich repeats (LRR domains), identify pathogen-associated molecular patterns (PAMPs) and trigger downstream signaling activities, which induce defense gene expression, and often result in cell wall reinforcement by callose deposition and SNARE-mediated secretion of anti-microbial compounds (Collins et al., 2003). This is referred to as PAMP or pattern-triggered immunity (PTI). Successful pathogens have evolved virulence factors (effectors) that act in the apoplast or inside the host cell to overcome PTI (Zipfel, 2008).

As a second line of host defense, plants evolved intracellular R-proteins of the NB-LRR type that detect specific virulence factors, either directly or via their effects on host targets (Chinchilla et al., 2006). Plants containing a particular R-gene product are resistant to a pathogen that produces the cognate effector gene product (avirulence factors encoded by *Avr* genes) contributing to gene-for-gene resistance (van der Biezen & Jones, 1998). This

is referred to as effector-induced immunity (ETI). Rounds of ETI and effector-triggered susceptibility (ETS) due to novel Avr genes on the pathogen side may result in an evolutionary arms-race, producing a "zig zag zig" pattern of host resistance and susceptibility (Feng & Tang, 2019). In breeding lettuce for disease resistance, it is critical to keep up with the evolution of pathogens to develop resistant cultivars.

Genes determining resistance phenotypes have been shown by classical genetics to often be clustered in the genomes of multiple species. Such loci may be organized either as clusters of genetically separable loci or as apparent multiallelic series. In lettuce, 30 of the 52 mapped resistance specificities to seven diseases are located in ten clusters in the genome (Wise et al., 2008). Such clusters are enriched for genes encoding NLRR resistance proteins (Reyes-Chin-Wo et al., 2017). The use of pangenome-based analysis will enhance our understanding of the structural variation and trait evolution underlying these large resistance gene clusters.

1.5.1 R gene function and biological significance

R-genes play a vital role in protecting crops from infection by microorganisms, and therefore are of great interest to plant breeders. In potato, for example, R-proteins of the NBS-LRR type confer resistance to the oomycete *Phytophthora infestans*, a hemibiotrophic pathogen that causes late blight (Ballvora et al., 2002). In Arabidopsis, R-proteins of the NB-LRR type have been studied extensively in terms of molecular function, structural organization, sequence evolution, and chromosomal distribution (Meyers et al., 2003). The NBS-LRR superfamily is encoded by multiple gene families per genome and is subdivided into two main classes 1) TIR-domain-containing (for TOLL/INTERLEUKIN LIKE RECEPTOR/RESISTANCE PROTEIN; TIR-NB-LRR or TNL) and 2) non-TIR-domain-

containing (NB-LRR or NL), including coiled-coil domain-containing (CC-NB-LRR or CNL) R-protein subfamilies (McHale et al., 2006).

The NB domain is suggested to have NTP-hydrolyzing activity (ATPase or GTPase), regulating signal transduction through conformational changes. The LRR domain contains tandem array repeats in the carboxy-terminal region of R-genes and its predicted biochemical function is to mediate protein–protein interactions involved in the specific recognition of pathogen effectors. Both TIR and CC domains are assumed to be involved in protein–protein interactions and signal transduction.

Recently, with improved deep learning techniques, such as RoseTTAFold and AlphaFold, we can predict the structure of proteins even in the absence of structural homologs (Outram et al., 2022). With the release of AlphaFold2, DeepMind's machine-learning protein structure prediction program, the structure of several resistance proteins has been resolved (Goulet & Cambillau, 2022). AlphaFold2 has predicted more than 200 million proteins. AlphaFold was trained on hundreds of thousands of known protein structures and learned the relationships between the constituent amino acids and the final overall shapes. Given an arbitrary input amino acid sequence, the model can predict a 3D protein structure. Now, the model has predicted nearly all protein structures known to science. This has revolutionized our understanding of several resistance proteins. The structure of these proteins can provide an understanding of the mechanism of plant diseases and be used to unravel complex structure–function relationships in the plant system.

1.5.2 R gene cluster and evolution

R genes encoding NBS-LRR proteins constitute one of the largest and most complex, diverse gene families found in plants, with most plant genomes containing several hundred

family members. NBS-LRR genes are unevenly distributed in plant genomes and are primarily organized in multi-gene clusters (Yang et al., 2008). Furthermore, significant numbers of nucleotide polymorphisms were observed in NBS-LRR genes, which possibly evolved in response to shifts in the populations of pathogens (Kuang et al., 2004; Meyers et al., 2003). The clustered distribution of R-genes is assumed to provide a reservoir of genetic variation from which new pathogen specificity can evolve via gene duplication, unequal crossing-over, recombination or diversifying selection (Michelmore & Meyers, 1998).

Several comparative sequence analyses of R-gene clusters have been performed across haplotypes or related genomes in different plant species including Arabidopsis (Meyers et al., 2003), wild potato (Kuang et al., 2005), tomato (Seah et al., 2007), Brassicaceae (Xiao et al., 2004), wheat (Wicker et al., 2007), rice (Wicker et al., 2007), soybean (Innes et al., 2008), and common bean (David et al., 2009). Regions containing resistance genes may show high levels of structural variation and R genes can follow strikingly different evolutionary trajectories. Kuang et al., (2004) divided NBS-LRR-genes into two evolutionary categories: Type I and Type II. Type I includes genes with accelerated evolution by frequent sequence exchange among paralogs. Therefore, the sequences of Type I genes have chimeric structure, clear allelic/orthologous relationships between different genotypes, and their lineages cannot be easily established. Type II includes slowly evolving genes with sequence evolution primarily occurring through the accumulation of amino acid substitutions. Orthology relationships are highly conserved among these genes (Kuang et al., 2008). The evolutionary rate of each domain of individual NBS-LRR-encoding genes has been shown to be heterogeneous (Kuang et al., 2004). The NBS domain appears to be subject to purifying selection, whereas the LRR region tends to be highly variable. Nucleotide

polymorphisms found in the LRR region of R genes have been shown to be responsible for pathogen specificity. In particular, codons encoding solvent-exposed residues in the LRR domain are hypervariable among different R proteins and show significantly elevated ratios of non-synonymous to synonymous substitutions, suggesting that the LRR domain is subject to positive selection for amino acid diversification (Michelmore & Meyers, 1998).

Analysis of variability across plant pan-genomes reveals that variable regions are enriched for disease resistance genes (Badet & Croll, 2020). NLRs are under extreme selection pressure; therefore, two accessions from the same species can display great NLR copy number and sequence variation due to duplications, deletions, and unequal crossing over. Such variability of disease resistance genes in pan-genomes is documented in wheat (Bayer et al., 2022), *B. napus*, *B. oleracea* (Song et al., 2020), and tomato (Alonge et al., 2020)(Table 1.2). In *A. thaliana,* just 37 out of 64 accessions were sufficient to recover 90% of the predicted NLR gene repertoire.

1.6 Lettuce Diseases

There is limited genetic diversity within domesticated lettuce. The three major species sexually compatible with *L. sativa—L. serriola, L. saligna,* and *L. virosa*—have been sources of disease resistance genes, particularly *L. serriola* (Parra et al., 2016); however, they remain a rich potential source of variation that has not been accessed systematically (Kuang et al., 2008).

Lettuce is grown as a monoculture in which several crops per season are planted. With such intensive production, the crop is susceptible to major epidemics and vulnerable to several pests and diseases. A combination of genetic resistance, cultural practices, and chemical protection with the use of over 1.6 million pounds of insecticides and fungicides

control these pests. With the availability of genetically modified crops, breeding is the most affordable, cleanest, safest, and reliable crop protection method available. Because pathogens are constantly changing and new diseases and pests appear periodically, it is necessary to continually breed for new resistances.

Lettuce downy mildew, caused by *Bremia lactucae*, is the most important disease affecting lettuce in California and worldwide. Several other diseases are problematic in lettuce (Davis et al., 2011) including fungal diseases, such as wilts caused by *Verticillium dahliae* and *Fusarium oxysporum*, lettuce drop caused by *Sclerotinia minor* and *S. sclerotiorum*, lettuce anthracnose caused by *Microdochium panattonianum*, grey mold (*Botrytis cinerea*), and bacterial diseases such as corky root caused by *Rhizorhapis suberifaciens* and bacterial spot caused by *Xanthomonas campestris* pv. *vitians*. Verticillium wilt is of particular concern because it threatens to devastate lettuce production in the western U.S. (History and Economic Importance of Lettuce, 2011). There are also several viral diseases of varying importance such as lettuce mosaic virus, lettuce dieback, lettuce big vein, beet western yellows, and tomato bushy stunt. Recently, Impatiens Spotted Wilt Virus (INSV) has emerged as a pathogen that is devastating lettuce production in the western U.S. Other pathogens, such as powdery mildew (*Erysiphe cichoracearum*), lettuce infectious yellows, turnip mosaic virus, and tomato spotted wilt virus (TSWV) are present but currently rarely cause significant losses.

The interaction between lettuce and *B. lactucae* is one of the most extensively characterized gene-for-gene plant–pathogen relationships (Hulbert & Michelmore, 1985; Michelmore & Wong, 2008). Over 50 major *Dm* genes and resistance factors are now known that provide resistance against specific isolates of *B. lactucae* in a gene-for-gene manner

(Parra et al., 2016; Wood et al., 2019). As in other plants, resistance genes are clustered in the lettuce genome; most *Dm* genes are located in three major resistance clusters (MRCs) along with genes determining resistance to other diseases (Figure 1.4; Christopoulou et al., 2015).

The major cluster on Chromosome 1 contains over nine genetically separable *Dm* specificities. MRC1 contains *Dm*5/8, *Dm*10, *Dm*17, *Dm*25, *Dm*36, *Dm*37, *Dm*43, and *Dm*45, as well as *Tu* and *Mo*2 for resistance to Turnip Mosaic Virus (TuMV) and Lettuce Mosaic Virus (LMV), respectively, and q*FUS*1.1 and q*FUS*1.2 for resistance to wilt caused *by Fusarium oxysporum* f. sp. *lactucae*. MRC2 includes *Dm*1, *Dm*2, *Dm*3, *Dm*6, *Dm*14, *Dm*15, *Dm*16, *Dm*18, *Dm*50, and *DMR*2.2, along with *Tvr* for resistance to Tomato Bushy Stunt Virus (TBSV), Ra for root aphid resistance, and q*ANT*1 for resistance to anthracnose (Parra et al., 2016).

Similarly, MRC4 contains *Dm*4, *Dm*7, *Dm*11, *Dm*24, *Dm*38, *Dm*44, and *Dm*48 as well as q*FUS*4.1 for resistance to Fusarium wilt. MRC9A contains q*DMR*9.1, q*DMR*9.2 q*DMR*9.3, and q*VERT*9.1 for resistance to wilt caused by *Verticillium dahlia* (Christopoulou et al., 2015).

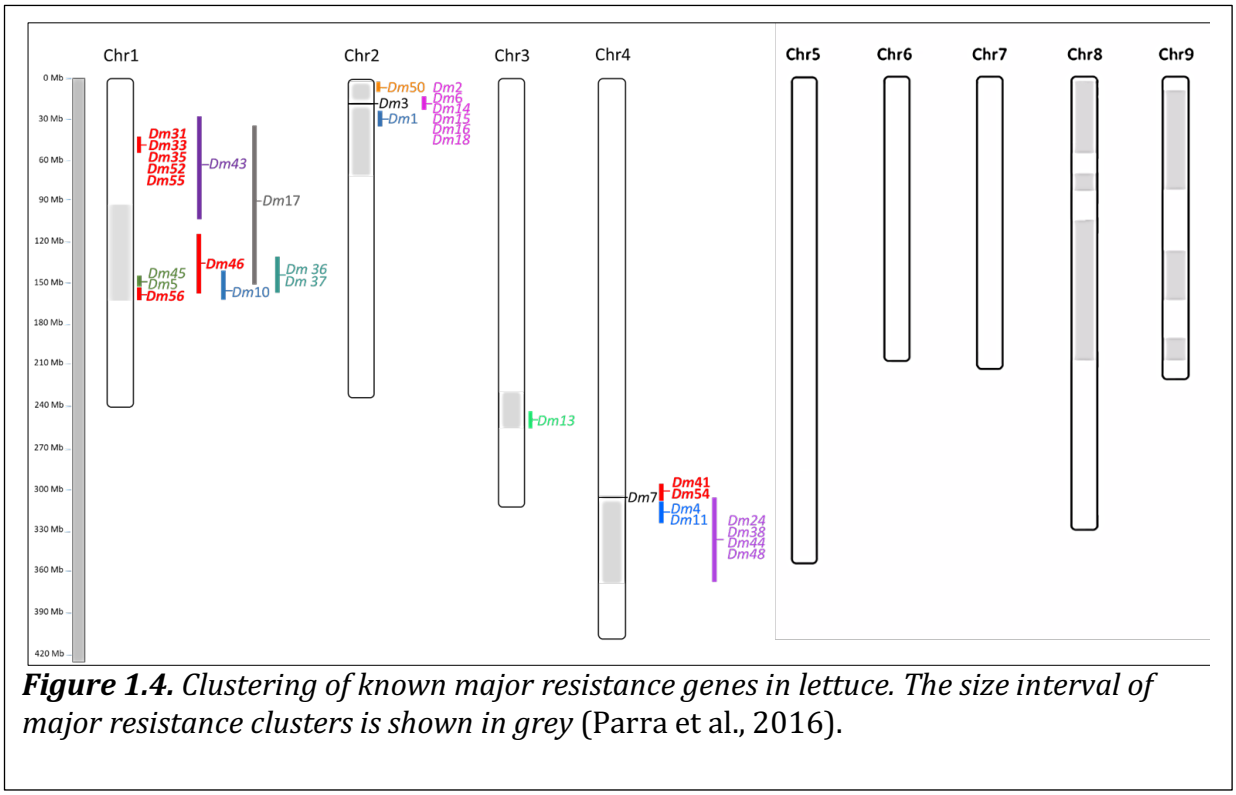

Figure 1.4. *Clustering of known major resistance genes in lettuce. The size interval of major resistance clusters is shown in grey* (Parra et al., 2016).

1.7 Introduction to this Thesis

In this dissertation, I utilized a pangenome-based approach to characterize variation within core and dispensable genomes among the different lettuce types, particularly with regard to disease resistance genes. In order to achieve this, I first generated a high-quality telomere-to-telomere, annotated reference genome assembly of *L. sativa* cv. Salinas using several long-read sequencing and scaffolding approaches (Chapter 2). I then generated additional annotated genome assemblies for six domesticated and wild accessions (Chapter 3). These seven genome assemblies were used to assemble a pan-genome of lettuce that was analyzed for structural variants and presence/absence variation of gene content (Chapter 4). Finally, I focused on variation in the major clusters of resistance genes (Chapter 5). This research lays the foundation for multiple studies of consequence for lettuce improvement

(Chapter 6). Additional accessions will be sequenced and assembled. These data will be used to mine for structural and functional variations in core genes that are shared by all *Lactuca* spp. and dispensable genes that are partially shared or specific to individual lettuce cultivars. A comprehensive understanding of the underexplored role of SVs in genotype-to-phenotype relationships and their widespread importance to lettuce improvement will be generated. The availability of the lettuce pangenome on multiple high quality genome assemblies of *Lactuca* species provides opportunities to explore the impact of SVs on many agronomic traits, such as tolerance to abiotic and biotic stress, disease resistance, flowering time, non-shattering, and changes in plant architecture, in a non-reference-biased manner. It will also lead to better characterization of the repertoire of NLR diversity within the lettuce genepool that will enhance breeding for disease resistance in lettuce.

References

Agricultural Statistics Service, N. (2020). *United States Department of Agriculture National Agricultural Statistics Service*.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *in press*, 1–17. https://doi.org/10.1016/j.cell.2020.05.021

Argyris, J., Truco, M. J., Ochoa, O., Knapp, S. J., Still, D. W., Lenssen, G. M., Schut, J. W., Michelmore, R. W., & Bradford, K. J. (2005). Quantitative trait loci associated with seed and seedling traits in *Lactuca*. *Theoretical and Applied Genetics*, *111*(7), 1365–1376. https://doi.org/10.1007/s00122-005-0066-4

Badet, T., & Croll, D. (2020). The rise and fall of genes: origins and functions of plant pathogen pangenomes. *Current Opinion in Plant Biology*, 56, 65–73. https://doi.org/10.1016/j.pbi.2020.04.009

Ballvora, A., Ercolano, M. R., Wei, J., Meksem, K., Bormann, C. A., Oberhagemann, P., Salamini, F., & Gebhardt, C. (2002). The R1 gene for potato resistance to late

blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant Journal*, *30*(3), 361–371. https://doi.org/10.1046/j.1365-313X.2001.01292.x

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6(8), 914–920. https://doi.org/10.1038/s41477-020-0733-0

Bayer, P. E., Petereit, J., Durant, É., Monat, C., Rouard, M., Hu, H., Chapman, B., Li, C., Cheng, S., Batley, J., & Edwards, D. (2022). Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. *Plant Genome*. https://doi.org/10.1002/TPG2.20221

Chin, C.-S., Peluso, P., Sedlazeck, F., Nattestad, M., Concepcion, G., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G., Delledonne, M., Luo, C., Ecker, J., Cantu, D., Rank, D., & Schatz, M. (2016). Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *Nature Methods, 13*, 1050-1054.

Chinchilla, D., Bauer, Z., Regenass, M., Boller, T., & Felix, G. (2006). The Arabidopsis receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception. *Plant Cell*, *18*(2), 465–476. https://doi.org/10.1105/tpc.105.036574

Christopoulou, M., Wo, S. R. C., Kozik, A., McHale, L. K., Truco, M. J., Wroblewski, T., & Michelmore, R. W. (2015). Genome-wide architecture of disease resistance genes in lettuce. *G3: Genes, Genomes, Genetics*, *5*(12), 2655–2669. https://doi.org/10.1534/g3.115.020818

Collins, N. C., Thordal-Christensen, H., Lipka, V., Bau, S., Kombrink, E., Qiu, J. L., Hückelhoven, R., Steins, M., Freialdenhoven, A., Somerville, S. C., & Schulze-Lefert, P. (2003). SNARE-protein-mediated disease resistance at the plant cell wall. *Nature*, *425*(6961), 973–977. https://doi.org/10.1038/nature02076

D'andrea, L., Felber, F., & Guadagnuolo, R. (2008). Hybridization rates between lettuce (*Lactuca sativa*) and its wild relative (*L. serriola*) under field conditions. *Environ. Biosafety Res*, *7*, 61–71. https://doi.org/10.1051/ebr:2008006

David, P., Chen, N. W. G., Pedrosa-Harand, A., Thareau, V., Sevignac, M., Cannon, S. B., Debouck, D., Langin, T., & Geffroy, V. (2009). A Nomadic subtelomeric disease resistance gene cluster in common bean. *Plant Physiology*, *151*(3), 1048–1065. https://doi.org/10.1104/pp.109.142109

de Vries, I. M. (1997). Origin and domestication of *Lactuca sativa* L. *Genetic Resources and Crop Evolution*, 44. Kluwer Academic Publishers.

della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-

genome is changing crop genomics and improvement. In *Genome Biology* (Vol. 22, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-020-02224-8

Feng, B., & Tang, D. (2019). Mechanism of plant immune activation and signaling: Insight from the first solved plant resistosome structure. *Journal of Integrative Plant Biology*, *61*(8), 902–907. https://doi.org/10.1111/jipb.12814

Golicz, A. A., Batley, J., & Edwards, D. (2016). Towards plant pangenomics. In *Plant Biotechnology Journal* (Vol. 14, Issue 4, pp. 1099–1105). Blackwell Publishing Ltd. https://doi.org/10.1111/pbi.12499

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H. R., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, *7*, 1–8. https://doi.org/10.1038/ncomms13390

Gordon, S. P., Priest, H., des Marais, D. L., Schackwitz, W., Figueroa, M., Martin, J., Bragg, J. N., Tyler, L., Lee, C. R., Bryant, D., Wang, W., Messing, J., et al. (2014). Genome diversity in Brachypodium distachyon: Deep sequencing of highly diverse inbred lines. *Plant Journal*, *79*(3), 361–374. https://doi.org/10.1111/TPJ.12569

Goulet, A., & Cambillau, C. (2022). *Present Impact of AlphaFold2 Revolution on Structural Biology, and an Illustration With the Structure Prediction of the Bacteriophage J-1 Host Adhesion Device*. https://doi.org/10.3389/fmolb.2022.907452

Harlan, J. R. (1986). Lettuce and the Sycomore: sex and romance in ancient Egypt. *Economic Botany*, *40*(1), 4–15. https://doi.org/10.1007/BF02858936

History and Economic Importance of Lettuce. (2011). https://doi.org/10.1094/PDIS-01-11-0075

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. v, Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). *An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps*. https://doi.org/10.1101/767764

Hulbert, S. H., & Michelmore, R. W. (1985). Linkage analysis of genes for resistance to downy mildew (*Bremia lactucae*) in lettuce (*Lactuca sativa*). In *Theor Appl Genet* (Vol. 70). Springer-Verlag.

Innes, R. W., Ameline-Torregrosa, C., Ashfield, T., Cannon, E., Cannon, S. B., Chacko, B., Chen, N. W. G., Couloux, A., Dalwani, A., Denny, R., Deshpande, S., Egan, A. N., Glover, N., Hans, C. S., et al. (2008). Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions

following polyploidy in the ancestor of soybean. *Plant Physiology*, *148*(4), 1740–1759. https://doi.org/10.1104/pp.108.127902

Jeuken, M. J. W., Pelgrom, K., Stam, P., & Lindhout, P. (2008). Efficient QTL detection for nonhost resistance in wild lettuce: Backcross inbred lines versus F2 population. *Theoretical and Applied Genetics*, *116*(6), 845–857. https://doi.org/10.1007/s00122-008-0718-2

Jeuken, M., van Wijk, R., Peleman, J., & Lindhout, P. (2001). An integrated interspecific AFLP map of lettuce (*Lactuca*) based on two *L. sativa* × *L. saligna* F2 populations. *Theoretical and Applied Genetics*, *103*(4), 638–647. https://doi.org/10.1007/s001220100657

Jiao, W. B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E. M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Hümann, U., Reinhard, R., Koch, M. A., Swan, D., Clavijo, B., Coupland, G., & Schneeberger, K. (2017). Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.*, *27*(5), 778–786. https://doi.org/10.1101/gr.213652.116

Jiao, Y. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*, 524–527.

Johnson, W. C., Jackson, L. E., Ochoa, O., van Wijk, R., Peleman, J., st. Clair, D. A., & Michelmore, R. W. (2000). Lettuce, a shallow-rooted crop, and *Lactuca serriola*, its wild progenitor, differ at QTL determining root architecture and deep soil water exploitation. *Theoretical and Applied Genetics*, *101*(7), 1066–1073. https://doi.org/10.1007/s001220051581

Jones, J., Dangl, J. (2006) The plant immune system. Nature 444, 323–329 https://doi.org/10.1038/nature05286

Kesseli, R., Ochoa, O., & Michelmore, R. (1991). *Variation at RFLP loci in Lactuca spp. and origin of cultivated lettuce (L. sativa)* Genome 34.3 (1991): 430-436.

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., & Varshney, R. K. (2020). Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends in Plant Science*, 25(2), 148–158). https://doi.org/10.1016/j.tplants.2019.10.012

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, *23*, 110–120. https://doi.org/10.1016/J.MIB.2014.11.014

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., et al. (2018).

High-resolution comparative analysis of great ape genomes. *Science*, *360*(6393). https://doi.org/10.1126/SCIENCE.AAR6343

Kuang, H., Caldwell, K. S., Meyers, B. C., & Michelmore, R. W. (2008). Frequent sequence exchanges between homologs of RPP8 in Arabidopsis are not necessarily associated with genomic proximity. *Plant Journal*, *54*(1), 69–80. https://doi.org/10.1111/j.1365-313X.2008.03408.x

Kuang, H., Wei, F., Marano, M. R., Wirtz, U., Wang, X., Liu, J., Shum, W. P., Zaborsky, J., Tallon, L. J., Rensink, W., Lobst, S., Zhang, P., Tornqvist, C. E., Tek, A., et al. (2005). The R1 resistance gene cluster contains three groups of independently evolving, type I R1 homologues and shows substantial structural variation among haplotypes of Solanum demissum. *Plant Journal*, *44*(1), 37–51. https://doi.org/10.1111/j.1365-313X.2005.02506.x

Kuang, H., Woo, S. S., Meyers, B. C., Nevo, E., & Michelmore, R. W. (2004). Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell*, *16*(11), 2870–2894. https://doi.org/10.1105/tpc.104.025502

Lebeda, A., Křístková, E., Kitner, M., Mieslerová, B., Jemelková, M., & Pink, D. A. C. (2013). Wild Lactuca species, their genetic diversity, resistance to diseases and pests, and exploitation in lettuce breeding. *European Journal of Plant Pathology, 138*, 597-640. https://doi.org/10.1007/s10658-013-0254-z

Lindqvist, K. (1960). On the origin of cultivated lettuce. *Hereditas*, *46*(3–4), 319–350. https://doi.org/10.1111/j.1601-5223.1960.tb03091.x

Matvienko, M., Kozik, A., Froenicke, L., Lavelle, D., Martineau, B., Perroud, B., & Michelmore, R. (2013). Consequences of Normalizing Transcriptomic and Genomic Libraries of Plant Genomes Using a Duplex-Specific Nuclease and Tetramethylammonium Chloride. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0055913

McHale, L., Tan, X., Koehl, P., & Michelmore, R. W. (2006). Plant NBS-LRR proteins: Adaptable guards. In *Genome Biology* (Vol. 7, Issue 4). Genome Biol. https://doi.org/10.1186/gb-2006-7-4-212

Meyers, B. C., Kozik, A., Griego, A., Kuang, H., & Michelmore, R. W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell*, *15*(4), 809–834. https://doi.org/10.1105/tpc.009308

Michelmore, R., Marsh, E., Seely, S., & Landry, B. (1987). Plant Cell Reports Transformation of lettuce (*Lactuca sativa*) mediated by *Agrobacterium tumefaciens*. In *Plant Cell Reports* (Vol. 6).

Michelmore, R. W., & Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. In *Genome Research* (Vol. 8, Issue 11, pp. 1113–1130). Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.8.11.1113

Michelmore, R., & Wong, J. (2008). *Classical and molecular genetics of Bremia lactucae, cause of lettuce downy mildew*. https://doi.org/10.1007/s10658-008-9305-2

Miiro, G., Serwanga, J., Pozniak, A., McPhee, D., Manigart, O., Mwananyanda, L., Karita, E., Inwoley, A., Jaoko, W., DeHovitz, J., Bekker, L. G., Pitisuttithum, P., et al. (2009). *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome*. https://doi.org/10.1126/science.1178746

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H. T., Chan, C. K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*(5), 1007–1013. https://doi.org/10.1111/TPJ.13515

Mou, B. (2011). Review Article Mutations in Lettuce Improvement. *International Journal of Plant Genomics*, *2011*. https://doi.org/10.1155/2011/723518

Nagarajan, N., Read, T. D., & Pop, M. (2008). Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, *24*(10), 1229–1235. https://doi.org/10.1093/BIOINFORMATICS/BTN102

Outram, M. A., Figueroa, M., Sperschneider, J., Williams, S. J., & Dodds, P. N. (2022). Seeing is believing: Exploiting advances in structural biology to understand and engineer plant immunity. *Current Opinion in Plant Biology*, *67*. https://doi.org/10.1016/J.PBI.2022.102210

Parra, L., Maisonneuve, B., Lebeda, A., Schut, J., Christopoulou, M., Jeuken, M., McHale, L., Truco, M. J., Crute, I., & Michelmore, R. (2016). Rationalization of genes for resistance to *Bremia lactucae* in lettuce. In *Euphytica* (Vol. 210, Issue 3, pp. 309–326). Springer Netherlands. https://doi.org/10.1007/s10681-016-1687-1

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O., Truco, M. J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14953

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., … Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, *326*(5956), 1112–1115. https://doi.org/10.1126/science.1178534

Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J., & Wang, Y. K. (1993). Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping. *Science*, *262*(55), 110–114. https://doi.org/10.1126/SCIENCE.8211116

Schwember, A. R., & Bradford, K. J. (2010). Quantitative trait loci associated with longevity of lettuce seeds under conventional and controlled deterioration storage conditions. *Journal of Experimental Botany*, *61*(15), 4423–4436. https://doi.org/10.1093/jxb/erq248

Seah, S., Telleen, A. C., & Williamson, V. M. (2007). Introgressed and endogenous Mi-1 gene clusters in tomato differ by complex rearrangements in flanking sequences and show sequence exchange and diversifying selection among homologues. *Theoretical and Applied Genetics*, *114*(7), 1289–1302. https://doi.org/10.1007/s00122-007-0519-z

Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. In *Nature Reviews Genetics* (Vol. 21, Issue 4, pp. 243–254). Nature Research. https://doi.org/10.1038/s41576-020-0210-7

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W. Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q. Y., Chen, L. L., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, *6*(1), 34–45. https://doi.org/10.1038/s41477-019-0577-7

Thompson, Ross C., and Kosar, W. "INTERSPECIFIC GENETIC RELATIONSHIPS IN *LACTUCA* (1941)." Journal of Agricultural Research 63 (1941): 91.

van der Biezen, E. A., & Jones, J. D. G. (1998). Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences*, *23*(12), 454–456. https://doi.org/10.1016/S0968-0004(98)01311-5

Wei, T., van Treuren, R., Liu, X., Zhang, Z., Chen, J., Liu, Y., Dong, S., Sun, P., Yang, T., Lan, T., Wang, X., Xiong, Z., Liu, Y., Wei, J., Lu, H., Han, S., Chen, J. C., Ni, X., Wang, J., … Liu, H. (2021). Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. *Nature Genetics*. https://doi.org/10.1038/s41588-021-00831-0

Wei, Z. (2016.). *Genetic diversity and evolution in Lactuca L. (Asteraceae) from phylogeny to molecular breeding*. Wageningen University and Research, 2016.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., …

Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wicker, T., Yahiaoui, N., & Keller, B. (2007). Contrasting rates of evolution in Pm3 loci from three wheat species and rice. *Genetics*, *177*(2), 1207–1216. https://doi.org/10.1534/genetics.107.077354

Wise, R. P., Moscou, M. J., Bogdanove, A. J., & Whitham, S. A. (2008). Transcript profiling in host-pathogen interactions. In *Annual Review of Phytopathology* (Vol. 45, pp. 329–369). https://doi.org/10.1146/annurev.phyto.45.011107.143944

Wood, K., Nur, M., Gil, J., Fletcher, K., Lakeman, K., Gothberg, A., Khuu, T., Kopetzky, J., Pandya, A., Pel, M., & Michelmore, R. (2020). Effector prediction and characterization in the oomycete pathogen *Bremia lactucae* reveal host-recognized WY domain proteins that lack the canonical RXLR motif. PLoS pathogens, 16(10), e1009012.

Xiao, S., Emerson, B., Ratanasut, K., Patrick, E., O'Neill, C., Bancroft, I., & Turner, J. G. (2004). Origin and maintenance of a broad-spectrum disease resistance locus in Arabidopsis. *Molecular Biology and Evolution*, *21*(9), 1661–1672. https://doi.org/10.1093/molbev/msh165

Yang, S., Zhang, X., Yue, J. X., Tian, D., & Chen, J. Q. (2008). Recent duplications dominate NBS-encoding gene expansion in two woody species. *Molecular Genetics and Genomics*, *280*(3), 187–198. https://doi.org/10.1007/s00438-008-0355-0

Zhang, L., Su, W., Tao, R., Zhang, W., Chen, J., Wu, P., Yan, C., Jia, Y., Larkin, R. M., Lavelle, D., Truco, M. J., Chin-Wo, S. R., Michelmore, R. W., & Kuang, H. (2017). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02445-9

Zhang, X., Liu, T., Wang, J., Wang, P., Qiu, Y., Zhao, W., Pang, S., Li, X., Wang, H., Song, J., Zhang, W., Yang, W., Sun, Y., & Li, X. (2021). Pan-genome of Raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes. *Molecular Plant*, *14*(12), 2032–2055. https://doi.org/10.1016/J.MOLP.2021.08.005

Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. In *Current Opinion in Immunology* (Vol. 20, Issue 1, pp. 10–16). https://doi.org/10.1016/j.coi.2007.11.003

Zohary, D. (1991). The wild genetic resources of cultivated lettuce (*Lactuca sativa* L.). *Euphytica*, *53*(1), 31–35. https://doi.org/10.1007/BF00032029

Chapter 2: Telomere-to-telomere, high-quality reference assemblies and annotation of domesticated lettuce (*Lactuca sativa* cv. Salinas) using Oxford Nanopore (ONT) and PacBio HiFi long-read technologies

Contributions:

I performed the majority of the work described in this chapter, especially the *de novo* construction of Oxford Nanopore based (v10) and PacBio HiFi based (v11) reference assemblies. I also performed the Bionano scaffolding for v11, *de novo* repeat analysis and *de novo* annotation for the v11 assembly. In addition, I performed assembly evaluation, including coverage and gap analysis, variant analysis, and orthogroup clustering and classification on the v10 and v11 assemblies and annotations.

Alexander Kozik contributed to Hi-C scaffolding and telomere and centromere analyses. Keri Cavanaugh prepared the plant material for sequencing by the UC Davis DNA Technologies core. Dean Lavelle and Kyle Fletcher provided computational support and contributed to Iso-seq analysis and base-calling of raw ONT reads. Rong Tao contributed to Bionano analysis of the v10 assembly along with Ting Ting. Mingchen Luo provided access to BioNano Compute On Demand for analysis.

2.1 Abstract

Lettuce is a commercially important crop worldwide with an annual farm-gate value of more than $3 billion in the United States. Domesticated lettuce (*Lactuca sativa* L.) displays enormous morphological diversity. Genomic resources based on short-read technologies are available; however, the large genome size, high repeat content, and family-specific whole-genome triplication has made it challenging to resolve complex regions of this genome using short-reads. Recently, long-read technologies have become available for generating high-quality reference assemblies and are especially helpful for complex plant genomes like lettuce. When paired with long-range scaffolding technologies, long-read technologies can reveal the architecture of complex genomic regions like centromeres or rDNA clusters. However, producing a telomere-to-telomere assembly remains a challenge that requires the use of several technologies and appropriate software. In this chapter, I evaluated the contemporary long-read and long-range scaffolding technologies, PacBio HiFi/Oxford Nanopore technology (ONT), BioNano, and Hi-C and generated multiple assemblies that were then compared for contiguity and accuracy. I then generated a high-quality telomere-to-telomere, highly contiguous, chromosome-scale annotated assembly of *L. sativa* cultivar Salinas. The final reference assembly that was based on PacBio HiFi sequencing has nine near-complete, telomere-to-telomere chromosomes, is 2.58 Gb, with a contig N50 of 12.5 Mb, consisting of 393 contigs and 98.5% complete for BUSCOs. This reference assembly resolves complex regions of the chromosome, including centromeres, telomeric repeats, and resistance gene clusters with great precision. Full-length transcripts generated by PacBio Iso-Seq along with Illumina-based RNA-seq data were used to validate the accuracy of the assembly and to annotate 44,241 protein-coding genes. This assembly provides the

foundation for developing a pan-genome for lettuce and building an extensive catalog of resistance genes.

2.2 Introduction

Lettuce (*Lactuca sativa*) is one of the world's most widely domesticated vegetable crops. It is one of the most profitable vegetable crops in the United States with an annual yield of over 8 billion pounds and a farm gate value of over $3.4 billion (USDA-NASS, 2020). Lettuce is a member of the Compositae (Asteraceae) family, which has a high number of flowering plants in terms of species that colonize diverse habitats. Domesticated lettuce has a wide range of morphologies. Based on their physical qualities, there are five principal varieties of lettuce: crisphead (iceberg), loose-leaf, romaine, butterhead, and stem(C. Yu et al., 2020). Despite their diverse morphologies, different types of lettuce share domestication characteristics such as leaf shape, lack of spines on leaves and stems, and loss of shattering seed. However, the molecular mechanism of domestication and diversification among its numerous horticultural forms remains under-studied. Each of these categories is comprised of multiple cultivars characterized by their morphology, acclimatization to their environment, and disease resistance. Understanding the genetic and genomic architecture underlying these cultivars is vital for lettuce improvement. Chromosome-scale reference genomes provide the foundation for understanding plant domestication and to understand the underlying molecular mechanism governing important traits (Ross-Ibarra et al., 2007; Yuan et al., 2017).

Lettuce genomics has advanced considerably over the past decade. Lettuce is a self-fertilizing diploid crop with 2n = 2x =18 chromosomes. The current publicly available reference genome (v8) of *L. sativa* cv. Salinas covers 2.4 Gb of the total estimated 2.7 Gb lettuce genome (Reyes-Chin-Wo et al., 2017). This genome assembly was constructed mostly using Illumina (short-read) and PacBio single-molecule real-time, continuous long read

(SMRT)(CLR) sequencing data. This reference genome is composed of 165,501 contigs with a contig N50 of 28.4 kb and a scaffold N50 of 1.8 Mb. The lettuce genome assembly provided the first chromosome-scale reference genome for the Compositae family and revealed whole-genome triplication events that are unique to this family. The v8 reference assembly of lettuce adds to the fundamental knowledge of the underlying genomic architecture of the lettuce genome; however, the high repeat content and family-specific triplication event has made it challenging to resolve complex regions of the genome using short-read technology (Claros et al., 2012).

The recent advancements in genome sequencing and assembly methods have permitted the near completion of high-quality genome assemblies (Berlin et al., 2015), particularly for plant genomes. The two long-read sequencing technologies, Single Molecular Real Time (SMRT) sequencing from Pacific Biosciences (PacBio) and Nanopore sequencing from Oxford Nanopore Technologies (ONT) have revolutionized the generation of highly contiguous genome assemblies (Goodwin et al., 2016). These two long-read technologies can provide average sequencing read lengths of 20 kb. This enables long reads to span most individual repeats, which otherwise cause thousands of fragmented contigs when using only short reads for genome assemblies. Long-read methods are under constant improvement. The PacBio SMRT sequencing platform released the Sequel II system and the updated SMRT cell enabled high-throughput HiFi reads using the circular consensus sequencing (CCS) mode to provide base-level resolution with >99% single-molecule read accuracy. ONT upgraded its PromethION platform, which can now yield >7 Tb per run, and its ultralong sequencing facilitates assembly of highly contiguous genomes. Despite the fact that long-read technologies have altered the way chromosome size assemblies are constructed, they are

often utilized in conjunction with other long-range data, such as optical mapping and/or chromosomal conformation capture sequencing (Hi-C) (Ghurye et al., 2019) for scaffolding and validating assemblies. The recent telomere-to-telomere (Kapustová et al., 2019) human genome (Miga et al., 2020) was driven by a combination of multiple contemporary technologies: ONT, SMRT, linked read sequencing from 10X Genomics (10X), and optical mapping from Bionano Genomics (BNG) (https://bionanogenomics.com/). The cost and effort required to achieve a T2T assembly have decreased dramatically over the past few years as sequencing reads have become longer and more accurate and as robust scaffolding methods, such as optical genome mapping and Hi-C (Miiro et al., 2009), have been developed. Until now, only a few crop genomes have been assembled telomere-to-telomere (T2T) including maize (Liu et al., 2020), barley (Navr et al., 2022), watermelon (Deng et al., 2022), and banana (Belser et al., 2022)

In this chapter, I describe the generation and comparison of two T2T chromosome-scale assemblies with very few gaps generated using ONT (v10) and PacBio single-molecule HiFi sequences (v11) for *de novo* assembly followed by scaffolding using Bionano optical maps and Hi-C. After several quality checks, the more accurate v11 assembly was selected as the new annotated reference assembly for lettuce. This T2T assembly is highly collinear with the previous publicly available v8 lettuce assembly but contains fewer gaps and improved gene structure.

2.3 Materials and Methods

2.3.1 Plant material collection and extraction of nucleic acids

High molecular weight (HMW) genomic DNA of *L. sativa* cv. Salinas was extracted using a modified method incorporating a sorbitol pre-wash combined with a high salt CTAB extraction as described in detail under "Method variations" by Ingles et al. (2018). Briefly, sterile, week-old seedlings were grown at 15°C in Magenta™ GA-7 boxes wrapped in aluminum foil to produce dark grown etiolated plant tissue (1–2 g) to minimize chloroplast formation (Sigma-Aldrich, Inc., St. Louis, MO). Additional modifications were made to this protocol as outlined in the "Method variations" section including substituting sodium metabisulfite (1% W/V) for beta-mercapto-ethanol in both the sorbitol pre-wash and lysis extraction buffers and lowering the lysis temperature from 65°C to 50°C. The integrity of the DNA samples was evaluated using the Femto Pulse (Agilent Technologies, Inc., Santa Clara, CA). Quantification and purity were assayed using the Qubit and NanoDrop (Thermo Fisher Scientific, Waltham, MA). DNA extraction was performed by the UC Davis Genome Center DNA Technologies Core (https://dnatech.genomecenter.ucdavis.edu/ Davis, CA).

2.3.2 Illumina sequencing and genome size estimation

Illumina short read datasets were used during the assembly analysis for nanopore read polishing, genome size estimation, assembly scaffolding, and variant detection. The Illumina data used are part of the two Hi-C libraries prepared by Dovetail Genomics. Libraries were sequenced in two lanes on an Illumina HiSeq 2500 in rapid run mode to generate 313 and 357 million 100 bp read pairs. This provided a total of 72x physical coverage.

In order to estimate the genome size and heterozygosity of *L. sativa* cv. Salinas, a k-mer based analysis was performed using the Jellyfish software version 2.2.10, and the k-mer

distribution was plotted with GenomeScope version 2.0 (k=27) (Vurture et al., 2017). The k-mer profile indicates the iteratively partitioned nucleotides of sequencing reads.

2.3.3 ONT PromethION library preparation and sequencing

Prior to library preparation, HMW DNA for one flow-cell was sheared to 50 kb to improve the ligation efficiency using a Megaruptor® 3 (Diagenode Inc., Denville, NJ). The remaining three libraries were not sheared, and directly used for library construction. 1 µg of purified genomic DNA was input into the Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, OX4 4DQ, UK), according to manufacturer recommendations, with the exception of end repair optimization.

The resulting sequencing libraries were sequenced on a R9.4.1 PromethION instrument (Oxford Nanopore Technologies, OX4 4DQ, UK). Four flow cells were used, with each flow cell receiving a nuclease flush every 20-24 hours. This flush removed long DNA fragments that could cause the pores to become blocked over time. Each flow cell received a fresh aliquot of the same library after the nuclease flush. In this way, a total of two outputs were obtained per flow cell, the initial data, followed by the post nuclease treatment data. The raw fast5 sequencing data was base-called with Guppy/v2.3.4 (Oxford Nanopore). The ONT DNA library construction and ONT DNA sequencing were performed by the UC Davis Genome Center DNA Technologies Core (https://dnatech.genomecenter.ucdavis.edu/ Davis, CA).

2.3.4 PacBio HiFi library preparation and sequencing

Long-read sequencing was performed using Circular Consensus Sequence (CCS) mode on a PacBio Sequel II instrument (Pacific Biosciences of California, Inc., Menlo Park,

CA). HMW DNA was sheared using Megaruptor® 3 (Diagenode Inc., Denville, NJ) to 15–18 kb for generation of PCR-free PacBio HiFi Libraries, while more highly sheared DNA (7–10 kb) was used in the construction of low and ultra-low input DNA libraries. Thus, three different library methods were used for four SMRT Cells: two PCR-free, one low (400 ng), and one ultra-low (7 ng) input each. Libraries were constructed using a SMRTbell Template Prep Kit 1.0 (Pacific Biosciences of California, Inc., Menlo Park, CA). Sequencing was performed using a 30-hour movie time with 2 hour pre-extension on a Sequel II instrument. The HiFi DNA library construction and HiFi DNA sequencing were performed by the UC Davis Genome Center DNA Technologies Core (https://dnatech.genomecenter.ucdavis.edu/). The resulting raw data was processed using either the CCS3.4 or CCS4 pipeline (GitHub, https://github.com/PacificBiosciences/ccs).

2.3.5 Genome assembly

2.3.5.a Draft lettuce assembly using Oxford Nanopore reads (v10)

Porechop v0.2.3 (https://github.com/rrwick/Porechop) was used to remove residual ONT adapters, and NanoFilt v2.7.1 (https://github.com/wdecoster/nanofilt) was used to select reads with an average quality score >Q10. NanoPlot v1.10.0 was used for visualization of ONT read qualities. To overcome the sequencing accuracy limitations of ONT reads, error-correction of ONT reads was performed using Canu v2.0 (Koren et al., 2017) with parameters "–correct stopOnReadQuality = false stopAfter = readCorrection." Nanopore reads have systematic errors in homopolymeric regions. Because a high-quality consensus assembly is needed for both aligning the optical map onto the contigs and for annotating genes, three rounds of iterative polishing were performed to improve the correctness of the assembly. The first round of polishing was done with Oxford Nanopore

reads as input to the Helen Marginpolish v0.01 software. The resulting self-corrected consensus assembly was polished again using Illumina reads as input to the Pilon v1.23 (Walker et al., 2014) tool. The final polishing was conducted using 12x PacBio HiFi reads as input to Hypo v1.0.3 software. Both Pilon v1.23 and Hypo v1.0.3 were used with default parameters and the consensus accuracy increased after each round (Goodwin et al., 2015).

2.3.5.b Draft lettuce assembly using PacBio HiFi reads (v11)

The highly accurate Q20 High Fidelity (HiFi) reads were generated with the Circular Consensus Sequencing (CCS) tool from PacBio ccs v6.0.0 (--min-passes 3 --min-length 10 --max-length 60000 --min-rq 0.99). To assess the impact of PCR libraries, adapter trimming and duplicate read removal was carried out using PacBio SMRT toolkit v 11.0.0.146107. The resulting 35x PacBio HiFi CCS reads were assembled with HiCanu (Nurk et al., 2020), Hifiasm v0.16.1, and IPA v1.3.1 assemblers. HiCanu was used through Canu/v2.0 with the following command: -assemble -pacbio-hifi genomeSize=3Gb maxInputCoverage=30. Simultaneously, the CCS reads were assembled using Hifiasm v0.16.1-r375 after purging of duplicated contigs using -lo option. PacBio's IPA/v1.3.1 was used in cluster mode (dist) and skipping phasing (--no-phase). The resulting draft assemblies from these assemblers were evaluated for contiguity, correctness, and completeness. Due to the low baseline error rate of HiFi Reads, no further polishing was done on these assemblies. Assembly completeness was checked by Benchmarking Universal Single Copy Orthologs (BUSCO) analysis. The embryophyta_odb10 data set comprises 57 species and 425 genes.

2.3.6 Bionano Genomics (BNG) Saphyr library preparation and fingerprinting

HMW DNA was extracted and stained according to the instructions provided by Bionano Prep Plant Tissue DNA Isolation Kit and Bionano Prep DLS Labeling Kit, respectively. Using the BNG Prep DLS DNA Labeling Kit (#80005), 750 ng of HMW gDNA was labeled with DLE-1 enzyme, followed by proteinase digestion and a membrane cleaning process. The tagged DNA was loaded onto a Saphyr Chip G2.3 (BNG #20366) and processed on a Saphyr system (BNG #60325) using the Saphyr Instrument Control Software (ICS version 3.1) to maximize the throughput of molecules. Using Saphyr ICS version 3.1, DNA raw images were transformed to digital molecular files. First, a genome map was created using Bionano Solve Pipeline/v3.1.1 and Bionano Access/v1.0. A *de novo* assembly was conducted using the following parameters: -i 0 -V 0 -A -z -u -m (pipelineCL.py). Next, ONT and PacBio HiFi draft assemblies were *in silico* digested to generate a sequence consensus map (CMAP). The sequence CMAP was aligned to the Bionano genome map using RefAligner tool with an initial alignment cutoff of P < 1 x 1-10. The generated CMAP was subsequently employed for hybrid scaffolding of v10 or v11 draft assembly contigs with BioNano maps. Molecules less than 180 kb and those with fewer than nine labeling sites were eliminated.

To resolve misassembles, conflict regions were checked manually to determine whether the consensus maps or the contig sequence in the conflicting alignments were misassembled (Figure 2.5) (Salzberg & Yorke, 2005). The assembly parameters were set as: -U -d -T 20 -j 4 -N 10 -i 5. Based on the following criteria, misassemblies were corrected in the draft assembly. First, if the chimeric score and coverage supported the Bionano genome map, a cut was made to the sequence map, otherwise the genome map was cut. Next, raw reads were mapped back to the assembly using minimap2; if the reads supported the sequence map, then a cut was made to the Bionano genome map, otherwise the sequence

map was cut. To further improve the conflicts found by the hybrid scaffold pipeline, several regions of the optical map were inspected where two contigs were joined or overlapped, but the overlap was not supported by the hybrid scaffolding workflow. Conflict resolution was carried out as follows. If there was a conflict in the alignment 10 kb from the conflicting site, within which the chimeric quality score of Bionano genome map, labels were examined for a minimum chimeric quality score of 35% and a minimum coverage threshold of 10x. Once all conflicts were resolved, hybrid scaffolds were generated by merging the sequence CMAP and Bionano genome map. Finally, sequences and Bionano genome maps were aligned back to hybrid scaffolds with an alignment cutoff of $P < 1 \times 10\text{-}10$. Positive gaps with lengths smaller than 23 bp were filled with 23 Ns, otherwise the gaps were filled with the number of Ns corresponding to the estimated length by genome map. Negative gaps were filled with 13 Ns. Negative gap sizes were manually checked and corrected using in-house script to avoid artificial genomic duplications (Fig. 2.1). With this workflow, potential misassembled sequences or maps were split at the misassembled regions, which further improved the contiguity of the draft assemblies (Shelton et al., 2015).

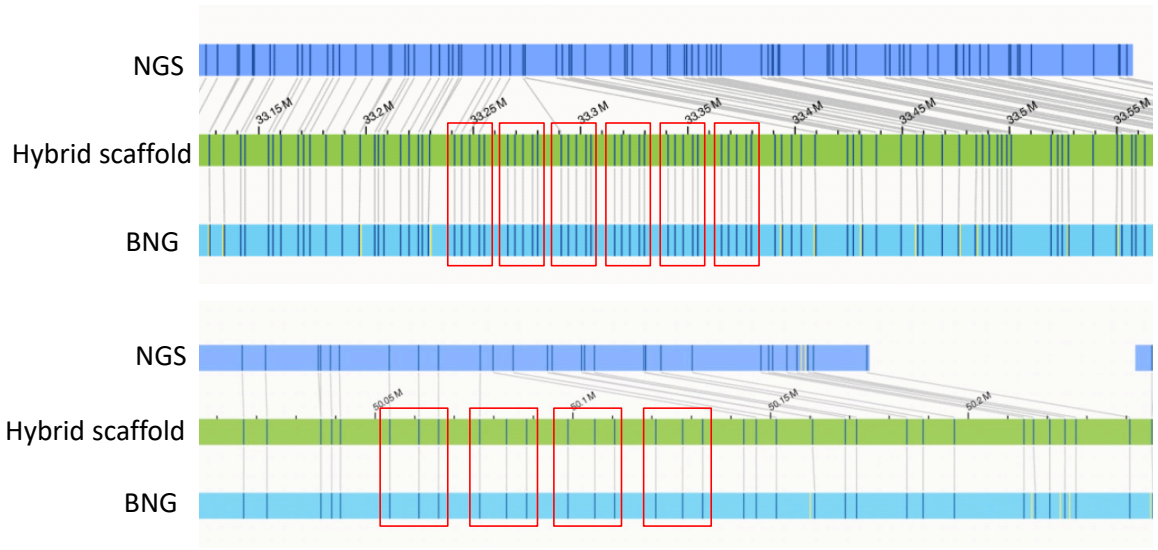

Figure 2.1.a. *Optical mapping-based assembly correction and scaffolding.*

Example showing a complex repeat region. NGS represents draft Salinas contig aligned to Bionano maps (BNG).

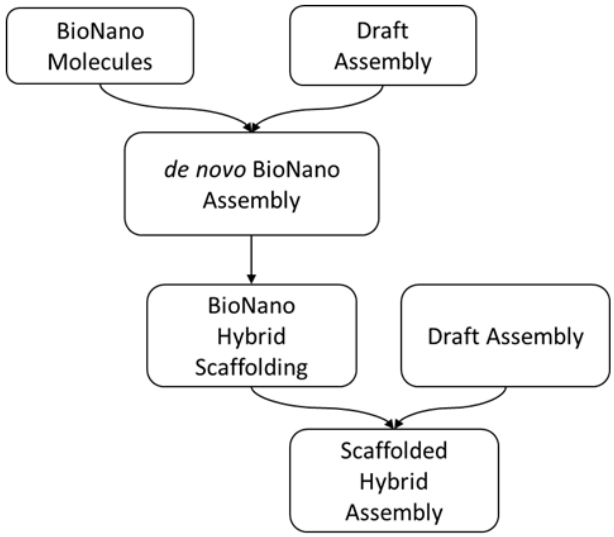

Figure 2.1.b. *Optical mapping-based assembly correction and scaffolding.*

Bionano hybrid scaffolding workflow.

2.3.7 Hi-C library construction and sequencing

The Hi-C library was prepared by Dovetail Genomics LLC (Santa Cruz, CA, USA) as described previously in Lieberman-Aiden et al. (2009). Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with *Dpn*II, the 5′ overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA was purified from protein. Biotin that was not internal to ligated fragments was removed from the purified DNA. Purified DNA was then sheared to ~350 bp mean fragment size. Sequencing libraries were generated using NEBNext® UltraTM enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were then sequenced on the Illumina HiSeq 4000 platform for 150 bp paired end reads, with a sequencing depth of approximately 72x coverage of the lettuce genome, with a total of 194 Gb.

Reads were mapped to BioNano scaffolds in draft assemblies using BWA with a mapping quality cutoff of 30 and then used for scaffolding using HiRise software. According to the orders and orientations provided by the alignment, those contigs were clustered into chromosomes using custom scripts. The resulting assemblies were manually corrected and validated by drawing contact maps with HiC-Explorer toolkit/v3.6 (Wolff et al., 2018).

2.3.8 PacBio Iso-seq library preparation and sequencing

For transcriptome sequencing, we isolated total RNA from developing cotyledons collected from *L. sativa* cv. Salinas, cv. Ninja, cv. ViAE, and cv. ViCQ. Each cultivar was grown for 5 days in 7hrs light/7hrs dark followed by 3 days in the dark, with and without infection

with *Bremia lactucae* (isolate 1326). Three biological replicates were generated for each condition (i.e., control/inoculated). For each sample, total RNA was extracted using a Qiagen RNeasy Plant Mini kit (cat 74903). cDNA synthesis and amplification were performed according to NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module (New England Biolabs Inc.) and purified using the standard workflow for samples composed primarily of transcripts centered around 2 kb. Each cDNA sample was uniquely barcoded with NEBNext Single Cell cDNA PCR Primer and Iso-Seq Express cDNA PCR Primer. These cDNA samples were pooled and constructed into a SMRT bell library as a single, 3-plexed sample according to PacBio "Procedure & Checklist – Iso-Seq Express Template Preparation for Sequel and Sequel II Systems (Version 02, October 2019)."

Total RNA was first quantified with a Nanodrop 8000 (avg. concentration = 1,596 ng/μL, avg. A260/280 = 2.13; and avg. A260/230 = 2.06). Next, the concentration of pure RNA was determined using a Qubit 2.0 and the Qubit RNA BR assay kit (avg. concentration = 1,318 ng/μL). Last, the RNA integrity was determined using the BIO RAD Experion™ automated electrophoresis station using the Experion™ RNA StdSens Analysis Kit (avg. RNA Quality Number = 8.4). The RNA samples were sequenced on a Sequel II instrument (Pacific Biosciences) at the UC Davis Genome Center.

The Iso-seq data were processed as described by PacBio here: https://github.com/PacificBiosciences/IsoSeq/blob/master/isoseq-clustering.md. SMRTlink software v7.0 was used to filter and process the raw sequencing subreads with the cutoff or read quality ≥0.8 minReadScore = 0.8). Subreads were merged to generate full-length circular consensus sequences (CCS). Primer artifacts were removed and the reads demultiplexed by library barcode. Polyadenylated tails and concatemers were trimmed and

removed. Finally, the CCS reads were clustered into partial and full-length transcripts. Additional precautions were taken with IsoSeq Polish to generate consensus for each read cluster by generating per base quality values (QVs) for transcript consensus sequences.

2.3.9 Genome annotation

Gene prediction for the v11 draft genome was carried out using the MAKER-P pipeline with the available high-quality Iso-seq transcripts and the expressed sequence tags (ESTs) available for lettuce. Prior to genome annotation, MAKER-P repeat masking workflow was used to mask repetitive and transposon-rich regions of the genome to avoid their annotation as protein-coding genes. Transposable elements were identified by combining homology-based and *de novo* techniques (Janicki et al., 2011).

Using RepeatMasker v4.0.7, Repbase (Bao et al., 2012), and a collection of custom repeat libraries as references, the genome was mined for repeat elements. RepeatScout v1.05 and RepeatModeler v1.0.9 was used to identify and classify *de novo* repeat families with default parameter settings. RepeatMasker/v4.07 was used to report different repeats (SINEs, LINEs, TE elements, DNA elements, interspersed repeats, small RNA, satellites, simple repeats, and low complex repeats) in the assembly. In addition, custom scripts and Tandem repeats finder trf v4.09 (Benson, 1999) were used to check the telomeric repeat ('TTTAGGG') on the lettuce assemblies. Custom scripts were used to screen long-terminal repeats to eliminate false positives and to avoid accidental masking of known R-genes in the lettuce genome. Finally, all collected repetitive sequences were compared to a BLAST database of plant proteins from SwissProt and RefSeq, where proteins from transposable elements are excluded.

Genome annotation was carried out by integrating evidence from *ab initio* prediction and homologous protein sequence alignment from 17 closely related plants including Arabidopsis, soybean, tomato, potato, carrot, tobacco, castor and Compositae species including sunflower and *Brassica rapa*. Exonerate v2.56.0 with the 'protien2genome' model was used to align protein sequences from these 17 species to predict gene structure. Two rounds of MAKER runs were conducted followed by extensive evaluation of gene models. In Round 1, high-quality transcript evidence, protein evidence, and *ab initio* gene predictors, like SNAP v-2013-11-29 and AUGUSTUS v3.3.2, were trained to generate a comprehensive set of protein coding genes. This was then followed by Round 2, which used the maker gff3 file containing predicted gene models with annotation edit distance (AED) score equal to 0 from round 1 as input, to train and polish *ab initio* gene predictors models. The two rounds of MAKER annotation were compared, and the better round was selected if their structures were better supported by homologous proteins or Iso-seq-assembled transcripts.

2.3.10 Orthology assignment

Orthofinder v2.1.2 was used using the default settings to identify clusters of orthologous genes. Using unigene sets, protein sets from closely related Compositae species, as well as *Brassica*, tomato, soybean, and Arabidopsis, were used for clustering OG assignment. The gene sequences from other plant species were acquired from NCBI GenBank. Using Bayesian and maximum likelihood approaches, the single copy orthologs retrieved from Orthofinder were utilized to generate a synteny plot and phylogenetic tree.

2.4 Results

2.4.1 Genome size estimation

The k-mer based analysis of Illumina reads estimated the genome size and heterozygosity of *L. sativa* cv. Salinas to be ~2.7 Gb, made up of at least 83% repetitive elements, and with a low level of heterozygosity (0.26%) (Figure 2.2). This size is consistent with previous reports based on estimates from Fleugen staining (Baranyi & Greilhuber, 1996).

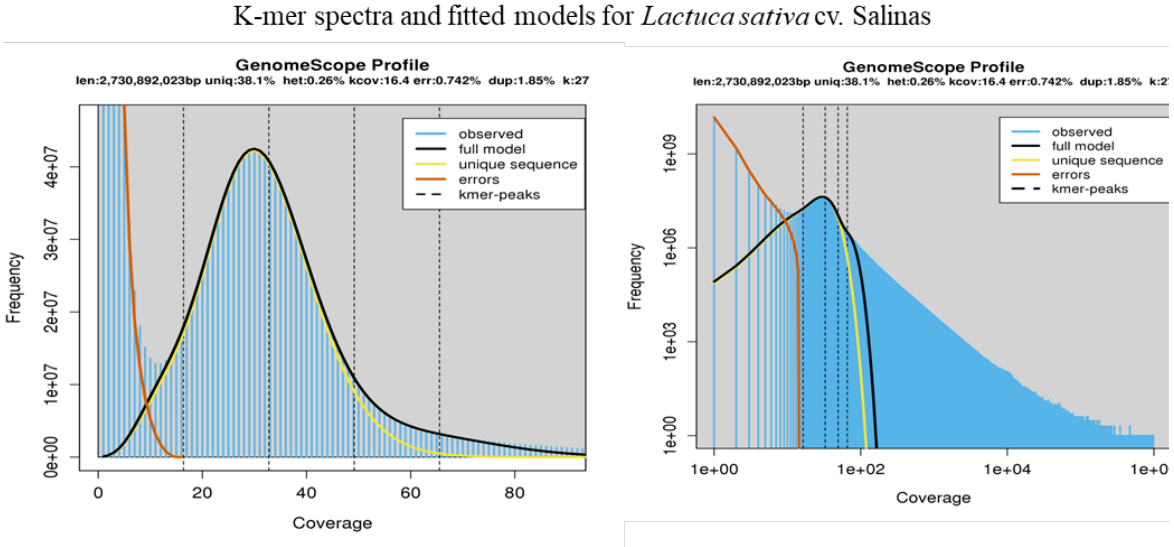

Figure 2.2. *GenomeScope plots for L. sativa cv. Salinas.*

Genome and read characteristics produced by GenomeScope. Len: haploid genome length; uniq: overall length of unique sequences; het: heterozygosity rate; kcov: mean k-mer coverage for heterozygous sequences; err: error rate of reads; dup: read duplication rate; k: k-mer length.

2.4.2 Long-read sequencing

Reads for genome assemblies were generated using ONT (https://nanoporetech.com/) or PacBio HiFi technologies (https://www.pacb.com/). Sequence data generated from four ONT PromethION flow cells resulted in a total of 251 Gb with an average yield of 31–79 GB per flow cell with reads averaging 9 to 23 kb and read length N_{50} ranging from 26 to 44 kb (Table 2.1). The PacBio HiFi sequence reads were generated from the PacBio Sequel II system. Sequencing of five PacBio SMRT cells generated a total of 91 Gb of high-quality (Q40) data, with an average read length of 10–16 kb. (Table 2.2).

Table 2.1*. Statistics for the Nanopore PromethION flow cells used to assemble the lettuce v10 genome.*

Metric	Flowcell - 01	Flowcell - 02	Flowcell - 03	Flowcell - 04
Mean read length (kb)	29,577	14,949	14,851	20,662
Mean read quality	9.4	9.4	9.6	9.3
Median read length (kb)	23,567	9,129	9,128	14,471
Median read quality	9.4	9.3	9.6	9.3
Number of reads	10,64,361	46,22,363	53,33,053	34,81,365
Read length N_{50} (kb)	44,921	26,638	28,699	32,744
Total bases (Gb)	31	69	79	72
Number, percentage, and megabases of reads above quality cutoffs				

>Q7	1,064,351 (100.0%) 31,480 Mb	4,622,343 (100.0%) 69,098 Mb	5,332,988 (100.0%) 79,203 Mb	3,481,200 (100.0%) 71,931 Mb
>Q10	318,037 (29.9%) 9464.1Mb	1,372,648 (29.7%) 20491.1Mb	2,062,965 (38.7%) 31690.0Mb	926,313 (26.6%) 18422.8Mb
>Q12	16,960 (1.6%) 332.8Mb	104,504 (2.3%) 988.9Mb	206,845 (3.9%) 1896.7Mb	62,191 (1.8%) 872.5Mb

Table 2.2*. Statistics for the PacBio raw polymerase reads and filtered subreads used to assemble the lettuce v11 genome.*

	SMRT CELL-1	SMRT CELL-2	SMRT CELL-3	SMRT CELL-4	SMRT CELL-5
Input DNA	STANDARD	5 ng	400 ng	STANDARD	STANDARD
PCR cycles	PCR free	13 cycles	6 cycles	PCR free	PCR free
Total Reads ≥Q20	760,072	1,799,634	654,802	1,666,111	1,871,160
Average Read Length (kb)	14	11	10	15	16
Average Read Quality	Q35	Q40	Q40	Q40	Q40
Average Yield ≥Q20 (Gb)	11	20	6	25	29

*UL – ultra low DNA input; *L – low DNA input

2.4.3 Genome assembly

The ONT reads were corrected using Canu/v2.0. Several genome assemblers, wtdbg2/v2.5 (https://github.com/ruanjue/wtdbg2), Canu/v2.0 (Koren et al., 2017), and Shasta/v0.5.0 (Shafin et al., 2020), were evaluated in parallel for construction of an ONT genome assembly. Based on the assembly metrics from the different assemblers, such as N50

for contiguity and BUSCO scores for completeness, the Shasta assembler produced the most contiguous and complete assembly and was therefore used to generate the ONT draft assembly (Table 2.3). The resulting draft genome assembly had a contig N_{50} of 8.1 Mb and a percentage BUSCO completeness of 75.3%. As the nanopore reads have systematic errors in homopolymer regions, in order to improve the accuracy of the ONT based assembly, we performed three rounds of iterative polishing using both Illumina and PacBio HiFi reads (Figure 2.3 Workflow). The polishing process significantly improved the number of complete BUSCOs detected; the percentage of complete BUSCO went from 75.3% to 98.2% in the final genome assembly.

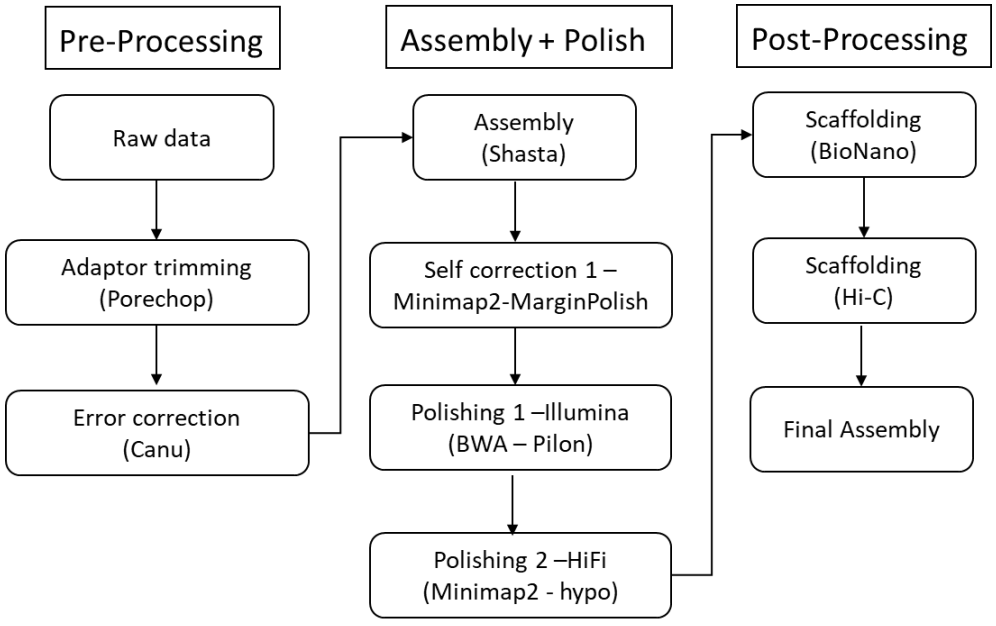

Figure 2.3. *Genome assembly workflow using Oxford Nanopore data*

Table 2.3. *Comparison of ONT based draft Lactuca sativa cv. Salinas assemblies using various assemblers.*

	wtdgb2/v2.5	Canu/v2.0	Shasta/v0.5.0
Assembly size (Mb)	2,317	2,584	2,572
Total contig	10,736	2,284	1,780
Total contig sequence (Mb)	2,317	2,569	2,572
Contig N/L50 (Mb)	1,079/0.623	97/0.008	97/0.008
Contig N/L90 (Mb)	3,886/0.154	252/3	251/3
Max contig length (Mb)	4.476	49.535	49.59

Subsequently, the assemblers HiCanu v2.0, HiFiasm v0.14, and IPA v1.5.0 were evaluated for assembling PacBio HiFi reads from three SMRT cells. Comparison of the resulting draft assembly using the default settings of the HiFiasm assembler resulted in the best assembly with high contiguity (Table 2.4; Fig. 2.4). More HiFi reads were generated using one additional SMRT cell, and a total of 6,751,779 reads were assembled using HiFiasm to generate a highly contiguous assembly with a contig number of 484 and N_{50} of 12.5 Mb. BUSCO assessment showed 98.5% completeness. No further polishing was necessary for this PacBio HiFi based assembly because such reads are 99.9% accurate. Based on the aggregated statistics, the HiFiasm-generated assembly was adopted as the HiFi draft genome assembly.

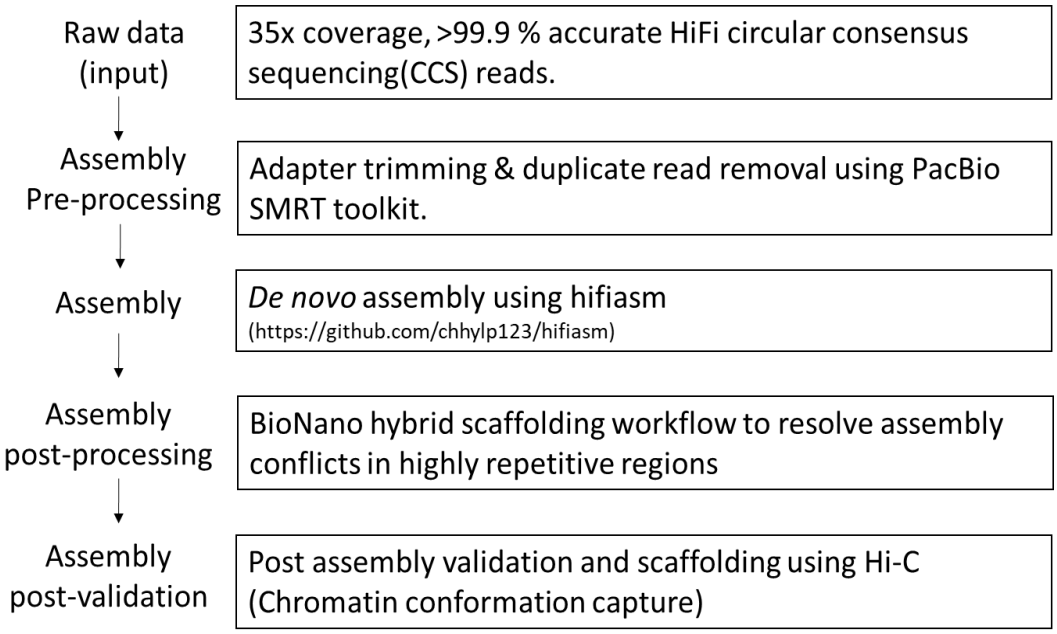

Figure 2.4. *Assembly workflow using PacBio HiFi data.*

Table 2.4. *Comparison of HiFi based L. sativa cv. Salinas assemblies using various assemblers.*

	HiCanu/v2.0	HiFiasm/v0.14	IPA/v1.5.0
Assembly size (Mb)	2,735	2,614	1,678
Total contig	7,784	1,764	15,165
Total contig sequence (Mb)	2735	2614	1678
contig N/L50 (Mb)	265/2.973	137/5.498	4506/0.122
contig N/L90 (Mb)	662/1.298	374/2.194	12162/0.062
contig N/L90 (Mb)	15.321	24.764	0.808

2.4.4 Analysis and integration of BioNano Optical Mapping data

To scaffold the contigs of the ONT and HiFi draft assemblies, *de novo* optical mapping data were generated using the Bionano genomics SAPHYR system. Fragments with an average length of 200 kb were collected and an optical map assembled *de novo*. Contigs from each draft assembly were mapped to the optical map to generate hybrid scaffolds. The resulting CMAP statistics for both draft assemblies are shown in Table 2.5.

Table 2.5. *BioNano consensus map (CMAP) input statistics of L. sativa cv. Salinas genome assemblies. The v10 assembly was generated from ONT reads and the v11 assembly was generated from PacBio HiFi reads.*

	Lsat_v10 (ONT based)	Lsat_v11 (PacBio-based)
Total number of molecules	19,627,828	20,670,345
Total length (Mb)	2,110,077.37	2,132,433.61
Average length (Mb)	0.107	0.103
Molecule N50 (Mb)	0.188	0.269
Label density (/100kb)	14.108	14.681

In the BNG workflow, *de novo* assembly of molecules was followed by hybrid scaffolding. The resulting assembly statistics after hybrid scaffolding is shown in Table 2.6.

Table 2.6. *Hybrid scaffolding statistics of L. sativa cv. Salinas genome assemblies.*

Lsat_v10 (ONT)	**Original BNG**	**Original NGS**	**NGS used in hybrid**	**Hybrid**	**Hybrid + not scaffolded NGS**

Number of maps	39.0	1714.0	574.0	24.0	1168.0
Min length (Mb)	0.1	0.0	0.0	0.2	0.0
Median length (Mb)	48.2	0.0	2.9	109.2	0.0
Mean length (Mbp)	70.4	1.5	4.5	107.2	2.2
N50 (Mb)	146.5	8.3	8.2	148.1	148.1
Max length (Mb)	344.3	49.6	49.6	323.9	323.9
Total length (Mb)	2744.8	2572.1	2563.3	2572.2	2580.7
Lsat_v11 (PacBio)	**Original BNG**	**Original NGS**	**NGS used in hybrid**	**Hybrid**	**Hybrid + not scaffolded NGS**
Number of maps	49.0	1705.0	402.0	16.0	1323.0
Min length (Mb)	0.1	0.0	0.0	1.1	0.0
Median length (Mb)	17.2	0.0	3.5	151.6	0.0
Mean length (Mb)	52.9	1.5	6.4	161.7	2.0
N50 (Mb)	121.3	12.5	12.5	206.7	172.9
Max length (Mb)	289.5	75.5	75.5	324.7	324.7
Total length (Mb)	2593.1	2633.5	2573.5	2586.6	2635.0

After resolving conflicts, the ONT based assembly had 574 contigs that were placed into 24 near-chromosome scale BioNano super scaffolds. The remaining 1,151 contigs were grouped together as unplaced scaffolds that spanned a length of 9.3 Mbp. Similarly, the PacBio based HiFi assembly had 402 contigs placed into 16 near-chromosome scale BioNano

super scaffolds. The remaining 91 contigs were grouped together as unplaced scaffolds that spanned length of 4.8 Mbp.

2.4.5 Analysis and integration of chromatin conformation capture data

In addition to optical mapping data from BioNano, chromatin conformation capture sequencing data for the *L. sativa* genome was also generated. Hi-C read pairs were mapped to the draft BioNano scaffolds. The Hi-C reads were used to correctly orient BioNano scaffolds in the draft assemblies. As anticipated, the majority of matched pairs within the same contigs were fewer than 25 kb apart. Nonetheless, some of them had linkage distances of up to several hundred kilobytes. Using read pairings mapped to distinct BioNano scaffolds, this scaffolding approach identified and divided possibly misassembled sequences and constructed the error-corrected scaffolds. By integrating the BNG optical mapping workflow and HiRise scaffolding, one minor inter chromosomal misassembly for the ONT draft assembly was corrected and no major misassembly for the HiFi draft were identified. By Hi-C integration, high repeat regions in Chromosomes 1, 4, and 5 were resolved mostly in the v11 assembly compared to the v10 assembly (Figure 2.5). The Hi-C analysis increased contig N_{50}s from 28 kb to 8.1 Mb and 12.5 Mb for the ONT and HiFi assemblies, respectively. Figure 2.6 shows a closer look at the better resolved centromeric region of v11 assembly compared to the v10 assembly.

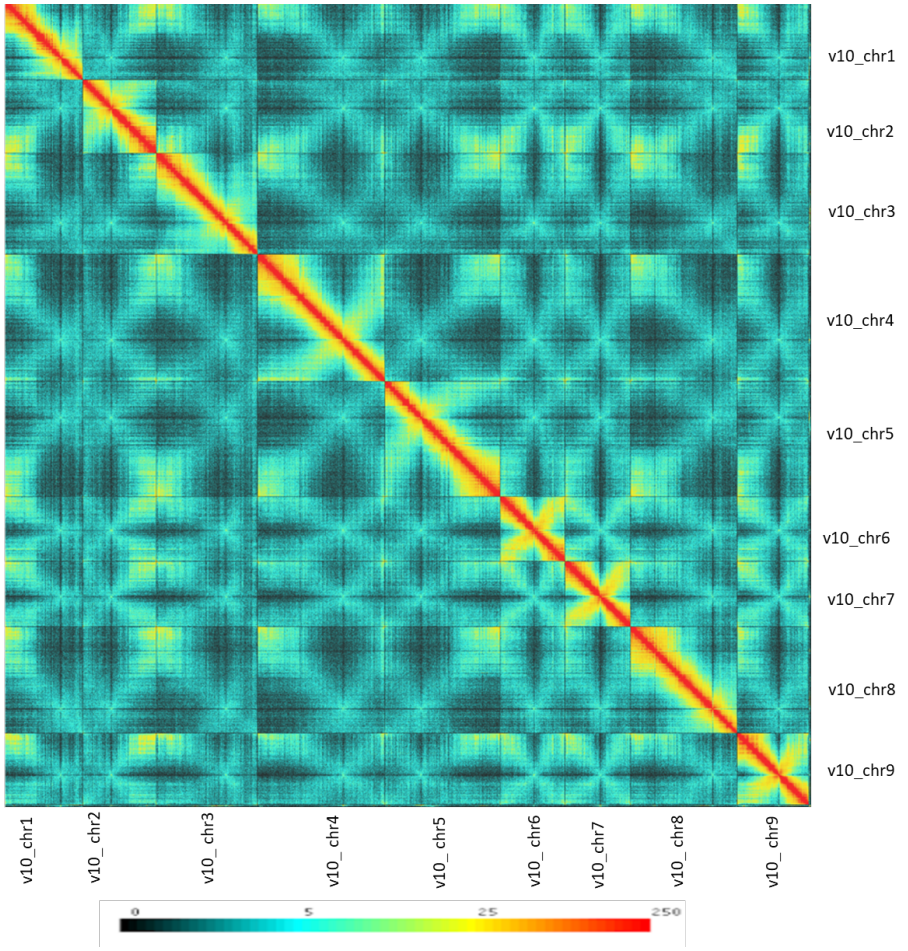

Figure 2.5.a. *Assembly scaffolding using chromatin capture data for L. sativa*_v10.

Hi-C contact frequencies for *L. sativa* (v10) nine chromosomes. The color gradient reflects the number of read-pairs from 0 (dark grey) to 250 and higher (red) in each 2-dimensional-1-MB bin. The total number of analyzed Hi-C read pairs was 100 million.

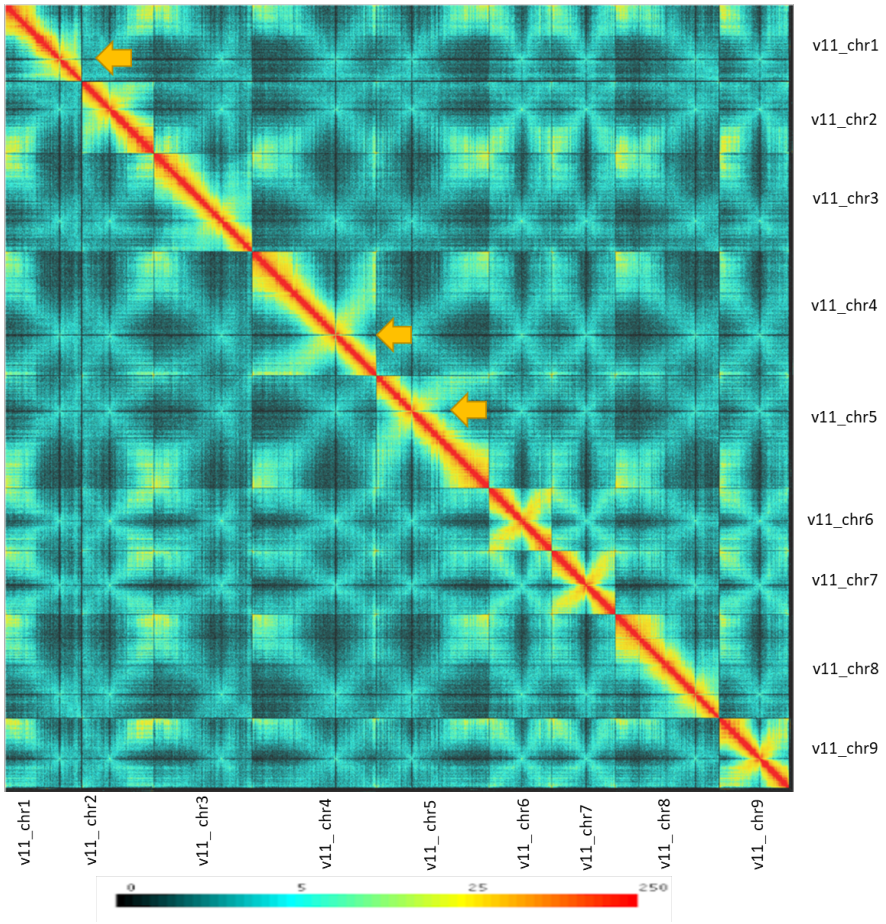

Figure 2.5.b. *Assembly scaffolding using chromatin capture data for L. sativa*_v11.

Hi-C contact frequencies for *L. sativa* (v11) nine chromosomes. The color gradient reflects the number of read-pairs from 0 (dark grey) to 250 and higher (red) in each 2-dimensional-1-MB bin. The total number of analyzed Hi-C read pairs was 100 million. The yellow arrows indicate the regions that are better resolved in the v11 assembly: chr 01, chr 04, and chr 05 after Bionano analysis and genetic orientation into the nine near-complete chromosomes.

Lactuca sativa cv. Salinas v10 – (chr-5) *Lactuca sativa* cv. Salinas v11– (chr-5)

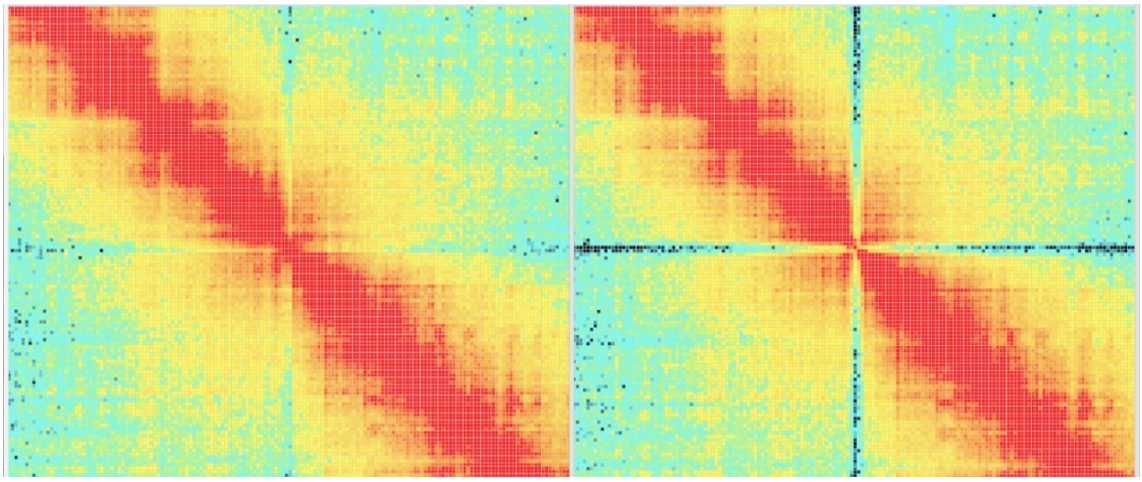

Figure 2.6. *Hi-C contact frequencies matrix of Chromosome 5 in the v10 and v11 assemblies showing the characteristic pattern of centromeric repeat in the middle.*

By first using the Bionano optical maps followed by the hybrid scaffolding workflow, major misassembles were identified and 526 v10 ONT contigs were placed into 29 near-chromosome scale Bionano scaffolds. Similarly, 393 v11 HiFi contigs were placed into 15 BioNano super scaffolds. The combination of the approaches resulted in highly contiguous assemblies. The final ONT assembly had a size of 2,566 Mb, a contig N50 of 8.13 Mb and a scaffold N50 of 323 Mb and was designated v10. The final HiFi assembly had a size of 2.588 Mb, a contig N50 of 12.52 Mb, and a scaffold N50 of 324 Mb and was designated v11.

2.4.6 Assembly evaluation and validation

The v11 and v10 assemblies were then compared to the previously published, publicly available *L. sativa* v8 reference genome of cv. Salinas. Whole genome synteny comparison between v8/v11 and v8/v10 were plotted with Assemblytics v1.2.1 with whole genome DNA alignments performed using NUCmer from MUMmer v4.0 (Delcher et al., 2003; Marçais et al., 2018) with the command --maxmatch -l 1000 -c 500. Both v10 and v11 assemblies are collinear with v8. Assembly contiguity expressed as N_{50} and L_{50} of both assemblies was greatly increased compared to v8, which has a N_{50} of 2.8 kb. For the completeness of the assembly, BUSCO analyses were performed on the v10 and v11 assemblies. BUSCO completeness for both v10 and v11 assemblies were almost complete (both 98.5%) with a slight improvement on the v8 assembly (Table 2.5) (Figure 2.3).

Table 2.7. *Comparison of Lactuca sativa cv. Salinas genome assemblies.*

	Illumina based	**Nanopore based**	**PacBio HiFi**
	Lactuca sativa cv. Salinas_v8	*Lactuca sativa* cv. Salinas_v10	*Lactuca sativa* cv. Salinas_v11
Assembly size (Mb)	2,391	2,566	2,588
# Contigs	168,554	1,793	484
Contig N50 (Mb)	0.028	8.135	12.52
Contig size N90 (Mb)	0.007	3.482	4.68
Largest contig (Mb)	0.363	31.735	75.542
BUSCO % Complete	98.2	98.4	98.5
BUSCO % Duplicate	2.9	3.3	3.3

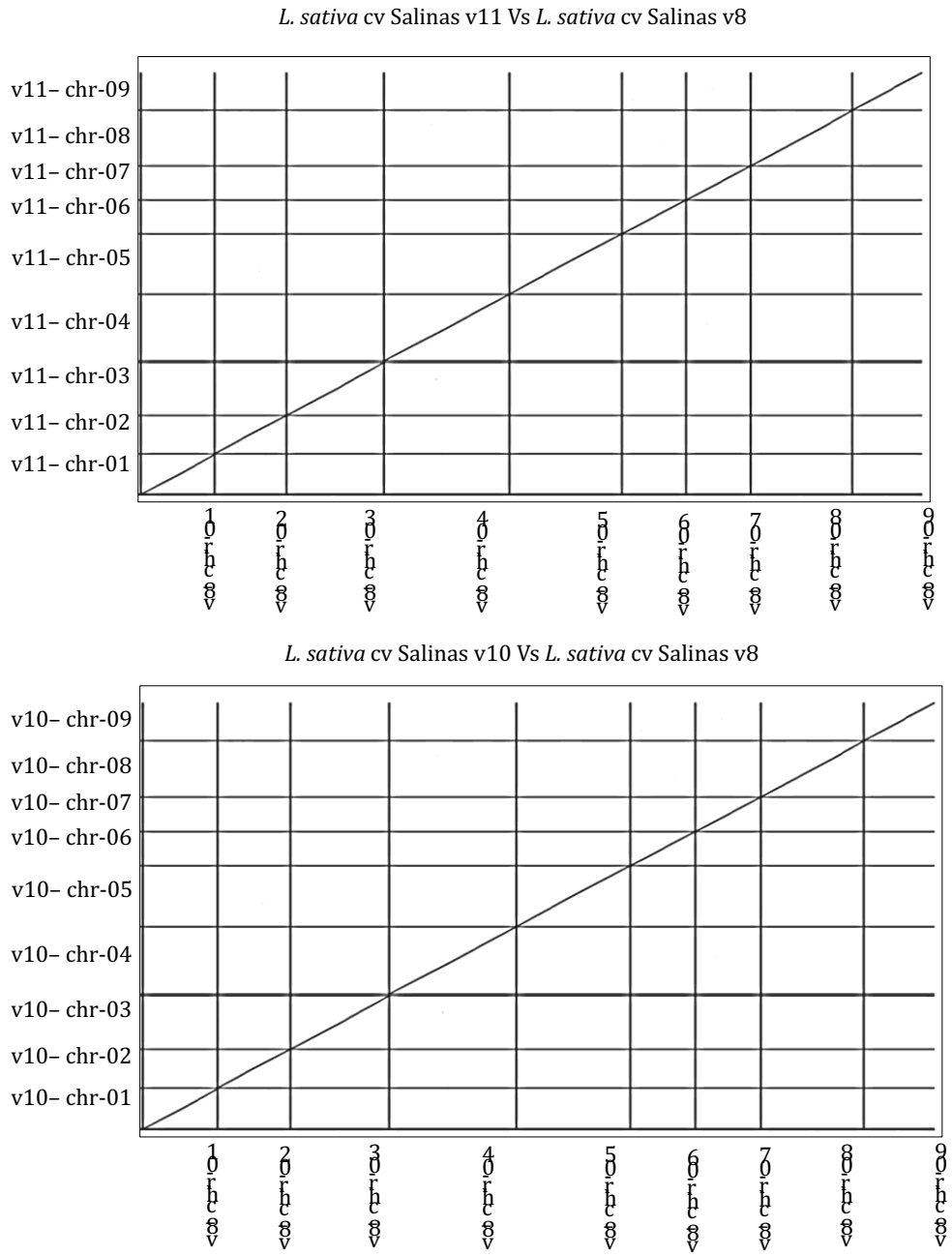

Figure 2.7*. MUMmerplot comparison of v10 (ONT based)/v11 (HiFi based) with v8).*

The *Lactuca sativa* cv. Salinas v8 assembly is plotted along the x-axis and v10/v11 assembly is plotted along the y-axis.

2.4.7 Assembly validation and estimation of assembly error rate

The error rate of the v10 and v11 assemblies was estimated in several ways. The single nucleotide errors and InDels were quantified based on alignments of Illumina paired-end reads to each assembly genome wide, in open reading frames (ORFs), and 2 and 10 kb upstream of ORFs (Table 2.8). Assembly errors were also estimated by alignments of Pac-Bio Iso-seq reads. v11 has greater concordance with Illumina and Isoseq data than v10. There were fewer discrepancies between Illumina reads and genome assembly in genic regions. There were only 62 single nucleotide polymorphisms (SNPs) between Isoseq reads and the v11 assembly. Therefore, the v11 assembly was judged to be more accurate than v10.

Table 2.8. *Variant discovery in Lactuca sativa cv. Salinas genome assemblies.*

	Lsat_v10		**Lsat_v11**	
	SNPs	Indels	SNPs	Indels
Total	60,680	59,006	11,535	9,244
Repeat-region	40,816	40,947	6,461	4,667
ORFs + flanking 2,000 bp	2,158	3,338	399	855
ORFs + flanking 10,000 bp	11,147	12,831	1,540	2,274
Iso-seq transcripts	246	366	62	186

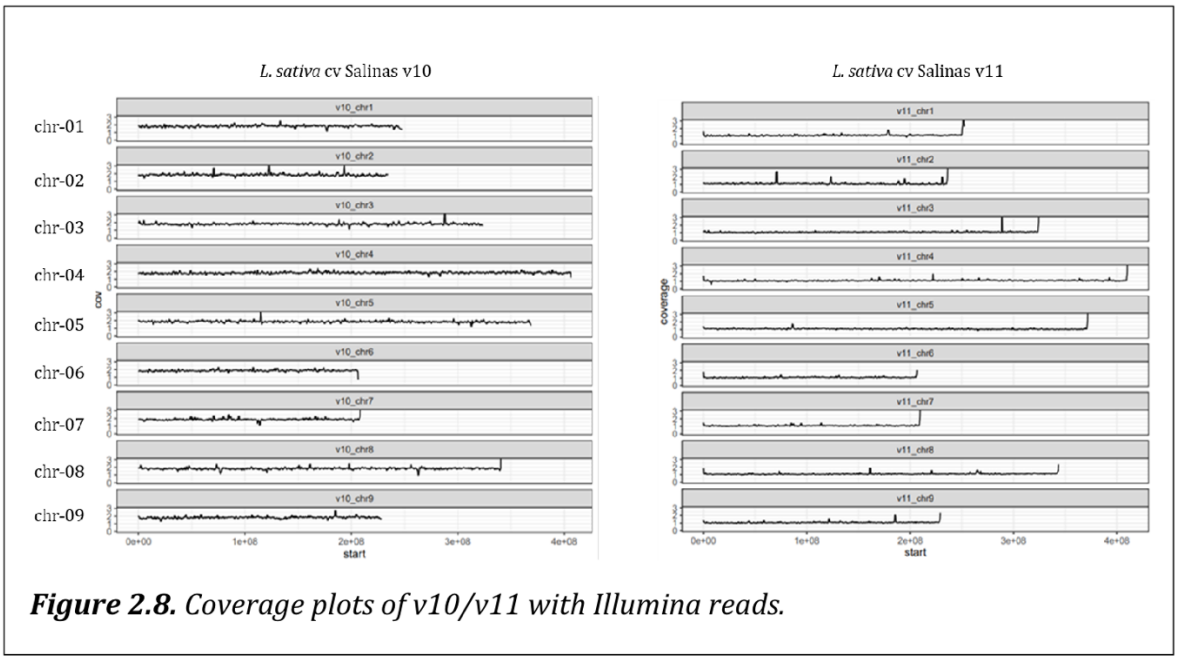

Figure 2.8. *Coverage plots of v10/v11 with Illumina reads.*

To further evaluate the completeness and correctness of the reference assemblies, coverage analysis and gap analysis were performed on the v10 and v11 assemblies (Figure 2.8). Short read Illumina reads were mapped to the respective assemblies, and reads were normalized across the whole genome and plotted using ggplot2. The lettuce v11 assembly had a more even distribution of reads across the whole genome when compared to the v10 assembly (Figure 2.8), confirming fewer mis-assemblies in collapsed repeat regions and structural modifications. Gap regions across the genome were compared between the v10 and v11 assemblies. The v10 assembly had more (545) gaps compared to the v11 assembly (384).

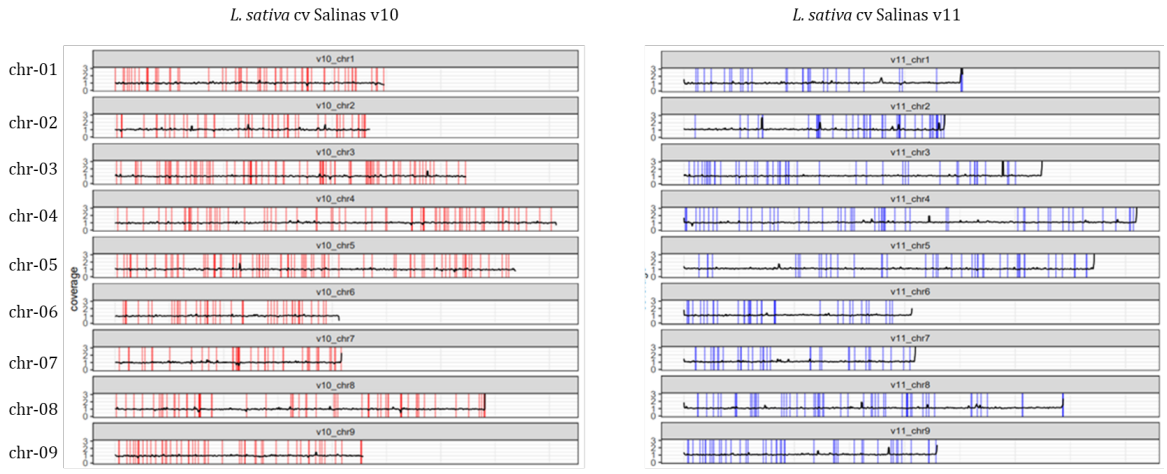

Figure 2.9. *Gap positions in v10 and v11 genome assemblies of L. sativa cv. Salinas.*

2.4.8 Centromeric and telomeric regions and rDNA clusters

Centromeres frequently contain tandem repeat arrays of short sequences with sizes up to one hundred kb (Talbert & Henikoff, 2020). To determine if assembly scaffolds encompassed centromeric areas, we first looked for tandem repeats with a high frequency, since such repeats are widely considered to be centromeric repeat candidates. In addition, the repeat and gene density along each scaffold were calculated to understand their typical distribution. This was determined to see whether they reflect the common architecture of plant chromosomes, which includes high repeat and low gene densities near the centromere and low repeat and high gene density at the ends of euchromatic regions (Lermontova et al., 2014) (Figure 2.10).

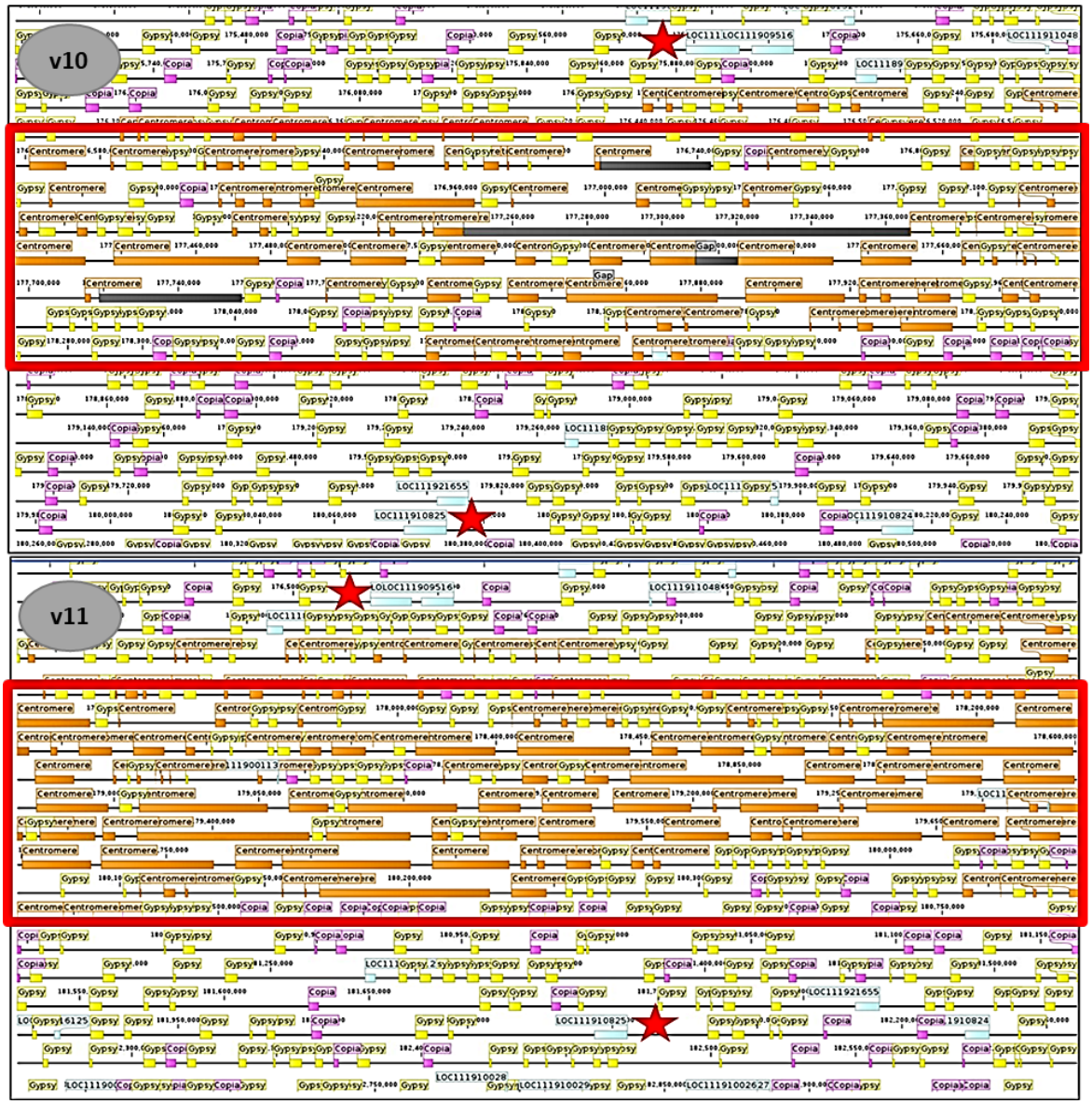

Figure 2.10: *Centromere comparison of v10 (ONT) vs. v11 (HiFi) genome assemblies.*

Shown is the sequence comparison of the centromeric repeat region in the v10 and v11 assemblies of Chromosome 1. The centromeric repeat regions are boxed in red. The dark grey regions show the gaps in the ONT-based v10 assembly when compared to the HiFi based v11 assembly that had no gaps and an additional 1.0 Mb resolved in the centromeric region. The purple and yellow regions represent high copy sequences. The low density of genes near the centromere is captured in both assemblies that are shown in blue.

High coverage regions were identified by mapping raw Illumina, ONT, and HiFi reads. Fourteen 12 kb segments were extracted that had high coverage in the genome. These were analyzed using LAST (Local Alignment Search Tool), followed by dot plots that revealed tandem repeat structures characteristic of centromeric regions in both the v11 and v10 reference genomes (Figure 2.11). Single putative centromeres were identified on all chromosomes except Chromosome 4. The centromere locations were further confirmed using Hi-C contact matrices (Figure 2.5). The distribution of centromeric repeat array regions across the genome is shown in Table 2.9.

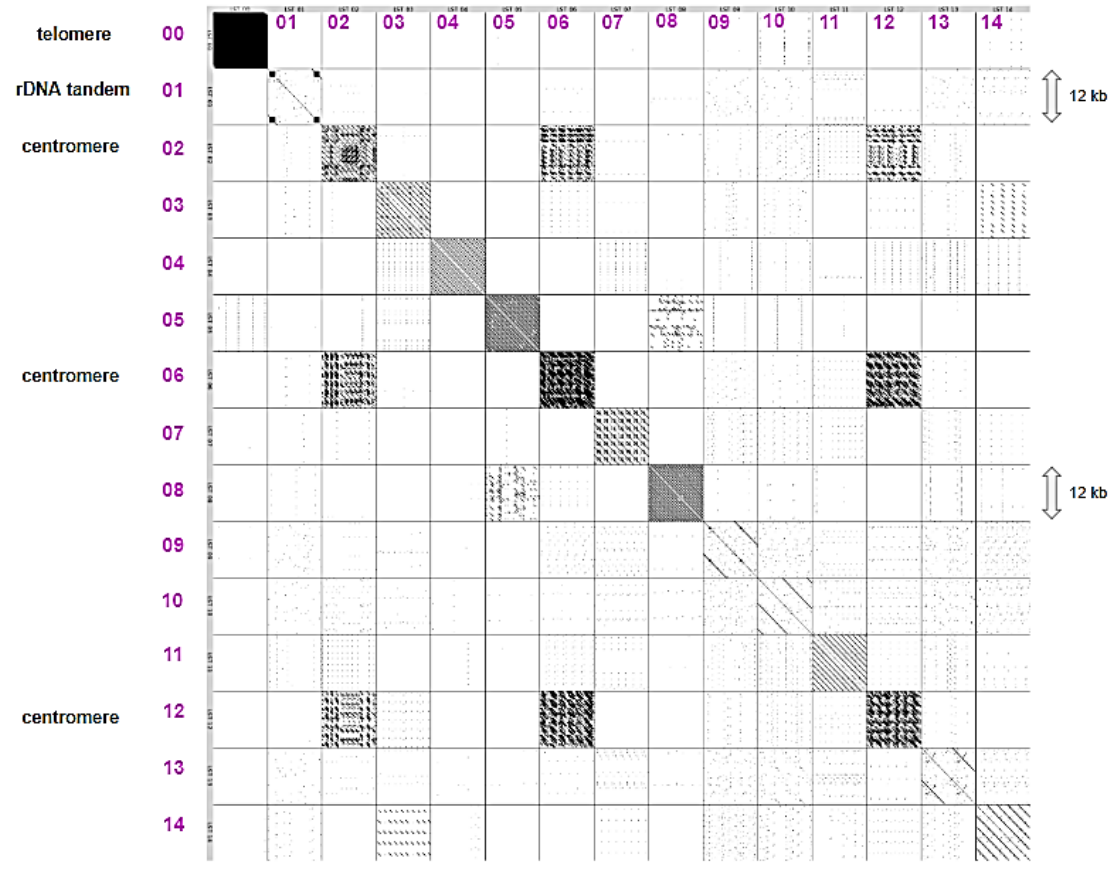

Figure 2.11. *Cross similarity analysis of selected tandem repeats using dot-plot.*

Table 2.9. *Location and size of centromeric repeat arrays in the v10 and v11*

	Nanopore based			
	Lactuca sativa cv. Salinas v10			
Chromosome	**Array start**	**Array end**	**Chromosome length**	**Centromere length**
Chr_01	176,155,649	178,445,199	247,928,466	2,289,550
Chr_02	90,474,772	92,010,621	234,982,731	1,535,849
Chr_03	222,987,471	223,218,671	323,094,788	231,200
Chr_04	273,470,743	273,507,277	406,360,200	36,534
Chr_05	115,005,078	117,251,740	368,878,906	2,246,662
Chr_06	107,793,867	108,847,473	206,250,288	1,053,606
Chr_07	111,888,392	114,795,254	208,571,662	2,906,862
Chr_08	262,271,931	264,796,986	340,764,772	2,525,055
Chr_09	132,045,287	134,675,397	228,713,657	2,630,110
	PacBio HiFi based			
	Lactuca sativa cv. Salinas v11			
Chromosome	**Array start**	**Array end**	**Chromosome length**	**Centromere length**
Chr_01	177,187,095	180,563,166	250,695,514	3,376,071
Chr_02	90,929,850	92,420,952	236,227,250	1,491,102
Chr_03	223,429,447	223,660,649	324,070,766	231,202

Chr_04	274,634,994	274,784,476	408,033,323	149,482
Chr_05	115,491,489	119,873,453	372,060,267	4,381,964
Chr_06	108,019,522	109,074,895	206,580,037	1,055,373
Chr_07	112,252,996	115,239,959	209,360,802	2,986,963
Chr_08	264,041,250	267,142,186	343,347,203	3,100,936
Chr_09	132,202,655	134,898,556	229,398,248	2,695,901

The HiFi based v11 assembly had more centromeric repeats captured compared to the ONT based v10 assembly. For example, in Chromosome 1, the v11 assembly resolved 3,376,071 bp of centromeric region, whereas the v10 assembly spanned a length of 2,289,550 bp only with gaps between the repeats. Similarly for Chromosomes 5 and 8, v11 resolved 4,381,964 bp and 3,100,936 bp, whereas in v10, only 2,246,662 bp and 2,525,055 bp repeat structure was captured, respectively. This was also noted in other chromosomes where the HiFi based assembly resolved centromere like repeat regions much better than the ONT based assembly (Figure 2.12).

To identify telomeres, the plant telomeric motif sequences (CCCTAAA) were found using BLASTn (Camacho et al., 2009). All but two of the expected 18 telomeres were identified at each end of the nine chromosomes in the v11 assembly (Table 2.10; Figure. 2.12). Sixteen telomeric repeat arrays were identified in the v10 assembly, but they were less complete.

Table 2.10. *Location of telomeric repeat arrays in lettuce v10 and v11 genome assemblies.*

	Nanopore based			
	Lactuca sativa cv. Salinas v10			
Chromosome	**Array start**	**Array end**	**Telomere length (bp)**	**Chromosome length (bp)**
Chr_01	17,741	18,200	460	247,928,466
Chr_02	29,451	29,922	472	234,982,731
	30,210	30,646	437	
	234,964,066	234,964,380	315	
Chr_03	213,903,690	213,904,435	745	323,094,788
	213,980,113	213,980,707	594	
Chr_04	7,477	8,698	1,222	406,360,200
	221,177,978	221,179,459	1,482	
	368,822,662	368,823,109	447	
	406,321,586	406,321,968	382	
	406,329,616	406,329,998	382	
Chr_05	368,822,662	368,823,119	457	368,878,906
	368,840,336	368,840,682	346	
Chr_06	8,816	9,475	660	206,250,288
	14,871	15,667	797	
Chr_07	208,562,426	208,563,225	800	208,571,662
Chr_08	17,907	18,272	365	340,764,772
Chr_09	10,101	10,756	656	228,713,657
	228,696,718	228,697,145	428	

	228,705,927	228,706,276	350	
	PacBio HiFi based			
	Lactuca sativa cv. Salinas v11			
Chromosome	**Array start**	**Array end**	**Telomere length (bp)**	**Chromosome length (bp)**
Chr_01	1	2,856	2,855	256,831,252
	11,470	13,342	1,872	
	251,449,925	251,456,799	6,874	
	252,184,824	252,198,739	13,915	
Chr_02	1	8,736	8,735	240,317,896
	236,356,641	236,356,944	303	
	236,375,645	236,378,258	2,613	
Chr_03	2	12,005	12,003	330,069,441
	324,646,841	324,658,466	11,625	
Chr_04	1	21,726	21,725	417,134,073
	31,118	32,333	1,215	
	410,279,523	410,295,809	16,286	
Chr_05	291	5,459	5,168	378,039,656
	371,824,962	371,842,284	17,322	
Chr_06	1	15,115	15,114	210,101,232
	206,649,944	206,656,949	7,005	
Chr_07	3	7,921	7,918	213,396,264
	209,875,080	209,875,874	794	
	209,884,601	209,897,964	13,363	
Chr_08	436	8,884	8,448	349,242,339
Chr_09	8	13,605	13,597	233,247,094

	23,823	24,492	669	
	229,402,348	229,423,371	21,023	

***Figure 2.12**. Location of telomeric repeat arrays in the 5' end of the chromosomes in the lettuce v10 and v11 genome assemblies.*

Shown in the red box is the 300 kb segment of the 5' end of all nine chromosomes across the v10 and v11 assemblies. The v11 assembly has more resolved telomere repeat arrays, shown in the black box, compared to the v10 assembly.

We also screened the v10 and v11 assemblies for rDNA clusters by sequence similarity to 18S. Two nucleolus organizer regions (NORs) were identified at one end of Chromosomes 1 and 8. In the v11 assembly, the NOR on Chromosome 1 and Chromosome 8 constituted approximately 10.4 Mb of 18S and 5S rDNA repeats. The whole sequence of

either NOR was not fully assembled because the majority of rDNA clusters were present in Chromosome 0; these were distributed over practically the entire sequence of the short scaffolds or located towards the end of the scaffolds.

Figure 2.13 shows the final architecture of the v11 genome assembly of *L. sativa* cv. Salinas. This T2T PacBio HiFi based assembly is the most comprehensive lettuce reference assembly with fewer gaps than prior lettuce assembly versions.

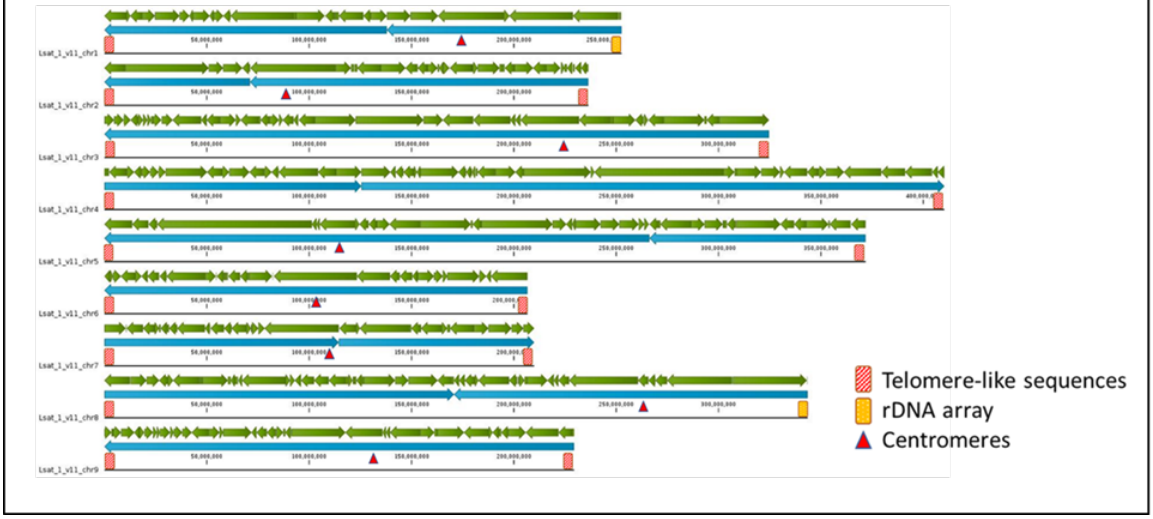

Figure 2.13. *Schematic of the v11 reference genome assembly of L. sativa cv. Salinas.*

The green boxes are 393 contigs totaling 2.58 Gb after assembly of the HiFi reads. The dark blue boxes are 15 scaffolds after Bionano analysis and genetic orientation into the nine near-complete chromosomes. The red triangles are repeated regions representing likely centromeres. The red dotted squares represent telomere-like sequences. The yellow dotted squares are rDNA repeat-like clusters.

2.4.9 Genome annotation

Much of the lettuce genome (81.13%) was identified as repetitive using RepeatModeler and RepeatMasker to search the v11 assembly for repetitive DNA sequences.

A total of 2.1 Gb was identified as transposable elements and unclassified repeats (Table 2.11). This is very similar to the previous v8 assembly (74.2%). Interspersed repeats were the most prevalent of repeats (79.83%), consisting of retrotransposons at 65.03% and DNA transposons at 0.27%.

Table 2.11. *Repeat content in the L. sativa v11 genome assembly.*

Types of repeats	**Number of elements**	**Bases (Mb)**	**Percentage of the assembly**
DNA transposons	14,259	7.3	0.27%
Retrotransposons			
LINE	5,738	3.7	0.14%
LTR: *Copia*	270,319	306.1	11.62%
LTR: *Gypsy*	341,556	314.1	11.93%
LTR: Others	3,384	2.7	0.10%
Other interspersed repeats	2,582,105	1,086.0	41.24%
RC : Helitron	3,455	1.0	0.04%
Simple sequence repeats	46,549	11.6	0.44%
Unclassified elements	1,396,452	405.2	15.38%
Total	4,663,817	2137.6	81.16%

Long-terminal retrotransposons (LTRs) were the most prevalent repeat element. The most prevalent classes of LTR were *Gypsy* and *Copia*. Even though broadly dispersed across the genome, they were distributed differentially throughout the nine chromosomes (Figure

2.16). The *Gypsy* LTRs were more abundant in proximity to the centromeres, while the *Copia* LTRs were less frequent near the centromeres. Long-interspersed nuclear elements (L1/LINE), which include a poly(A) tail and two OFRs for autonomous retro-transposition, comprised 0.14% of the genome. Using this custom repeat database created by RepeatModeler and careful manual curation for resistance locus, about 81% of the lettuce genome was soft masked for annotation.

The v11 assembly was then annotated *de novo* using the MAKER (Campbell et al., 2014) workflow with additional information compared to v8. To increase the accuracy of gene prediction in *Lactuca sativa*, PacBio Iso-Seq full-length transcripts of up to 10–12 kb were included, which allowed us to accurately define the exon–intron structure of the predicted genes. For Salinas, of the 1,093,806 full-length non-chimeric transcripts, 73,556 were classified as high quality full-length consensus transcript sequences, with an average length of 1,357 bp for annotation. For homology-based gene prediction, proteins from 17 closely related Compositae species were used. For *ab initio* prediction, two training sets from AUGUSTUS and SNAP were utilized for the prediction. For high-confidence genes, sequences having an annotation edit distance score of >0.5 were chosen. Two rounds of MAKER initially annotated 50,668 protein-coding genes in the v11 assembly. The relationship between the annotations of v8 and v11 are described in Chapter 3.

2.4.10 Refinement of genome annotation

The MAKER-P analysis resulted in a set of protein-coding sequences (CDSs) putatively consisting of 50,668 gene loci and 44,254 complete gene models (Table 2.12). Several gene features were explored to validate the MAKER gene model predictions, such as

completeness (genes with canonical start and stop codons), transcriptome support, protein evidence (similarity to plant proteins especially Compositae proteins), and recognizable protein domains (PfamA and/or the Panther database). Based on these criteria, the final gene models resulted in 44,241 protein coding genes that were complete with a start and stop codon and had at evidence from at least one or more proteins or transcripts. The number of annotated protein-coding genes increased from the previous v8 annotation by approximately 35,841 to 44,241 gene loci in v11 assembly. Therefore, the complete number of annotated protein-coding genes improved by approximately 8,400 genes from the v8 reference assembly.

Table 2.12. *Annotation statistics of the Lactuca sativa cv. Salinas (v11) assembly.*

	Lactuca sativa cv. Salinas_v11
Number of predicted gene models	44,254
Total gene length (Mb)	107,580,123
Average gene size (bp)	2,431
Number of exons	230,974
Average number of exons / genes	5
Total exon length (Mb)	60,385,231
Average exon length (bp)	261
Number of introns	186,720
Average number of introns / genes	4
Total intron length (Mb)	47,568,332
Average intron length (bp)	255

The resulting genes incorporated a total of 230,974 exons and 186,720 introns. The total exon length was 60,385,231 Mbp, and the total intron length was 47,568,332 Mbp. On average, genes and transcripts spanned 107,580,123 Mbp with the longest gene being up to 59,205 kbp. BUSCO was used for evaluation of annotation completeness. BUSCO analysis using 2121 conserved genes on 44,254 protein coding genes in the annotated v11 reference genome, among which 1991(93.9%) were complete and 108 (5.1%) duplicated BUSCO genes.

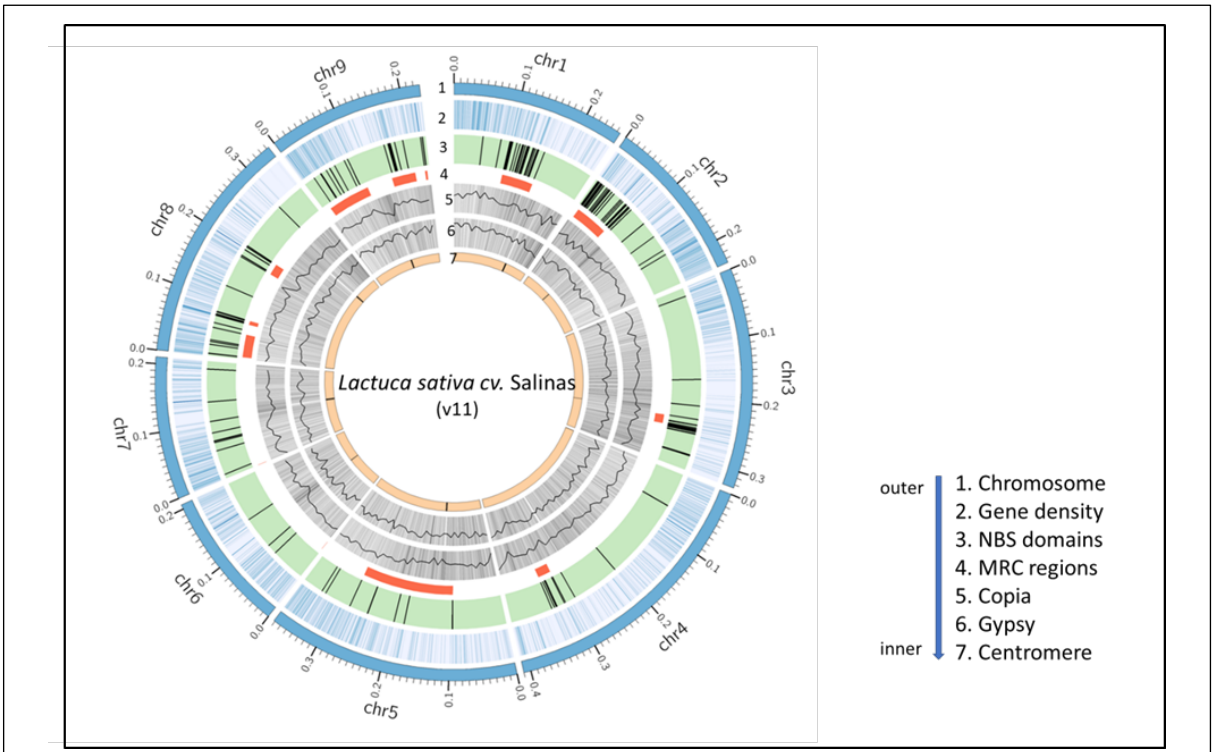

Figure 2.14 *Circular genomic visualization of the v11 reference genome assembly and annotation.*

1. The outer circle in blue has the nine chromosomes of lettuce. 2. The second track in blue is the gene density and is scaled to 1 Mb bins. 3. The location of canonical NBS-LRR genes across the genome. 4. Regions that are phenotypically classified as MRC. Tracks 5, 6, and 7 show the distribution of *Copia* and *Gypsy* elements and the centromeric regions, respectively.

2.4.11 Functional annotation

In order to describe protein functions and enhance annotation, homology information has been used to assign functions to genes and proteins. There are 110,616 protein sequences from species in the Uniprot/Swiss-Prot database. In addition, predicted functional domains (InterproScan) were presented as supplementary evidence for current gene annotations. This included a combination between databases (Dbxref) and gene ontology (GO) keywords.

In the functional gene annotation, 95.6% of the protein-coding genes of lettuce showed significant homology to entries of known protein databases. Using a conserved protein domain search, 11,222 (25.3%) of the protein-coding genes showed significant hits. With these search results, 42,294 genes were assigned to at least one term in the Gene Ontology.

2.5 Discussion

In this chapter, I describe the generation of two new, highly contiguous, T2T reference assemblies of lettuce using either Oxford Nanopore or PacBio HiFi sequencing technologies. The assemblies produced using these technologies were compared, and the v11 PacBio-based assembly when combined with long-range Bionano optical mapping and Hi-C chromosome conformation capture technologies were selected as the highest quality based on assembly completeness, accuracy, and contiguity. This compares well with few T2T currently available for other plant species (Table 2.13).

Table 2.13. *Summary of T2T assemblies across plant genomes.*

	Sequencing technology	T2T assemblies	Assembly size (Mb)	# of Contigs	Contig N50 (Mb)	Reference
Lactuca sativa cv. Salinas_v11 (lettuce)	PacBio	Yes with few gaps	2,588	484	12.52	N/A
Arabidopsis thaliana	PacBio / ONT (merged)	Yes with few gaps	133	NA	26.1	(Hou et al., 2022)
Oryza sativa (rice)	PacBio / ONT (merged)	Yes with no gaps	397	12	N/A	(Zhang et al., 2022)
Citrullus lanatus (watermelon)	PacBio / ONT (merged)	Yes with no gaps	369.3	11	32.5	(Deng et al., 2022)
Zea mays (maize)	PacBio / ONT (merged)	Yes with few gaps	2,365	63	162	(Liu et al., 2020)
Musa acuminata (banana)	ONT	Yes with no gaps	484	124	32	(Belser et al., 2021)

The v11 assembly is a big improvement compared to the previous publicly available v8 reference genome. The v11 assembly has 400-fold increase in N_{50} from 28 kb to 12.5 Mb. The assembly size also increased from 2,391,578,241 to 2,588,783,166 bp. The v11 assembly is less fragmented than the v8 assembly with a reduced number of contigs from 168,554 to 484. BUSCO assessment of the v11 assembly had a slight improvement to 98.5% completeness when compared to the v8 assembly. Long read sequencing provided a

powerful approach to improve the assembly of repetitive regions of the genome, especially in centromeric and telomeric repeat regions of the genome (Figure 2.16); gapless centromeric repeat arrays ~4 Mb are assembled in v11 assembly. With only 91 unplaced contigs and fewer gaps (384) in the assembly, the v11 lettuce reference genome has captured telomeres in all nine chromosomes. This HiFi based v11 assembly will be adopted as the current reference genome for lettuce research.

These assemblies were enabled by recent long-read sequencing techniques along with long-range technologies. ONT and PacBio have advantages and disadvantages. The average read length for PacBio HiFi sequencing is 15 kb, while the average read length for Nanopore sequencing is 30 kb. The PacBio read length is dictated by the need to have multiple reads of the circular DNA template. The ONT read length is limited by the input material, which has resulted in the need for better DNA extraction protocols that preserve the HMW molecules. PacBio HiFi was inherently more accurate (>99 percent single-molecule read accuracy) than the ONT reads due to the CCS nature of data generation. However, improvements in sequencing chemistry and base-calling algorithms were continually being made during the project. For the v10 assembly build, several base-calling algorithms (flappie v2.0.0, guppy v2.1 guppy 3.x) were used; base calling accuracy increased dramatically with the progression of assembly algorithms. However, rebase-calling the same reads with improved algorithms, increased the time, and compute requirements and resulted in redundant effort. The current genome assembly projects are based on reads from a single sequencing platform, either PacBio or ONT platform; however, driven by efforts to sequence human genomes, algorithms are being developed that utilize both types of reads to provide

gapless assemblies (Miga et al., 2020). These will be applied to plant genomes, including lettuce, to routinely provide gapless genomes in the future.

For both the v10 and v11 draft assemblies, the separate integrations of optical mapping and chromatin conformation capture data yielded comparable enhancements. These two technologies did not supply duplicate scaffolding information, but rather they addressed different scaffolding challenges and their combination enhanced assembly scaffolding. Due to the poor alignments of Illumina short reads, Dovetail Hi-C scaffolding did not aid in the resolution of tandem repeat regions. In contrast, optical mapping accurately resolved these regions. Optical maps were challenged by regions with closely linked restriction sites; however, Dovetail Hi-C data were not impacted by such regions. Consequently, we scaffolded first using optical mapping and then Dovetail Hi-C data.

Genome annotation was improved using additional RNAseq and high quality PacBio Iso-Seq transcriptomic data. The full-length Iso-Seq data revealed predicted genes with a higher mean length than the gene annotations in v8. Of the 50,668 putative genes in the v11 assembly, 44,231 were predicted to be protein coding genes with complete gene models. Only 36 annotated genes were in Chromosome 0 (unscaffolded contigs) and the rest of the genes were contained in the nine chromosomes of the v11 assembly. Gene clustering with closely related plants assigned 42,294 genes to orthogroups and only 1,937 genes (4%) were unassigned to orthogroups, which is comparable to other plant species with high quality annotations (Table 2.14).

***Table 2.14**. Summary of orthogroup clustering statistics for plant genomes.*

	Lactuca sativa cv. Salinas (v11) (lettuce)	*Arabidopsis thaliana*	*Glycine max* (soybean)	*Solanum lycopersicum* (tomato)
GenBank ID	GenBank GCA_002870075.4	GCF_000001735.4	GCF_000004515.5	GCF_000188115.4
Number of genes	44,231	48,265	71,219	37,658
Number of genes in orthogroups	42,294	46,117	68,488	36,297
Number of unassigned genes	1,937	2,148	2,731	1,361
% of genes in orthogroups	95.6	95.5	96.2	96.4
% of unassigned genes	4.4	4.5	3.8	3.6
Number of orthogroups containing species	19,473	18,410	16,774	14,855
% of orthogroups containing species	48.1	45.5	41.4	36.7
Number of species-specific orthogroups	539	1,263	2,831	995
Number of genes in species-specific orthogroups	2,259	4,458	13,837	4,537
% of genes in species-specific orthogroups	5.1	9.2	19.4	12

	Brassica napus (rapeseed)	*Helianthus annuus* (sunflower)	*Lactuca sativa* var. *angustana* (stem lettuce)	
GenBank ID	GCF_000686985.2	GCF_002127325.1	Stem_lettuce_v1.0	
Number of genes	123,465	73,839	40,341	
Number of genes in orthogroups	117,890	71,067	36,854	
Number of unassigned genes	5,575	2,772	3,487	
% of genes in orthogroups	95.5	96.2	91.4	
% of unassigned genes	4.5	3.8	8.6	
Number of orthogroups containing species	22,034	18,055	18,596	
% of orthogroups containing species	54.4	44.6	46	
Number of species-specific orthogroups	4,503	2,279	1,225	
Number of genes in species-specific orthogroups	24,305	21,713	5,264	
% of genes in species-specific orthogroups	19.7	29.4	13	

In conclusion, new third-generation genomic technologies have enabled the generation of a high-quality reference genome assembly. Integrating multiple long-read sequencing and long-range scaffolding technologies promises chromosome-scale and near

complete assemblies for large and repetitive genomes like lettuce. This provides the foundation for generating multiple genome assemblies within the same species, which will allow research to overcome potential biases due to use of a single reference assembly. In the subsequent chapters, multiple chromosome scale assemblies for lettuce are *de novo* assembled and annotated to understand the complex genomic regions of the lettuce genome, such as the major resistance cluster regions and facilitate comparative genomics to understand genomic diversity in lettuce.

References:

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

Baranyi, M., & Greilhuber, J. (1996). Flow cytometric and Feulgen densitometric analysis of genome size variation in Pisum. *Theor Appl Genet*, *92*, 297–307.

Belser, C., Baurens, F.-C., Noel, B., Martin, G., Cruaud, C., Istace, B., Yahiaoui, N., Labadie, K., Hřibová, E., Doležel, J., Lemainque, A., Wincker, P., D'hont, A., & Aury, J.-M. (2022). *Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing*. https://doi.org/10.1038/s42003-021-02559-3

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/NAR/27.2.573

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology, 33*, 623-630. https://doi.org/10.1038/nbt.3238

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformations, 10*(421). https://doi.org/10.1186/1471-2105-10-421

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Curr. Protoc. Bioinform*. https://doi.org/10.1002/0471250953.bi0411s48

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., & Fernández-Pozo, N. (2012). Why Assembling Plant Genome Sequences Is So Challenging. *Biology 2012, 1(2), 439-459*. https://doi.org/10.3390/BIOLOGY1020439

Delcher, A. L., Salzberg, S. L., & Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics*, *10*(1), 10.3.1-10.3.18. https://doi.org/10.1002/0471250953.bi1003s00

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y., Li, B., Xu, Y., Deng, X. W., He, H., & Zhang, X. (2022). A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Molecular Plant*. https://doi.org/10.1016/J.MOLP.2022.06.010

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology*, *15*(8). https://doi.org/10.1371/journal.pcbi.1007273

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., & McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, *25*(11), 1750–1756. https://doi.org/10.1101/GR.191395.115

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Janicki, M., Rooke, R., Yang, G., Gregory, R., Bainard M Janicki, J. D., Rooke, R., Yang, G., & Janicki, M. (2011). Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Res*, *19*, 787–808. https://doi.org/10.1007/s10577-011-9230-7

Kapustová, V., Tulpová, Z., Toegelová, H., Novák, P., Macas, J., Karafiátová, M., Hřibová, E., Doležel, J., & Šimková, H. (2019). The Dark Matter of Large Cereal Genomes: Long Tandem Repeats. *International Journal of Molecular Sciences*, *20*(10). https://doi.org/10.3390/IJMS20102483

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/GR.215087.116

Lermontova, I., Sandmann, M., & Demidov, D. (2014). Centromeres and kinetochores of Brassicaceae. *Chromosome Research*, *22*(2), 135–152. https://doi.org/10.1007/s10577-014-9422-z

Liu, J., Seetharam, A. S., Chougule, K., Ou, S., Swentowsky, K. W., Gent, J. I., Llaca, V., Woodhouse, M. R., Manchanda, N., Presting, G. G., Kudrna, D. A., Alabady, M., Hirsch, C. N., Fengler, K. A., Ware, D., Michael, T. P., Hufford, M. B., & Dawe, R. K. (2020). Gapless assembly of maize chromosomes using long-read technologies. *Genome Biology*, *21*(1), 1–17. https://doi.org/10.1186/S13059-020-02029-9/TABLES/1

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1). https://doi.org/10.1371/JOURNAL.PCBI.1005944

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/S41586-020-2547-7

Miiro, G., Serwanga, J., Pozniak, A., McPhee, D., Manigart, O., Mwananyanda, L., Karita, E., Inwoley, A., Jaoko, W., DeHovitz, J., Bekker, L. G., Pitisuttithum, P., Paris, R., Allen, S., Lieberman-Aiden, E., van Berkum, N. L., Williams, L., et al. (2009). *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome*. https://doi.org/10.1126/science.1178746

Navr Atilov A, P., Toegelov, H., Tulpov, Z., Kuo, Y.-T., Stein, N., Dole Zel, J., Houben, A., Simkov, H., & Mascher, M. (2022). Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the MorexV3 reference genome. *Plant Biotechnology Journal.* https://doi.org/10.1111/pbi.13816

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., & Koren, S. (2020). HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, *30*(9), 1291–1305. https://doi.org/10.1101/GR.263566.120/-/DC1

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O., Truco, M. J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14953

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/J.GPB.2015.08.002

Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(SUPPL. 1), 8641–8648. https://doi.org/10.1073/PNAS.0700643104

Salzberg, S. L., & Yorke, J. A. (2005). *Beware of mis-assembled genomes*. *21*(24), 4320–4321. https://doi.org/10.1093/bioinformatics/bti769

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., et al. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*. https://doi.org/10.1038/s41587-020-0503-6

Shelton, J. M., Coleman, M. C., Herndon, N., Lu, N., Lam, E. T., Anantharaman, T., Sheth, P., & Brown, S. J. (2015). Tools and pipelines for BioNano data: Molecule assembly

pipeline and FASTA super scaffolding tool. *BMC Genomics*, *16*(1), 1–16. https://doi.org/10.1186/S12864-015-1911-8/FIGURES/9

Talbert, P. B., & Henikoff, S. (2020). What makes a centromere? *Experimental Cell Research*, *389*(2). https://doi.org/10.1016/J.YEXCR.2020.111895

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics, 33*(14), 2202–2204. https://doi.org/10.1093/BIOINFORMATICS/BTX153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., Manke, T., Backofen, R., Ramírez, F., Bj¨, B., Grüning, B. A., & Grüning, G. (2018). Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Research*, *46*, 11–16. https://doi.org/10.1093/nar/gky504

Yu, C., Yan, C., Liu, Y., Liu, Y., Jia, Y., Lavelle, D., An, G., Zhang, W., Zhang, L., Han, R., Larkin, R. M., Chen, J., Michelmore, R. W., & Kuang, H. (2020). Upregulation of a KN1 homolog by transposon insertion promotes leafy head development in lettuce. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(52), 33668–33678. https://doi.org/10.1073/PNAS.2019698117/-/DCSUPPLEMENTAL

Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2017). Improvements in Genomic Technologies: Application to Crop Genomics. *Trends in Biotechnology*, *35*(6), 547–558. https://doi.org/10.1016/J.TIBTECH.2017.02.009

Chapter 3: *De novo* assembly, annotation, and comparative analysis of seven chromosome-scale assemblies of wild and domesticated lettuce

Contributions:

I performed majority of the work described in this chapter, including the *de novo* assembly and annotation of five domesticated (*L. sativa*, cv. La Brillante, cv. Ninja, VIAE, and PI251246) and two wild (*L. serriola* US96UC23 and Armenian 999) genotypes of lettuce using Oxford Nanopore and PacBio HiFi data. I completed the Bionano scaffolding for the two wild accessions and reference guided scaffolding for the other four domesticated genotypes. I performed assembly and annotation evaluation and orthogroup clustering and classification for all seven assemblies.

Keri Cavanaugh prepared the plant material for sequencing by the UC Davis DNA Technologies core. Dean Lavelle and Kyle Fletcher provided computational support and base-calling of raw ONT reads. Mingchen Luo provided access to Bionano Compute On Demand for analysis and generated *de novo* BNG assembly for the two wild (*L. serriola* US96UC23 and Armenian 999) genotypes of lettuce.

3.1 Abstract

Domesticated lettuce (*Lactuca sativa* L.) is one of the most popular leafy vegetables in the United States. Genetic diversity is an important resource in crop breeding to improve cultivars with desirable traits. High-quality genomes provide insight into gene content, genomic diversity, and the genetic basis of phenotypic traits. Phylogenetic and population genetic analyses have indicated substantial genetic divergence among the five horticultural types of lettuce: butterhead, crisphead, leaf, romaine, and stem. To capture the genomic diversity in lettuce, it is desirable to generate high-quality assemblies of diverse accessions in the lettuce gene pool. In this chapter, I describe individual *de novo* genome assemblies and annotations for an additional six diverse lettuce genotypes to complement the v11 reference genome assembly. Both short (Illumina) and long-read (Oxford Nanopore and PacBio HiFi) sequencing technologies were used to construct and annotate chromosome-scale assemblies of five domesticated (*L. sativa*, La Brillante, Ninja, VIAE, and PI251246) and two wild (*L. serriola* US96UC23 and Armenian 999) genotypes of lettuce. Investigation of orthologs in these assemblies revealed 37,223 orthologous gene families, of which 18,042 were highly conserved or core gene families and 19,181 were dispensable/variable gene families. The differences in the orthologous gene content between these seven assemblies are the foundation for study of the lettuce pangenome. The presence or absence of gene content variations can be used in marker-assisted selection, which can be used to breed domesticated cultivars that have desirable traits from wild species; gene content variations are also the basis for targets of gene editing applications.

3.2 Introduction

Lettuce (*Lactuca sativa* L.) is the most valuable, widely consumed, fresh leaf crop in the U.S.; it has an annual farm gate value of ~$3.4 billion (Agricultural Statistics Service, 2020). It is a diploid (2n = 18) species within the Compositae (Asteraceae) family. There are four well-established species within subsection *Lactuca*, domesticated *L. sativa* and three wild species, *L. serriola*, *L. saligna*, and *L. virosa* (Michelmore et al., 1994). Wild species, particularly *L. serriola*, have been sources of disease resistance genes (Farrara et al., 1987), and they remain a rich potential source of variation that has not been accessed systematically (Kesseli et al., 1991; Lindqvist, 1960).

The release of the v8 reference genome of *Lactuca sativa* cv. Salinas has enabled several studies of lettuce genetics at a genome-wide scale (Atkinson et al., 2013; Christopoulou et al., 2015; Wei et al., 2021). With the v8 annotation and the prediction of 36,136 gene models, it was possible to analyze the gene space, providing the basis to understand important horticultural traits (Reyes-Chin-Wo et al., 2017b). The availability of a reference genome marked the beginning of a genomics phase in lettuce research, allowing whole-genome resequencing, development of high-density genotyping tools, and the genetic dissection of important agronomical traits in lettuce. Although genetically organized into nine chromosomal superscaffolds, the v8 assembly is fragmented into 11,454 scaffolds and 168,554 contigs, which compromises the accuracy of candidate gene prediction.

The recent development of long-read sequencing and long-range scaffolding methods has enabled chromosome-scale assembly for several plant species (Belser et al., 2021; Li et al., 2014; Su et al., 2021). Long-read technologies use single DNA molecules as templates

without using PCR amplification. The PacBio Sequel Il system can generate high-throughput HiFi reads using circular consensus sequencing (CCS) mode. These reads provide base-level accuracy of >99%, similar to Sanger sequencing (Hon et al., 2020). Oxford Nanopore Technologies (ONT) sequences by quantifying changes in electrical conductivity as DNA passes through a protein nanopore. The PromethION platform from ONT can yield >7 Tb of data per run and can generate reads as long as 2 Mb, although the read accuracy is not as high as PacBio HiFi. Long-read sequencing by either technology has an advantage over short-read sequencing in generating high quality genomes because they facilitate resolving complex repeat regions of the genomes. As described in Chapter 2, I used these sequencing technologies to generate highly contiguous, near complete telomere to telomere v10 and v11 assemblies for *L. sativa*. However, a single reference is not enough to capture the genetic diversity of lettuce. The availability of high-quality genomes with fully assembled chromosomes is required to provide the foundation for understanding domestication and evolution as well as the mechanisms governing important traits (e.g., flowering time, disease resistance).

In this chapter, I describe the assembly of an additional four domesticated and two wild lettuce accessions of lettuce to complement the v11 assembly. I present an optimized workflow to construct chromosome scale assembly (Vaser et al., 2017) using Oxford Nanopore or PacBio HiFi sequencing data. Following this, *de novo* annotation and comparisons of *L. sativa*, La Brillante, Ninja, VIAE, and PI251246 and *L. serriola* US96UC23 and Armenian 999 were carried out, resulting in approximately 94,187 and 2,884 protein coding genes, respectively, in the core and accessory genomes.

3.3 Materials and Methods

3.3.1 Plant material and DNA isolation

To provide DNA for *de novo* sequencing of four domesticated and two wild genotypes of lettuce, the seedlings were grown in the dark for seven days. High molecular weight (HMW) DNA was extracted from *L. sativa* cvs. Ninja, VIAE, PI251246, and *L. serriola* accessions US96UC23 and Armenian 999 for ONT sequencing and from *L. sativa* cv. La Brillante for PacBio SMRT HiFi sequencing.

Table 3.1. *Six additional lettuce accessions for de novo assembly and annotation.*

Accession ID	**Accession Name**	**Seed Source**
16G313-3	*L. sativa* cv. Salinas	Univ. of California, Davis
GBS-543	*L. sativa* cv. La Brillante	Univ. of California, Davis
17G712-1	*L. sativa* cv. Ninja	Univ. of California, Davis
17G853-1	*L. sativa* VIAE	Univ. of California, Davis
12G504	*L. sativa* PI251246	Univ. of California, Davis
16G692-1	*L. serriola* US96UC23	Univ. of California, Davis
12G239-2	*L. serriola* Armenian	Univ. of California, Davis

DNA was extracted by the UC Davis Genome Center DNA Technologies Core (https://dnatech.genomecenter.ucdavis.edu/) from cotyledons of each cultivar using a

modified method incorporating a sorbitol pre-wash combined with a high salt CTAB extraction as described in detail under "Method variations" by Ingles *et al*. (2018). Modifications were made to this protocol by substituting sodium metabisulfite (1% W/V) for beta-mercapto-ethanol in both the sorbitol pre-wash and lysis extraction buffers and lowering the lysis temperature from 65°C to 50°C. The integrity of the DNA samples was evaluated using a Femto Pulse (Agilent Technologies, Inc., Santa Clara, CA). Quantification and purity were assayed using a Qubit and Nanodrop, respectively (Thermo Fisher Scientific, Waltham, MA).

3.3.2 ONT PromethION library preparation and sequencing

The ONT library construction and sequencing were performed by the UC Davis Genome Center DNA Technologies Core. HMW DNA of Ninja, VIAE, PI251246, US96UC23, and Armenian was used as input for library preparation. ONT libraries were prepared using 1 µg of purified genomic DNA as input into the Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, UK), according to manufacturer recommendations.

The two or three R9.4.1 flow cells were sequenced for each genotype on an ONT PromethION instrument. Each flow cell received a nuclease flush every 20 to 24 hours. This flush removed long DNA fragments that could cause the pores to become blocked over time. Each flow cell received a fresh aliquot of the same library after the nuclease flush. In this way, a total of two outputs were obtained per flow cell: the initial data and the post nuclease treatment data. The raw fast5 sequencing data was base-called with Guppy v4.5 or Guppy v5.1 (ONT).

3.3.3 PacBio HiFi library preparation and sequencing

The HiFi DNA library construction and HiFi DNA sequencing were performed by the UC Davis Genome Center DNA Technologies Core (https://dnatech.genomecenter.ucdavis.edu/). La Brillante HMW DNA was sheared using Megaruptor® 3 (Diagenode Inc., Denville, NJ) to 15–18 kb for generation of three PCR-free PacBio HiFi Libraries. Libraries were constructed using a SMRT bell Template Prep Kit 1.0 (Pacific Biosciences). Long-read sequencing was performed using the Circular Consensus Sequence (CCS) mode on a PacBio Sequel II instrument (Pacific Biosciences of California, Inc., Menlo Park, CA) using a 30-hour movie time with 2-hour pre-extension. The resulting raw data was processed using either the CCS3.4 or CCS4 pipeline (GitHub, https://github.com/PacificBiosciences/ccs) to generate HiFi reads.

3.3.4 Bionano sequencing for *L. serriola* acc. US96UC23 and acc. Armenian 999

According to Bionano Prep Plant Tissue DNA Isolation Protocol, high-molecular-weight DNA was extracted as explained earlier. Then, endonuclease DLE1 was used for digestion. The labeling and staining processes were implemented according to the Bionano Prep Direct Label and Stain Protocol. Bionano Saphyr chip (Bionano Genomics) was used for sequencing at the UC Davis Genome Center DNA Technologies Core.

3.3.5 Genome assembly for *L. sativa* cv. Ninja, cv. VIAE, PI251246, *L. serriola* acc. US96UC23, and acc. Armenian

All Nanopore reads were base called using Guppy v5.0.7. Before draft assembly construction, Porechop v0.2.3 (https://github.com/rrwick/Porechop) was used to remove residual ONT adapters and NanoFilt v2.7.1 (https://github.com/wdecoster/nanofilt) was

used to select reads with an average quality score >Q10. The trimmed reads were used for draft assembly construction using Shasta v0.5.0 (Shafin et al., 2020). Two rounds of iterative polishing were performed to improve the accuracy of the assembly. The first round of polishing was done with Oxford Nanopore reads as input to the Pepper v0.01 software (https://github.com/kishwarshafin/pepper). The resulting self-corrected consensus assembly was polished again using Illumina reads as input to Pilon v1.23 (Walker et al., 2014). Both Pepper v0.01 and Pilon v1.23 were used with default parameters and the consensus accuracy increased after each round. The resulting ONT based draft assemblies of Ninja, VIAE, and PI251246 was scaffolded with Ragtag v.2.1.0 (https://github.com/malonge/RagTag) (Alonge et al., 2019) using the v11 reference assembly. Genome scaffolding of US96UC23 and Armenian 999 was carried out using Bionano optical mapping data followed by scaffolding with v11 reference guided placement of scaffolds using Ragtag v.2.1.0. A detailed scaffolding protocol using Bionano optical mapping data is described in the Materials and Methods section of Chapter 2.

3.3.6 Genome assembly for *L. sativa* cv. La Brillante

The highly accurate >q20 HiFi reads were generated with the Circular Consensus Sequencing (CCS) tool from PacBio ccs v6.0. The La Brillante genome assembly was constructed using the PacBio CCS reads in the Hifiasm v0.16.1-r375 (Cheng et al., 2021) assembler using default settings. The resulting draft assembly was evaluated for contiguity, correctness, and completeness. Due to the low error rate of HiFi Reads, no further polishing was done on this assembly.

3.3.7 Genome annotation

First, repetitive elements were annotated using Tandem Repeat Finder v.4.09 (Benson, 1999) before gene model prediction. LTR_FINDER v1.07 was used to build an LTR-retrotransposon library and RepeatModeler v.1.0.10 was used to build a *de novo* repetitive element library. The above libraries and Repbase (Bao et al., 2015) were used by RepeatMasker to annotate repetitive elements.

After repetitive sequences were masked, annotation of putative protein-coding genes was performed utilizing *ab initio*, homology, and Iso-seq-based methods. Augustus v3.3 (Stanke & Morgenstern, 2005) and SNAP v2013-11-29 (Korf, 2004)were used for *ab initio* gene prediction. For homology-based annotation, protein sequences from 17 RefSeq species, *Glycine max* (soybean), *Arabidopsis thaliana*, *Ricinus communis* (castor), *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Prunus persica* (peach), *Brassica napus* (rape), *Nicotiana tabacum* (tobacco), *Solanum pennellii* (wild tomato), *Cynara cardunculus* (cardoon), *Daucus carota* (carrot), *Helianthus annuus* (sunflower), *L. sativa* (lettuce), *Ipomoea triloba* (morning glory), *Pistacia vera* (pistachio), and *Cannabis sativa* (hemp) were obtained from NCBI and aligned to each of the six genomes using TBLASTN. Exonerate v2.2.0 was used to build gene structures based on the BLAST results. For Iso-seq-based gene prediction, reads were mapped to each assembly as part of the PacBio Isoseq workflow to filter for high quality gene models. Lastly, a consensus gene set was generated by integrating gene annotations from each method using MAKER-P (Campbell et al., 2014).

For the functional annotation, InterProScan v.5.51-85.0 (P. Jones et al., 2014) was run for the predicted protein sequences and InterPro ID, PFAM domains, and Gene Ontology (GO) terms. BLASTp with the Uniprot database was used to assign gene descriptors to each transcript based on the best BLAST hit.

3.3.8 BUSCO evaluation of genome completeness and annotations

BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão et al., 2015) evaluation was performed on the genome assemblies and gene annotations using BUSCO v.3.0.2 with the embryophyta_odb10 dataset.

3.3.9 Whole genome alignments and synteny analysis

Whole-genome comparisons between wild and domesticated lettuce genomes were performed using nucmer from the MUMmer v4.0 package and visualized with mummerplot from MUMmer (Marçais et al., 2018). Large structural variations (>50 kb) were identified based on synteny alignment blocks, using Assemblytics v1.2.1 (Nattestad & Schatz, 2016a).

3.3.10 Clustering of the predicted proteome data

The OrthoFinder v2.3.7 (Emms & Kelly, 2019) pipeline was used to cluster predicted proteome datasets across all seven domesticated and wild lettuce genotypes using default settings throughout. This clustering procedure determined which sequences shared similarities, grouped those sequences into phylogenetically related clusters (orthogroups), and left those sequences that did not share any similarities with any other protein sequences as independent sequences (singletons). With an expected value cut-off of $1e10^{-3}$, this pipeline employed a BLASTP search to derive the pair-wise sequence similarity score across each pair of proteome datasets. Then, Markov Clustering Algorithm (MCL) (Li et al., 2003) was applied to the BLAST results to generate protein clusters according to their similarity based on their bit score. This created primary result files, as well as tabular result files with orthogroups (rows) containing sequence IDs from species allocated to each cluster (columns).

3.4 Results

3.4.1 ONT and PacBio SMRT sequencing data for *de novo* assembly

Three domesticated (*L. sativa* Ninja, VIAE, and PI251246) and two wild (*L. serriola* acc US96UC23 and Armenian 999) genotypes of lettuce were selected as diverse genotypes for ONT PromethION sequencing. Nanopore sequencing resulted in 237 Gb of data with a read N_{50} of 36 kb for Ninja, 173 Gb of data with a read N_{50} of 34 kb for VIAE, and 175 Gb of data with a read N_{50} of 32 kb for PI251246. For the wild accessions US96UC23 and Armenian 999, there were 191 and 162 Gb of data with a read N_{50} of 32 and 36 kb, respectively (Table 3.2). After base calling with Guppy v4.5 or Guppy v5.1, removing adaptors, and filtering for reads over 20 kb in length from the "pass" folder, which had a Q score of >7, the resulting data for each assembly had a coverage of approximately 64 to 87X and with reads ranging from 17 to 22 kb.

Table 3.2. *Statistics on the ONT PromethION flow cells used to sequence wild and domesticated genotypes of L. sativa.*

*In blue are the wild accessions of lettuce.

General summary	*L. sativa* cv. Ninja	*L. sativa* cv. VIAE	*L. sativa* cv. PI251246	*L. serriola* acc. US96UC23	*L. serriola* acc. Armenian
Number of flow cells	3	2	3	3	2
Mean read length (kb)	22,728	21,823	17,643	19,161	24,359
Mean read quality	10.5	11.8	11.9	12	12
Median read length (kb)	18,189	16,867	10,768	12,841	20,875

Median read quality	10.5	11.6	11.6	11.7	11.8
Number of reads	10,446,819	7,963,046	9,919,781	9,988,715	6,681,780
Read length N50 (kb)	36,416	34,117	32,984	32,315	36,560
Total bases (Gb)	237	173	175	191	162
Number, percentage, and megabases of reads above quality cutoffs					
>Q7	10,440,850 (99.9%) 237341.8Mb	7,963,046 (100.0%) 173774.5Mb	9,919,757 (100.0%) 175011.9Mb	9,988,715 (100.0%) 191397.9Mb	6,681,780 (100.0%) 162763.8Mb
>Q10	6,755,743 (64.7%) 159063.4Mb	7,962,816 (100.0%) 173774.1Mb	9,914,227 (99.9%) 175004.2Mb	9,988,439 (100.0%) 191397.4Mb	6,681,615 (100.0%) 162763.6Mb
>Q12	1,310,359 (12.5%) 25253.5Mb	2,999,281 (37.7%) 63220.8Mb	3,800,140 (38.3%) 56739.2Mb	4,144,053 (41.5%) 72508.6Mb	2,945,103 (44.1%) 68604.8Mb

When the high quality of PacBio HiFi assemblies became apparent, efforts were switched to HiFi sequencing. Three PacBio SMRT flow-cells for *L. sativa* La Brillante resulted in 85 Gb of data with an average subread length of 9 kb and an N_{50} of 21 kb with a coverage equal to 34X (Table 3.3).

Table 3.3. *Statistics of PacBio raw reads from the sequencing of L. sativa cv. Salinas and cv. La Brillante.*

	L. sativa cv. Salinas	*L. sativa* cv. La Brillante
No of SMRT cells	5	3
Total Reads ≥Q20	6,751,779	8,668,212
Average Read Length (kb)	14	9

Average Read Quality	Q40	Q40
Average Yield ≥Q20 (Gb)	92.2	85.6
Coverage	35x	34x

3.4.2 ONT-based long-read genome assembly

An ONT-based *de novo* assembly of the five genotypes were assembled using the Shasta v0.6.0 assembler (Shafin et al., 2020). One of the major limitations of Nanopore sequencing is the high error rate, which can range between 5% and 15%. To overcome this limitation, the initial *de novo* assembly was error corrected using two rounds of polishing, one with raw ONT reads and the other with Illumina short reads. The resulting contigs of Ninja, VIAE, and PI251246 were then placed to nine chromosomes with a reference-guided approach using the chromosome-scale v11 assembly.

The final ONT assemblies of *L. sativa* cvs. Ninja, VAIE, and PI251246 resulted in assembly sizes of 2.531, 2.538, and 2.515 Gb, respectively. Additional assembly metrics are shown in Table 3.4. Use of the updated Guppy v5.1 base caller rather than Guppy v4.5 for *L. sativa* VIAE and PI251246 resulted in fewer contigs (230 and 214, respectively) and greatly improved contig N_{50} relative to Ninja (Table 3.4).

Bionano data from *L. serriola* US96UC23 and Armenian 999 were used to assemble ONT contigs into super-scaffolds. A consensus map (CMAPS) consisting of 29 and 61 consensus map counts was *de novo* assembled, yielding genome sizes of 2.557 and 2.637 Gb with N_{50} sizes of 224.8 and 161.8 Mb, respectively (Table 3.5). In the BNG workflow, *de novo* assembly of molecules was followed by hybrid scaffolding. Then the Bionano super-scaffolds were further oriented and placed into nine chromosomes using the reference guided

approach with RagTag v.2.1.0. The final assembly size was 2.494 and 2.569 Gb, respectively, with a contig N_{50} of 6.8 and 8.6 Mb in length.

Table 3.4. *Wild and domesticated lettuce genome assembly and BUSCO statistics of seven lettuce genotypes.* Marked in blue are the wild accessions of lettuce.

	L. sativa cv. Salinas v11	*L. sativa* cv. La Brillante	*L. sativa* cv. Ninja	*L. sativa* VIAE	*L. sativa* PI251246	*L. serriola acc.* US96UC23	*L. serriola* acc. Armenian 999
Sequencing technology	PacBio - HiFi	PacBio - HiFi	ONT	ONT	ONT	ONT	ONT
	Bionano and HiC					Bionano	Bionano
Sequencing coverage	35x	34x	87x	64x	64x	77x	84x
Assembly size (Mb)	2,588	2,616	2,531	2,538	2,515	2,494	2,569
No. of contigs	484	1204	1291	230	214	1,300	2,044
Contig N50 (Mb)	12.52	45.059	12.059	49.454	33.33	6.894	8.679
Contig N90 (Mb)	3.482	8.2	4.488	16.823	14.946	2.41	3.325
Largest contig size (Mb)	75.542	101.93	53.663	138.484	181.178	45.832	38.103
BUSCO % (Complete)	98.5	98.4	98.4	96.7	97.7	98.6	98.6
BUSCO % (Duplicate)	3.3	3.3	3.2	3	3.1	3.2	3.4

***Table 3.5**. Bionano consensus map counts (CMAP) data and assembly statistics.*		
	L. serriola US96UC23	*L. serriola* Armenian 999
CMAP	29	61
Total genome map length (Mb)	2,557.01	2,637.98
Genome map N50 (Mb)	224.865	161.887
Total reference length (Mb)	2,494.93	2,569.39
Number of consensus maps aligned (Fraction)	24 (0.83)	59 (0.97)
Total unique aligned length (Mb)	231.848	466.299
Total unique aligned length / reference length	0.093	0.181

3.4.3 PacBio HiFi-based genome assembly

For *L. sativa* La Brillante, three SMRT cells generated 85.6 Gb (34x) HiFi data with an average read length of 9,253 bps. An assembly size of 2.616 Gb was generated with a contig N_{50} of 45.0 Mb (Table 3.4). The time taken to assemble these HiFi reads with Hifiasm was much faster than assembling the ONT reads with SHASTA and subsequent error correction programs (~200 cpu hour/~5 hours wall time versus a month of computation). The contigs were then oriented and scaffolded to nine chromosomes using the v11 reference.

3.4.4 Evaluation of genome quality

The quality of the final assemblies was evaluated for correctness, completeness, and contiguity by mapping short-read Illumina sequences to the assemblies, the BUSCO scores, and synteny with the reference. Nearly all, 96.65% to 98.89%, of the Illumina short reads could be aligned to the genome sequence, indicating that the genome assemblies are high quality. BUSCO scores indicated at least 96.7% completeness in all assemblies. The genomes

of the wild and domesticated background were largely co-linear with the v11 reference assembly (Figure 3.1).

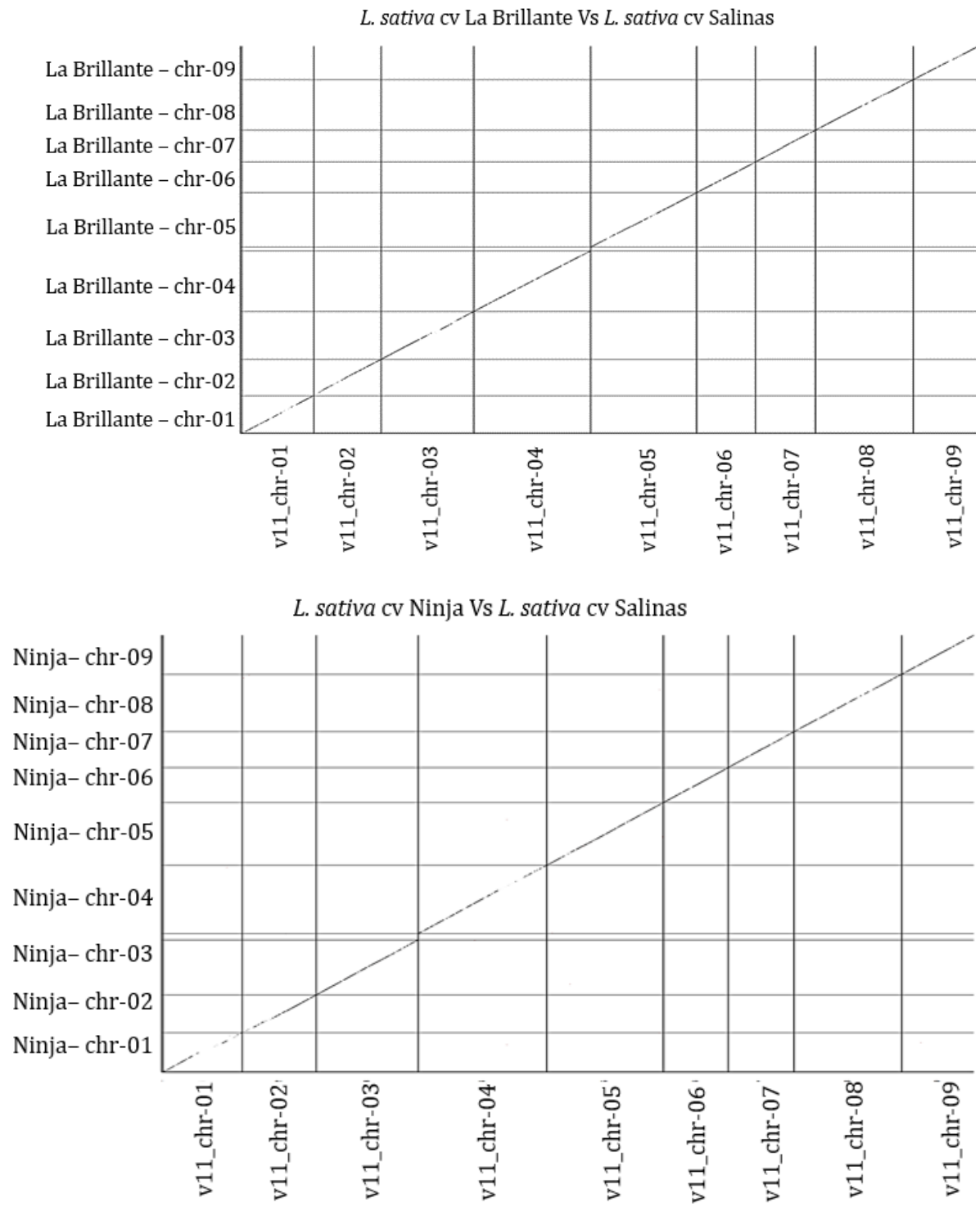

Figure 3.1.a *Collinearity analysis of wild and domesticated lettuce lines with the v11 reference assembly.*

The v11 assembly is plotted along the x-axis and the assembly of each new genotype is shown along the y-axis.

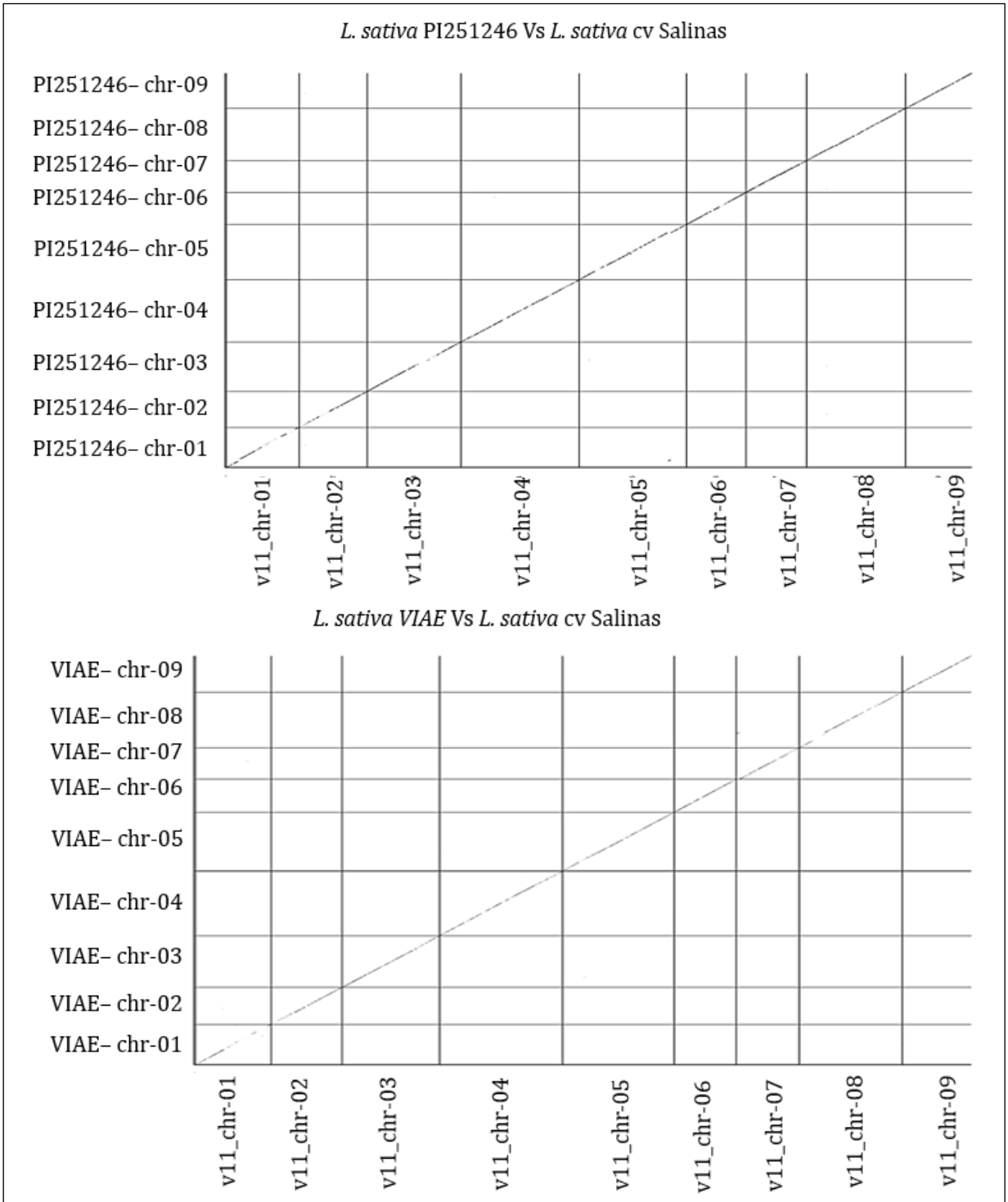

Figure 3.1.b *Collinearity analysis of wild and domesticated lettuce lines with the v11 reference assembly.*

The v11 assembly is plotted along the x-axis and the assembly of each new genotype is shown along the y-axis.

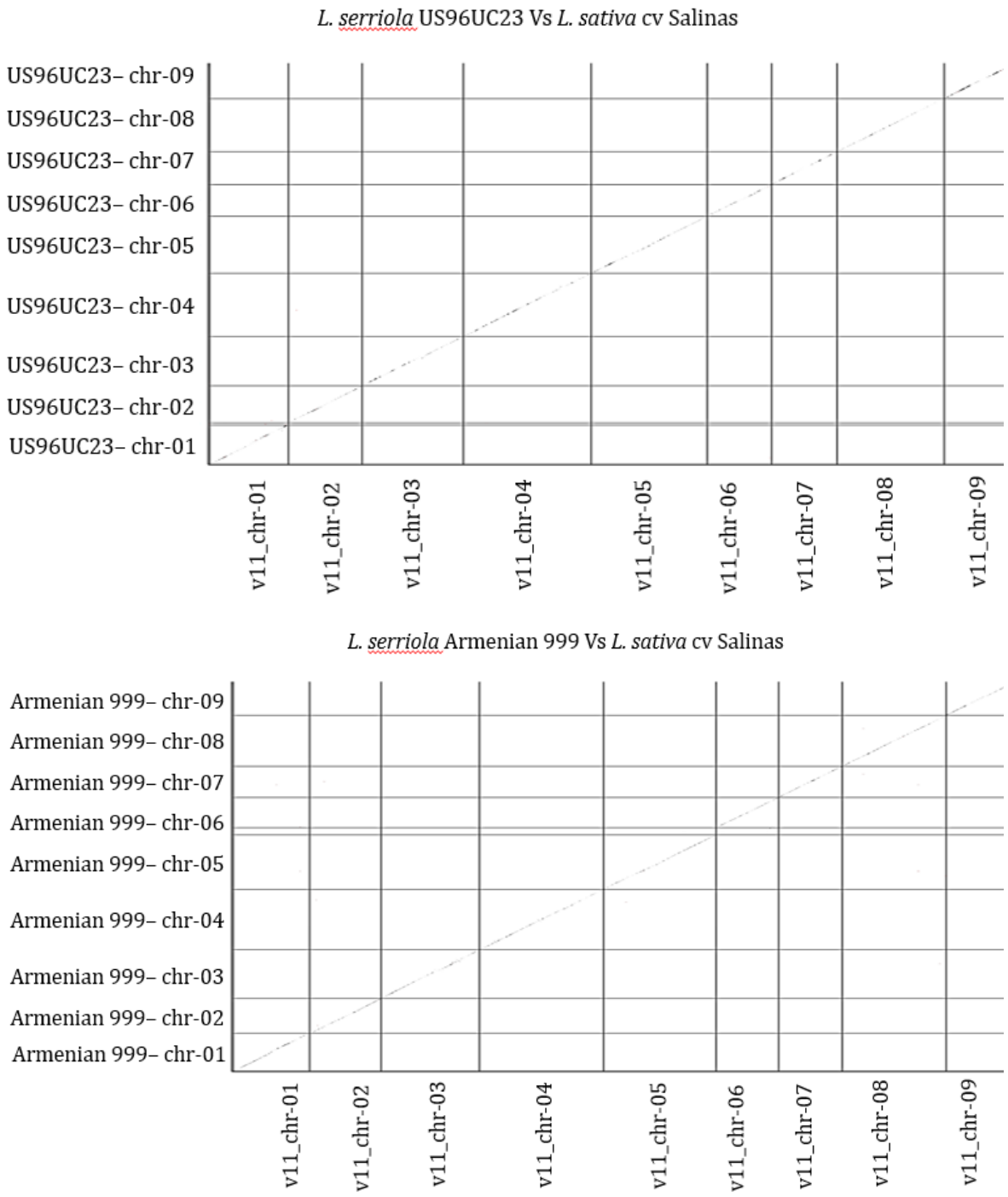

Figure 3.1.c *Collinearity analysis of wild and domesticated lettuce lines with the v11 reference assembly.*

The v11 assembly is plotted along the x-axis and the assembly of each new genotype is shown along the y-axis.

3.4.5 Genome repeat identification and classification

Repeated sequences were identified in all seven assemblies. The combined results of the homology based and *de novo* predictions indicated that repeat sequences account for >80.0% of these lettuce genomes. Long Terminal Repeats (LTRs) accounted for the greatest portion of the repeat content (46 to 50%) (Table 3.6). The LTR elements were mainly *Copia* and *Gypsy* elements. DNA transposons were approximately 1.2% of the whole genome, except for Salinas, which had half of the other genotypes.

Table 3.6. *Repeats identified in seven wild and domesticated lettuce backgrounds.*

	L. sativa cv. Salinas	*L. sativa* cv. La Brillante	*L. sativa* cv. Ninja	*L. sativa* PI251246	*L. serriola* acc. US96UC23	*L. serriola acc.* Armenian 999
LTR-Retro-transposons (%)	46.82	50.85	50.05	48.82	48.53	50.83
LINE (%)	0.26	1.45	0.36	0.34	0.66	0.52
SINE (%)	0	0	0	0	0	0
DNA	0.55	1.21	1.11	1.1	1.33	1.08
Transposons (%)	32.2	29.53	31.91	33.51	33.05	30.88
Satellites (%)	0	0.24	0.17	0	0.03	0
Simple repeats (%)	2.11	1.06	1.02	0.95	0.96	1.01
Total (%)	81.94	84.34	84.62	84.72	84.56	84.32

*The wild lettuce genotypes are indicated in blue.

3.4.6 Genome annotation

Protein-coding genes were annotated for each genome by integrating homology, PacBio Isoseq transcript evidence, and *ab initio* predictions using the MAKER workflow. After correction for gene fragments, an average of 40,000 genes were estimated per genome (Table 3.7). The number of predicted genes varied from 40,915 in the HiFi-based assembly of La Brillante to 47,262 in Ninja. Fewer gene models are shown for VIAE because this annotation was the result of lift-over annotations from the v8 reference using Liftoff v1.6.3 (https://github.com/agshumate/Liftoff); *de novo* annotation of this genome is yet to be completed. The average predicted gene length ranged from 2,250 bp in Ninja to 2,916 in VIAE. The gene density was approximately 4.1 of the genome in all genotypes. Further, the assessment of the gene models using BUSCO shows a high percentage of presence of single copy orthologs in all these annotations. Additional metrics are shown in Table 3.7.

Table 3.7. *Summary of genome annotation statistics per genotype.*
The wild lettuce genotypes are indicated in blue.

	L. sativa cv. Salinas	*L. sativa* cv. La Brillante	*L. sativa* cv. Ninja	*L. sativa* PI251246	*L. sativa* VIAE	*L. serriola* US96UC23	*L. serriola* Armenian
Total sequence length (Mb)	2,590	2,616	2,531	2,515	2,538	2,495	2,569
Number of genes	44,231	40,915	47,262	41,138	36,206	44,010	42,371
Average gene length (bp)	2,431	2,371	2,250	2,310	2,916	2,299	2,470
Average CDS length (bp)	1,151	1,077	1,020	1,050	1,639	995	1,077
Number of exons	230,832	209,870	238,745	216,600	261,545	229,966	244,860

Average exon number	5	5	5	5	7	5	6
Average exon length	261	251	243	240	297	233	224
Number of introns	186,601	168,955	191,483	175,462	214,817	185,956	202,489
Average intron length	255	264	255	248	351	258	248
% of genome covered by genes	4.2	3.7	4.2	3.8	4.2	4.1	4.1
% of genome covered by CDS	2	1.7	1.9	1.7	2.3	1.8	1.8
Complete BUSCOs (%)	1991 (93.9)	1974 (92.9)	2050 (96.7)	1959 (92.3)	1859 (87.6)	1969 (92.9)	1980 (93.3)
Fragmented BUSCOs (%)	39 (1.8)	25 (1.2)	31 (1.5)	28 (1.3)	178 (8.4)	36 (1.7)	36 (1.7)
Missing BUSCOs (%)	91 (4.3)	125 (5.9)	40 (1.8)	134 (6.4)	84 (4.0)	116 (5.4)	105 (5.0)

3.4.7 Orthology assignment and gene family analysis

Orthofinder was used to cluster protein coding genes of six genotypes plus stem lettuce that became available (https://www.lettucegdb.com/); VIAE was not included in this analysis because *de novo* annotations were not available. A total of 287,966 genes from the seven genotypes were clustered into 37,223 orthogroups containing 287,966 genes (96.8% of genes were in orthogroups). A total of 18,042 orthogroups were shared by all seven genotypes. A total of 1,618 genotype-specific orthogroups contained 6,705 genes. A total of 9,082 single copy orthogroups were identified (Table 3.8). The analysis assigned ~97% of all *de novo* annotated genes to orthogroups, suggesting the high quality of our annotations.

Table 3.8. *Summary of gene clustering statistics per genotype.*

The wild lettuce genotypes are indicated in blue.

	L. sativa cv. Salinas	*L. sativa* cv. La Brillante	*L. sativa* cv. Ninja	*L. sativa* PI251246	*L. sativa* Angustana	*L. serriola* US96UC23	*L. serriola* Armenian
# Genes	44,231	36,296	45,400	39,173	40,341	42,151	40,374
# Genes in orthogroups	43,187	35,827	44,120	38,493	37,256	40,749	39,252
# Unassigned genes	1044	469	1,280	680	3,085	1,402	1,122
% Genes in orthogroups	97.6	98.7	97.2	98.3	92.4	96.7	97.2
% Unassigned genes	2.4	1.3	2.8	1.7	7.6	3.3	2.8
# Orthogroups containing species	29,143	26,222	30,988	27,602	24,691	28,149	27,825
% Orthogroups containing species	78.3	70.4	83.2	74.2	66.3	75.6	74.8
# Genotype-specific orthogroups	120	41	139	63	1024	137	94
# Genes in genotype-specific orthogroups	710	142	385	259	4401	420	388
% Genes in genotype-specific orthogroups	1.6	0.4	0.8	0.7	10.9	1	1

Orthogroup clustering revealed that most (20,868) gene families were conserved across the two wild accessions and the *L. sativa* v11 reference assembly. A total of 254 genes were unique to the *L. sativa* assembly; 3,592 genes were specific to *L. serriola*. Each *L. serriola* genotype had 981 or 747 unique genes reflecting their diverse origins (Figure 3.2). Within the four *L. sativa* genotypes, only a few hundred gene families were genotype specific (Fig. 3.3).

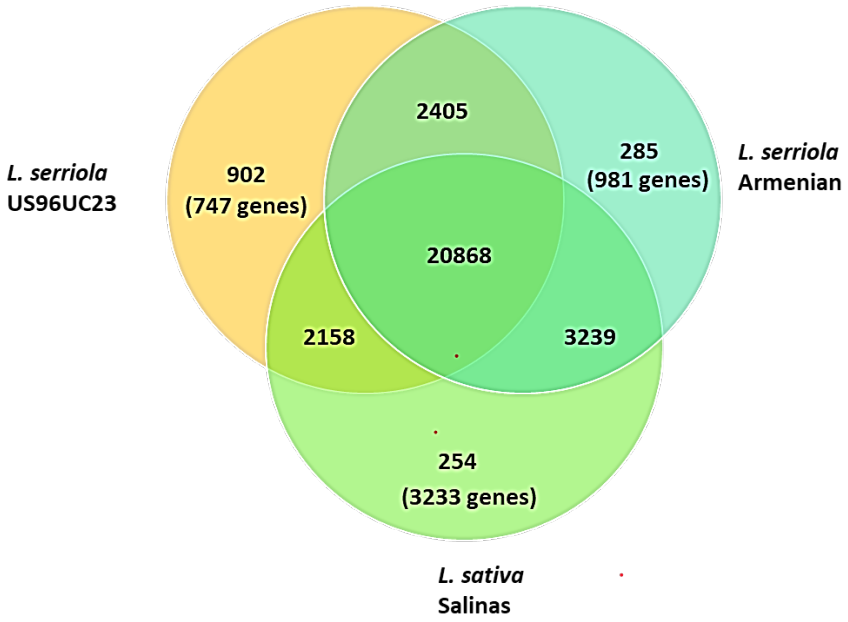

Figure 3.2 *Venn diagram displaying shared and unique orthogroups between two wild and one domesticated lettuce genotype.*

The number of genes within the orthogroups are shown in parentheses.

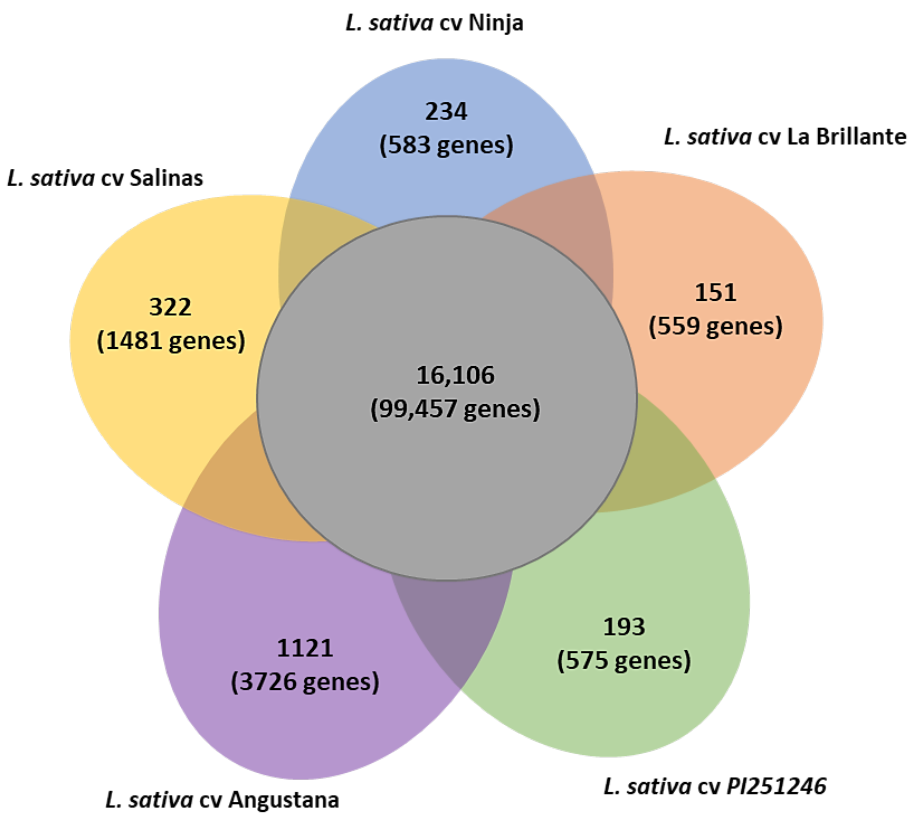

Figure 3.3. *Annotated gene families of the core and dispensable genomes of domesticated lettuce.*

The number of genes is shown in parentheses.

3.5 Discussion

In this chapter, I described the *de novo* assembly of additional chromosome-scale genomes for two wild (*L. serriola* acc US96UC23 and Armenian 999) and four domesticated (La Brillante, Ninja, VIAE, and PI251246) lettuce genotypes, using long-read sequencing technologies. This is a significant addition to genetic resources for lettuce. At this time, there is only one other chromosome scale genome assembly for lettuce—that of stem lettuce, which has been released on the Lettuce Genome Database (https://www.lettucegdb.com/).

The genome size of stem lettuce is comparable to the additional lettuce genome assemblies, with 2.597 Gb. The stem lettuce assembly was sequenced with 105x PacBio SMRT sequencing data along with Hi-C and Bionano data for scaffolding. However, the contig N_{50} of the v11 reference genome along with the six additional assemblies ranges from 6.8 Mbs to 49 Mbs, which is greater than the contig N_{50} of stem lettuce at 4.7 Mbs in length. The genome assembly of stem lettuce is slightly more fragmented, with 2,053 contigs compared to 484 contigs in the v11 assembly, covering the whole genome. With improvements in long-read technologies, we were able to construct a chromosome-scale assembly of *L. sativa* PI251246 with just 214 contigs. As long-read data improves, we are now able to develop highly contiguous chromosome-scale assemblies. The number of telomere-to-telomere assemblies is likely to increase as the long-read technologies become cheaper and more accurate.

Long-read and long-range technologies have greatly improved larger genome assembly and annotation projects. In this chapter, the first five genotypes were sequenced with ONT. During the project, PacBio HiFi technology became available and, as described in Chapter 2, we determined that PacBio HiFi data was the more accurate technology. Nonetheless, ONT resulted in good assemblies. By implementing an improved ONT workflow including re-base calling with the latest (at the time) Guppy v5 algorithm and using Bionano optical mapping data, I was able to resolve most assembly conflicts and greatly increase the contiguity of the assemblies. The ONT and HiFi-based assemblies have similar high BUSCO scores, indicating that comparative analyses can be conducted.

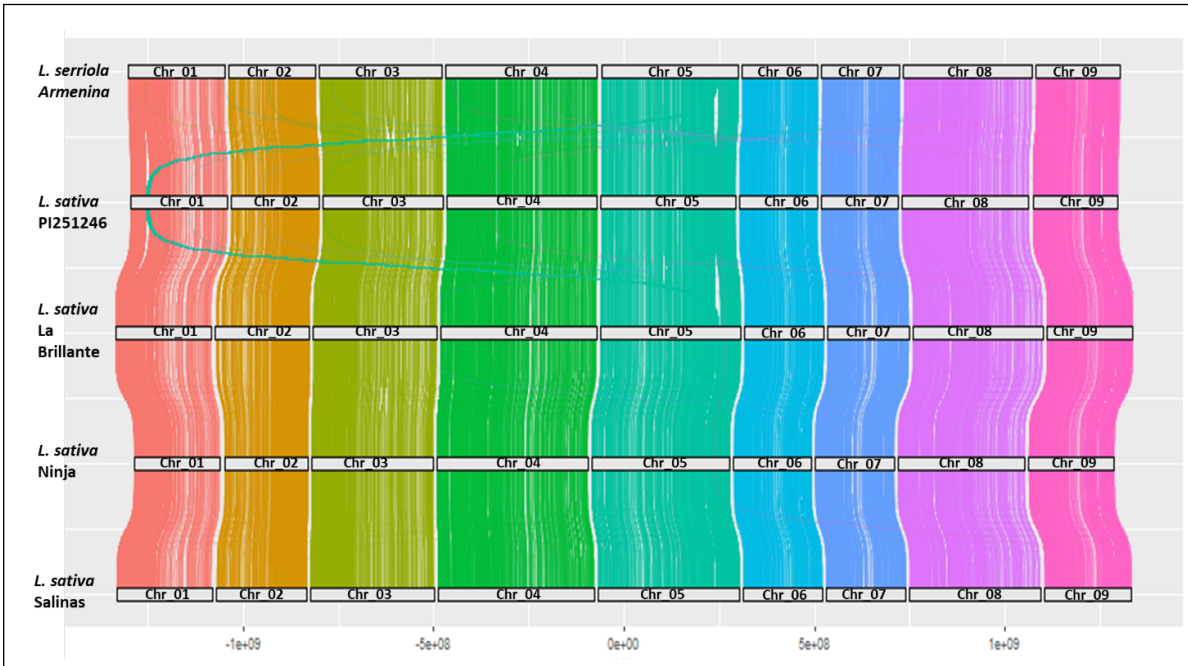

Figure 3.4 *Genome-wide synteny across single copy orthologs between lettuce genotypes.*

Genome annotation and orthology analysis grouped almost more than 97% of the annotated genes into orthogroups across all assemblies. Figure 3.4 shows the genome-wide synteny of the annotated genes across different lettuce genotypes. However, there are some striking differences between the assemblies. In particular, the number of genes annotated varied from 40,915 to 47,262. It is unclear to what extent this is real and to what extent it is noise in the annotation pipeline. Future work will investigate the reasons for these differences by screening for broken genes and for sequences that are present but not annotated for some reason. Also, VIAE could not be used for all the analyses because the final round of polishing could not be done because Illumina reads were not available, and it was only annotated using lift-off from v8 of Salinas rather than *de novo*. Further time-consuming

improvement of the ONT assembly of VIAE was halted because it was decided to generate a HiFi assembly instead, which is underway.

The availability of these six additional chromosome-scale genome assemblies, plus the v11 reference and stem lettuce assemblies provides the foundation for understanding the presence/absence variations among genotypes in the context of the lettuce pangenome as described in Chapter 4.

References:

Agricultural Statistics Service, N. (2020). *United States Department of Agriculture National Agricultural Statistics Service*.

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1829-6

Atkinson, L. D., Mchale, L. K., Truco, M. J., Hilton, H. W., Lynn, J., Schut, J. W., Michelmore, R. W., Hand, P., & Pink, D. A. C. (2013). An intra-specific linkage map of lettuce (*Lactuca sativa*) and genetic analysis of postharvest discolouration traits. *Theor Appl Genet*, *126*, 2737–2752. https://doi.org/10.1007/s00122-013-2168-8

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1). https://doi.org/10.1186/s13100-015-0041-9

Belser, C., Baurens, F.-C., Noel, B., Martin, G., Cruaud, C., Istace, B., Yahiaoui, N., Labadie, K., Hřibová, E., Doležel, J., Lemainque, A., Wincker, P., D'hont, A., & Aury, J.-M. (2021). *Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing*. https://doi.org/10.1038/s42003-021-02559-3

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/NAR/27.2.573

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Curr. Protoc. Bioinform*. https://doi.org/10.1002/0471250953.bi0411s48

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods, 18*. https://doi.org/10.1038/s41592-020-01056-5

Christopoulou, M., Wo, S. R. C., Kozik, A., McHale, L. K., Truco, M. J., Wroblewski, T., & Michelmore, R. W. (2015). Genome-wide architecture of disease resistance genes in lettuce. *G3: Genes, Genomes, Genetics*, *5*(12), 2655–2669. https://doi.org/10.1534/g3.115.020818

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1832-y

Farrara, B. F., Ilott, T. W., & Michelmore, R. W. (1987). Genetic analysis of factors for resistance to downy mildew (*Bremia lactucae*) in species of lettuce (*Lactuca sativa* and *L. serriola*). *Plant Pathology*, *36*(4), 499–514. https://doi.org/10.1111/j.1365-3059.1987.tb02267.x

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data, 7(399).* https://doi.org/10.1038/s41597-020-00743-4

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. v, Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. https://doi.org/10.1101/767764

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kesseli, R., Ochoa, O., & Michelmore, R. (1991). Variation at RFLP loci in Lactuca spp. and origin of cultivated lettuce *(L. sativa).* Genome 34(3): 430-436.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5(59).* http://www.biomedcentral.com/1471-2105/5/59

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S. S., Zuo, Q., Shi, X. H., Li, Y. F., Zhang, W. K., Hu, Y., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology 2014 32:10*, *32*(10), 1045–1052. https://doi.org/10.1038/nbt.2979

Lindqvist, K. (1960). On the origin of lettuce. *Hereditas*, *46*(3–4), 319–350. https://doi.org/10.1111/j.1601-5223.1960.tb03091.x

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1). https://doi.org/10.1371/JOURNAL.PCBI.1005944

Michelmore, R. W., Kesseli, R. v., & Ryder, E. J. (1994). Genetic mapping in lettuce. In Advances in Cellular and Molecular Biology of Plants, vol 1. pp. 223–239. https://doi.org/10.1007/978-94-011-1104-1_12

Michelmore, R., & Wong, J. (2008). Classical and molecular genetics of Bremia lactucae, cause of lettuce downy mildew. *European Journal of Plant Pathology, 122*, 19-30. https://doi.org/10.1007/s10658-008-9305-2

Nattestad, M., & Schatz, M. C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics*, *32*(19), 3021–3023. https://doi.org/10.1093/bioinformatics/btw369

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O., Truco, M. J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14953

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., et al. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*. https://doi.org/10.1038/s41587-020-0503-6

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. v., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, *33*, W465. https://doi.org/10.1093/NAR/GKI458

Su, X., Wang, B., Geng, X., Du, Y., Yang, Q., Liang, B., Meng, G., Gao, Q., Yang, W., Zhu, Y., & Lin, T. (2021). A high-continuity and annotated tomato reference genome. *BMC Genomics*, *22*(1). https://doi.org/10.1186/S12864-021-08212-X

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wei, T., van Treuren, R., Liu, X., Zhang, Z., Chen, J., Liu, Y., Dong, S., Sun, P., Yang, T., Lan, T., Wang, X., Xiong, Z., Liu, Y., Wei, J., Lu, H., Han, S., Chen, J. C., Ni, X., Wang, J., … Liu, H. (2021). Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. *Nature Genetics*. https://doi.org/10.1038/s41588-021-00831-0

Chapter 4: Testing approaches for developing a pangenome of lettuce

4.1 Abstract

The overall objective of this study is to understand genomic diversity in *Lactuca* spp. by using a graph-based data structure as a reference and to enhance analysis of difficult genomic regions that are currently missed by using single linear reference genomes. Pangenome development is becoming essential as an increasing number of genomes are sequenced from the same species. Using a single reference genome limits the study of genetic diversity, evolution, and domestication of a species. The capture of variation using a single reference genome poses many challenges when analyzing individuals whose genetic background is not the same as the reference genome. To resolve the limitations imposed by mapping sequence reads against a single reference genome or serially mapping them against multiple reference genomes, pangenome-based methods allow simultaneous comparison against multiple high-quality reference genomes to make downstream analysis more efficient. In this chapter, seven high quality, chromosome scale assemblies from five diverse? lettuce cultivars (*Lactuca sativa* cvs. Salinas, La Brillante, Ninja, PI251246, VIAE) and two wild accessions *(L. serriola* US96UC23, Armenian 999) were used to construct a lettuce pangenome. Several current pangenome methods were investigated for constructing a pangenome to determine the core and dispensable gene/sequencing differences across lettuce genotypes. Using this computationally driven comparative analysis approach, we will be able to predict haplotype blocks that are unique for each genotype and better understand the role of structural variation in the determination of agronomic traits.

4.2 Introduction

Pangenomes provide fundamental resources for functional genomics and crop breeding. With the recent advances in long-read sequencing technologies and the availability of an increasing number of high-quality genome assemblies, there is an urgent need to integrate multiple genomes of the same species to understand genetic diversity. Using a single reference genome can result in biases due to substantial structural variation (SV) exhibited within a species (Yu et al., 2014; Ho et al., 2020.) A pangenome represents the genetic repertoire of a species rather than a single genotype (Tettelin et al., 2005). It describes a set of core sequences that are shared among all individuals and a set of variable or dispensable sequences that are unique to one or more genotypes. A pangenome minimizes reference bias in genomic research and enables more accurate prediction of traits.

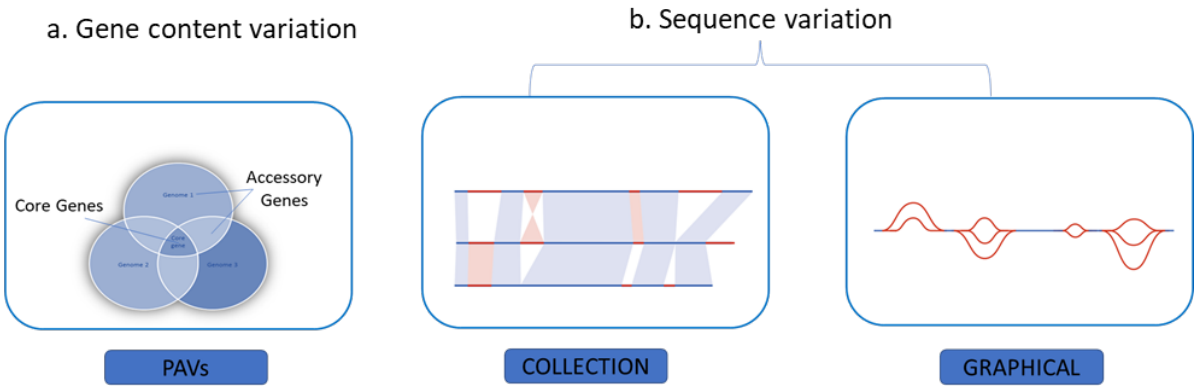

Figure 4.1 *Pan-genome approaches for lettuce* (from Bayer et al., 2020).

Currently, there are several strategies for developing and representing pangenomes. The most popular crop pangenome approaches can be broadly classified into two categories: gene content based (presence/absence variation; PAV) and sequence based that use assembly alignments or graphs (Figure 4.1) (Golicz et al., 2020a; Li et al., 2022a). Most

previously studied crop pangenomes mainly focused on the presence or absence of gene content due to the lack of high-quality genome assemblies. Sequence-based approaches can be subdivided into *de novo* assembly and comparison of each genotype and iterative mapping back to a reference assembly and they are complementary to each other (Bayer et al., 2020). *De novo* assembly requires the assembly of multiple individual genomes independently, followed by the comparison of each genome to identify sequence or gene variation. The iterative mapping and assembly strategy entails mapping reads from several genotypes to a reference genome, assembling the unmapped read to new contigs, and then adding the novel contigs to the reference to construct a pangenome (Sherman et al., 2019). As more high-quality, chromosome-scale assemblies are generated, it has become feasible to develop graphical based pangenomes (Llamas et al., 2019; Sherman & Salzberg, 2020b).

Each of these different pangenome approaches has advantages and disadvantages. Iterative assembly does not distinguish between extreme sequence divergence at a locus and the structural modifications caused by insertion or deletion of sequences. The whole genome assembly approach cannot distinguish between genome diversity between individuals and carries the errors and variations observed in assembly and annotation methods. Based on the complexity of the genome and the sequence data quality and availability, multiple pangenome approaches can be adopted to construct and study the pangenome of a crop.

Construction of plant pangenomes is challenging due to the large size and high repeat content of many crop genomes (Hübner, 2022). For crops like lettuce with high repeat content, repetitive regions form a major portion of the genome. SVs may occur as a result of whole-genome duplication as well as tandem and segmental duplication of genomic areas. Both copy number variation (CNV) and PAV of gene content are results of SV duplication and

fragmentation (Ho et al., 2020; Zhao et al., 2021). CNV and PAV may also be caused by transposable element insertion, *de novo* gene birth, unequal crossing-over, introgression of SV from closely related species, and horizontal gene transfer, all of which can impact important phenotypic traits (Golicz et al., 2016; Golicz et al., 2020b; Li et al., 2022b; Zanini et al., 2022). Understanding the CNV and PAV structure across the genome is crucial for constructing a lettuce pangenome.

Early crop pangenomes described PAVs. The first crop pangenome was published in 2014, representing seven wild soybean genotypes (Li et al., 2014). Since then, pangenomes in crop species such as maize (*Zea mays*) (Bradbury et al., 2022), rice (*Oryza sativa*) (Qin et al., 2021; Zhao et al., 2018), and Brassica (*Brassica* spp.) (Golicz et al., 2016) have led to the identification of genes linked with disease resistance and yield components. Recently, as more high-quality genome assemblies became available, graphical pangenomes have been adopted more widely to understand sequence variation across many genotypes in tomato (*Solanum lycopersicum*) and soybean (*Glycine max*) (Li et al., 2014). Graph-based pangenomes will be major features of plant pangenomics in the future; however, the huge computational memory requirements limit their current use to population scale studies. Improvements in graph-based algorithms are currently being developed.

In this chapter, seven *de novo* assemblies of lettuce genotypes along with the v11 reference genome of lettuce were used to develop a lettuce pangenome. As graphical pangenome construction method combines the benefits of both *de novo* assembly and iterative mapping approaches, this chapter primarily focuses on graphical pangenome construction for lettuce. The primary objective of this chapter was to evaluate the graphical pangenome approaches for lettuce. Therefore, I focused mainly on specific regions of interest

including gene families and highly repeated regions. For most of the analyses presented in this chapter, I used representative assemblies of lettuce and in some cases analysis of only one chromosome. This chapter lays the foundation for future larger-scale population-based pangenome studies and genome-wide analysis.

4.3 Materials and Methods

4.3.1 *De novo* assembly of seven lettuce genotypes

Six representative *de novo* assemblies of lettuce genotypes and the v11 reference genome of lettuce were used to develop a comprehensive lettuce pangenome assembly as described; the assemblies were described in detail in Chapters 2 and 3. Two of the seven *de novo* assemblies (*L. sativa* cvs. Salinas, a crisphead type, and La Brillante, a Latin/romaine type) are PacBio HiFi based and five assemblies (*L. sativa* cv. Ninja, a butterhead type with introgression from *L. saligna*, VIAE, a complex pedigree with introgressions from *L. virosa*, and PI251246, an oil seed type from Egypt; *L. serriola* US96UC23 collected from California, Armenian 999 collected from Armenia) are Oxford Nanopore (ONT) based assemblies. These represent five diverse domesticated and two geographically distinct wild genotypes of *Lactuca* spp. For the ONT-based *de novo* assembly, the Shasta v0.5.0 assembler was used to construct the draft assembly, followed by Pepper v0.01 to self-polish the draft assemblies with raw ONT reads and Pilon v1.23 for further polishing with Illumina. Scaffolding of the assemblies was carried out using Bionano data for the wild accessions and then the super scaffolds were oriented using Ragtag v.2.1.0. The PacBio HiFi based assemblies were constructed using Hifiasm v0.16.1-r375 (Cheng et al., 2021). For the v11 *L. sativa* cv. Salinas reference assembly, both Bionano and Hi-C data were used to further scaffold the draft

assemblies. La Brillante was scaffolded using a reference guided approach with Ragtag v.2.1.0. All assemblies were quality checked for correctness, completeness, and contiguity as described in Chapters 2 and 3.

In addition, v12, a *de novo* assembly constructed using the PacBio CCS reads in the Hifiasm v0.16.1-r375 assembler, was constructed from a subset of reads that are from PCR-free libraries that were used in the *L. sativa* cv. Salinas v11 reference assembly. The main purpose of this assembly version was to use it as a control in the downstream analysis.

4.3.2 Repeat analysis and genome annotation

RepeatScout v1.0.6 and RepeatModeler v.1.0.10 were used to identify and classify *de novo* repeat families. RepeatMasker v4.0.7 was used to report different repeats (SINEs, LINEs, TE elements, DNA elements, interspersed repeats, small RNA, satellites, simple repeats, and low complexed repeats) in the assembly as described in Chapters 2 and 3.

For gene annotation, a workflow combining both *ab initio* gene finding, and homology-based gene prediction was used as described previously. Augustus v3.3 (Stanke & Morgenstern, 2005) and SNAP v2013-11-29 (Korf, 2004) were used in *ab initio* gene finding. cDNA sequences from 17 plant species downloaded from NCBI were used to predict homologous genes. MAKER-P (Campbell et al., 2014) was used to combine all these gene predictions to construct the main structure of protein coding genes. PASA v2 was used to predict alternative splicing types. Finally, mRNAs that encoded peptides less than 10 amino acids or that did not start with methionine were filtered out.

4.3.3 Pangenome approaches

Several pangenome approaches and toolkits were evaluated using six *de novo* assemblies of lettuce genotypes and the v11 reference genome of lettuce (Fig. 4.1). Below, I will focus on the software that proved useful for the construction of the lettuce pangenome.

4.3.3.a PAV gene content using gene clustering

Gene clustering and identification of PAVs of gene content was performed using OrthoFinder v2.3.12 (Emms & Kelly, 2019; Li et al., 2003). Inferred protein sequences from five assemblies (*L. sativa* cvs. Salinas, Ninja, Angustana, and *L. serriola* US96UC23, Armenian 999) were used as the input data for the clustering. VIAE was not included because the assembly was not polished by error correcting with Illumina reads; its inclusion would have artificially inflating the numbers of PAVs. The La Brillante assembly only became later. Gene clustering methods are further described in Chapter 3. Using this tool, gene sequences that were shared among phylogenetically related orthogroups were clustered, and the sequences with no similarity—those that were specific to an accession—were considered singletons.

4.3.3.b Whole genome alignment and comparison for identification of structural variations

To analyze the distribution of SVs, the six ONT and HiFi based assemblies were mapped on to the v11 reference of cv. Salinas using minimap2 v with default parameters (Li, 2021). Two long-read SV callers were used: Sniffles v1.0.12 (Sedlazeck et al., 2018) and CuteSV v.1.0.10 (Jiang et al., 2020) with default parameters. Assemblytics (Nattestad & Schatz, 2016b) was also applied to the genome alignments generated using Mummer v4.0.0rc1 with default parameters. All of the SVs from the three SV callers were merged using SURVIVOR v.1.0.6; only calls supported by at least two callers and where the callers agreed

regarding the type of variant were reported. Boxed in dotted lines in Figure 4.2 is the short-read SV caller workflow that is currently being evaluated, and results not available for this chapter.

Code for calling of structural variants:

```
        ## Make alignment with minimap:
        minimap2 -t 32 -ax asm5 ref.fa ${g} | samtools view -bS -@ 32 - | samtools sort -o
${g}.sorted.bam -
        samtools index -@32 ${g}.sorted.bam

        ##Calling SV with sniffles:
        sniffles --input ${g}.sorted.bam --vcf ${g}.sv.sniff.vcf
        bcftools view -Ov -i 'FILTER="PASS"' ${g}.sv.sniff.vcf > ${g}.sv.sniff.pass.vcf

        ##Calling SV with CuteSV:
        cuteSV ${g}.sorted.bam ref.fa ${g}.sv.CSV.vcf CSV_output/
        bcftools view -Ov -i 'FILTER="PASS"' ${g}.sv.CSV.vcf > ${g}.sv.CSV.pass.vcf

        ##Calling SVs using Assemblytics:
        nucmer --maxmatch -l 40 -c 90 -t 32 ref.fa > ${g}.delta
        Assemblytics ${g}.delta> ${g} 10 1 10000 > ${g}.sv.assm.vcf
        bcftools view -Ov -i 'FILTER="PASS"' ${g}.sv.assm.vcf > ${g}.sv.assm.pass.vcf

        ##Merge SVs using SURVIVOR:
        SURVIVOR merge all.vcf 1000 2 1 1 0 30 all_merged.vcf
```

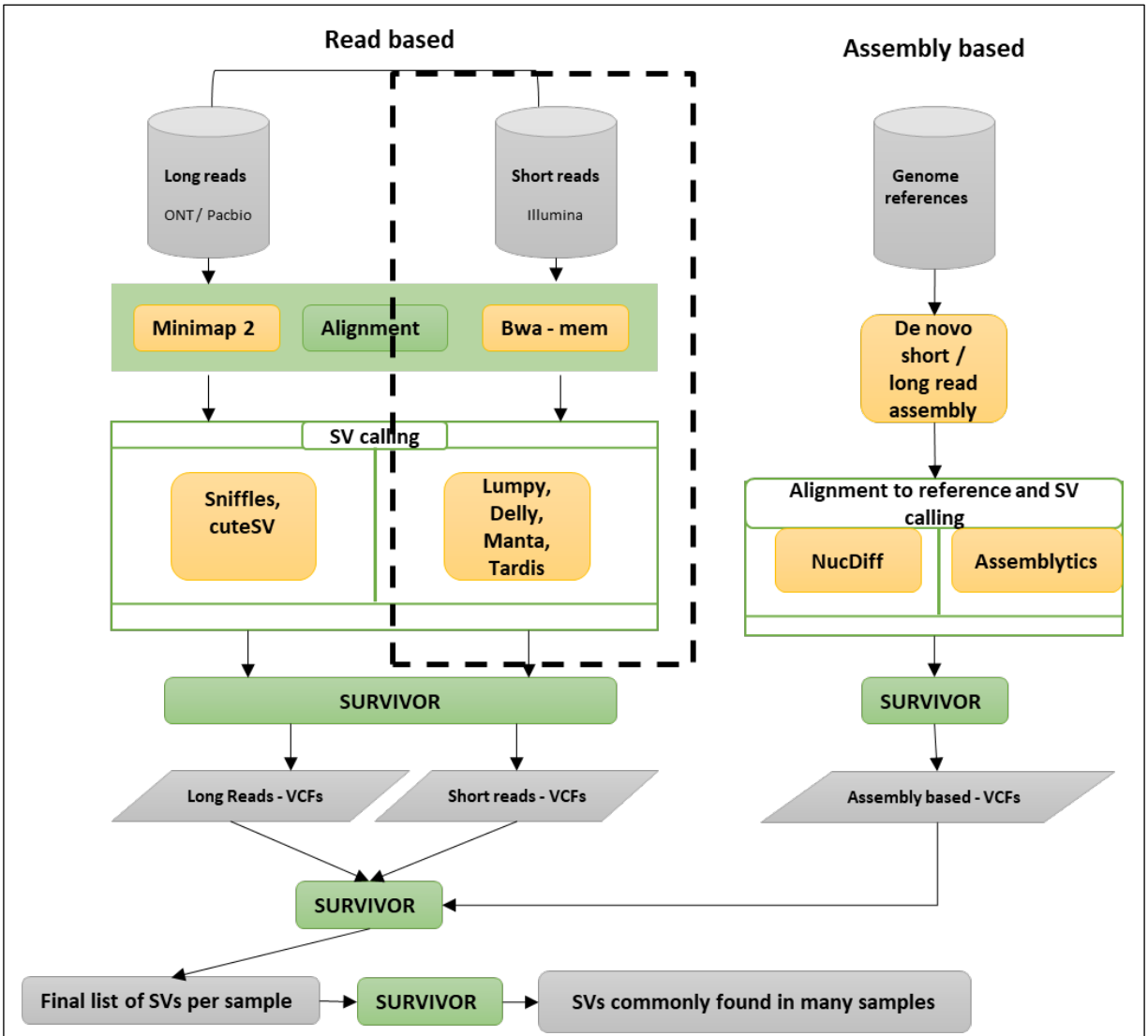

Figure. 4.2. *Structural variation workflow between wild and domesticated lettuce species.*

***Boxed in dotted lines is the short-read based SV analysis workflow that is currently being run.*

4.3.3.c. Graphical pangenome approaches

For the construction of a graph-based pangenome, I evaluated the graph-based tools minigraph (Li et al., 2020), variation graph (vg) (Garrison et al., 2018), and Pangenome Graph Builder (pggb) (Guarracino et al., 2022; Hickey et al., 2022) using the six available high

quality lettuce assemblies (Figure 4.3). To limit the computational complexity, most evaluations were based on one chromosome.

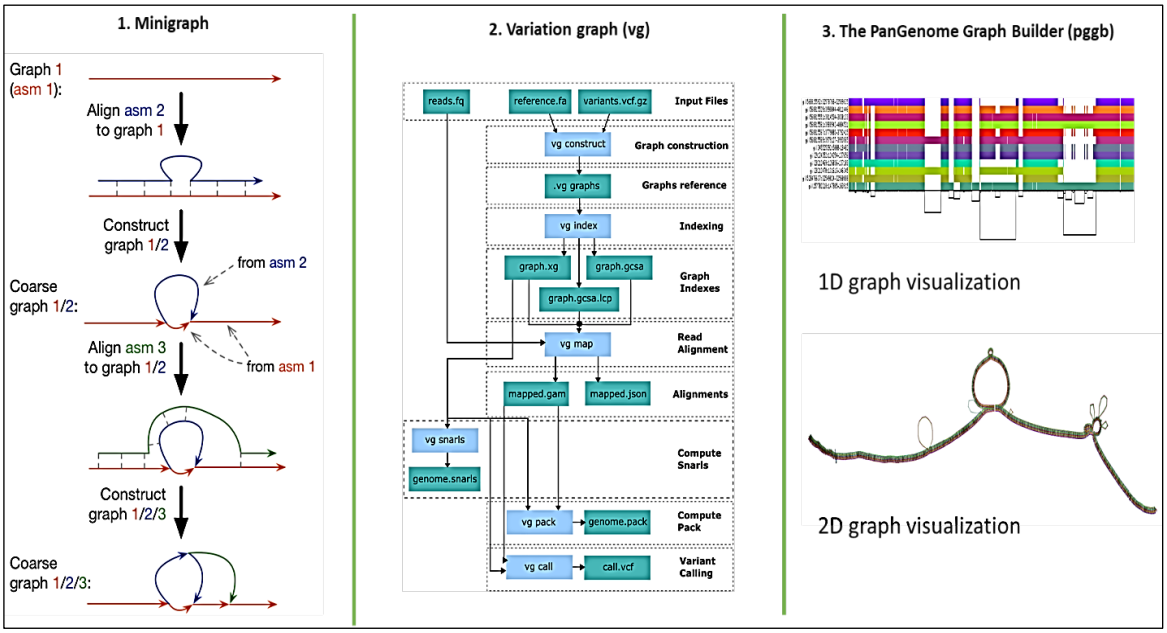

Figure 4.3. *Various pangenome graph methods for lettuce.*
Software sources for figures:
1. https://github.com/lh3/minigraph; 2. https://github.com/vgteam/vg;
3. https://github.com/pangenome/pggb.

a. Minigraph: multi-assembly graph for structural variation analysis

Minigraph v 0.12 with option -xggs was used to integrate six chromosome-scale genome assemblies into a multi assembly graph starting with the reference assembly v11 as the backbone to the graph. Minigraph extends the minimizer-mapping of Minimap2 to graphs and is computationally efficient because it avoids base-level alignment. Graphs generated by minigraph were visualized using bandage (Wick et al., 2015). The bubble popping algorithm of gfatools v0.8 (https://github.com/lh3/gfatools) was used to extract the SVs from the multi-assembly graph. A bubble is the branching region in the graph for which the start and end node are the reference sequences. A path traversing the start and

end nodes represents an allele of an SV. gfatools reports the shortest and longest path for each bubble.

Code for generating graph pangenome using minigraph:

```
##Graph pangenome with minigraph
minigraph -xggs ref.minigraph.gfa asm_*.fa > lettuce_asm.minigraph.gfa
##Graph statistics
gfatools stat lettuce_asm.minigraph.gfa
##Graph to fasta
gfatools gfa2fa -s lettuce_asm.minigraph.gfa lettuce_asm.minigraph.stable.fa
## Call structural variants with gfatools
gfatools bubble lettuce_asm.minigraph.gfa > lettuce_asm.minigraph.structural.bed
```

b. pggb workflow for syntenic and haploblock analysis

pggb generates an all-to-all alignment of input assembly sequences using wfmash, an aligner for pangenomes, with sparse homology and wavefront inception (https://github.com/waveygang/wfmash). Graph induction and normalization of graphs results in a graph pangenome of the input sequences. Visualization of the graphs is performed by odgi (Guarracino et al., 2022). The resulting graph was used to call both small and large variants.

Code for generating pangenome using pggb:

```
## Graph pangenome with pggb
singularity exec --bind /usr/lib/locale/ -H
/share/rwmalt/Sagaya/pangenome_WS/pggb $pggb_IMG_PATH/pggb-latest.simg pggb \
 -i chr02_all.fa \
 -o pggb/output_chr2_95_all/ \
 -t 42 -p 95 -s 100000 -V 'Lsat_1_v11_chr2' -n 90 -k 311

 ## Visualization of pggb graphs using ODGI for a ROI:
 odgi build -g pggb/output_chr2_95_all/chr02_all.fa.c50e82f.04f1c29.seqwish.gfa -o
chr02_all_95.og
 odgi sort -i chr02_all_95.og -P -Y -o chr02_all_95.sort_PY.og
 odgi viz -i chr02_all_95.sort_PY.og -o chr02_all_95.sort_PY.MRC.png -x 500 -r
Lsat_1_v11_chr2:5423607-73645167
```

c. Genotyping application with the vg toolkit

The vg toolkit was the first openly available variation graph tool to scale multi-gigabase genomes. For the lettuce pangenome, vg was used for read-mapping, variant calling, and pangenome visualization. Vg can build graphs both from variants in vcf format and from assembly alignments. Short reads from diverse lettuce lines are mapped to the reference graph to extract the genotype information for each background.

Code for generating graph structure using the vg toolkit:

```
## Pangenome graph using vg for 200 whole genome reseq data:
vg construct -r ref.fa -v all.sv.vcf.gz -S -a -f -p -t 112 > all.vcf.vg

## Collapse all nodes:
vg mod -u all.sv.vcf.vg -t 112 > all.vcf.unchop.vg

## To view graphs in Bandage:
vg view all.sv.vcf.unchop.vg -g > all.vcf.gfa

##index vg graph:
vg index -x all.sv.vcf.xg -g all.sv.vcf.gcsa all.sv.vcf.vg -t 112

## read mapping:
vg map -d all.sv.vcf -f all.illumina.fastq.gz -t 112 > all.illumina.gam

## mapping statistics:
vg stats -a all.illumina.gam

## creating bams for viewing:
vg surject -x all.sv.vcf.xg -b all.illumina.gam -t 112 > all.illumina.bam

## Call variants:
vg pack -x all.sv.vcf.xg -g all.illumina.gam -Q 5 -s 5 -o all.illumina.pack -t 112

## Generate vcf:
vg call all.sv.vcf.xg -k all.illumina.pack -t 112 > all.illumina.graph_calls.vcf
```

4.4 Results

4.4.1 PAV between wild and domesticated lettuce species

Orthogroup clustering of 212,497 protein sequences from five representative lettuce backgrounds, three *L. sativa* and two *L. serriola*, resulted in 36,959 orthogroups. Protein clustering indicated that 36,959 orthogroups contained 97.8% of the total number of input genes. Only 3,344 (9%) were unique to one genotype, and 18,824 orthogroups had representatives in all genotypes. Of the 36,959 orthogroups, 23,751 represent the core gene cluster of the lettuce genome that were present in at least four out of the five genotypes; allowance for absence in one genotype is conventional for defining core gene sets to take into account annotation artifacts (Song et al., 2020). A total of 9,864 gene clusters were variable or dispensable gene clusters that are not present in all the lettuce backgrounds (Fig. 4.4).

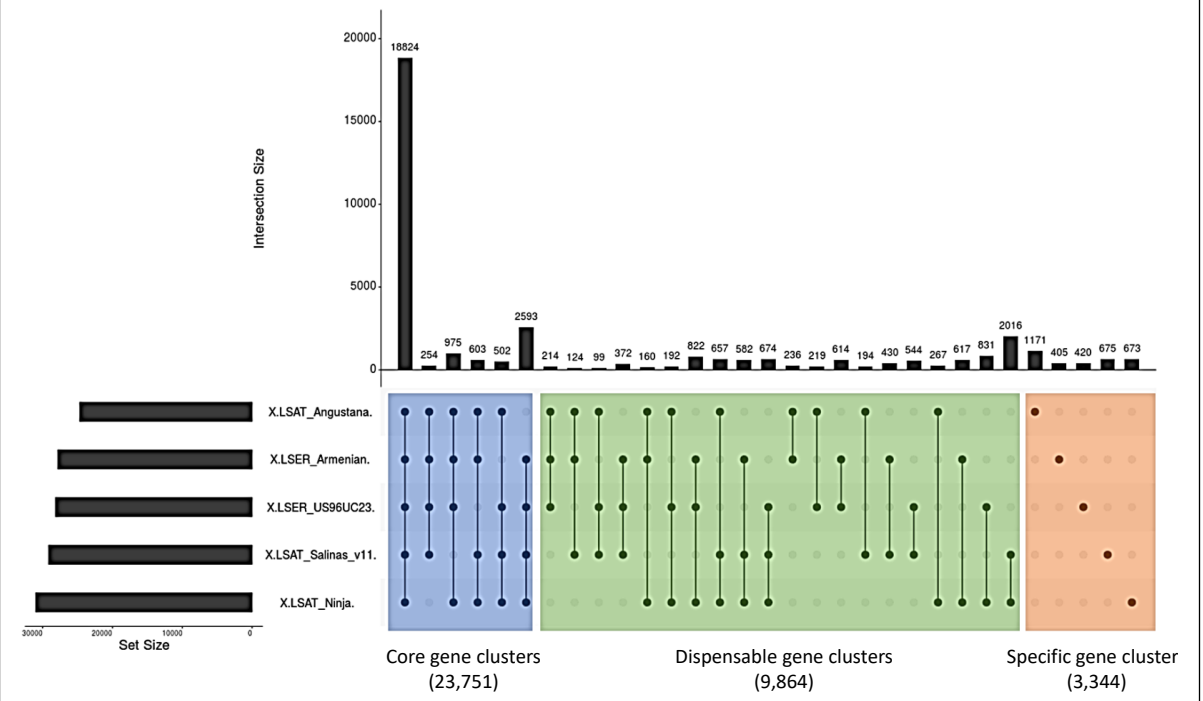

Figure 4.4. *Core and dispensable orthogroup clustering*

4.4.2 Whole genome alignment and comparison of SVs between wild and domesticated lettuce species

The five genomes analyzed were highly syntenic, as shown by the distribution of 1:1 ortholog groups (single copy in each genotype when present) (Fig. 4.5). Synteny analysis revealed a translocation event in Chromosome 1 of *L. sativa* PI2521246, the oil seed type, relative to the other genomes. Also, *L. sativa* cvs. PI251246 and La Brillante share an inversion in Chromosome 3 relative to the other genotypes.

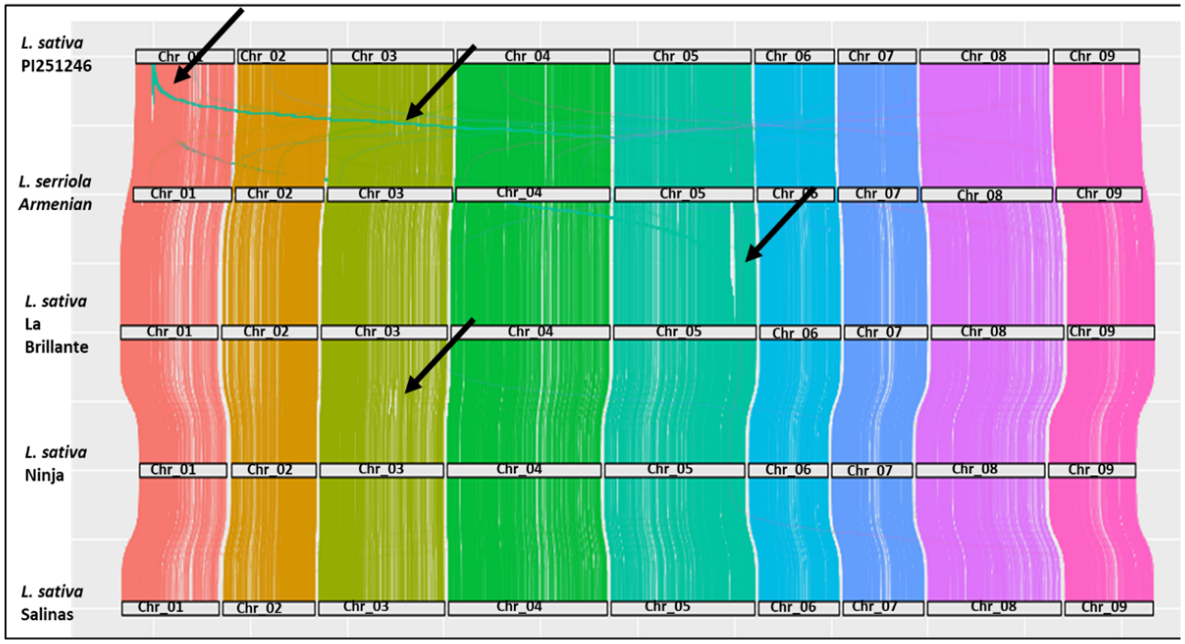

Figure. 4.5 Genome wide synteny across lettuce pangenome

4.4.3 Genome-wide distribution of SVs

To identify different types of genetic variation between the assembled lettuce genomes, five of the additional assembled genomes were aligned to the v11 reference genome; VIAE was not included because its assembly was not fully polished due to the lack of Illumina reads. Numerous SVs were detected as distributed across all chromosomes; however, there tended to be an increase in regions enriched with repeats and in regions of segmental duplications. More deletions were detected than insertions (Table 4.1). More genomic regions (bps) affected by SVs were detected in US96UC23. Of the domesticated genotypes, Ninja had slightly lower SVs compared to the other genotypes.

Table 4.1. *Frequency of structural variations (SVs) between wild and domesticated lettuce genotypes and the L. sativa (v11) reference assembly.*

Marked in blue are the wild accessions of lettuce

Genotype	*L. sativa* cv. La Brillante	*L. sativa* cv. Ninja	*L. sativa* cv. PI251246	*L. serriola* US96UC23	*L. serriola* Armenian
SVs					
Total SVs	5,758	4,079	5,350	6,212	10,893
Insertions	5,689	4,059	5,339	6,182	10,858
Deletions	62	19	9	29	32
Total indels	5,751	4,078	5,348	6,211	10,890
Translocations	5	0	1	1	3
Inversions	2	1	1	0	0

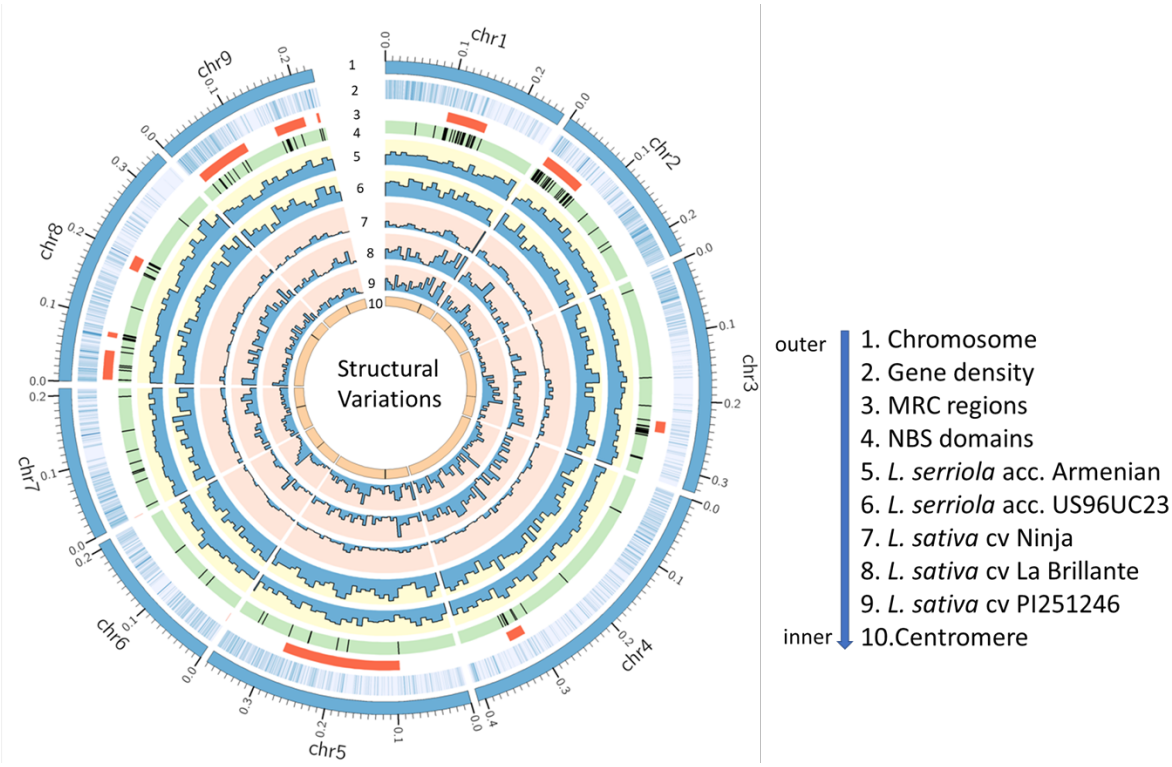

Figure 4.6 *Density and chromosome distribution of structural variations.*

4.4.4 Evaluation of methods for generating pangenome graphs of lettuce

Three graph-based approaches, minigraph, pggb, and vg, were evaluated for generating a pangenome of lettuce. The v11 reference assembly was used as the backbone of the graphs, and the order for including the six other assemblies was determined based on the mash distance between the genotypes. To limit the computational complexity and increase the ease of analysis, all graphical approaches were evaluated using one chromosome of lettuce. The outputs of the graphs were visualized using Bandage, IGV, or odgi.

Table 4.2 *Comparison of three pangenome graphs constructed for lettuce Chromosome 2.*

	minigraph	pggb	vg
Total nodes	102,749	30,361,801	37,095,225
Node length	332,256,189	745,218,484	793,152,311
Reference node	66,867	14,342,234	14,792,779
Reference node length (bp)	236,378,258	236,378,258	229,547,282
Non-Reference node	35,882	16,019,567	22,302,446
Non-Reference node length (bp)	95,877,931	508,840,226	563,605,028
CPU time (hrs)	5.2	96.37	104.2
Wall-clock time (82 threads)	0.06	7.3	8.8

Minigraph generated pangenome graphs with less computational wall-time (5.2 hours) compared to the pggb and vg pipelines (7.3 hours and 8.8 hours, respectively) because minigraph avoids base-level alignments of sequences and instead uses minimizer-

mapping for alignment extensions in minimap2. Both the pggb and vg pipelines use base-level alignments; thus, they demand more time for graph generation. The pggb and vg pipelines produced pangenomes that contained a total of 16 M and 22 M of non-reference nodes, respectively, which carry mostly SVs longer than 50 bp. They also contained 508 Mb in pggb and 563 Mb in vg of non-reference bases due to the presence of single nucleotide polymorphisms (SNPs) and indels <50 bp. The pangenome graphs produced using pggb and vg contained a larger number of variations compared to the graph constructed with minigraph. Most of these variants were missing in the minigraph pangenome and were present in highly repetitive regions of the genome. However, SVs larger than 50 bp identified by all pangenome approaches were mostly similar.

4.4.5 Minigraph-based comparison of ONT and HiFi-based assemblies of lettuce

Minigraph was used to generate graphs from the v10 and v11 assemblies of *L. sativa* to examine differences between these two assemblies. The HiFi-based assembly v12, an assembly developed from a subset of PCR-free reads used in the construction of the v11 assembly, was included in this analysis. The alignment of v11 and v12 revealed very few indels outside of the rDNA region, reflecting the high accuracy of the independent HiFiasm assemblies with an average segment length of 1,427,637 bp (Fig. 4.7a). In contrast, the v10 ONT-based and v11 HiFi-based graphs differed much more frequently throughout the length of the chromosome with an average segment length of 856,847 (Fig. 4.7b). This reflects the higher rate of indel errors from nanopore sequencing.

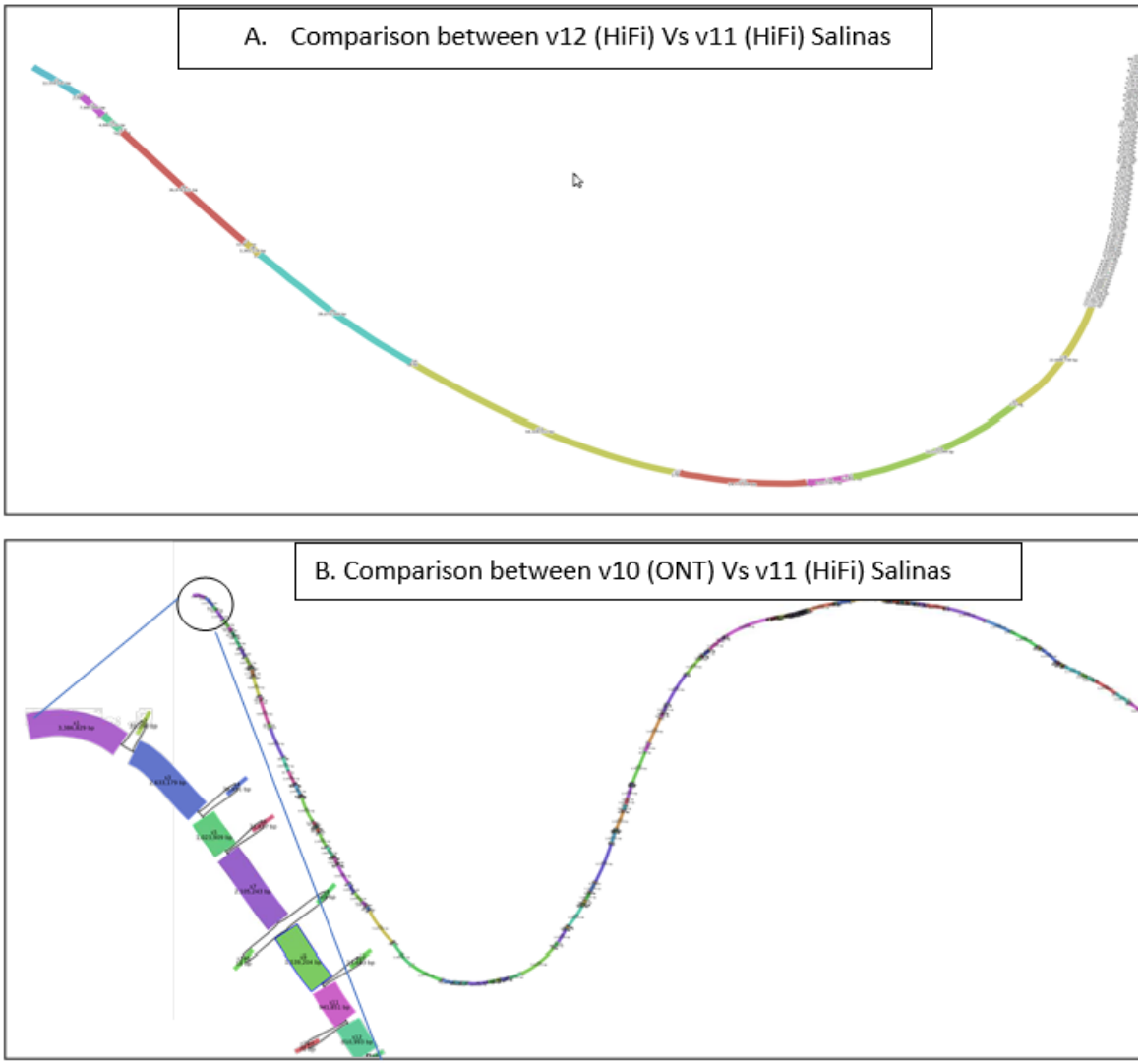

Figure 4.7. *Graphs of HiFi and ONT-based assemblies of chromosome 1 of cv. Salinas generated using minigraph.*

A. The HiFi v11 assembly aligned against the HiFi v12 assembly. B. The HiFi v11 assembly aligned against the ONT v10 assembly. The right terminus of the chromosome contains the nucleolar organizer region, which segments the alignment in A but was not detected in B.

4.4.6 Graph-based SVs across lettuce genotypes and visualization

Minigraph was used to identify SVs within the graphical structure. Aligning Chromosome 1 of the *L. sativa* cvs. La Brillante and Ninja assemblies to the v11 reference revealed a total of 7,336 and 5,225 SV events, respectively. Most of these SV events were due to insertions and deletions. Minigraph captured two inversions in La Brillante; these two inversions were visualized at sequence level using Bandage (Figure 4.8).

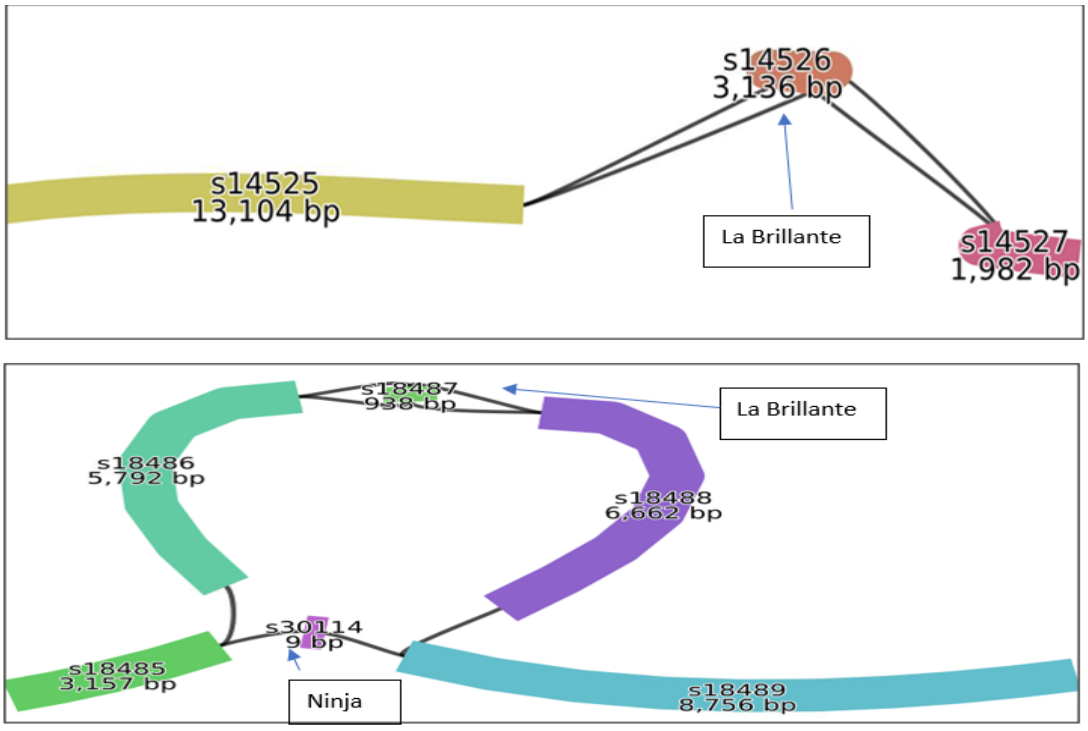

Figure 4.8. *Bandage plots showing inversions in La Brillante and a deletion in Ninja relative to Salinas.*

Inversions are depicted by the two black lines indicating the inversion of a 3,136 bp segment in figure A and 938 bp in figure B. In addition, Ninja has a 9 bp segment in place of a 13,392 bp segment in Salinas and La Brillante.

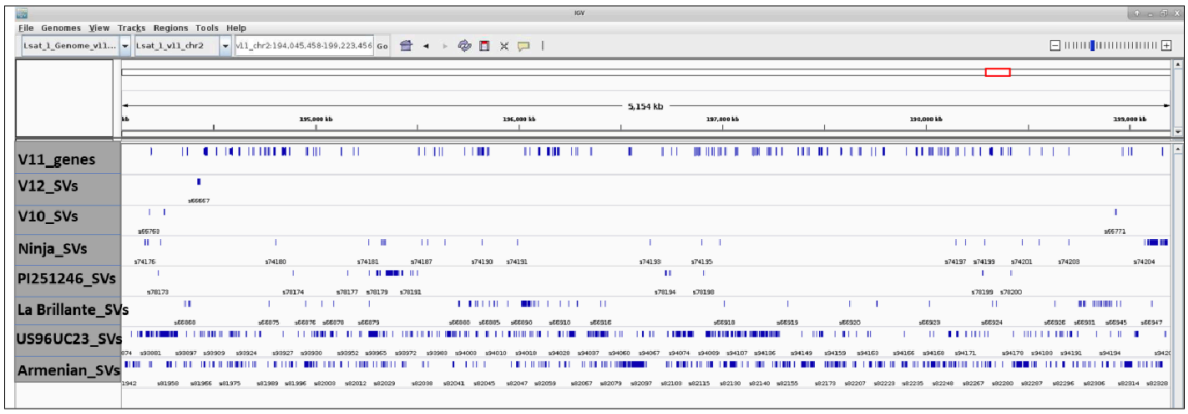

Figure 4.9. *Distribution of structural breakpoints visualized using IGV (chr2: 194,045,458 - 199,223,456)*

The open red box in the top track shows the region of Chromosome 2 displayed below. The first track shows the distribution of predicted genes in the region. The tracks below show the positions of SV breakpoints in the listed genotypes relative to the v11 Salinas assembly.

The breakpoints of SVs detected between minigraph could also be displayed using Integrative Genomics Viewer (IGV). Breakpoints in Chromosome 1 of *L. sativa* cvs. La Brillante, Ninja, and PI251246; *L. serriola* US96UC23 and Armenian 999 relative to the v11 Salinas reference were visualized using IGV. As expected, very few variants were detected between v11 and v12 of cv. Salinas (Fig. 4.9). More SVs were detected between v10 and v11, but variants were still infrequent compared to HiFi and ONT-based assemblies of other genotypes.

4.4.7 pggb-based comparison of wild and domesticated assemblies of lettuce and haploblock detection

The pggb workflow was evaluated to build pangenome graphs using Chromosome 2 of *L. sativa* cv. Salinas v11, La Brillante, Ninja, PI251246 and *L. serriola* US96UC23 and Armenian. In contrast to minigraph, pggb builds reference-free pangenome graphs and performs base level alignments. The workflow for pangenome graph construction with pggb

involves base-level alignment with wfmash, graph induction with seqwish, and normalization with smoothxg (Figure 4.10).

Graphs from pggb were visualized using odgi (Guarracino et al., 2022). Base level comparison on Chromosome 2 between the six lettuce assemblies clearly shows that the wild accession *L. serriola* Armenian 999 has many structural differences at the sequence level. In the region shown in Figure 4.11 even though overall domesticated *L. sativa* genotypes are similar, there are major differences in few regions. La Brillante has additional sequences present in the far-left region in the figure below compared to other domesticated lines. In some regions, La Brillante and Ninja have deletion sequences compared with other lettuce assemblies.

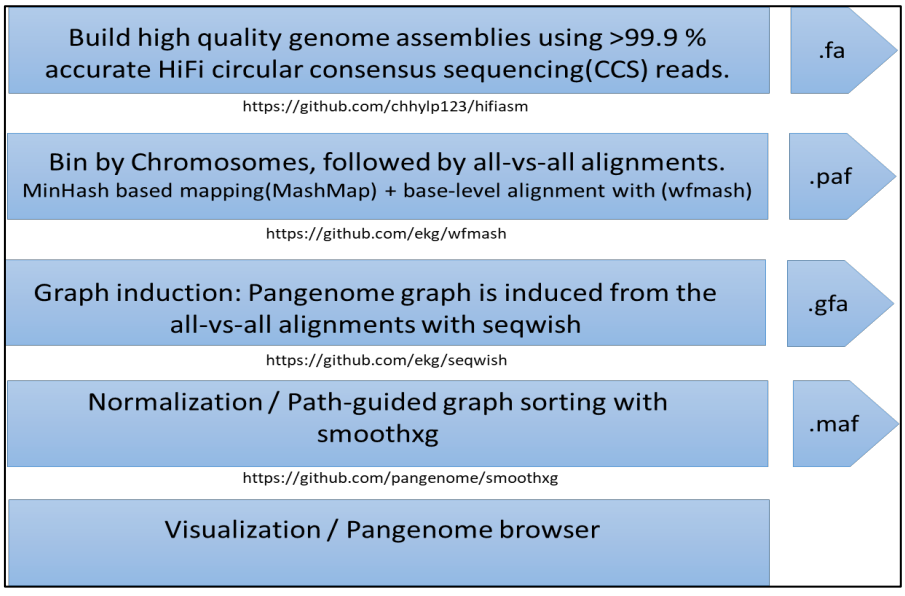

Figure 4.10. *Workflow for constructing a graph pangenome using pggb.*

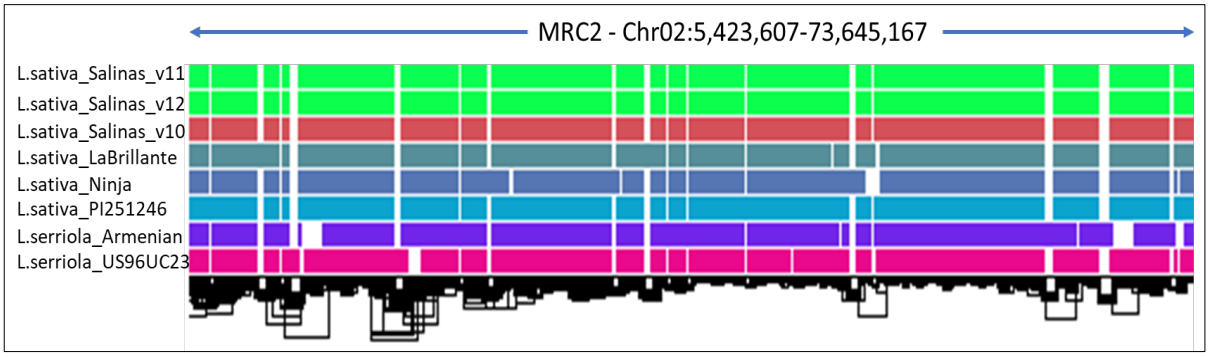

***Figure 4.11**. Comparison of haploblocks across lettuce genotypes using the pggb tool.*

Major resistance cluster region across lettuce genotypes Chromosome 2: 5,423,607 - 73,645,167.

4.5 Discussion

There are multiple approaches to analyzing variation across multiple genome assemblies. This is becoming increasingly important as more complete genomes are assembled and computational approaches to generate and visualize pangenomes are evolving rapidly. In this chapter, I evaluated pairwise presence/absence variation and three graph-based approaches for generating pangenomes of plants including those with large genomes like lettuce.

Direct comparison between multiple linear references were used to compare and contrast the gene and sequence variations between the different genotypes of lettuce. Gene clustering using orthogroup classification enabled us to identify genes that are unique or present in one specific genotype of lettuce and absent in others. Understanding the importance and function of core and variable gene content broadens genomic studies at scale. Similarly, genome-wide alignment and synteny comparison revealed an inversion in Chromosome 3 shared between *L. sativa* cvs. PI251246 and La Brillante relative to the other

genotypes. The next steps will be to corelate these PAVs with phenotypic variation by segregation analyses or GWAS. Causal relationship between a PAV and a phenotype will be the basis for altering the phenotype using genome editing.

The different multi-assembly genome graph approaches (i.e., minigraph, pggb, and vg) have their own advantages and disadvantages. minigraph is highly efficient and robust to build graph genomes, but it avoids base-level alignment across the whole genome and requires a reference back-bone for graph construction. minigraph took a few minutes to analyze one chromosome of lettuce on an eight-core compute node. In contrast to reference-free pangenome approaches, it depends on a high-quality reference and an order of input of the other assemblies. pggb integrates reference-free, base-level alignment to build the genome graphs; however, it is computationally intensive for large repeat genomes like lettuce. pggb took ~10 hours for the same analysis. The vg toolkit, developed by the human pangenome consortium, uses bi-directional graphs and is extremely robust in identifying and capturing all variants from SNPs to large SVs; however, it is computationally intensive, and the resulting graphs are more convoluted and difficult to interpret, especially if the assembly is not of the highest quality. Therefore, given the current state of the field, minigraph will be used to provide an overall view of each chromosome, and then pggb will be used to focus on variation in specific regions of interest.

I also considered other graph pangenome tools, including Pantools (Jonkheer et al., 2022), Cactus (https://github.com/glennhickey/progressiveCactus), and Graphtyper (Eggertsson et al., 2017), each of which has pros and cons. The current version of Pantools did not seem to scale, although it has the potential to be very useful as it evolves. Cactus uses a complicated hierarchical alignment method that results in complex graphs that are difficult

to visualize and interpret. Graphtyper uses a reference and did not have obvious advantages over vg. These and other approaches are evolving rapidly, driven in part by the needs of human genomics. Recently, the Human Pangenome Consortium has shown that a combination of minigraph and Cactus to integrate 350 diverse human genome assemblies increased base-level accuracy in a consensus graph pangenome (Hickey et al., 2022). Currently, deep learning models are being developed to autonomously detect and learn from patterns in the training data.

There is a scarcity of tools available to visualize and conduct downstream analyses using graph genomes. This is also a rapidly advancing area of research. Pangenome visualization can be broadly classified to gene-centric tools and sequence-centric tools (https://pangenome.github.io/). Most gene-centric tools focus on the presence or absence of gene content and grouping genes based on their function such as PanViz (Pedersen et al., 2017). Sequence-centric tools include SNPs, inversions, translocations, tools like Pantograph, Sequence Tube Map, Panache (Durant et al., 2021), MoMI-G (Yokoyama et al., 2019), and Bandage to capture small/large scale sequence variations. I used Bandage, IGV (Thorvaldsdóttir et al., 2013), and odgi. These provided several levels of resolution from the whole chromosome level to the sequence level. Visualization tools are being actively developed to handle the growing number of genomes to assimilate in the consensus graph genome and will be necessary as more lettuce genomes are assembled from HiFi reads.

Plants pose more complex challenges for pangenome construction than humans because of their high repeat content and higher levels of variation at the sequence and structural levels, as well as, in some cases, cytological states such as polyploidy; therefore, pangenome approaches tailored to plants may be needed. Consequently, it will be important

to continuously evaluate the various approaches as they evolve and apply the current versions of programs to tackle the questions being addressed. The graph-based approaches investigated here facilitate future population-scale pangenome variation graph studies in lettuce rather than a linear reference genome. Analysis of specific regions of interest, such as clusters of disease resistance genes, using minigraph and pggb is described in Chapter 5.

References

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6(8) 914–920. Nature Research. https://doi.org/10.1038/s41477-020-0733-0

Bradbury, P. J., Casstevens, T., Jensen, S. E., Johnson, L. C., Miller, Z. R., Monier, B., Romay, M. C., Song, B., & Buckler, E. S. (2022). The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics*, *38*(15), 3698–3702. https://doi.org/10.1093/bioinformatics/btac410

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Curr. Protoc. Bioinform*. https://doi.org/10.1002/0471250953.bi0411s48

Durant, É., Sabot, F., Conte, M., & Rouard, M. (2021). Panache: A web browser-based viewer for linearized pangenomes. *Bioinformatics*, *37*(23), 4556–4558. https://doi.org/10.1093/bioinformatics/btab688

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. v. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, *49*(11), 1654–1660. https://doi.org/10.1038/ng.3964

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1832-y

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*. https://doi.org/10.1038/nbt.4227

Golicz, A. A., Batley, J., & Edwards, D. (2016). Towards plant pangenomics. In *Plant Biotechnology Journal*, 14(4) 1099–1105. Blackwell Publishing Ltd. https://doi.org/10.1111/pbi.12499

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H. R., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms13390

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics*, 36(2), 132–145. Elsevier Ltd. https://doi.org/10.1016/j.tig.2019.11.006

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics, 38,* (13), 3319-3326*.* https://doi.org/10.1093/bioinformatics/btac308

Hickey, G., Monlong, J., Novak, A., Eizenga, J. M., Pangenome, H., Consortium, R., Li, H., & Paten, B. (2022). Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus. Pre-print. https://doi.org/10.1101/2022.10.06.511217

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. In *Nature Reviews Genetics* (Vol. 21, Issue 3, pp. 171–189). Nature Research. https://doi.org/10.1038/s41576-019-0180-9

Hübner, S. (2022). Are we there yet? Driving the road to evolutionary graph-pangenomics. *Current Opinion in Plant Biology*, 66. https://doi.org/10.1016/j.pbi.2022.102195

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02107-y

Jonkheer, E. M., van Workum, D.-J. M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J. R., Berke, L., van der Lee, T. A. J., de Ridder, D., & Smit, S. (2022). PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics*, *38*(18), 4403–4405. https://doi.org/10.1093/bioinformatics/btac506

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5*(59). http://www.biomedcentral.com/1471-2105/5/59

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *BMC Genome Biology, 21*(265). https://doi.org/10.1186/s13059-020-02168-z

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Li, W., Liu, J., Zhang, H., Liu, Z., Wang, Y., Xing, L., He, Q., & Du, H. (2022). Plant pan-genomics: recent advances, new challenges, and roads ahead. In *Journal of Genetics and Genomics* (Vol. 49, Issue 9, pp. 833–846). Institute of Genetics and Developmental Biology. https://doi.org/10.1016/j.jgg.2022.06.004

Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S. S., Zuo, Q., Shi, X. H., Li, Y. F., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*(10), 1045–1052. https://doi.org/10.1038/nbt.2979

Llamas, B., Narzisi, G., Schneider, V., Audano, P. A., Biederstedt, E., Blauvelt, L., Bradbury, P., Chang, X., Chin, C.-S., Fungtammasan, A., Clarke, W. E., et al. (2019). A strategy for building and using a human reference pangenome. *F1000Research*, *8*, 1751. https://doi.org/10.12688/f1000research.19630.1

Nattestad, M., & Schatz, M. C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics*, *32*(19), 3021–3023. https://doi.org/10.1093/bioinformatics/btw369

Pedersen, T. L., Nookaew, I., Wayne Ussery, D., & Månsson, M. (2017). PanViz: Interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics*, *33*(7), 1081–1082. https://doi.org/10.1093/bioinformatics/btw761

Qin, P., Lu, H., Chen, X., Liang, C., & Correspondence, S. L. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, *184*, 3542-3558.e16. https://doi.org/10.1016/j.cell.2021.04.046

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51(1), 30–35. https://doi.org/10.1038/s41588-018-0273-y

Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics*, 21(4), 243–254. https://doi.org/10.1038/s41576-020-0210-7

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W. Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q. Y., Chen, L. L., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of

Brassica napus. *Nature Plants*, *6*(1), 34–45. https://doi.org/10.1038/s41477-019-0577-7

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, *33*, W465. https://doi.org/10.1093/NAR/GKI458

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. v., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13950–13955. https://doi.org/10.1073/PNAS.0506758102/SUPPL_FILE/06758TABLE2.PDF

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., & Kasahara, M. (2019). MoMI-G: Modular multi-scale integrated genome graph browser. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-3145-2

Yu, J., Tehrim, S., Zhang, F., Tong, C., Huang, J., Cheng, X., Dong, C., Zhou, Y., Qin, R., Hua, W., & Liu, S. (2014). Genome-wide comparative analysis of NBS-encoding genes between Brassica species and Arabidopsis thaliana. *BMC Genomics*, *15*(1), 1–18. https://doi.org/10.1186/1471-2164-15-3/FIGURES/5

Zanini, S. F., Bayer, P. E., Wells, R., Snowdon, R. J., Batley, J., Varshney, R. K., Nguyen, H. T., Edwards, D., & Golicz, A. A. (2022). Pangenomics in crop improvement—from coding structural variations to finding regulatory variants with pangenome graphs. *Plant Genome*, 15(1). https://doi.org/10.1002/tpg2.20177

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*(2), 278–284. https://doi.org/10.1038/s41588-018-0041-z

Zhao, X., Collins, R. L., Lee, W. P., Weber, A. M., Jun, Y., Zhu, Q., Weisburd, B., Huang, Y., Audano, P. A., Wang, H., Walker, M., Lowther, C., Fu, J., et al. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *American Journal of Human Genetics*, *108*(5), 919–928. https://doi.org/10.1016/J.AJHG.2021.03.014

Chapter 5: Analysis of structural rearrangements and gene diversity in genomic regions encoding clusters of resistance genes

5.1 Abstract

Resistance (R) genes play a key role in plant defense. R-genes directly or indirectly detect pathogens and elicit the activation of innate immune responses. The largest R-gene family in plant genomes encodes nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins. Another large gene family encodes receptor-like proteins (RLPs). R-genes are often found in clusters as tandem arrays across the genome and are prone to structural rearrangements. In this chapter, seven chromosome-scale annotated assemblies of both wild and domesticated lettuce genotypes were analyzed for NBS-encoding genes and their relationship to the major clusters of phenotypic resistance genes. Using resources and approaches described in previous chapters, I identified and characterized between 454 and 485 canonical NBS-LRR encoding genes in *L. sativa* cvs. Salinas, Ninja, VIAE, and PI251246 as well as *L. serriola* accessions Armenian and US96UC23. NBS-LRR encoding genes and structural modifications within eight major resistance cluster (MRC) regions of the wild and domesticated assemblies were classified and visualized using graph-based pangenome approaches. Variation in MRCs 1, 2, and 9 were analyzed in detail. In addition, structural variation at the *Verticillium* (*Ve*) locus that confers resistance to *V. dahliae* was analyzed relative to genes conferring RLPs. These data provide a catalog of variation at MRCs, which are dynamic regions of the genome, and will inform breeding for disease resistance in lettuce.

5.2 Introduction

Plant disease resistance (R) genes encoding NLRs and RLPs play an important role in providing resistance to plant pathogens. NLR and RLP proteins directly or indirectly recognize pathogen effectors and trigger downstream immune responses (reviewed in Chapter 1). The NLR superfamily is encoded by diverse genes within a genome and is subdivided into two main classes: the TIR-domain-containing (for TOLL/INTERLEUKIN LIKE RECEPTOR/RESISTANCE PROTEIN; TIR-NB-LRR or TNL) and the non-TIR-domain-containing (NB-LRR or NL), including coiled-coil domain-containing (CC-NB-LRR or CNL) R-protein subfamilies (Kuang et al., 2004; Zipfel, 2008). NLRs are under strong selection pressure. Consequently, accessions from the same species can display a large NLR copy number and sequence variation due to duplication, deletion, and unequal crossing-over (Thind et al., 2018). Resistance genes with NBS or NBS-LRR domains are significantly enriched in structural variations (SVs) as shown in wheat, tomato, *Brassica*, and lettuce (Plocik et al., 2004; Alonge et al., 2020; Bayer et al., 2020, 2022; Montenegro et al., 2017). TNL and CNL encoding genes are clustered in lettuce within eight major resistance clusters (MRCs) of phenotypic disease resistance genes, which include many *Dm* genes (Parra et al., 2016). While there is significant genetic variation in NLR-encoding genes within domesticated lettuce, pathogens, particularly *B. lactucae*, have evolved to overcome resistance, rendering many resistance genes ineffective. The loss of resistance highlights the need for new NLR genes and an understanding of the complexity of SVs at MRCs.

Developing lettuce cultivars resistant to diseases is one of the most important objectives in lettuce breeding. There are several important diseases of lettuce caused by a variety of pathogens. Downy mildew caused by the oomycete *Bremia lactucae* can result in

significant economic losses (Parra et al., 2016). Lettuce immunity against specific races of *B. lactucae* is often obtained by introducing dominant resistance genes encoding nucleotide-binding site leucine-rich repeat (NLR) receptor proteins from wild relatives (Christopoulou et al., 2015), such as *L. serriola*, *L. saligna*, and *L. virosa* (Treuren et al., 2013). Verticillium wilt is another important disease that is caused by *V. dahlia* (Hayes et al., 2011). Resistance to race 1 of *V. dahliae* is conferred by the *Ve1* gene, which encodes a receptor-like protein (RLP) with extracellular leucine-rich repeats (Inderbitzin et al., 2019).

In this chapter, the resources and approaches described in previous chapters were used to characterize SVs at MRCs at several levels of resolution. I first identified and classified NLRs and R gene candidate families. The availability of seven high quality reference genomes allowed a survey of NLR-related SV landscapes in both wild and domesticated lettuce. Pangenome graphs were developed for three MRCs. The development of haploblocks using the lettuce pangenome enabled the discovery of conserved vs. divergent clusters of NLR genes. The *Ve* resistance locus was also examined using a graph-based approach. This revealed extensive structural variation at the locus but conservation of *Ve* paralogs. This pangenome approach helped us to identify and characterize a more complete repertoire of NLR genes and understand variation in diverse lettuce genomes.

5.3 Materials and Methods

5.3.1 Genome assembly and annotation

Six representative *de novo* assemblies and annotations of lettuce genotypes and the v11 reference genome of lettuce were used to study the structural rearrangements and distribution of NLR genes; details of the assemblies and annotations were described in detail

in Chapters 2 and 3. Two of the seven *de novo* assemblies (*L. sativa* cvs. Salinas and La Brillante) are PacBio HiFi based and five assemblies (*L. sativa* cvs. Ninja, VAIE, and PI251246; *L. serriola* US96UC23 and Armenian 999) are Oxford Nanopore based assemblies. These represent the domesticated and wild genotypes of lettuce used for this study.

5.3.2 NLR gene identification and classification

HMMER (Wheeler & Eddy, 2013) was used to search Hidden Markov Models (HMMs) for identifying domains for NLR encoding genes. All HMMs were attained from the Pfam website (http://pfam.xfam.org/) or NIBLRRS (http://niblrrs.ucdavis.edu/At_RGenes/). The models included PF00931 for the NBS domain, PF01582 and PF13676 for TIR, PF05659 and PF18052 for CC, and eight HMMs for the LRR domain (PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13855, PF14580). In addition to HMM searches, genes with NB domains were identified by InterProScan, and CC motifs predicted by Paircoil (McDonnell et al., 2006) were integrated with the HMMER output.

All identified candidate NLR genes from each lettuce genotype were further validated and classified into different categories by structures, determined through protein translation of genomic sequences, protein–protein searches, and alignment in the Plant Resistance Genes database (PRGdb) using the Disease Resistance Analysis and Gene Orthology tool (DRAGO 2).

DRAGO (prgdb.org/prgdb/drago2) was executed with peptide sequence file of *L. sativa* cv. Salinas (v11) as an input to define the normalization value and the minimum score thresholds. DRAGO 2 detects LRR, kinase, NBS, and TIR domains using hidden Markov models (HMMs) with the HMMER v3 package web service (ebi.ac.uk) and computes the

alignment score of the different hits based on a BLOSUM62 matrix. The normalization value was the absolute smallest similarity score found among the input sequences considering all domains. The minimum score thresholds were calculated from the smallest similarity score reported in a specific domain among the input sequences. DRAGO 2 generated files with numeric matrix that represents the similarity score of every single protein input to their HMM profile, the domain name, start position, end position, resistance class and identification for every putative plant resistance protein.

5.3.3 SV analysis and comparative genomics across wild and domesticated lettuce genotypes

The SV workflow is described in detail in Chapter 4. To analyze the distribution of SVs, the six assemblies were mapped onto the v11 reference of cv. Salinas using minimap2 with default parameters (Li, 2021). Two long-read SV callers were used: Sniffles v1.0.12 (Sedlazeck et al., 2018) and CuteSV v.1.0.10 (Jiang et al., 2020) with default parameters. Assemblytics (Nattestad & Schatz, 2016b) was also applied to the genome alignments generated using Mummer v4.0.0rc1 with default parameters. All the SVs from the three SV callers were merged using SURVIVOR v.1.0.6; only calls supported by at least two callers and where the callers agreed regarding the type of variant were reported. Distribution of SVs were visualized using IGV (Thorvaldsdóttir et al., 2013). A pangenome graph-based method to call SVs was also implemented. Minigraph v0.12 was used to call SVs from pangenome alignments of the assemblies and the resulting distribution of SVs were visualized using Bandage (Wick et al., 2015) as described in Chapter 4.

5.3.4 Graph-based pangenome construction for three major resistance cluster regions of lettuce

As described in detail in Chapter 4, for the construction of a graph-based pangenome, minigraph (Li et al., 2020) and Pangenome Graph Builder (pggb) (Guarracino et al., 2022; Hickey et al., 2022) were primarily used to construct graphs. To limit the computational complexity, most evaluations were based on individual chromosomes. The graphs were extracted based on the coordinates of the MRCs defined phenotypically (Christopoulou et al., 2015) to focus visualization on the MRC regions of specific chromosomes.

5.4 Results

5.4.1 NLR genes relative to MRCs in wild and domesticated lettuce genotypes

Phenotypic resistance genes to multiple pathogens are located in MRCs in the lettuce genome (as discussed in Chapter 1). A genome-wide analysis of NBS-encoding genes was conducted across wild and domesticated lines of lettuce using a combination of several approaches, HMM-based search, drago2 analysis, and manual curation. The clusters of predicted NLR genes matched with the MRCs on Chromosomes 1, 2, 3, 4, 5, 8, and 9. A total of 380,431 protein coding genes were used for this analysis. Between 435 and 485 putative NBS-encoding genes were identified across all genotypes of lettuce (Table 5.1). The wild genotypes, *L. serriola* Armenian 999 and US96UC23 had slightly more predicted NLR loci, 470 and 485, respectively, in comparison with the domesticated cultivars analyzed. The

number of RLK-encoding genes ranged from 248 to 295, and putative RLP genes lacking a kinase domain numbered between 131 and 427.

Table 5.1. Classification of R genes in different genotypes of lettuce

Lettuce genotype	Number of NBS-encoding genes					Total NBS	RLK-Kinase	RLP-LRR
	NLR	CNL	TNL	CN	TN			
L. sativa cv. Salinas	263	57	133	220	240	460	282	368
L. sativa cv. La Brillante	194	40	89	186	134	435	278	216
L. sativa cv. Ninja	303	75	151	236	242	456	295	327
L. sativa VIAE	246	36	115	226	383	459	248	427
L. sativa PI251246	197	55	78	199	108	454	248	131
L. serriola US96UC23	220	49	101	214	154	485	269	224
L. serriola Armenian 999	262	47	162	216	268	470	287	247

NBS-encoding genes were present in all chromosomes and enriched but not exclusively located within MRCs (Figures 5.1 and 5.2). A contingency chi-square confirmed that the MRC regions were enriched for NBS-encoding genes ($X^2 = 585.29$, $p < 2.2x10^{-16}$). The degree of enrichment varied between MRCs; large numbers of NBS-encoding genes were located within MRC1 and MRC2 with 108 and 73 respectively, while there were only a small number of NBS-encoding genes predicted within MRC5 with only 14 NBS-encoding genes. Genes containing both NBS and LRR domains had a similar distribution to genes identified as containing an NBS domain. In contrast, genes identified as encoding LLRs were more common and did not show obvious enrichment within MRCs (confirmed by $X^2 = 30.222$, $p = 0.3852$). Genes encoding CC and TIR domains showed intermediate distributions.

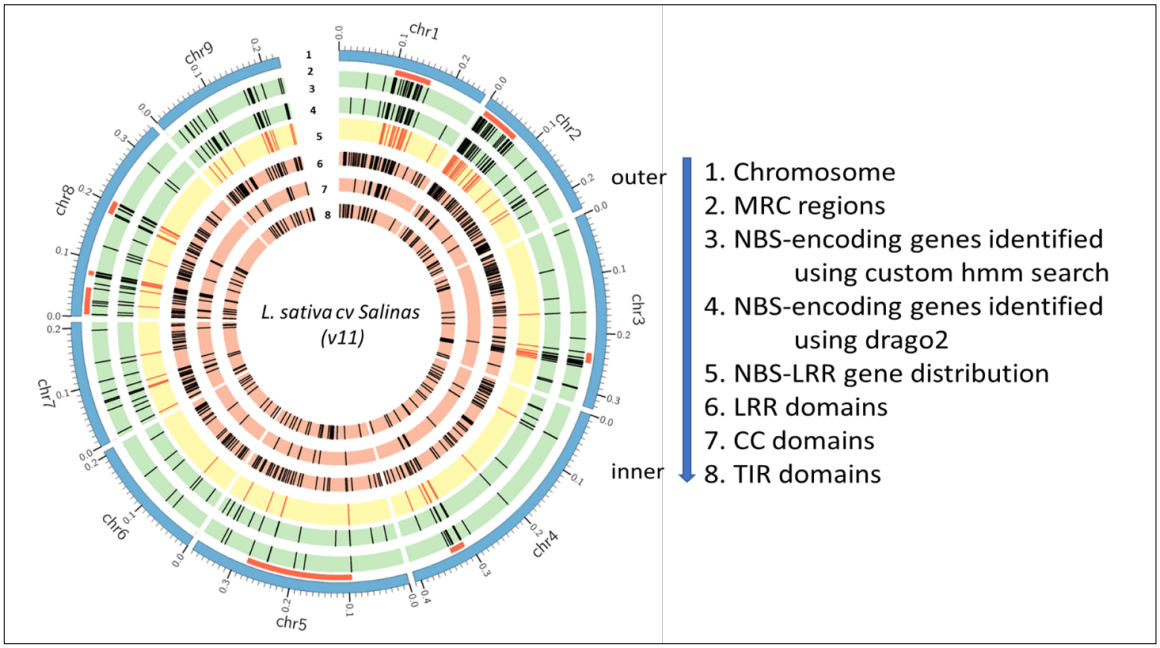

Figure 5.1 *Distribution of putative NLR genes in the v11 reference assembly of L. sativa cv. Salinas.*

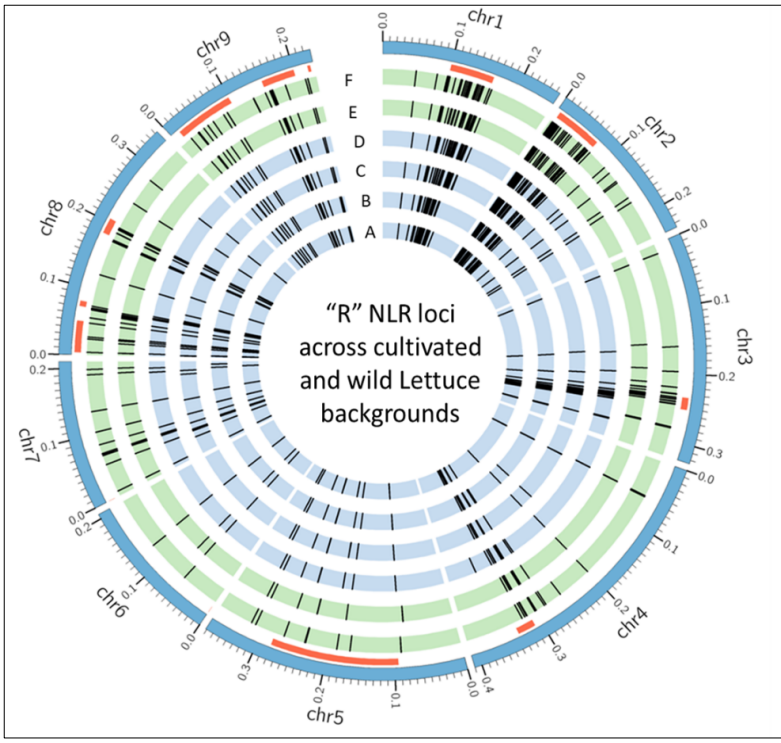

Figure 5.2 *Distribution of predicted NLR loci across the nine chromosomes of wild and domesticated lettuce genotypes.*

Vertical black lines in each track indicate the location of NBS-encoding genes in the different genotypes of lettuce. Blue tracks: domesticated species. Green tracks: wild species. The locations of the phenotypically defined major resistance clusters are shown in red.

Even though the NLR genes showed generally conserved distribution patterns across chromosomes among the cultivars (Figure 5.2), differences were evident in the number of predicted genes and distinct patterns unique for a genotype when viewed at higher resolution as illustrated for *L. sativa* cv. Ninja and cv. Salinas (Figure 5.3).

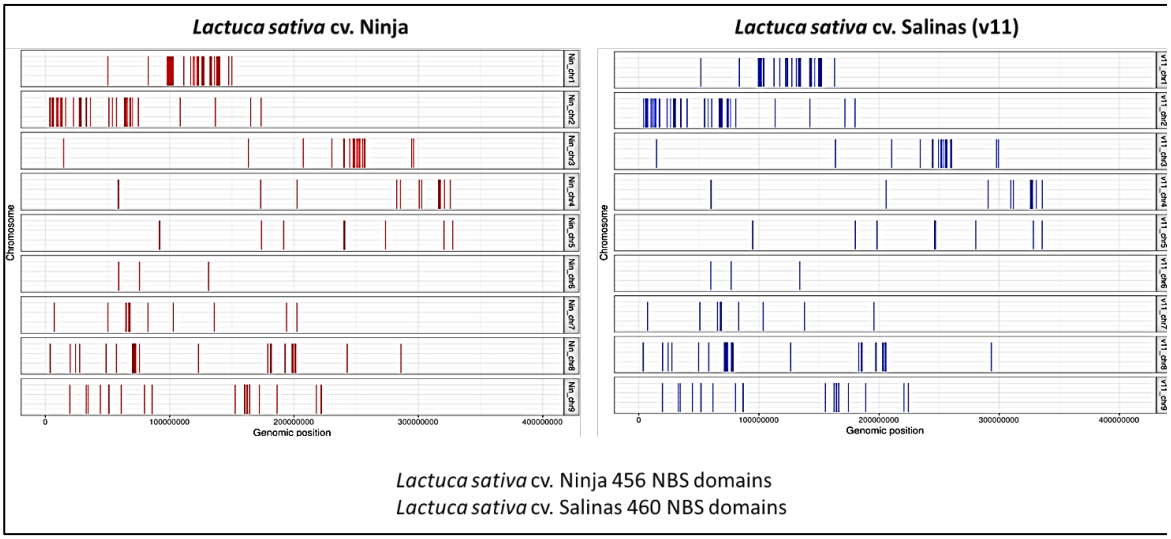

Figure 5.3 *Predicted NBS-encoding gene distribution across two domesticated lettuce genotypes, L. sativa cvs. Ninja and Salinas.*

Vertical red and blue lines indicate the location of NBS-encoding genes along the nine chromosomes of cvs. Ninja and Salinas, respectively.

5.4.2 SVs underlying MRCs in wild and domesticated lettuce genotypes

SVs were a major type of variation within MRCs. The genome-wide distribution of SVs in five wild and domesticated genotypes in comparison with the v11 reference assembly was described in Chapter 4. Of the total 32,292 SVs detected across the five genotypes, 6,979 large SVs (>100 bp) were present in the MRC regions, the majority of which were insertions (99%). This is similar to the genome-wide average of 12 per Mb, assuming a total combined size of the eight MRCs of 629 Mb.

Table 5.2. *SVs underlying MRC regions in five genomes of Lactuca spp. and the number of genes affected by SVs***.**

Lettuce genotype	Deletion	Duplication	Insertion	Inversion	Translocation	Total		Total Genes in SV region
L. sativa cv. La Brillante	10	0	1,461	0	0	1,471		26
L. sativa cv. Ninja	7	0	750	0	0	757		19
L. sativa PI251246	3	0	1,259	0	0	1,262		24
L. serriola US96UC23	8	0	2,190	0	1	2,199		23
L. serriola Armenian	7	0	1,283	0	0	1,290		18

5.4.3 NLR gene and SV distribution in MRC 1

MRC1 contains *Dm5/8, Dm10, Dm17, Dm25, Dm36, Dm37, Dm43*, and *Dm45*, as well as *Tu* and *Mo2* for resistance to Turnip Mosaic Virus (*TuMV*) and Lettuce Mosaic Virus (LMV), respectively. It also contains two QTLs, qFUS1.1 and qFUS1.2, for resistance to wilt caused by *Fusarium oxysporum* f. sp. *lactucae*. Prediction of domains/motifs in MRC1 that encode NBS genes across the assemblies is shown in Figure 5.4. The v10, v11, and v12 assemblies of cv. Salinas were identical, indicating that these potentially challenging areas of the genome had been assembled well using both ONT and PacBio HiFi sequences. The distribution of NLRs in Ninja and VIAE are distinct, reflecting introgressions from *L. saligna* and *L. virosa*, respectively. PI251246, the oil seed accession, is distinct, reflecting an ancient separation of the lineages from the leafy types. The two *L. serriola* lines also have unique patterns.

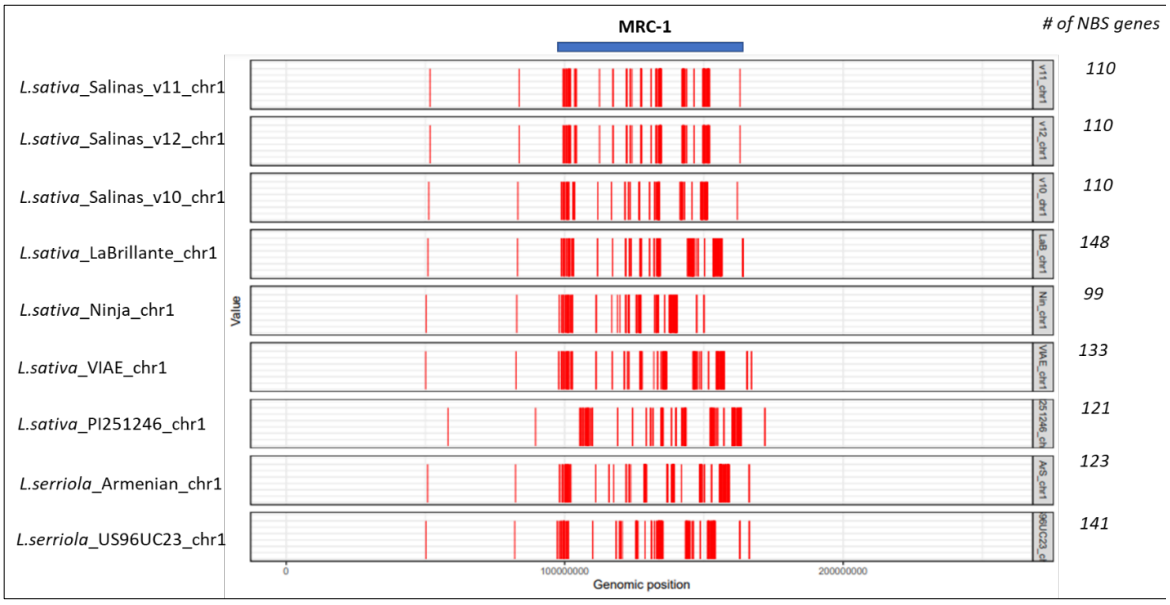

Figure 5.4 *Predicted NBS gene distribution in the MRC1 region.*

Vertical red lines indicate the location of NBS-encoding genes across the different genotypes of lettuce in Chromosome 1. Dark blue box shows the MRC 1 region (Chr01:99,435,657-162,953,765) as defined by the phenotypic resistance genes.

Pangenome graphs were used to visualize the structural variations and SNPs in the MRC1 region (Figure 5.5). Again, the v10, v11, and v12 assemblies of cv. Salinas were highly similar, although not identical at one end. Comparison of the genotypes clearly shows a distinct pattern for Ninja compared to the other domesticated lines, which results from the introgression of resistance from *L. saligna* that maps to this region (McHale et al., 2009; Wroblewski et al., 2007). There was also considerable variation between the two *L. serriola* genotypes and cv. Salinas but less than that observed with cv. Ninja. VIAE was not included in this analysis because the assembly had not been polished and its inclusion would have artificially inflated the number of nodes.

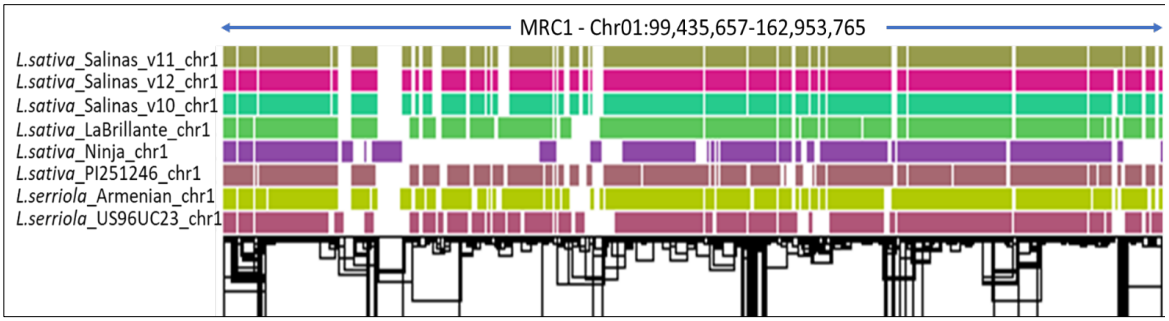

Figure 5.5 *Pangenome graph using pggb workflow showing the haplotype blocks across MRC1 – Chr01:99,435,657 to 162,953,765 bp.*

Alignment of all assemblies was performed using pggb. Each color represents one haplotype in the pangenome graph. Eight haplotypes are shown including three assemblies of cv. Salinas. Coordinates are relative to the v11 reference genome. Each column represents one panBlock or haploblock. A colored haploblock represents presence; no color represents absence. Black lines at the bottom represent the topology of the graph with edges connecting the nodes.

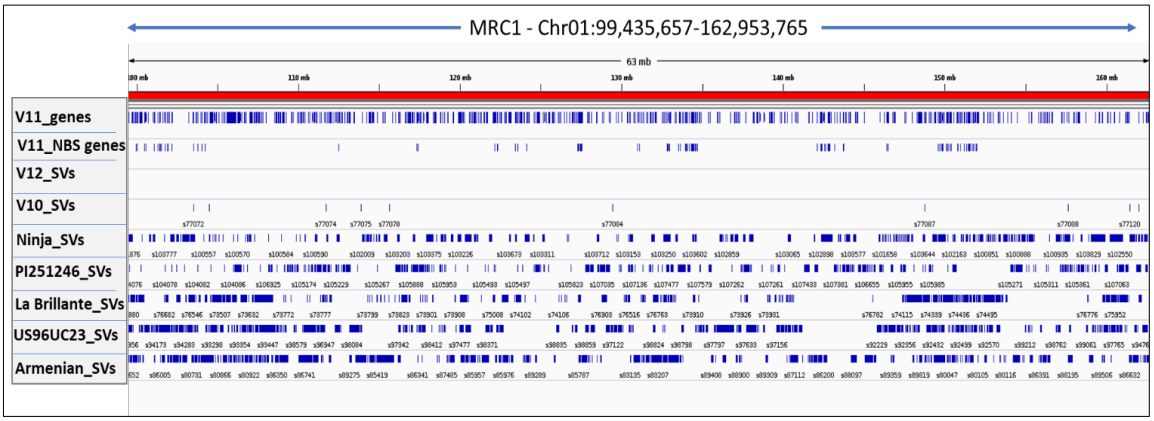

Figure 5.6 *Pangenome graph-based SV distribution with putative NLR genes in MRC1 region Chr01:99,435,657-162,953,765.*

The red bar shows the MRC1 region. The first track shows the distribution of all the predicted genes in the region. The track below shows the location of NBS-encoding genes. The tracks below show the positions of SV breakpoints in the listed genotypes relative to the v11 assembly of cv. Salinas.

The distribution of SVs relative to NLRs was investigated using minigraph (Figure 5.6). SV distribution in the MRC1 region was high in wild accessions compared to the domesticated lines analyzed

5.4.4 NLR gene distribution and SV density relative to MRC2

MRC2 includes *Dm1*, *Dm2*, *Dm3*, *Dm6*, *Dm14*, *Dm15*, *Dm16*, *Dm18*, *Dm50*, and *DMR2.2,* along with *Tvr* for resistance to Tomato Bushy Stunt Virus (TBSV), *Ra* for root aphid resistance, and qANT1 for resistance to anthracnose (Christopoulou et al., 2015). In the MRC-2 region, 81, 103, 72, 77, 84, 79, and 92 NLR genes were predicted in *L. sativa* cv. Salinas (which spans 68.2 Mb as defined by the phenotypic resistance genes), La Brillante, Ninja, VIAE, PI251246, and *L. serriola* Armenian 999 and US96UC23, respectively (Figure 5.8). La Brillante carries a greater number of predicted NLR genes in MRC2 compared to the other genotypes included in this genome-wide comparison.

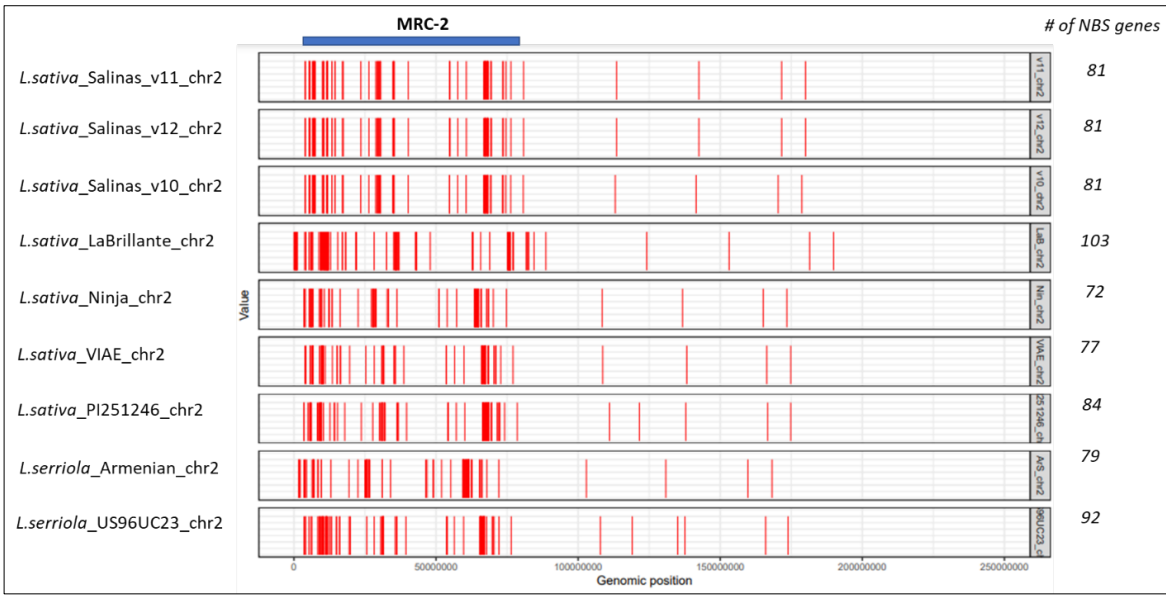

Figure 5.7. *Distribution of predicted NBS-encoding genes in the MRC2 region.*

Vertical red lines indicate the location of NBS-encoding genes across the different genotypes of lettuce in Chromosome 2. Dark blue box shows the MRC 2 region (Chr02: 5,423,607 to 73,645,167) as defined by the phenotypic resistance genes.

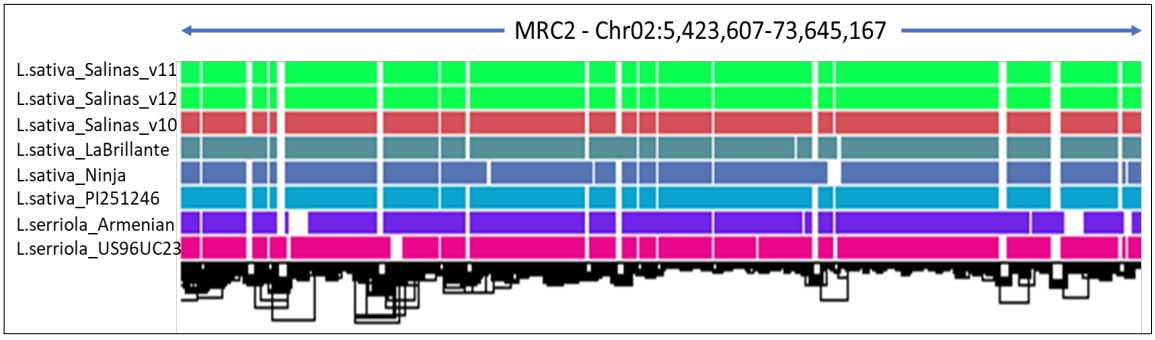

Figure 5.8 *Pangenome graph using pggb workflow showing the haplotype blocks across Chromosome 2-MRC, Chr02: 5,423,607 to 73,645,167 bp.*

Alignment of all assemblies was performed using pggb. Each color represents one haplotype in the pangenome graph. Eight haplotypes are shown including three assemblies of cv. Salinas. Coordinates are in reference to the v11 assembly of Salinas. Each column represents one panBlock or haploblock. Colored haploblocks represents presence; empty represents absence. Black lines in the bottom represents the topology of the graph with edges connecting the nodes.

The distribution of SVs and NLRs in the MRC2 region were again analyzed using minigraph (Figure 5.9). SV distribution in the wild genotypes Armenian 999 and US96UC23 shows high levels structural modifications, compared to the domesticated genotypes. This pangenome approach again revealed that cv. Ninja has more SVs in the MRC2 region in comparison to the other domesticated genotypes analyzed.

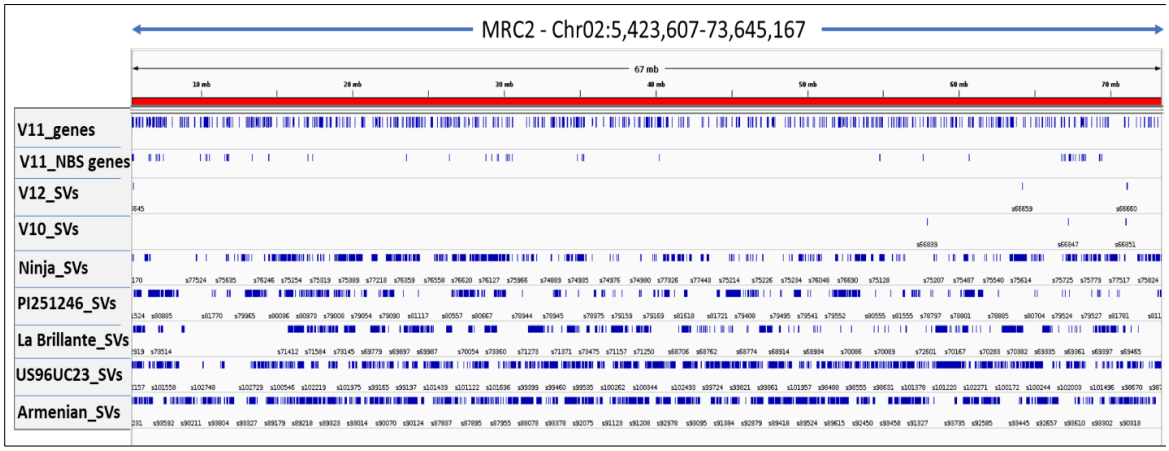

Figure 5.9 *Pangenome graph-based SV distribution with putative NLR genes in MRC2 region Chromosome 2-MRC – Chr02:5,423,607-73,645,167.*

The red bar shows the MRC2 region that is displayed below. The first track shows the distribution of all predicted genes in the region. The track below shows the location of NBS-encoding genes. The tracks below show the positions of SV breakpoints in the listed genotypes relative to the v11 assembly of cv. Salinas.

5.4.5 SV and *Ve* gene distribution in MRC9

MRC9 contains qDMR9.1, qDMR9.2, and qDMR9.3 for resistance to *B. lactucae* (Christopoulou et al., 2015). MRC9 also encodes resistance to *V. dahlia* that had been mapped to the single dominant Verticillium resistance 1 (*Vr1*) locus in cv. La Brillante (Hayes et al., 2011). The *Vr1* locus is located within MRC9, which in addition to NLR genes contains three genes with sequence similarity to the two *Ve* genes in tomato that encode RLPs, which interact to provide resistance against *V. dahlia* (Kawchuk et al., 2001). Three *Ve* genes (*LsVe*1, *LsVe*2, and *LsVe*3) were identified in the MRC9 region on Chromosome 9 of cv. Salinas, which is susceptible to *V. dahlia* (Sandoya et al., 2021). The high-quality assembly of the resistant cv. La Brillante allowed a multi-level comparison of structural variation between a resistant and a susceptible cultivar. Graph-based approaches were applied to build the haploblocks in the MRC9 region (Figure 5.10). At this level of resolution, MRC9 had fewer major SVs than MRC1 and MRC2. When plotted at greater resolution, many small SVs were revealed, and PI251246 was totally lacking the *Ve* locus (Figure 5.11). The RLP-encoding *Ve* genes are flanked by NLR-encoding genes (Figure 5.12). The Bandage viewer allowed pairwise polymorphisms to be displayed at high resolution (Figure 5.13); this revealed that although there were multiple large indels between cvs. Salinas and La Brillante, both genotypes retained all three *Ve* paralogs, even though there were differences in gene content in between these paralogs.

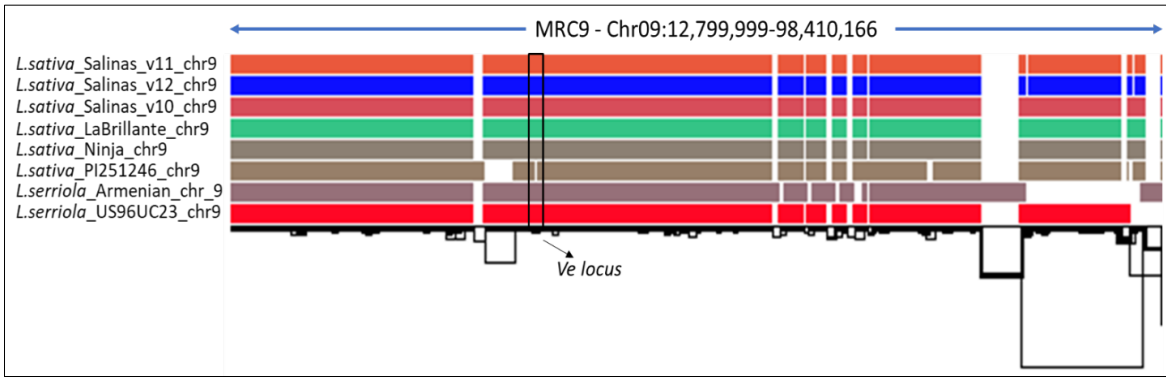

Figure 5.10 *Structural rearrangements at the Ve locus within MRC9, Chr09: 40,940,398 to 41,074,984 bp.*

Alignment of all assemblies was performed using pggb. Each color represents one haplotype in the pangenome graph. Eight haplotypes of lettuce are shown including three assemblies of cv. Salinas. Each column represents one panBlock or haploblock. Colored haploblocks represent presence; empty represents absence. Black lines at the bottom represent the topology of the graph with the edges connecting the nodes. The black box locates the *Ve* locus.

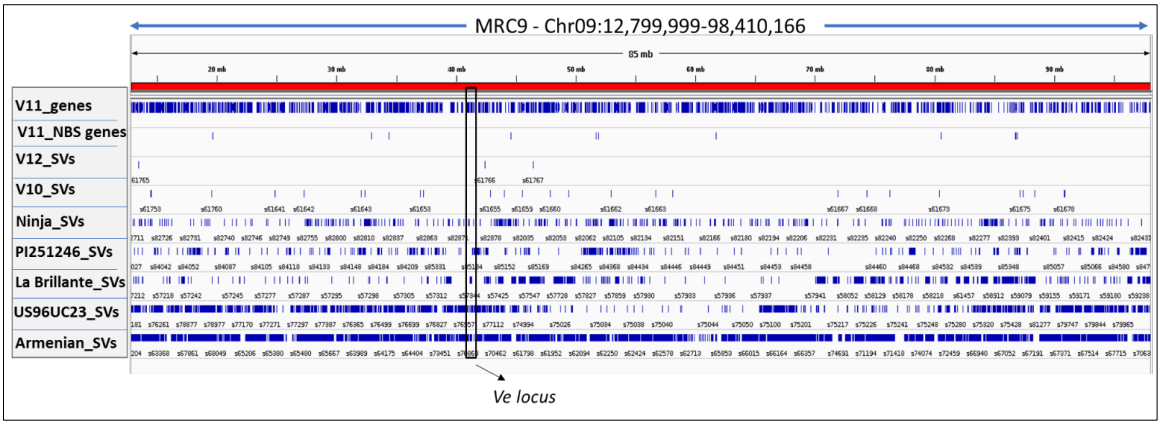

Figure 5.11 *Pangenome graph-based SV distribution with predicted NLR-encoding genes in the MRC9 region, Chr09: 12,799,999 to 98,410,166 bp with Ve locus.*

The open red bar shows the MRC9 region that is displayed below. The first track shows the distribution of all the predicted genes in the region. The track below shows the location of NBS-encoding genes. The tracks below show the positions of SV breakpoints in the listed genotypes relative to the v11 assembly of cv. Salinas. The vertical black box shows the *Ve* locus across all the assemblies.

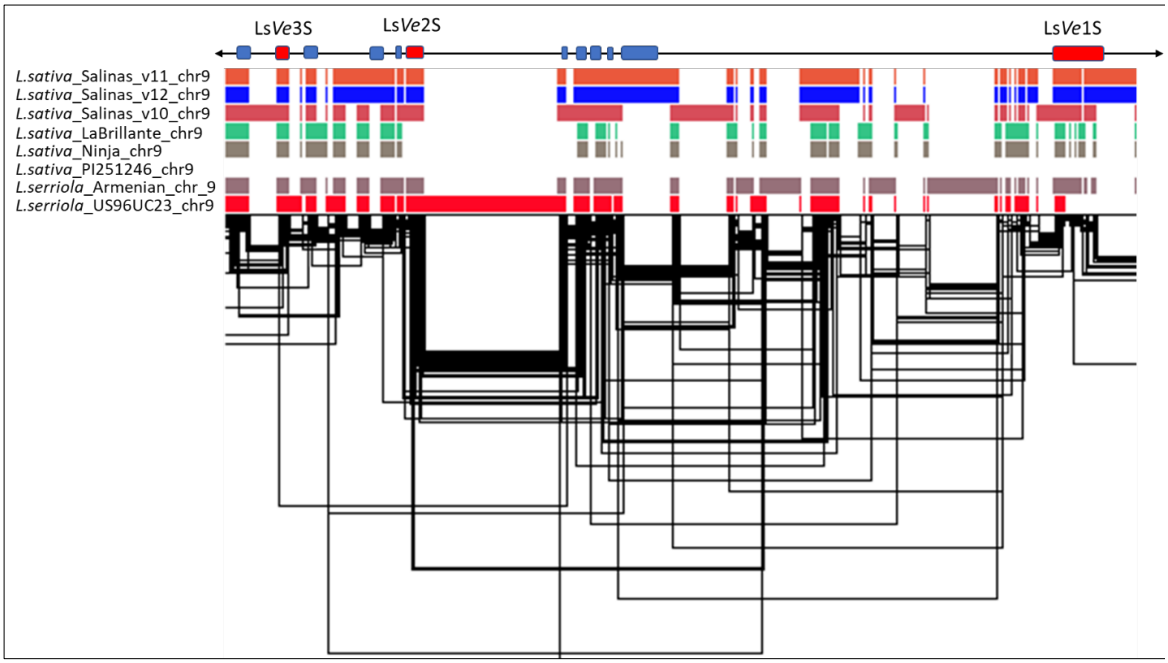

Figure 5.12 *Structural rearrangements at the Ve locus within MRC9, Chr09: 40,940,398 to 41,074,984 bp.*

Schema as for Figure 5.10. The three paralogous *Ve* loci in cv. Salinas are shown as red bars on the top.

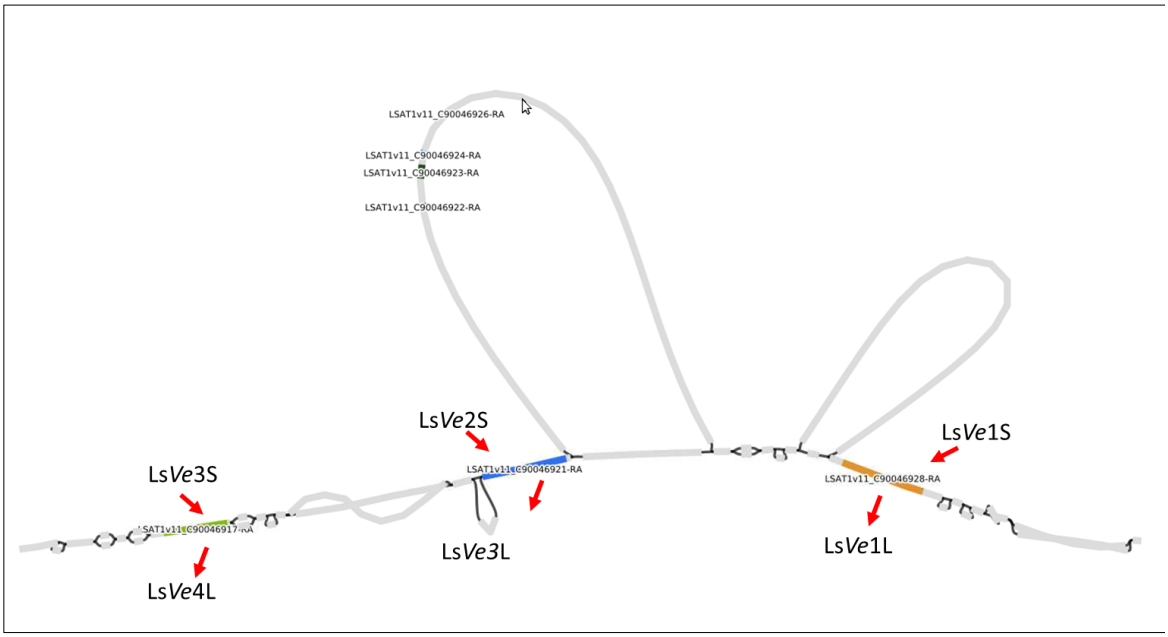

Figure 5.13 *Structural rearrangements between cvs. Salinas (Verticillium susceptible) and La Brillante (Verticillium resistant) genotypes at the Ve locus, Chr09: 40,940,398 to 41,074,984 bp.*

Bandage plots showing the structural differences between the cvs. La Brillante and Salinas genomes. The three conserved paralogs, LsVe1, LsVe2, LsVe3, are shown in brown, blue, and green, nodes, respectively. The large node forming the loops shows the SVs causing absence or divergence of sequence between cvs. Salinas and La Brillante. Marked in the large loop are four genes that are present in cv. Salinas but missing in cv. La Brillante.

5.5 Discussion

The sequencing and assembly of several wild and domesticated *Lactuca* genomes allowed the identification and classification of candidate R-genes within MRCs. I characterized NLR genes by a combination of methods using hmmer-based prediction and protein–protein search with the drago2 (Disease Resistance Analysis and Gene Orthology) tool. With this tool I was able to identify and classify NLR encoding genes before and after repeat masking from a diverse set of lettuce accessions. This consensus approach allowed

me to identify putative NLR encoding genes with high confidence. From this analysis, I predicted 278 genes encoding both nucleotide binding site (NBS) and leucine-rich repeat domains in cv. Salinas, which is similar to the previously reported 294 NBS-LRR genes (McHale et al., 2009). More genes were predicted that encoded the NBS domain but not an LRR domain.

Variable genes caused by SVs are often associated with useful agronomic traits (Ho et al., 2020a; Yuan et al., 2021). Previous studies have reported that NBS-LRR encoding genes that confer different resistance specificities are often clustered in plant genomes (Christopoulou et al., 2015; Fernandez-Gutierrez et al., 2021; Sekhwal et al., 2015). My approach using the graphical pangenome based methods to construct and visualize one to all and all to all alignments of multiple genomes (Hameed et al., 2022) clearly shows the level of sequence divergence between the different genotypes of lettuce in the MRC regions at multiple levels of resolution. The MRC1 region of cv. Ninja had significant differences compared to the other genotypes, reflecting a region of introgression from *L. saligna* (Wroblewski et al., 2007).

The results presented in this chapter align with previous reports of graph-based plant pangenomes that have combined reference and nonreference haplotypes into single graph genomes. In tomato, 238,490 SVs were found in 100 accessions that showed significant expression changes in fruit flavor, size, and yield (Alonge et al., 2020). A graph pangenome was used to map the large chromosomal rearrangements in cucumber that are linked to warty fruits, flowering times, and root growth; this was based on HiFi genome assemblies of 12 cucumber accessions (Li et al., 2022). In maize, pangenomic analysis using high-quality genome assemblies of 66 inbred lines, showed large inversions in regions of disease

resistance genes (Schwartz et al., 2020). A pangenome based on 53 lines of *Brassica napus* showed that nearly 70% of the variation was in dispensable regions (Dolatabadian et al., 2020). Resistance gene analogs in 50 accessions of *B. napus* revealed 753 variable genes out of a total of 1,749 genes were related to disease resistance (Golicz et al., 2016a). My results also showed clusters of NBS encoding genes underlying previously phenotypically-defined MRC regions in both wild and domesticated lines of lettuce. In addition, the NBS-LRR genes were spread in different patterns across the genomes that are unique to each genotype. This is consistent with variation/PAVs in NBS-LRR gene content between different genotypes playing an important role in the evolution of resistance or susceptibility to disease.

The region surrounding the *Ve* locus in the MRC9 region is also highly variable. From previously published data, the *Ve* locus in cv. La Brillante has significant sequence divergence from cv. Salinas (Sandoya et al., 2021). Graph comparisons of cvs. Salinas and La Brillante at the *Ve* locus revealed several large indels affecting the gene content in the region, even though the three *Ve* paralogs are retained in both genotypes. The nomenclature of the paralogs in Sandoya et al., (2021) is inconsistent with the minigraph analysis and warrants further investigation.

This study shows the prevalence of SVs in *Lactuca* germplasm. Because SVs can be important in determining phenotypic diversity in crops (Gao et al., 2019; Li et al., 2022; Qin, et al., 2021), there is the potential of pangenomics to assist lettuce breeding. Further analysis of single nucleotide polymorphisms and SVs relative to phenotypic variation is necessary for lettuce improvement. Breeding disease resistant cultivars would particularly benefit from this because MRC regions are hotspots for disease resistance genes in wild germplasm of lettuce (McHale et al., 2009).

References:

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, *182*(1), 145-161.e23. https://doi.org/10.1016/j.cell.2020.05.021

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6(8), 914–920. https://doi.org/10.1038/s41477-020-0733-0

Bayer, P. E., Petereit, J., Durant, É., Monat, C., Rouard, M., Hu, H., Chapman, B., Li, C., Cheng, S., Batley, J., & Edwards, D. (2022). Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. *Plant Genome*. https://doi.org/10.1002/TPG2.20221

Christopoulou, M., Wo, S. R. C., Kozik, A., McHale, L. K., Truco, M. J., Wroblewski, T., & Michelmore, R. W. (2015). Genome-wide architecture of disease resistance genes in lettuce. *G3: Genes, Genomes, Genetics*, *5*(12), 2655–2669. https://doi.org/10.1534/g3.115.020818

Dolatabadian, A., Bayer, P. E., Tirnaz, S., Hurgobin, B., Edwards, D., & Batley, J. (2020). Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. *Plant Biotechnology Journal*, *18*(4), 969–982. https://doi.org/10.1111/pbi.13262

Fernandez-Gutierrez, A., Gutierrez-Gonzalez, J. J., & Gutierrez-Gonzalez, J. J. (2021). Bioinformatic-Based Approaches for Disease-Resistance Gene Discovery in Plants. *Agronomy.* https://doi.org/10.3390/agronomy11112259

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H. R., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, *7*, 1–8. https://doi.org/10.1038/ncomms13390

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics, 38(13).* https://doi.org/10.1093/bioinformatics/btac308

Hameed, A., Poznanski, P., Nadolska-Orczyk, A., & Orczyk, W. (2022). Graph Pangenomes Track Genetic Variants for Crop Improvement. *International Journal of Molecular Sciences*, *23*(21), 13420. https://doi.org/10.3390/ijms232113420

Hayes, R. J., Mchale, L. K., Vallad, G. E., Truco, M. J., Michelmore, R. W., Klosterman, S. J., Maruthachalam, K., & Subbarao, K. v. (2011). The inheritance of resistance to Verticillium wilt caused by race 1 isolates of *Verticillium dahliae* in the lettuce cultivar La Brillante. *Theoretical and Applied Genetics, 123,* 509-517. https://doi.org/10.1007/s00122-011-1603-y

Hickey, G., Monlong, J., Novak, A., Eizenga, J. M., Pangenome, H., Consortium, R., Li, H., & Paten, B. (2022). Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus. Pre-print. https://doi.org/10.1101/2022.10.06.511217

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, *21*(3), 171–189. https://doi.org/10.1038/s41576-019-0180-9

Inderbitzin, P., Christopoulou, M., Lavelle, D., Reyes-Chin-Wo, S., Michelmore, R. W., Subbarao, K. v, & Simko, I. (2019). The LsVe1L allele provides a molecular marker for resistance to *Verticillium dahliae* race 1 in lettuce. *BMC Plant Biology, 19.* https://doi.org/10.1186/s12870-019-1905-9

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02107-y

Kawchuk, L. M., Hachey, J., Lynch, D. R., Kulcsar, F., van Rooijen, G., Waterer, D. R., Robertson, A., Kokko, E., Byers, R., Howard, R. J., Fischer, R., & Prüfer, D. (2001). Tomato Ve disease resistance genes encode cell surface-like receptors. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(11), 6511–6515. https://doi.org/10.1073/PNAS.091114198

Kuang, H., Woo, S. S., Meyers, B. C., Nevo, E., & Michelmore, R. W. (2004). Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell*, *16*(11), 2870–2894. https://doi.org/10.1105/tpc.104.025502

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., Feng, X., & Chu, C. (2022). The design and construction of reference pangenome graphs with minigraph. *Genome Biology, 21.* https://doi.org/10.1186/s13059-020-02168-z

Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H., Xu, Y., Lin, S., Chen, X., Yao, Z., Yang, Q., Fei, Z., Huang, S., & Zhang, Z. (2022). Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature Communications, 13.* https://doi.org/10.1038/s41467-022-28362-0

McDonnell, A. v., Jiang, T., Keating, A. E., & Berger, B. (2006). Paircoil2: Improved prediction of coiled coils from sequence. *Bioinformatics*, *22*(3), 356–358. https://doi.org/10.1093/bioinformatics/bti797

Mchale, L. K., Truco, M. J., Kozik, A., Wroblewski, T., Ochoa, O. E., Lahre, K. A., Knapp, S. J., & Michelmore, R. W. (2009). The genomic architecture of disease resistance in lettuce. *Theor Appl Genet*, *118*, 565–580. https://doi.org/10.1007/s00122-008-0921-1

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H. T., Chan, C. K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*(5), 1007–1013. https://doi.org/10.1111/TPJ.13515

Nattestad, M., & Schatz, M. C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics*, *32*(19), 3021–3023. https://doi.org/10.1093/bioinformatics/btw369

Parra, L., Maisonneuve, B., Lebeda, A., Schut, J., Christopoulou, M., Jeuken, M., McHale, L., Truco, M. J., Crute, I., & Michelmore, R. (2016). Rationalization of genes for resistance to *Bremia lactucae* in lettuce. In *Euphytica* (Vol. 210, Issue 3, pp. 309–326). Springer Netherlands. https://doi.org/10.1007/s10681-016-1687-1

Plocik, A., Layden, J., & Kesseli, R. (2004). Comparative analysis of NBS domain sequences of NBS-LRR disease resistance genes from sunflower, lettuce, and chicory. *Molecular Phylogenetics and Evolution*, *31*(1), 153–163. https://doi.org/10.1016/S1055-7903(03)00274-4

Schwartz, C., Lenderts, B., Feigenbutz, L., Barone, P., Llaca, V., Fengler, K., & Svitashev, S. (2020). CRISPR–Cas9-mediated 75.5-Mb inversion in maize. *Nature Plants*, *6*(12), 1427–1431. https://doi.org/10.1038/s41477-020-00817-6

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Sekhwal, M. K., Li, P., Lam, I., Wang, X., Cloutier, S., & You, F. M. (2015). Disease resistance gene analogs (RGAs) in plants. *International Journal of Molecular Sciences*, 16(8), 19248–19290. MDPI AG. https://doi.org/10.3390/ijms160819248

Thind, A. K., Wicker, T., Müller, T., Ackermann, P. M., Steuernagel, B., Wulff, B. B. H., Spannagl, M., Twardziok, S. O., Felder, M., Lux, T., Mayer, K. F. X., Keller, B., & Krattinger, S. G. (2018). Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1477-2

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

van Treuren, R., van der Arend, A. J. M., & Schut, J. W. (2013). Distribution of downy mildew (*Bremia lactucae* Regel) resistances in a genebank collection of lettuce and its wild relatives. *Plant Genetic Resources: Characterisation and Utilisation*, *11*(1), 15–25. https://doi.org/10.1017/S1479262111000761

Wheeler, T. J., & Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*(19), 2487–2489. https://doi.org/10.1093/bioinformatics/btt403

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wroblewski, T., Piskurewicz, U., Tomczak, A., Ochoa, O., & 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 Journal*, *51*(5), 803–818. https://doi.org/10.1111/j.1365-313X.2007.03182.x

Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2021). Current status of structural variation studies in plants. *Plant Biotechnology Journal*, *19*(11), 2153–2163. https://doi.org/10.1111/PBI.13646

Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. In *Current Opinion in Immunology*, 20(1), 10–16. https://doi.org/10.1016/j.coi.2007.11.003

Chapter 6: Conclusions and perspectives for future research

Overall, this dissertation provides a foundation for developing telomere-to-telomere (T2T) assemblies with enhanced annotations, which can be used to characterize structural differences in diverse lettuce accessions. In addition, T2T assemblies can be used to understand the core and dispensable gene content as well as evaluate workflows for the construction of a comprehensive lettuce pangenome as more genome assemblies become available.

To accomplish these objectives, I developed a workflow for generating a T2T reference assembly of lettuce, using a combination of long-read and long-range technologies as described in Chapter 2. I compared the two widely used sequencing technologies, Oxford Nanopore (ONT) and Pacific Biosciences-HiFi (PacBio-HiFi) and generated 96x (ONT) and 35x (HiFi) data to construct two T2T assemblies. With the integration of both BioNano and Dovetail Hi-C data, I was able to scaffold and resolve many tandem repeat regions in the genome because these two technologies complemented each other. My results showed PacBio HiFi reads and ONT ultra-long reads have their own merits. Comparison of the two assemblies showed the assembly using PacBio HiFi reads had fewer errors at the single nucleotide level, and the highly repetitive regions like telomeres and centromeres were better resolved in these assemblies. ONT-based assembly due to the high error rates struggled to resolve complex repetitive regions, but the ultralong reads delivered higher contiguity. Therefore, the HiFi based v11 assembly with a contig N_{50} of 12.5 Mb consisting of 393 contigs and few gaps was selected as the reference assembly for subsequent studies; it was released on NCBI Genbank in November 2022.

The v11 assembly is not gapless and a more complete genome assembly will be generated in the future. Continued improvements in sequencing chemistry and plant-based base-caller models are expected to greatly improve ONT-based assemblies. Even during this study there were big improvements in base calling accuracy; the updated Guppy v5 base caller improved the metrics of the ONT-based assemblies we generated later relative to other genotypes. Also, improvement in the accuracy of long-read sequencing may make scaffolding technologies obsolete in the future (van Rengs et al., 2022), as read lengths increase to span the complex repetitive regions, making more complete assemblies possible without scaffolding and minimal error correction. For the future, I suggest the use of HiFi reads for primary assembly construction and error-corrected ONT reads to verify and gap-fill the assembly to generate a gapless T2T assembly; this is the approach being adopted for the human genomes (Nurk et al., 2022.)

The use of PacBio Isoseq data enabled us to identify full-length transcripts and define precise gene models for the new assemblies. With the integration of both Iso-Seq and RNA seq data, 44,214 complete protein coding genes were identified in the v11 assembly. Additional annotation is underway for the v11 reference assembly including annotation of microRNA genes and tRNAs. In the longer term, there is a need for the definition of long non-coding RNAs, accessible promoter regions under different physiological and developmental states, and topologically associating domains (TADs).

The technological improvements in long-read sequencing made it feasible and cost effective to create six additional chromosome-scale genome assemblies of domesticated and wild genotypes (Chapter 2). I was able to generate *de novo* annotations and classify genes into core and dispensable gene sets based on orthogroup clustering. The presence and

absence of gene content identified between the different genomes lays the foundation for more comparative analyses and a foundation for building a lettuce pangenomic data structure. This knowledge of dispensable genomes and identifying novel genes of lettuce are understudied at present and have a direct impact on agronomic traits.

As in other crops, wild genotypes of lettuce are often the source of beneficial variation that is introgressed to the domesticated lines. Understanding the genetic diversity between the genotypes will further improve our understanding of evolutionary divergence and help identify causal genes controlling important traits. *De novo* assembly and annotations of many more wild and domesticated lettuce lines will be generated based on HiFi reads. Multiple genotypes are currently being sequenced or in different stages of assembly. This will be accelerated by the upcoming availability of the PacBio Revio sequencer that will be 15x more efficient than the current PacBio Sequel II. A genetic resource of numerous high-quality genome assemblies will benefit many downstream functional analyses and comparative genomics studies.

As an increasing number of genomes are sequenced, a pangenome-based approach has become essential for understanding the genetic diversity within domesticated and wild lettuce. In Chapter 4, I used seven *de novo* assemblies to explore the core and dispensable gene content. Synteny analysis between these assemblies revealed translocation regions between genotypes. Further, I included evaluation of a structural variation (SV) calling workflow with both long and short-read SV calling tools and with graph pangenome-based methods. In particular, I evaluated several graph-based tools that are being developed and deployed by the Human Pangenome Consortium (Eizenga et al., 2020; Hickey et al., 2022.). Constructing the pangenome by whole genome assembly and comparison, or by utilizing

graph-based method, for large genomes like lettuce, still needs refinement, but preliminary results are informative. The generation of population-scale variation graphs of 200 whole genome resequenced lines of diverse lettuce using the vg toolkit is currently underway. These short reads are being mapped to the reference graph to extract the genotype information for each background. Multiple tools will be used to identify and precisely characterize SVs and SNPs within the lettuce genepool. On completion, this comprehensive lettuce pangenome database will help us to correlate SVs and SNPs with phenotypic variation by segregation analyses and/or GWAS. Future work will focus on quantitative trait loci (QTL) and genome-wide association studies (GWAS) to identify graph-based haplotype markers for traits of interest. This will better explain the under-explored role of SVs in genotype-to-phenotype relationships and their importance and utility in crop improvement. Moreover, understanding the impacts of SVs on chromatin conformation, epigenetic variation, and the effects of SVs and SNPs on gene expression, have yet to be explored in lettuce. Also, additional data types, such as proteomic and metabolomic data, as well as phenotypic data combined with network analysis are necessary to fully understand the impacts of genomic variations on phenotype.

Graph-based pangenomes are the future to understanding genetic diversity across multiple genomes. However, pangenome tools are constantly being improved. For complex genomes like lettuce, it is computationally very intensive and demands high-cost infrastructure to store and run these analyses. The continued algorithmic improvements will make these approaches more accessible in the future.

The ultimate goal of this dissertation was to build a comprehensive repertoire of disease resistance genes as a resource for lettuce breeding. As more pan-genome and pan-

NLRome studies become available, the knowledge of Nucleotide-binding domain leucine-rich repeat (NLR) sequence diversity has greatly improved. In Chapter 5, I identified and classified NLR genes and their distribution across the multiple reference genomes of lettuce. We noticed distinct distributions of NBS-encoding genes specific to each species, confirming that accessions from the same species showed differences in nucleotide-binding site leucine-rich repeat (NLR) copy number and sequence variation due to duplication, deletion, and unequal crossing-over. Different approaches were used to identify SVs underlying the major resistance cluster (MRC) regions of lettuce using pangenome graph-based analyses and workflows including short- and long-read tools. These approaches will be extended to the other MRCs and will include additional genome assemblies as they are generated.

In conclusion, lettuce pangenome analyses will be the foundation to a better understanding of genetic processes such as directional selection, divergence, and neofunctionalization that are directly linked to phenotypic traits. Deciphering the role of SVs will provide insights into MRCs that will inform future breeding of disease resistant lettuce. Graph-based pangenomes will provide resolution at the nucleotide level rather than at the gene presence/absence level. This will impact future lettuce breeding with precise insight into the basis of variation of agronomically important traits. In addition, the high-quality reference genome and other assemblies developed in this project will be key resources for diverse researchers of multiple aspects of lettuce biology, including comparative genomics, gene expression network analysis, and functional genomics, which are currently ongoing in the global lettuce research community.

REFERENCES:

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., & Garrison, E. (2020). *Pangenome Graphs*. https://doi.org/10.1146/annurev-genom-120219

Hickey, G., Monlong, J., Novak, A., Eizenga, J. M., Pangenome, H., Consortium, R., Li, H., & Paten, B. (2022). Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus. Pre-print. https://doi.org/10.1101/2022.10.06.511217

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. v, Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome*. Science, 376(6588).* https://doi.org/10.1101/2021.05.26.445798

van Rengs, W. M. J., Schmidt, M. H. W., Effgen, S., Le, D. B., Wang, Y., Zaidan, M. W. A. M., Huettel, B., Schouten, H. J., Usadel, B., & Underwood, C. J. (2022). A chromosome scale tomato genome built from complementary PacBio and Nanopore sequences alone reveals extensive linkage drag during breeding. *Plant Journal*, *110*(2), 572–588. https://doi.org/10.1111/tpj.15690