ABSTRACT

LIU, SU. Improved Hybrid de novo Genome Assembly, Resistance Gene Prediction and Annotation of Carrot (*Daucus carota*). (Under the direction of Dr. Massimo Iorizzo, Dr. Penelope Perkins-Veazie).

Carrot (*Daucus carota* L.) is one of the most economically important crops in the *Apiaceae* family (Grzebelus *et al*. 2014), *Daucus* genus. Carrots are grown on more than one million hectares in temperate climate regions (Grzebelus *et al*. 2014) worldwide and provide pro-vitamin A, fiber, and other antiulcer, anti-aging and antioxidant properties. Black carrot is of interest as source of natural colorants and fibers. The taste, high nutritive value, good storage life and relatively low cost has made carrot a popular vegetable with consumers (Simon et al. 2019).

The relatively slow growth of carrots in the field restricts the breeding cycle to one season per year. Moreover, carrot is susceptible to several disease-causing pathogens such as bacteria, fungi, viruses and nematodes, which significantly reduce the yields. By combining multiple favorable traits using molecular markers, trait selection can be facilitated for faster progress in breeding programs. Additionally, using natural or innate resistance via resistance (R) genes is an economical and sustainable method to prevent or manage carrot disease. To advance markers assisted selection (MAS) and incorporation or gene editing of R genes, a high quality genome is critical (Simon, Philipp W. 2019a).

The first carrot genome assembly ‘Double Haploid Orange Nantes Type (DH1) carrot genome v2’, was published in 2016 (Iorizzo et al., 2016), and was developed primarily using second sequencing technologies such as Illumina. While robust, it needs improvement to address the low contig level and to increase the fraction of the genome anchored to chromosome level. Third generation sequencing and scaffolding technologies can generate long DNA reads and span long physical distances, providing an opportunity to improve the carrot genome assembly.

The objectives of this study were to use third generation sequencing and scaffolding technologies to improve and create version 3 of the DH1 carrot genome, and to predict R genes in both v2 and v3 genomes for comparison of results between genome versions.

With the v3 carrot assembly, 692 contigs with total length of 438,921,773bp and N50 of 4,945,074bp were found, accounting for 92.8% of the estimated genome size. An additional 24 Mb sequences anchored to chromosomes level were found, and the contig N50 was increased by 159-fold compared with the published DH1 v2 genome assembly. With less but longer contigs, scaffolds and super-scaffolds, the accuracy and continuity of the whole genome increased significantly. In the v3 genome, over 300 more R genes and over 3,500 more R gene domains were predicted than in the v2 genome. The percent of R genes located on chromosome nine of carrot also increased slightly, from 98.3 % to 99.4 % and short R genes were predicted. The R genes in the v3 genome contained more domains on average and had better continuity than those in the v2 genome. Finally, using a pairwise comparison, the v3 genome was found to contain several new R genes in the v3 genome.

Improved Hybrid de novo Genome Assembly, Resistance Gene Prediction and Annotation of Carrot (*Daucus carota*).

by
Su Liu

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Horticulture Science

Raleigh, North Carolina

2020

APPROVED BY:

_____	_____
Dr. Massimo Iorizzo Committee Chair	Dr. Penelope Perkins-Veazie Committee Co-chair
_____	
Dr. Amanda Hulse-Kemp	

BIOGRAPHY

Su Liu was born in Tai'an, Shandong Province, China, in 1995. She found her passion of plants and gardening in elementary school. In 2013, she made the decision to major in horticulture and was admitted to Beijing Forestry University. After four years of undergraduate studies, she determined to concentrate on plant genetics and breeding. Then in 2017, she became a master student in horticulture science and began her new life at NC State University.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisor and committee chair, Dr. Massimo Iorizzo, who took me into the world of genomics and bioinformatics. During these two years, he supported me, guided me, trained me and gave feedback to me patiently. Thank you for sharing the knowledge and experience with me selflessly.

I would like to express appreciation to my committee members, Dr. Penelope Perkins-Veazie and Dr. Amanda Hulse-Kemp, who gave me advice on my slides, presentation and thesis in each committee meeting.

I would like to say thanks to members in Dr. Iorizzo's lab. You are my colleagues as well as my sincere friends. Thanks Dr. Hamed Bostan for giving me so much guidance in bioinformatics and cooperating with me in polishing as well as gap filling steps in v3 genome construction. Without your help, it is not possible for me to make such progress in short time. Thanks Dr. Marti Pottorff for giving me suggestions on my thesis. Thanks Ashley Yow for providing me with lab work supports. I cannot forget the help from you all.

Finally, I would give special thanks to the supporters of this project – United States Department of Agriculture National Institute of Food and Agriculture (Hatch project 1008691) and National Institute of Food and Agriculture, United States Department of Agriculture (under award number 2016-51181-25400, project "Identifying phenotypes, markers, and genes in carrot germplasm to deliver improved carrots to growers and consumers").

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 Literature Review

1.1 Taxonomy and domestication history

Carrot (*Daucus carota* L.) is one of the most economically important crops in the Apiaceae (Umbelliferae) family (Grzebelus *et al*. 2014). The Apiaceae family is a large family of angiosperms (Liu, J. *et al*. 2014) which includes 466 genera and 3,820 species (Simon et al. 2019). Other economically-relevant plants in the Apiaceae family include celery and other plants used as spices, such as parsley, angelica, anise, caraway, and cilantro (Rubatzky *et al*. 1999).

Carrot belongs to the *Daucus* genus, which includes about 20 herbaceous species. The genus center of origin for *Daucus* is the Mediterranean region. In addition, *Daucus* species also occur in Australia and on the American continent (Simon et al. 2019). The origin or center of diversity for carrot is in Central Asia as suggested by Vavilov (1992) and this is supported by molecular studies (Iorizzo *et al*. 2013; Iorizzo *et al*. 2016). The Himalayan-Hindu Kush region (Kashmir-Afghanistan) is the origin of eastern cultivated carrots, and the Anatolian region of Asia Minor (Turkey) is the center of diversity for western carrots (Simon, Philipp W. 2010). As carrots have shown adaptation to tropical and sub-tropical climates, cultivars have been developed for those regions (Simon, P. W. *et al*. 2008). Currently, carrot cultivation occurs throughout the world.

The first evidence of carrot used as a storage root crop is from modern-day Afghanistan and Turkey in 900 C.E. (Clotault *et al*. 2010), who described the roots as purple or yellow. In the Middle Ages, carrot cultivation spread throughout Middle East, North Africa and Europe, following the trade routes. In the beginning of the 1600's, white- and orange-rooted carrots began to appear in Europe, and gradually replaced the former purple and yellow variants. Concomitantly, in the 1700's, red carrots were described in Asia, particularly in China and India.

Since the 19th century, orange-rooted carrots spread from Europe and became the main carrot type in the market (Soufflet-Freslon *et al*. 2013).

Eastern carrots, which are commonly grown in Asia, have pubescent leaves, early flowering, more branches and always produce conical roots that are much thicker and shorter than western carrots (Grzebelus *et al*. 2014). Eastern carrot roots are poor in provitamin A carotenoids, but contain lutein, anthocyanin and lycopene, which imparts a yellow, purple or red coloration. Compared to eastern carrots, western carrots contain less phenolic compounds and larger amounts of carotenoids. The improvement of economical benefits, such as the accumulation of carotenoids, is a hallmark of carrot domestication (Ellison. 2019).

1.2 Economical and nutritional importance

Carrots are a popular vegetable due to their taste, high nutrition value, good storage life, and relatively low cost (Simon et al. 2019). They are also of interest to industry for their use as natural colorants providing pigments used to color cosmetic products and foods such as yogurts, ice-cream, beverages, and candy. Carrot is predominantly grown under conventional (non-organic) management system but the portion of the carrot crop grown under organic production management practices has increased in the North American and European markets, accounting for 11% of the U.S. market and 25 – 30% of Danish and German markets in 2016 (Simon et al. 2019).

Carrots are grown on more than one million hectares in temperate climate regions (Grzebelus *et al*. 2014). According to Food and Agriculture Organization (FAO) of the United Nations, the world area harvested and world production of carrots and turnips, most of which are

carrots, had a 4 times and 9 times increase respectively during the last 50 years (Fig.1.1) (FAO. 2019).

Both fresh and processed carrots are popular as they are used in a variety of processes and are made into many different food products. Baby carrot is a very popular product in North America and is made from carrots that are very long that are peeled and cut into a smaller size. Carrots can also be processed into juice, concentrate, dried powder, canned, preserve, candy and pickle. Carrot pomace can also be used in cake, bread and biscuits (Sharma *et al*. 2012). Carrots may even be dried at 130 °C and pulverized into a powder, prized for its high nutritional content, including high fiber and vitamin concentrations.

Bioactive metabolites

Carrots have a high nutritional value and low caloric content. These nutritional related factors have contributed to increased consumption over the past few years (Wrzodak *et al*. 2012). Among the most important compositions in carrots are bioactive metabolites, most of which are pigments. Bioactives include phenolic compounds, nitrogen compounds, carotenoids and ascorbic acid, which are commonly found in fruits and vegetables (Gajewski et al. 2007). Bioactives with antioxidant activity are of special interest in human health as they can inhibit the oxidation of other molecules and protect cells against reactive oxygen.

Orange carrots contain β-carotene, as well as some α-carotene. The typical orange-red color of carrots is derived from β-carotene, which increases with the age and size of carrot root (Wrzodak *et al*. 2012). β-carotene is a major source of pro-vitamin A, which humans cannot synthesize and must uptake from a dietary source like carrot. Vitamin A has such health benefits such as anti-ulcer, anti-aging, antioxidant properties, as well as promoting an increased immune

response (Pinheiro Sant'Ana *et al*. 1998). Yellow carrots contain lutein, which has anti-cancer properties and also helps defend the eye against diseases like macular degeneration and maintains optical health (Rock. 1997). The main pigment present in red carrots is lycopene, which helps to fight against heart disease and some cancers. Carotene, lutein and lycopene all belong to the carotenoid family (Rock. 1997). Experimental studies suggest that a higher dietary intake of carotenoids protects against certain cancers, optical deterioration, and other health conditions involved with oxidative or free radical damage (Rock. 1997).

Purple carrots contain higher amount of phenolics, mainly anthocyanins, which provides the purple coloration. Anthocyanins act as very powerful antioxidants, binding onto harmful free radicals in the body and reducing their reactive potential. Anthocyanins can also help prevent heart attacks by slowing blood clotting (Alasalvar *et al*. 2005).

Due to the numerous colors of carrots, they are good sources of natural dietary and industrial pigments. Food colorants and industrial pigments from natural sources like carrot are gaining popularity among consumers. This is due to the numerous human health and environmental benefits over the synthetic dyes (Ersus and Yurdagel. 2007).

Other components

In addition to the aforementioned components, carrots contain 12 % of dry matter, 4.5 % of sugars, 2 % of dietary fiber and 5.9 % of vitamin C (Holden *et al*. 1999). As the main storage compounds in carrot roots, soluble sugars are stored in vacuoles, accounting for 35-70 % dry weight of the root. Sucrose is the predominant storage sugar at root maturity with the concentration reaching 3.6 % (Warman and Havard. 1997). Western carrots are sweeter, having on average 18 % higher sugar content than Eastern carrots (Grzebelus *et al*. 2014). The content

of these nutrients can be affected by cultivar, growing conditions and storage (Alasalvar *et al*. 2005).

Carrot is also a moderate source of dietary fiber (Grzebelus *et al*. 2014), and supplementary products can be produced from carrot roots (Chantaro *et al*. 2008). Moreover, it is a source of alternative carbon and fibers used for material development, such as composite films (Idrovo Encalada *et al*. 2016). Due to the economic importance stated above, there is a bright future of improving carrot production and expanding basic carrot research, which is based on the availability of a complete carrot genome (Simon, Philipp W. 2019b).

1.3 Breeding methods

The growth of carrots has two phases, a vegetative phase, including vernalization and stem elongation, as well as a generative phase, including flower development, pollination, fertilization, embryo and seed development (Simon et al. 2019). During flower induction, the apical meristem of a carrot becomes uplifted and inflorescences are initiated. Then, several dozen umbels develop to produce thousands of flowers per plant. Carrot is an outcrossing species and pollination depends on insects (Simon, Philipp W. 2010).

Carrot breeding was first based on the open pollination method. Phenotypic recurrent selection was a primary breeding approach, which is a method involving reselection based on phenotypes generation after generation with interbreeding of selections to provide genetic recombination (Simon, Philipp W. 2010). Then a hybrid breeding method, the method producing offspring from two genetically different parent lines, that takes advantage of cytoplasmic male sterility was applied, which improved carrot nutritional quality. Because of the relatively slow growth of carrots in the field, a restricted breeding cycle of once per year and the need to

improve cultivars by combining multiple favorable traits, molecular markers are important in breeding programs to facilitate selection, assuring further breeding progress. The study to develop high quality genetic and genomic resources, such as genomes, linkage maps and their integration with phenotyping data in quantitative trait locus (QTL, a section of DNA correlating with variation of a quantitative trait in the phenotype of a population of organisms), are a critical step to routinely and effectively apply marker assisted selection (MAS) or gene editing technologies in breeding programs (Simon, Philipp W. 2019a).

1.4 Carrot molecular genetics and genomics

Given the increasing production and interest in developing carrot cultivars with improved production and nutritional attributes, carrot genetic and genomic research has been significantly advancing. Indeed, the number of DNA based markers, physical and genetic maps and whole genome studies have increased (Simon et al. 2019), opening new perspectives for carrot research and breeding (Budahn *et al*. 2014).

1.4.1 Carrot genetics

Carrots are a diploid outcrossing species with 2n = 2x = 18 chromosomes (Iovene, Cavagnaro *et al*. 2011a) and have a relatively small genome, estimated at 473 Mb (Arumuganathan and Earle. 1991). The estimated genome size represents the amount of DNA in an unreplicated gametic nuclear genome, and is also known as the C-value (Bennett and Leitch. 2011). The chromosomes of carrot are small but morphologically distinguishable (Iorizzo *et al*. 2016). Besides carrot, a few other *Daucus* species, including *Daucus capillifolius, D. sahariensis* and *D. syrticus,* are 2n = 18, while most *Daucus* species are 2n = 20 or 22.

1.4.2 Genetic markers

With genetic markers, different features in DNA sequences can be identified. As a result, differences between individuals can be found. Individuals between varieties or cultivars can be classified using genetic markers. Inheritance of different regions of the genome can also be tracked. DNA sequences can be anchored on chromosomes. Genetic markers are crucial to modern genetics and genome construction, helping to answer the important questions in population genetics, ecological genetics and evolution (Hohenlohe *et al*. 2011), also providing a basis for modern breeding methods.

Isozyme markers

An isozyme is an enzyme with varying forms that differ in amino acid sequence, each of which catalyze different chemical reactions. They usually display different kinetic parameters or regulatory properties. Isozymes can be used as biochemical markers and since they are often codominant, they can be used to distinguish heterozygous and homozygous individuals. The first carrot linkage map was constructed with 14 isozyme markers (Westphal and Wricke. 1991). Eight of these isozyme markers were also used to assess the genetic variability within the *D. carota* complex (St. Pierre *et al*. 1990).

RFLP, RAPD, AFLP markers

Restriction Fragment Length Polymorphisms (RFLP), Randomly Amplified Polymorphic DNA (RAPD), and Amplified Fragment Length Polymorphisms (AFLP) are three types of DNA-based markers (Vos *et al*. 1995). Most of the carrot molecular genetic studies have used DNA-based markers. They are abundant in the genome, easy to detect, and are usually not

influenced by the environment. RAPD markers are DNA fragments produced throughpolymerase chain reaction (PCR, a method making several copies of a specific DNA segment), amplification of random segments of genomic DNA with single primer of arbitrary nucleotide sequence. RFLP uses restriction enzymes to digest DNA, separates them with gel electrophoresis and hybridizes them with probes. Hence, the length of different fragments can determine DNA polymorphisms. AFLP, which is a combination of RAPD and RFLP (Vos *et al*. 1995), combines restriction enzymes, adaptor ligation and amplification methods. RFLP are codominant while RAPD and AFLP are dominant (Simon et al. 2019).

These three types of DNA-based markers, RFLP, RAPD and AFLP, have been used in carrot linkage map construction (Table 1.1), QTL mapping and phylogenetic studies (Simon et al. 2019). Because the polymorphic rate of AFLP is higher than those of RFLP and RAPD, AFLPs are more efficient in generating markers for linkage maps (Bradeen and Simon. 2007). In phylogenetic studies, the AFLP system can yield more useful makers per reaction, while RFLPs can benefit from their codominant property (Simon et al. 2019).

DArT markers

Diversity Arrays Technology (DArT) is a high-throughput genotyping technology that does not require knowing the genome sequence (Kilian *et al*. 2012). It uses cloned fragments that can be sequenced by Sanger sequencing. Hence, the non-model species, including carrot, can benefit from it significantly (Simon et al. 2019).

A DArT array containing 7,680 DArT clones was generated from 169 wild and cultivated germplasm genotypes, which contributed to carrot population genetics and linkage map construction (Grzebelus *et al*. 2014). In 2017, a DArT marker connected with carrot root

domestication was cloned and named as DcAHLc1. This marker is in the gene family involved in carrot root tissue patterning (Macko-Podgorni *et al*. 2017).

SSR markers

Simple sequence repeats (SSR), otherwise called microsatellites, are abundant in plants. All higher plants contain repetitive DNA sequences that can account for up to 90% of the genome in some species. Because the number of repeated units is highly diverse, sequence repeats can represent major differences across genomes. Microsatellites are unique tandemly repeated genomic sequences, which are abundantly distributed across plant genomes and demonstrate high levels of allele polymorphism (Vieira *et al*. 2016).

In carrot, about 300 SSR markers were developed, including 144 SSRs detected on BAC-end sequences (BSSR) (Cavagnaro *et al*. 2009) and 156 SSRs from an enriched repetitive sequence library (GSSR) (Cavagnaro *et al*. 2011b). SSR markers can also be derived from expressed sequence tags (EST), which is ESSR. 114 computationally polymorphic ESSRs were identified based on computational preselection (Iorizzo *et al*. 2011).

SNP markers

A single DNA building block is a nucleotide; single nucleotied polymorphism (SNP) represents a difference in a nucleotide, and this group comprises a major portion of the DNA variants, and thus are the most used molecular markers in plant genetic studies (LaFramboise. 2009). SNP markers were used for the carrot linkage maps B9304×YC7262 (Bradeen and Simon. 1998), B1896×B7261 (Yildiz *et al*. 2013a), 70349 (Cavagnaro *et al*. 2014), 2569 (S. Ellison *et al*. 2017), 74146 (Ellison *et al*. 2017) and HM related maps (Parsons *et al*. 2015) (see Table 1.1).

Apart from linkage map construction and QTL studies, SNP markers were also used to characterize the genetic diversity, study carrot domestication and conduct phylogenetic studies (Iorizzo *et al*. 2019; Simon, Philipp *et al*. 2019).

1.4.3 Genetic linkage maps

The construction of genetic linkage maps is useful for ordering marker loci (Royaert *et al*. 2016), performing marker-trait associations studies and identifying candidate genes. Genetic linkage maps have multiple functions. First, they can be used to explore genome structure, function, and evolution (Niedzicka *et al*. 2017). Second, they are tools for improvement of genome assemblies (Kawakami *et al*. 2014). Third, the investigation of chromosomal segments conservation at various evolutionary time scales also relies on genetic linkage maps (Postlethwait *et al*. 2000). Finally, they are able to explore the framework for examining both the genomic architecture underlying the reproductive barriers between species or populations, and interactions between differentiated genes in hybrid genomes (Rieseberg *et al*. 2000). Once markers are linked with traits, these can be used as molecular markers for MAS and also be used in comparative genetic/genomic studies for use by breeders (Cassia da Silva Linge *et al*. 2018).

Linkage map construction

There is a well-established workflow for genetic map construction (Gunderson *et al*. 2006). Initially, genetic markers are mined from a small pool of individuals, followed by genotyping of selected mapping populations using sets of marker panels. Restriction site associated DNA (RAD) enables synchronous SNP marker discovery and genotyping using massively parallel sequencing. Once a panel of markers is established from initial SNP

discovery, samples from a selected population are then genotyped using oligo-extension or array-based platforms (Gunderson *et al*. 2006).

Currently, next-generation sequencing (NGS) based approaches include SLAF-seq (specific-locus amplified fragment sequencing), RAD genotyping and genotyping-by-sequencing (GBS). They have yielded huge number of markers, contributing to the construction of high-density genetic linkage maps. High-density linkage maps can assist in the discovery of functional genes, genome assembly and comparative analysis of genome structures (Liu, D. *et al*. 2014). NGS data still inevitably suffers from sequencing errors, especially when sequencing depths are low and genotypes are highly heterozygous. This occurrence is highly possible and leads to genotyping errors (Liu, D. *et al*. 2014).

Carrot genetic maps

The first carrot linkage map was constructed in 1993 and used isozymes, RFLPs and RAPDs, yielding 55 markers assigned to 8 linkage groups (Schulz *et al*. 1994). Then, Brasilia × HCM and B493 × QAL, two F_2 populations, were applied to the construction of carrot linkage map using AFLP and SCAR markers assigned to 9 linkage groups (Santos, Carlos A. F. and Simon. 2004a). This linkage map was later used to detect QTL related to carotenoid accumulation (Santos, C. and Simon. 2002). In 2007, SNP markers were integrated into the B493 × QAL AFLP linkage map (Just *et al*. 2007). In 2011, 49 SSR markers were added into the B493 × QAL map (Cavagnaro *et al*. 2011a) and the new map was anchored to carrot chromosomes by FISH mapping of 18 BACs (Iovene, Cavagnaro *et al*. 2011b). In 2013, 355 AFLP, RAPD, SCAR and SSR markers were assigned to 9 linkage groups and a linkage map was constructed in order to detect loci relevant to vernalization and male fertility restoration

(Alessandro *et al*. 2013a). In the same year, AFLP, SSR and SNPs were applied in B1896 × B7262 to construct a linkage map with 279 maker data points and map the anthocyanin and carotenoid pigmentation genes to an associated chromosome (Yildiz *et al*. 2013b). PI 652188 × B7262 was utilized to generate a linkage map in order to map root-knot nematode resistance genes to the genome with RAPD and SSR markers (Ali *et al*. 2014). In 2014, a linkage map aiming to detect fertility control and flower development loci was generated with 285 RAPD, AFLP, SCAR and SSR markers (Budahn *et al*. 2014). In 2016, SNPs from three mapping populations were integrated into a high-density linkage map to anchor the carrot genome (Iorizzo *et al*. 2016). (Table 1.1)

1.4.4 Carrot genome

In 2016, a high-quality genome assembly of Double Haploid Orange Nantes Type Carrot (DH1) carrot sequenced with Sanger, Roche 454 and Illumina HiSeq sequencing technologies was released (Iorizzo *et al*. 2016). The resulting DH1 v2 assembly contains 4,907 sequences with an N50 of 12.7 Mb, spanning 421.5 Mb and accounting for around 90 % of the 473 Mb estimated carrot genome size. The assembly is composed of 30,938 contigs with total length of 386.8 Mb and N50 of 31.2 kb. There are 60 scaffolds anchored to 9 chromosomes finally, including 50 anchored and oriented ones, with a total length of 361.1 Mb, accounting for approximately 86% of the assembled genome and 76% of the estimated genome size. In DH1, 46% of the genome assembly is represented by repetitive sequences, among which 98% (193.7 Mb) were transposable elements (TEs), especially class II TEs that account for a greater amount of the genome when comparing with similarly sized plant genomes (Iorizzo *et al*. 2016).

In total, 32,113 genes were predicted, among which 79% had substantial homology with known genes and 98.7% of gene predictions were supported by cDNA and/or EST evidence, which indicates the high accuracy of prediction. Besides coding genes, 564 tRNAs, 31 rRNA fragments, 532 small nuclear RNA (snRNA) genes, and 248 microRNAs (miRNAs) distributed among 46 gene families were also identified (Iorizzo *et al*. 2016).

The release of the carrot genome v2 represented a major milestone for carrot research and crop improvement. Since its publication in 2016, the DH1 genome assembly v2 has been used in multiple studies to identify candidate genes controlling metabolite accumulation and root development, to study the transcriptome profile associated with flowering, and carotenoid accumulation, and to perform phylogenetic studies (Simon et al. 2019). However, while this represented a major advance, this genome assembly inherited the challenges and problems associated with short-read based assembly obtained through first and second generation sequencing technologies. Indeed, the total length of gaps filled with unknown sequences in the assembly is 33.7 Mb, accounting for about 8 % of assembled sequences, and at chromosome level assembly the length of gaps is approximately 21.6 Mb (Iorizzo *et al*. 2016). Also, the fraction of known genome anchored to chromosome level, used for genetic studies is 340 Mb representing 71 % of the estimated genome size, indicating that a large portion of the genome and genes is still unknown and unused for DNA based crop improvement studies and projects.

Recent advances in sequencing technologies provides an opportunity to improve genome assemblies and gene predictions. To highlight the impact that the different sequencing technologies had in genome projects, in the next section an overview of the history and most recent advances on sequencing technologies will be reviewed.

1.5 Sequencing technologies

DNA sequencing is the technology that determines the order of four kinds of bases in DNA molecules, which provides us with information carried by DNA. There are three generations of sequencing technologies.

1.5.1 First generation sequencing

Sanger sequencing, is the first generation of sequencing technology and was first reported in 1975 (Ninomiya *et al*. 2014). The principle of this technology is that with primers, deoxyribonucleotide triphosphate (dNTPs) and dideoxynucleotides (ddNTPs) are used by DNA polymerase I to extend and copy of the template in the presence of four deoxyribotriphosphates, stopping at ddNTPs due to the lack of a 3'-OH group. Then ddNTPs will be fluorescently labeled for detection in automated sequencing machines (Sanger and Coulson. 1975). This technology produces reads with lengths up to ~1,000 bp and its accuracy is about 99.999 % per base (Wang, X. V. *et al*. 2012). Since it can only read one DNA fragment at a time, it is not considered a high-throughput sequencing method.

As the prevailing DNA sequencing method for over 30 years, Sanger requires less handling of toxic chemicals and radioisotopes (van Dijk *et al*. 2014). It can detect almost all the base substitutions, small insertions and deletions, while having a modest limit of detection (Tsiatis *et al*. 2010). As a result, it is still considered to be very high quality compared to other methods. This method has been applied in sequencing of the human genome (up to 3 billion bases) (Pareek *et al*. 2011), as well as several another model organisms, such as Arabidopsis (Kaul *et al*. 2000), rice (Matsumoto *et al*. 2005), Sorghum (Paterson *et al*. 2009), and maize

(Schnable *et al*. 2009), but compared with second and third generation sequencing technologies, it is an expensive sequencing technology (Waterston *et al*. 2002).

1.5.2 Second generation High-throughput next generation sequencing (HT-NGS)

After the emerging of Sanger sequencing, next-generation sequencing (NGS) became dominant, which includes 454 Life Science, Illumina and SOLiD (Reuter *et al*. 2015). A common character among them is before sequencing, preparing amplified libraries by amplifying DNA clones is required (Thompson and Milos. 2011). They can be applied to DNA sequencing, RNA sequencing, whole genome genotyping, de novo assembly and so on (Milos and Ozsolak. 2011).

454 Life Sciences

As the first NGS system, 454 Life Sciences was founded in 2000 (Pareek *et al*. 2011; van Dijk *et al*. 2014). GS20 was the first NGS sequencer put on the market by 454 Life Science (Ninomiya *et al*. 2014). In the following years, the new 454 version GS FLX titanium was developed (Pareek *et al*. 2011). It relies on the preparation of NGS libraries in a cell free system, producing thousands-to-many-millions of sequencing reactions in parallel and detecting without electrophoresis (van Dijk *et al*. 2014). When developed, 454 sequencing can produce double-stranded fragments ranging from 400 to 600 bp (Rothberg and Leamon. 2008) and the error rate was tested to be 0.49 % per base (Shao *et al*. 2013).

454 sequencing was the first NGS technology to sequence and *de novo* assemble bacterial genomes (Margulies *et al*. 2006). Barley bacterial artificial chromosome (BAC) clones were sequenced with 454 sequencing. Compared with Sanger sequencing, it provided a more even

coverage (Wicker *et al*. 2006). In the human genome project, 454 sequencing can achieve about the same coverage for approximately 1 % of the price (Wheeler *et al*. 2008).

Illumina/Solexa

Illumina/Solexa is the company whose sequencing technologies belong to the second generation HT-NGS, which can generate billions of bases in a single run (Pareek *et al*. 2011). Detection method of Illumina is based on fluorescent emission (Pareek *et al*. 2011). Through incorporation, washing, imaging and cleavage, the template strand will be sequenced one nucleotide at a time, using a cyclic reversible termination strategy (Reuter *et al*. 2015).

The DNA sample requirement is less than 1 g for single or paired-end libraries, and the template required is around 75 bp. The reads generated are short (150 bp) (Hackl *et al*. 2014), but are around 30 million (van Dijk *et al*. 2014). Apart from being very high-throughput, other advantages of the Illumina technology include the reductions in cost and error rates. Indeed the average error rate is $0.24 \pm 0.06\%$ per base (Pfeiffer *et al*. 2018) and the most common type of error is substitution (Reuter *et al*. 2015). However, it may produce systematic and sequence-specific substitution errors and coverage bias problems during the library construction and sequencing process (Hu *et al*. 2012). HiSeq2000/2500 and MiSeq are two sequencing system platforms Illumina released, producing paired end reads with length between 100 and 300 bp, respectively, but the later has a much shorter run time (Buermans and den Dunnen. 2014).

Seven paired-end libraries of flax (*Linum usitatissimum*) were sequenced with an Illumina genome analyzer (Wang, Z. *et al*. 2012). Illumina sequencing was also used in sequencing and *de novo* assembling the genomes of viruses (Khalifa *et al*. 2016). In the same

year, mitochondrial genome of brown algal (Xu *et al*. 2016) and chiton species (Veale *et al*. 2016) were also sequenced using Illumina platforms.

AB/ Life Technologies SOLiD

Sequencing by Oligo Ligation Detection (SOLiD) is a platform produced by Life Technologies in 2007, which can generate much larger numbers of reads than 454, almost 100 million; however, the reads are only about 35 bp long (van Dijk *et al*. 2014). The DNA sample requirement is < 2 μg for shotgun library or 5 - 20 μg for paired end. The detection method is also based on fluorescent emission from ligated dye-labelled oligonucleotides. Its total throughput bases are 10 - 20 Gb and its raw accuracy is 99.94 % (Pareek *et al*. 2011).

Vibrio vulnificus, the leading cause of death from seafood consumption, was sequenced using SOLiD sequencer (Gulig *et al*. 2010). In 2011, chloroplast genomes from three *Lemnoideae* (Duckweeds) were sequenced using SOLiD platform (Wang, W. Q. and Messing. 2011). In plants, ABI-5500xl SOLiD sequencer was used to sequence the whole genomes of six tomato lines (Shirasawa *et al*. 2013).

1.5.3 Third generation HT-NGS

While former sequencing technologies are based on PCR amplification, the third generation sequencing platform does not, resulting in fewer errors caused by PCR and less use of biochemicals, and most importantly increased read length (Pareek *et al*. 2011). The read length produced by second generation NGS technologies limits the production of highly contiguous genome assemblies. For example, currently the read length of 454 is only 400-500 bp, but is still longer than Illumina, as mentioned above (Pareek *et al*. 2011). However, third generation

sequencing technologies can overcome this problem by reading long sequences. Apart from allowing the detection of single molecules, third generation methods can also let sequencing occur in real time (van Dijk *et al*. 2014). Compared to second generation HT-NGS that can only perform cDNA sequencing, third generation HT-NGS can sequence RNA directly (Pareek *et al*. 2011).

1.5.3.1 PacBio sequencing

Single-molecule real-time sequencing (SMRT) by Pacific Biosciences of California, produces long reads with half the data in reads larger than 50 kb. The longest read length can be larger than 175 kb (Hackl *et al*. 2014). Their third generation sequencer, PacBio RS II, is able to generate up to 400 Mb per sequencing run (Hackl *et al*. 2014). Then, Sequel II System, a new sequencing instrument was released by the company, which has an 8-fold increase in throughput (Pacific Biosciences of California Inc. 2019). Due to the long read, it is suitable for unresolved problems in genome, transcriptome and epigenetics research, also an excellent tool to close gaps in assemblies and detect mutations. Besides, it is useful for the direct detection of base modification such as methylation (Rhoads and Au. 2015).

The template of PacBio sequencing is a closed, single-stranded circular DNA named SMRTbell. When loaded to a chip called a SMRT cell, the SMRTbells will diffuse into units that provides the smallest available volume for detection. In each unit, a polymerase is immobilized at the bottom. Four fluorescent-labeled nucleotides are added to SMRT cell so once a base is held by polymerase, a light pulse is produced to identify them (Rhoads and Au. 2015).

Comparing to second generation sequencing, the disadvantages of PacBio sequencing includes lower throughput, higher cost per base and a higher error rate. The accuracy of PacBio

RS II is only 80 - 85 %, but when sequencing reads are short (<1 kb), the accuracy can be increased due to the circular templates it uses (Hackl *et al*. 2014). Since reads with a high error rate (~15 %) are difficult to assemble, PacBio sequences can utilize autocorrection to reduce the error rate to 0.5 – 1 % (Chakraborty *et al*. 2016), a bioinformatics analysis also known as polishing. Since DH1 has a completely homozygous genome, which represents one haplotype, to achieve optimal coverage for PacBio reads by autocorrection, a 40 × depth coverage is needed.

Numerous plant genomes have been sequenced and de novo assembled with PacBio sequencing up until now, including a tropical maize (*Zea mays*) small-kernel inbred line (Yang, N. *et al*. 2019), a high-quality peanut (*Arachis hypogaea*) (Zhuang *et al*. 2019), the upland cotton (*Gossypium hirsutum* L.) (Yang, Z. E. *et al*. 2019) and so on. Apart from plants, the genomes of spotted lanternfly (*Lycorma delicatula*) (Kingan *et al*. 2019), goldfish (*Carassius auratus*) (Chen, Z. L. *et al*. 2019), and human (Wenger *et al*. 2019) were also sequenced by PacBio sequencing and the continuity of these genomes was improved significantly comparing to the previous genomes based on short reads only.

1.5.3.2 Nanopore sequencing

Nanopore technology has been used for almost a decade (Nivala *et al*. 2013). Different from former sequencing technologies, Nanopore DNA sequencer does not use nucleotide labeling and detection (Pareek *et al*. 2011) and can be used as a platform for estimating DNA biophysics, chemistry, and physiochemistry (Howorka and Siwy. 2009). The nanopores which are located between lipid bilayers are sensors for biological molecules and can identify and quantify molecules carrying electric potential individually when they pass by at a high throughput (Rhee. 2007; Wiggin *et al*. 2008). This requires that the molecules be unfolded and

gain a non-uniform charge (Nivala *et al*. 2013). These molecules include nucleic acids, peptides, proteins and biomolecular complexes (Howorka and Siwy. 2009).

There are two methods to prepare Nanopores (Rhee. 2007). The most commonly used organic Nanopores, such as α-hemolysin pores, have a 3.6 nm diameter vestibule and a 2.6 nm wide connection with β-barrel (Venkatesan and Bashir. 2011), whereas the synthetic solid-state Nanopores pore sizes are smaller (Rhee. 2007). Among them, α-hemolysin pores are quite stable and can remain functional even at the boiling point of water (Wiggin *et al*. 2008). There are some disadvantages of using Nanopore such as the relatively high current error rate (Laver *et al*. 2015), the mechanical instability of the lipid bilayer and the sensitivity to experimental conditions (Venkatesan and Bashir. 2011). Additional improvements could include reducing the rate at which the DNA passes by to increase the sensitivity of the method.

MinION is a new sequencing technology of Oxford Nanopore Technologies (ONT) that potentially offers read lengths of tens of kilobases (kb), limited only by the length of DNA molecules presented. The device has a low capital cost, is the most portable DNA sequencer available and can produce data in real-time. It has numerous applications including improving genome sequence assemblies and resolving repeat-rich regions (Laver *et al*. 2015). According to ONT, each consumable flow cell can now generate 10–30 Gb of DNA sequence data. Ultra-long read lengths are possible (hundreds of kb) since the fragment length can be chosen.

Nowadays, plant genomes like *Arabidopsis thaliana* (Michael *et al*. 2018), as well as the genome of bacterium, *Acinetobacter baylyi* ADP1 (Madoui *et al*. 2015), European eel (Jansen *et al*. 2017) and yeast (Fournier *et al*. 2017), were sequenced using MinION by Oxford Nanopore and assembled.

Sequencing technology is the basis of genome construction, providing materials to *de novo* assembly and scaffolding. Further, through a much complete genome, the objective of efficient breeding of carrots is achievable.

1.6 Scaffolding technologies

Despite the impressive recent progress in long-read DNA sequencing, it is still challenging to assemble a complete plant genome from sequence reads alone (Jiao and Schneeberger. 2017). After sequencing, large amounts of DNA fragments are obtained and can be assembled into contigs (Kaplan and Dekker. 2013). Then a scaffolding step will join the contigs into scaffolds, which contain gaps of unknown sequence between connected contigs. Usually scaffolding in plants is a challenge, not only due to repetitive sequences (Kaplan and Dekker. 2013), but also due to shotgun library construction that require DNA with large inserts (>15 kb) (Jibran *et al*. 2018). There are three generations of scaffolding technologies.

1.6.1 First generation scaffolding technologies (BAC-end, linkage maps, physical maps)

Bacterial artificial chromosome (BAC)-end sequences can be used in multiple ways, such as gene discovery, genome scaffolding and anchoring. With inserted DNA of the organism being sequenced, BAC-end sequences of individual clones are very informative in assessing the content and organization of genomes since they represent a significant portion of a genome (Paux *et al*. 2006). The length of insert is 150 - 350 kb (Chen, X. e. *et al*. 2017). Since a method to connect contigs is generating read pairs sequenced from the two ends of a molecule (Roach *et al*. 1995), BAC-end sequences can be used for scaffolding. However, BAC-end sequencing is very low throughput, expensive, tedious and time-consuming (Jibran *et al*. 2018).

Producing a genome physical map, involves restriction mapping through cleaving unknown DNA with restriction enzymes, then measuring the range of fragment sizes with gel-based methods. The spectrum of fragment sizes reveals the structure of the unknown DNA and can be viewed as a fingerprint or barcode of this sequence (Nathans and Smith. 1975). A physical map consists of a set of markers and a function that assigns each marker to a position in the DNA sequence, which is used to help scaffold contigs. Usually this was produced using BACs containing pieces of the whole genome in question. Later, physical map contig-specific sequences are produced by randomly distributing genome sequences in each physical contig. Then, after tagging these contigs with the information provided by BAC-end sequences, contigs can be anchored (Jiang *et al*. 2013).

High-density genetic linkage maps can also be used for scaffolding (Jibran *et al*. 2018). Due to next generation sequencing technology, attaining thousands of SNPs for creating high-density genetic maps has become attainable (Qi *et al*. 2014). This increases the ability to overlay the genetic map, which contains recombination events with the physical maps of a genome (Sim *et al*. 2012). However, due to the preferential or distorted recombination, sometimes markers can be placed in an incorrect order (Jibran *et al*. 2018). Since significant parts of genomes can be heterochromatic (like centromeric regions) and do not undergo meiotic recombination, contigs from such regions remain unordered.

1.6.2 Second generation scaffolding technology (NGS paired-end data)

An alternative way to connect contigs is to generate read pairs sequenced from the two ends of a molecule of an approximately known size, which can then be used to order and orientate contigs (Roach *et al*. 1995). A way to increase throughput of paired-end (PE) sequences

is to use NGS technology. Sequencing technologies like Illumina and 454 can produce PE reads (Xue *et al*. 2013).

Two types of PE libraries, short-insert PE and long-insert mate-pairs (MP), can act as an ideal input for scaffolding (Boetzer *et al*. 2011). Short-insert PE sequencing typically only allow for insert sizes of less than 500 bp, while most commonly MP sequencing can span 1 - 3 kb. The insert size of the NGS second generation scaffolding technologies range from 5 kb libraries, 8kb libraries, 20 kb libraries, up to 40 kb. Other parameters like sequenced pairs, non-duplicate pairs, PCR cycles and relative complexity also have some differences (van Heesch *et al*. 2013).

1.6.3 Third generation scaffolding technologies (Hi-C, optical mapping)

Similar to third generation sequencing technologies, third generation scaffolding technologies also work on long sequences, such as chromosome conformation capture and optimal mapping. They overcome the weakness of short-read scaffolding and largely increase the continuity of the whole chromosomes (Jiao and Schneeberger. 2017).

1.6.3.1 Hi-C technology

The biological functions of genomes are highly related to the three-dimensional folding of chromosomes (Battulin *et al*. 2015), which can make distant functional elements, such as promoters and enhancers, closer to each other. Technologies such as chromosome conformation capture (3C) and Hi-C aim to understand chromosome structure and spatial nuclear arrangement, which will help to make some genomic processes clearer, such as transcription and replication (Mifsud *et al*. 2015; Nagano *et al*. 2013; van Berkum *et al*. 2010).

3C can measure local protein–protein, RNA–RNA, and DNA–DNA interactions (Ramani *et al*. 2017), but its subsequent adaptations require millions of nuclei (Nagano *et al*. 2013) and the choice of a set of target loci, resulting in an impossible genome-wide study (van Berkum *et al*. 2010).

Hi-C, the extension of 3C, is a novel strategy combining capture of chromatin interaction and NGS (Jibran *et al*. 2018). In Hi-C, first cells are fixed, making interacting loci bound to each other through DNA-protein cross-links and remain linked when the DNA is fragmented. Next, ligation events between cross-linked DNA fragments result in a genome-wide library. At the site of the junction, each ligation product is marked with biotin. Then after the library is sheared, the junctions are pulled-down with streptavidin beads and analyzed using a high-throughput sequencer, which reveals the catalog of interacting fragments (Jibran *et al*. 2018; van Berkum *et al*. 2010).

Hi-C allows unbiased, genome-wide studies of long range interactions, the resulting data of which can be used to study chromosome territories, segregation of open and closed chromatin, as well as chromatin structure at the megabase scale (van Berkum *et al*. 2010). A major benefit of Hi-C is that the cross-linking process is random and will not be influenced by DNA sequences themselves. Near-complete pseudo-chromosome assemblies of complex genomes can also be developed by proximity-guided assembly (PGA) using Hi-C data (Jibran *et al*. 2018).

In 2015, Dovetail Genomics developed the Chicago approach (Putnam *et al*. 2016) which is based on Hi-C technology. Such data produces links between genomic regions that can be up to several hundred kb apart and thus are useful for long-range scaffolding (Jiao *et al*. 2017). Hi-C has been applied in both prokaryotes and some eukaryotes, including humans, some species in the Rosaceae family, diploid strawberry (*Fragaria vesca* L.), Asian and European pear (*Pyrus*

pyrifolia [Burm.] Nak. and *P. communis* L.), Chinese plum (*Prunus mume* Siebold & Zucc.) and China rose (*Rosa chinensis* Jacq.) (Jibran *et al*. 2018).

1.6.3.2 Optical mapping

Optical mapping is a variant of restriction mapping, aimed at mapping the location of specific landmarks along DNA. In both optical and restriction mapping, landmarks correspond to the recognition sites of specific restriction enzymes. Optical mapping extends this approach by providing extra information about the order of fragment occurrence in DNA (Samad *et al*. 1995). Generally, by imaging the locations of the restriction sites using fluorescent labels under light microscopes, it generates fingerprints of DNA sequences of several hundred kb (Lam *et al*. 2012). This can help to construct genome-wide maps and guide the order and orientation of sequence contigs.

Several algorithms for scaffolding using optical mapping have been released. For example, first, a score is created for each contig corresponding to each possible placement of the contig in the optical mapping. Second, with these scores, each contig is transformed into an ordered sequence of fragment sizes. A greedy scoring scheme is then applied to find a score for each contig for each possible placement of the contig in the optical restriction maps (ORM). Greedy placement algorithms are then used to place the contigs in a correct order by using the matching scores (Saha and Rajasekaran. 2014). Recently, BioNano Genomics released a commercial high-throughput platform, Irys system, which is based on optical mapping. Both BioNano Genomics optical mapping and Dovetail Genomics chromosome conformation capture data for genome scaffolding. Despite their technical differences, both technologies performed similarly and doubled N50 values (Jiao *et al*. 2017).

Nowadays, long-read sequencing has revolutionized *de novo* genome assembly, and Hi-C sequencing is becoming more economically feasible. However, some open-source scaffolding tools have limitations, such as higher error rates. The scaffolder named SALSA is good at exploiting the genomic proximity information using Hi-C data for long range scaffolding of *de novo* genome assemblies with higher accuracy (Ghurye *et al*. 2017), which is able to contribute to genome construction based on third generation sequencing and downstream improve the plant breeding methods.

1.7 Resistance genes

Plants are continuously exposed to pathogen attacks, but unlike animals, plants lack a circulatory system and antibodies, and cannot move to avoid pathogens (Staskawicz *et al*. 1995). However, plants do contain a type of immune system, the Resistance genes (R genes), which help them fight against pathogens when they are exposed to pathogen attacks (Staskawicz *et al*. 1995). As indicated in the 'gene-for-gene model', resistance in the host and parasite ability of pathogens to cause disease are controlled by pairs of matching genes -- one is the R gene of plants; the other is a parasite gene called the avirulence (Avr) gene. Plants producing a specific R gene product are resistant towards a pathogen that produces the corresponding Avr gene product (Flor. 1971), by direct or indirect interaction (Staskawicz *et al*. 1995).

1.7.1 Carrot diseases

Carrot plants can be affected by various plant pathogens. Alternaria leaf blight (ALB) is the most common and destructive disease in carrot caused by the fungus *Alternaria dauci* (Le Clerc *et al*. 2019). This disease is a world-wide problem (Pryor *et al*. 2002), with losses of 40-

60% (Ben-Noon *et al*. 2003). Fungal infection results in loss of photosynthetic tissues, weakening or even killing the leaves (Ben-Noon *et al*. 2003). Harvesting effectiveness is also negatively affected, as carrots are usually lifted from the ground by their leaves with mechanical harvesting (Soylu *et al*. 2005).

Carrots are also susceptible to diseases caused by bacteria and nematodes. For example, soft-rot, bacterial leaf blight, root-knot and rot are caused by *Pectobacterium carotovorum*, *Xanthomonas campestris* pv. *carotae*, *Meloidogyne* spp. and *Fusarium solani* respectively (Siddiqui *et al*. 2019). Nematodes (*Meloidogyne* spp.) can cause great damage to carrot roots, resulting in root malformation exhibited as forked, distorted, or stunted tap roots with excessive galling (Rao *et al*. 2017), including root-knot, affecting both quality and quantity of carrots (Gugino *et al*. 2006). In addition, the wounds caused by nematodes can be an entry point for other pathogens, especially the soil borne bacterium, *P. carotovorum*, which makes root tissues die and leads to soft-rot diseases (Rao *et al*. 2017).

In order to prevent yield loss due to diseases caused by pathogens, researchers must develop varieties highly resistant to these diseases (Le Clerc *et al*. 2019). Use of genetic linkage maps, QTL mapping and sequenced genomes has enabled the ability to identify R genes. Once R genes and the resistant alleles are identified, they can be targeted using molecular markers to select for the resistant allele and introgressed into susceptible cultivars. Integrated genetic and genomic resources, such as a sequenced genome with good continuity and accurate annotations are fundamental to the efficient study of any trait.

1.7.2 Class of R genes

R genes are characterized by the presence of specific domains. Leucine-rich repeats (LRRs), nucleotide-binding site (NBS), toll-interleukin region (TRR), coiled-coil (CC) and kinase domain (K) are the main domains encoded by R genes (Hulbert *et al*. 2001). The combination of domains defines the respective sub-families (Hulbert *et al*. 2001). The largest family is NBS-LLR, which carries LRRs and NBS domains, comprising an estimated 1% of the genes in the Arabidopsis genome. The NBS-LRR class of R genes can be further subdivided into two groups: the TNL group includes a TRR at the N terminal of a TIR-NB-LRR; non-TNL groups are mainly CNL, which contains a CC at the end to form the CC-NB-LRR R gene (Hulbert *et al*. 2001).

1.7.3 R gene studies in carrots

Researchers first applied genetic engineering methods as single and multi genes for carrot disease resistance. Enhanced resistance was achieved via transformation of a single pathogenesis-related (PR) protein, such as chitinase to protect against fungi (Jayaraj and Punja. 2007; Punja and Raharjo. 1996), thaumatin-like protein to produce herbicide- and disease-resistant carrot (Chen, W. P. and Punja. 2002), β-1,3 glucanase to protect against fungi and peroxidase to protect against necrotrophic pathogens. Others used a combination of two genes (Wally, Jayaraj and Punja. 2009a). In order to provide broad-spectrum resistance towards a wide range of pathogens, researchers ultimately manipulated a plant's innate defense signaling pathways by controlling a large number of induced genes, either directly or indirectly (Wally, Jayaraj and Punja. 2009b).

All the aforementioned methods can study only a limited number of R genes and provide little information about the structures of relations of these genes. After constructing the DH1 carrot genome v2 in 2016, relevant genetics studies could be carried out. Results of orthologous gene characterization showed that 26,320 carrot genes in 13,881 families are identified, with 10,530 of them unique to carrot. Then, 634 putative R genes in carrots were predicted, 206 of which were located in clusters. These putative R genes include classes containing NBS, CNL, NL, RLK and so on. However, in carrot, most R gene classes were under-represented (Iorizzo *et al*. 2016).

1.8 Objectives

Although second generation sequencing technology enabled significant progress in sequencing new crop genomes, the technology is not complete. The short length of the sequences produced by Illumina or 454 technologies, combined with the abundance of repetitive sequences that are longer than the read lengths in plant genomes, make the resulting assembly very fragmented. As a result, the assembly usually contains a large number of gaps of unknown sequences and a large fraction of the genome is not assembled. The current carrot genome v2, developed using second generation sequencing technologies, has such problems. It contains 30,938 contigs with a total length of 386.8 Mb and N50 of 31.2 kb. The total size of gaps is 21.6 Mb. Large amounts of ambiguous areas in the genome may influence future research in carrot genetics and breeding.

With the development of third generation sequencing and scaffolding technologies, new opportunities exist to overcome the challenges posed by assemblies based on second generation sequencing technologies. Read length of sequencing has been increased exponentially, and more

evidence is provided to connect reads in long distance. New software for long reads assembly and scaffolding such as CANU and SALSA have been produced. Some algorithms are also optimized to overcome former disadvantages. The objectives of this study are: (1) to use third generation sequencing and scaffolding technologies to improve the DH1 carrot genome by increasing the length of assemblies and filling more gaps, producing the v3 carrot genome; (2) to predict R genes in both v2 and v3 genomes to determine whether the new genome provides better results. The v3 genome, which has shown great progress in sequence length and continuity, may be more beneficial to R gene prediction than the previous one, ultimately leading to future yield increases of carrots.

Chapter 2: Genome improvement of DH1 carrot

Abstract

The first carrot genome assembly named DH1 carrot genome v2, was published in 2016 (Iorizzo et al., 2016). It was developed primarily using whole-genome sequences obtained from short sequencing technologies such as Illumina. Although the genome assembly achieved standard parameters of high quality, it has the typical problems that genomes assembled with short sequence technologies have. For example, the contiguity at the contig level is low, due to the large amount of inserted unknown sequences (Ns), which are used during the assembly process to fill gaps between contigs that are assembled at scaffold level. Nowadays, third generation sequencing and scaffolding technologies can generate long DNA reads and span long physical distances, which provide an opportunity to improve the carrot genome assembly in terms of contiguity, and to increase the fraction of the genome anchored to chromosome level.

In this study, PacBio, Nanopore and Hi-C sequencing data were used to generate an improved carrot genome assembly v3. The v3 assembly consists of 692 contigs with total length of 438,921,773bp and N50 of 4,945,074bp, accounting for 92.8% of the estimated genome size. Compared with the published DH1 v2 genome assembly, the v3 assembly represents a 159 fold increase in contig N50, and 24 Mb extra sequences anchored to chromosomes level. With less but longer contigs, scaffolds and super-scaffolds, the accuracy and continuity of the whole genome increased significantly.

Introduction

Carrot is an economically important member of the Apiaceae family (Grzebelus *et al*. 2014), which is grown worldwide for the edible taproot on more than one million hectares (Grzebelus *et al*. 2014). The consumption of carrot has increased over the past years, owing to its high nutritional value and low caloric content (Wrzodak *et al*. 2012). Especially, antioxidants and dietary fiber are two contents beneficial to human health. A breeding cycle of carrot is restricted to one per year due to the relatively slow growth in the field. In addition, the need to combine multiple traits into improved cultivars, which requires molecular markers to facilitate selection, is important to assure progress in breeding programs. The development of high quality genetic and genomic resources such as genomes, linkage maps and their integration with phenotyping data in QTL studies are a critical step to routinely and effectively apply marker assisted selection (MAS) or gene editing technologies in breeding programs. An accurate sequenced genome is a prerequisite to identify genes to clone and can increase the efficiency of obtaining functional transformants. Hence, to accelerate the development of improved cultivars, it is important to have access to a high-quality sequenced genome.

The basis of genome construction is obtaining qualified DNA that fulfils the requirements of sequencing library preparation. DNA extraction is the process of isolating and purifying DNA from samples with physical and chemical methods. In 1868, Miescher noticed the precipitation of DNA from pus cells (Dahm and Dahm. 2008). Today, DNA extraction has been a basic method in molecular biology and there are two types: isolation of recombinant DNA such as plasmids or bacteriophage and isolation of chromosomal or genomic DNA (Tan and Yiap. 2009). There have been multiple methods and kits for DNA extraction, using the similar techniques as Miescher's (Ayoib *et al*. 2017). In general, extraction of DNA includes four steps: 1) cell or

tissue disruption, 2)nucleoprotein complexes denaturation,3) nucleases inactivation and 4) contamination removing (Tan and Yiap. 2009). The output DNA should be free of contaminants such as protein, carbohydrate, lipids, or other nucleic acid (Buckingham and Flaws. 2007). Quality and integrity of the output DNA will affect the results of succeeding scientific research (Cseke *et al*. 2005).

The quality of DNA can be measured by 260/280 ratio and 260/230 ratio. The 260/280 ratio indicates whether the DNA sample is contaminated by RNA or protein (Glasel. 1995). According to technical support of Nanodrop (NanoDrop Technologies Inc. 2007), a pure DNA sample 260/280 ratio should be around 1.80 and a pure RNA sample is around 2.0; therefore, if the 260/280 ratio is greater than 1.80, it may indicate that there is an excess of RNA. If the ratio is lower than 1.80, this indicates that there is protein contamination. The 260/230 ratio is another measure of nucleic acid purity. The expected 260/230 ratio of DNA is in the range of 2.0-2.2. If the ratio is lower, it indicates there are contaminants which absorb at 230 nm, such as salts and organic matters. Thus, using the average Nanodrop 260/280 ratio and the 260/230 ratio of the DNA samples, the most efficient DNA extraction method can be selected.

Apart from obtaining qualified DNA, the preparation of the sequencing library also determines the quantity and quality of the DNA reads sequenced. In Nanopore library preparation (Oxford Nanopore Technologies. 2017), there are two AMPure XP bead binding steps. AMPure XP beads can achieve the functions of DNA purification (Beckman Coulter. 2016) and size selection (Beckman Coulter. 2012), utilizing the property of beads that can bind DNA molecules of specific size range in 70 % ethanol or washing buffer and release DNA in elution buffer or water. The concentration of DNA and contaminations in the sample will both influence the binding ability of beads.

In 2016, the genome of DH1 carrot was published and Illumina reads were used for the assembly (Iorizzo *et al*. 2016). The v2 assembly contains 4,907 sequences with N50 of 12.7 Mb, spanning 421.5 Mb and accounting for around 90 % of the estimated carrot genome size of 473 Mb (Iorizzo *et al*. 2016). Sixty scaffolds were anchored to 9 chromosomes, including 50 anchored and oriented ones, with a total length of 361.1 Mb, accounting for about 86 % of the assembled genome and 76 % of the estimated genome size (Iorizzo *et al*. 2016). Although the genome assembly achieved standard parameters of high quality at that time, it has the typical problems that genomes assembled with short sequence technologies have. For example, the total length of gaps in the assembly is 33.67 Mb, accounting for about 8 % of assembled sequences (Iorizzo *et al*. 2016). The objective of this experiment was to use third generation sequencing and scaffolding technologies to improve the DH1 carrot genome assembly in terms of contiguity at the chromosome level.

Materials and Methods

Plant materials

A double haploid orange Nantes type carrot (DH1) was used in this study. DH1 carrot plants were grown in a greenhouse at the NC Research Campus, Kannapolis, NC. In the daytime the temperature in the greenhouse was between 28 to 29.5 ℃, while at night and overcast days it was 25 ℃. There was no artificial light. Young, unexpanded leaves were collected and used for DNA extraction.

DNA extraction and purification

To extract high quality high molecular weight DNA required for Nanopore sequencing, six DNA extraction protocols were tested, including 'rapid CTAB', 'nuclei method', 'QIAGEN mini kit' and three versions of 'general CTAB' method – original 'general CTAB' method, 'CTAB with ethanol' method and 'CTAB without NaAc' method. The cut-tip micropipettes was used in all methods and all steps involving DNA molecules, to prevent DNA fragments from breaking.

Using the 'general CTAB' method, 30 mg of young DH1 carrot leaf tissue was taken as one sample and ground with a mortar and pestle and liquid nitrogen. 1 mL of CTAB (100 mL/L 1M Tris pH 8.0, 40 mL/L 0.5M EDTA, 280 mL/L 5M NaCl, 20 g CTAB (cetyltrimethyl ammonium bromide)), 0.04 g PVP (polyvinylpyrrolidone), 5µL 2-mercaptoethanol and 5 µL of proteinase K was added to each sample. After incubating at 65 °C for 2 hours and centrifuging at max speed (21,130 ×g for Eppendorf 5424 centrifuge) for 5 minutes at room temperature (RT), an equal volume of solution containing chloroform: isoamyl alcohol (24: 1) was added to the supernatant. This step was repeated until the aqueous phase was clear. Next, 0.08 volumes of 3 M sodium acetate and 100 % cold isopropanol was used to precipitate DNA at -20 °C for 2 hours. At RT, after centrifuging at 21,130 ×g for 5 min, the DNA was washed with 800-1000 µ L 75 % ethanol. Then, centrifuged at 21,130 ×g for 5 minutes and decant the supernatant. Finally, the samples were fried using the SpeedVac (Thermo DNA120-115) until all excess ethanol was evaporated and the DNA was re-suspended in 50 µ L nuclease free water.

Two other versions of this general CTAB were tested by modifying the purification steps. First, to better remove RNA, after incubation at 65 °C, 5 µL RNase (100 mg/mL) was added to each sample, followed by incubating at 37 °C for half an hour. Second, to better discard salt

contamination, in the DNA precipitation step, 2.5 volume of 100% ethanol was added rather than isopropanol ('CTAB with ethanol' method). In second general for the CTAB modified method, NaAc was not added to the solution ('CTAB without NaAc' method). Besides, before drying DNA with SpeedVac, washed DNA pellets with 70% ethanol for 3 times. Finally, to precipitate as much DNA possible, samples with NaAc and ethanol were kept at -20 °C overnight and then centrifuged using Eppendorf centrifuge 5417R at maximum speed (14,000 rpm) under 4 °C for 30 minutes to minimize loss of DNA.

'Rapid CTAB' is a CTAB based method that allowed to extract genomic DNA in less than one day (Allen *et al*. 2006). First, 200 mg fresh tissue sample was ground in the mortar containing liquid nitrogen. Then, 1.2 mL 65 ℃ preheated extraction buffer (0.1 M Tris, 1.4 M NaCl, 0.02 M EDTA and 0.02 g/ml of CTAB, 0.5–1 % (v/v) 2-Mercaptoethanol) was added to the frozen tissue powder. The mixture was incubated at 65 ℃ for 30 min in water bath, followed by centrifugation at 13,500 g for 10 min at room temperature (RT). Next, proteins were removed using 800 µL of phenol: chloroform: isoamyl alcohol, by inverting tubes for 20 min at RT. The aqueous (upper) layer was incubated in 800 µL -20 ℃ isopropanol for 10 min to precipitate DNA. After removing the supernatant by centrifuging at 13,500 g for 10 min, the pellet was resuspended in 250 µL TE buffer at RT. Finally, after incubating with 2.5 µL DNase-free RNase at 37 ℃ for 30 min, 25 µL 3 M NaAc and 600 µL -20 ℃ precooled ethanol for 20 min, the DNA pellet was washed with 70 % cold (-20 ℃) ethanol and finally resuspended in 25 µL water.

In 'nuclei method', nuclei were isolate first, and then the DNA was extracted from these nuclei, aiming to keep the DNA intact (Zhang, M. *et al*. 2012). The downside of this approach is that this method takes three days. First, around 100 g frozen or fresh tissue was ground with

liquid nitrogen in motar with pestle. After adding 800 – 1,000 mL NIB buffer (1× HB solution, 0.5% Triton X-100, 0.15% (vol/vol) 2-mercaptoethanol), the mixture was filtered through cheesecloth. Then, the suspension was centrifugated in a fixed-angle rotor at 3,110 g, 4 ℃ for 20 min. 1% (wt/vol) LMP agarose was mixed with 45 ℃ prewarmed nuclei or cells and solidified in LMP agarose plugs on ice. The plugs were incubated in lysis buffer (0.5 M EDTA (pH 9.0–9.4), 1% (wt/vol) sodium lauryl sarcosine, 0.3 mg ml^{-1} proteinase K) for 16 - 24 h at 50 ℃ and washed in 50 mM EDTA (pH 8.0) for 1 h on ice. Plugs were washed three times with 10 – 20 volumes of ice-cold TE (10 mM Tris-HCl (pH 8.0), 1.0 mM EDTA), ice-cold TE containing 0.1 mM PMSF and ice-cold TE, respectively, each wash for 1 h. Finally, the plugs containing DNA were digested and its quality was assessed.

QIAGEN 'mini kit' is a commercial DNA extraction kit (DNeasy Plant Mini Kit) available through QIAGEN (Cat. No. 69104), which enables extracting DNA in half a day. Manufacturer's instructions were included in the kit, which the extraction process based on. Less than 100 mg wet weight carrot leaf tissues were used according to instruction (QIAGEN. 2016).

All DNA extraction methods were repeated at least 3 times. All DNA samples were assayed for DNA concentration using Qubit (Invitrogen Q33216), the 260/280 ratio and 260/230 ratio was assessed using a Nanodrop 2000 spectrophotometer (Thermo Scientific).

Beads binding tests

These tests were carried out to address high DNA loss with Nanopore sequencing and AMPure XP beads (Beckman Coulter A63880) binding steps. After learning about the bead theory, two factors that were most likely to be influential were selected, which were salt contamination and initial DNA concentration.

For library preparation, the 1D Lambda Control Experiment (SQK-LSK108) (Oxford Nanopore Technologies. 2017) was used. In this protocol, AMPure XP beads binding step requires the DNA recycling rate to be greater than 70 %. DNA extracted with 'general CTAB' and 'CTAB without NaAc' were chosen as the former is supposed to contain more salt than the later. For each sample, several concentrations were selected. They were mixed with the same volume of AMPure XP beads, followed by magnetic rack exposure and separation of supernatant from beads. Following DNA bead binding, the concentration of DNA was tested by Qubit. After washing with 70% ethanol, being kept on the magnetic rack and drying, DNA was eluted from beads with ddH_2O. Then the DNA concentration of the elution was evaluated and the recycling rate could be calculated by dividing the amount of DNA in elution by that at the beginning.

Library preparation and sequencing

Nanopore library preparation and sequencing

Four Nanopore libraries were prepared and each was used to run a flow cell independently. Libraries were prepared according to the 1D Lambda Control Experiment (SQK-LSK108) protocol from the manufacturer. DNA was first diluted to the concentration of 26 ng/μL and its intactness was evaluated in an 0.8% agarose gel through electrophoresis, running for about half an hour. DNA samples were then spun twice to thoroughly fragment DNA using a filtered tube, Fisher Scientific Covaris g-TUBE (Covaris, Catalog No. NC0380758), at 6000 rpm with an Eppendorf 5424 centrifuge, each for 1 min. To evaluate the DNA fragment distribution, DNA samples were analyzed using Tape Station (Agilent TapeStation). The end-prep was performed using Ultra II End-prep reaction buffer and Ultra II End-prep enzyme mix (NEBNext, Catalog No. E7546S), followed by incubating for 5 minutes at 20 °C and 5 minutes at 65 °C in a

PCR thermal cycler. Agencourt AMPure XP beads (Beckman Coulter, Catalog No. A63880) were used for the bead clean up steps. Samples were mixed using an equal volume of eluted DNA and beads, incubated on a rotator mixer at RT for 5 minutes, centrifuged for 1 to 2 seconds to spin down the beads and solution. Then the tube was placed on a magnetic rack until the supernatant was clear. The supernatant was then removed using a pipette and the beads were washed twice with 200 µL of freshly prepared 70% ethanol without disturbing the pellet. Beads were then eluted with 31 µL of nuclease-free water. Next, Adapter Mix 1D in kit and NEB Blunt/ TA Ligase Master Mix (Catalog No. M0367) were applied for adapter ligation, following by a second bead clean-up step using 140 µL ABB Buffer instead of 70 % ethanol. The DNA was then re-eluted in 15 µL elution buffer. Finally, the DNA was mixed with running buffer with fuel mix (RBF), library loading beads (LLB) solutions and nuclease-free water and loaded into Nanopore sequencing MinION. To start the sequencing process the MinION was connected to the host computer and the software MinKNOW.

Hi-C library preparation and sequencing

The Hi-C library was prepared following the Phase Genomics Plant Hi-C Kit protocol. About 0.2 g of young carrot leaf tissues were harvested, kept frozen at -80 °C and used to prepare the library. Following the manufacturer's instructions, the crosslinking solution, quenching solution, Tris Buffered Saline (TBS), plant lysis buffer was used to cross-link the DNA in vivo and cell lysis. The DNA was then fragmented using endonucleases. The fragmented DNA was biotinylated and ligated using the Proximity Ligation Buffer, Proximity Ligation Enzyme and RX Enzyme. This step, named Proximity Ligation, created chimeric

junctions between adjacent sequences. Finally, after PCR and purification, the DNA was sent to Novogene genome sequencing company for paired-end sequencing.

Genome assembly

Genome assembly includes *de novo* assembly, polishing, scaffolding, gap filling, anchoring, orienting and mis-assembly. The bioinformatics tools and input data are shown in Fig. 2.1. The information of the reads listed in the picture is in Table 2.4. The PacBio reads and the PE BAC end reads are from our lab, which are already available. Illumina PE reads can be downloaded from NCBI (sequence read archive: SRP062113).

Assembly and polishing

DH1 PacBio reads from 8 SMRT cells corresponding to 17.47 Gb sequencing data (Table 2.5) were *de novo* assembled using CANU (Koren *et al*. 2017), a software designed to process PacBio sequences. The software performs three key steps: 1) reads auto-correction to reduce the high error rate typical of PacBio reads; 2) *de novo* assembly using corrected reads; 3) reconstruct phased haplotypes. The statistics of the reads and the statistics of *de novo* assembly are summarized in Table 2.6.

After *de novo* assembly, two polishing steps were performed to correct any remaining uncorrected sequencing errors. The first polishing step was performed using Quiver (PacBio GenomicConsensus) which used PacBio long reads raw data as the input. By aligning multiple PacBio reads, the bases with errors had a large probability to be corrected due to evidence given by most of the reads. A second polishing step was performed using Pilon (Walker *et al*. 2014) and Illumina reads (Hi-C data).

Scaffolding and gap filling

After the *de novo* assembly step was completed, contigs were connected using Hi-C paired end sequences, to construct scaffolds. For this process, Hi-C reads were aligned against the contigs using BWA MEM (version: 0.7.15-r1140) (Li. 2013). Then, using the input bed file from bwa, scaffolds were constructed using SALSA (Ghurye *et al*. 2017) (parameters: -e GATC –i 500). SALSA generates a fasta file containing scaffold and contig (un-scaffolded) sequences, with a list of coordinates where chimeric (assembly errors) sequences were predicted. When contigs are connected to produce a scaffold, gaps of unknown sequences ("Ns") are inserted between two contigs. PacBio and Nanopore reads were used to fill the gaps using PBJelly (English *et al*. 2012) with default parameters.

Super-scaffolding and mis-assembly

Super-scaffolds were constructed from two or more scaffolds, connecting with 1,000 'N's. After scaffolding and gap filling, PE BAC end reads were used to connect two or more scaffolds into super-scaffolds. If there were more than three PE BACs as supports, two scaffolds were connected with 1000 'N's in the middle.

Chimeric sequences are incorrect connections of sequences. Chimeric sequences can be identified as scaffolds containing sequences of markers mapped to different linkage groups or to distal locations of the same linkage group; or genomic regions where raw reads could not be mapped (zero coverage after mapping with bwa). To identify and correct chimeric sequences, five sources of sequences were aligned against the v3 carrot assembly: PacBio reads, Hi-C reads, Illumina 40k PE reads, Illumina 20k PE reads, Illumina 10k PE reads, PE BAC end reads (Table

2.5), as well as SSR and SNP markers. All the reads were mapped using software BWA MEM (version: 0.7.15-r1140) using the default parameters.

For each scaffold or contig, the aligned sequences were uploaded and visualized in Integrative Genomics Viewer (IGV) (Thorvaldsdóttir *et al*. 2011). First, scaffolds and contigs containing markers mapped to different chromosomes were recorded according to the number and distribution of markers. Then, the coverage of aligned sequences was checked. Zero coverage regions can help to reduce the possible mis-assembly areas. Finally, if the Illumina PE reads or BAC PE reads around the suspected areas had read pairs mapped to other sequences, this provides evidence of a mis-assembly. All these regions were manually inspected. The mid-points of the chimeric regions were defined as mis-assembly points, which were used to break the sequences. The corrected scaffolds were used to construct super-scaffolds, putting 1,000 'N's in between. Adjacent super-scaffolds in each chromosome were connected by 2000 'N's.

Anchoring and orienting

A total of 5,010 marker sequences extracted from 46 carrot linkage maps were mapped against the newly developed assembly (v3) using BWA MEM. The resulting sam file was then filtered using a custom python script to remove the hard clips. With the same methods, markers were mapped on carrot genome v2, which helps to find the connections between markers and chromosomes. Via these two connections, the number of markers on each sequence and in each chromosome was calculated. Hence, the connection of sequences and chromosomes were determined.

Then, bin markers, which represent unique recombination events, were mapped and uploaded to IGV. The bin map is the linkage map representing unique recombination events,

which can help to determine the order of scaffolds or contigs on chromosomes. Sequences including at least two bin markers can be oriented and sequences containing at least three bin markers can be confidently oriented. Sequences with only one marker could not be oriented.

In every assembly step, statistics including number of sequences, sequence size, N50, L50, N90, L90 and so on were calculated using software BBMap (Bushnell.) and custom python scripts.

Results

DNA extraction and purification

In term of quantity, the 'general CTAB' protocol and its two modifications were the most effective methods to extract DNA, especially the 'general CTAB without NaAc' method, which produced 16,579 ng DNA from 1 g of carrot tissue (Table 2.1). The 'Nuclei method' was the least efficient protocol, producing only 8.58 ng DNA per one g of carrot leaf tissue. The 'Rapid CTAB' and 'QIAGEN mini kit' extracted an average of 3,494.28 ng and 2,013.33 ng of DNA from every g of tissue, respectively, which was acceptable but not as good as the 'general CTAB without NaAc' protocol.

The 260/280 ratio showed that DNA samples from 'rapid CTAB' and 'nuclei method' had low 260/280 ratios due to protein contaminations. The 'nuclei method' had the lowest 260/280 ratio of only 1.32. DNA samples using the QIAGEN 'mini kit' and 'general CTAB' had 260/280 ratio higher than 2.0, which meant that there was RNA in the samples. 'General CTAB with ethanol' method produced DNA with 260/280 ratio of exactly 2.0; while DNA produced

from ‘general CTAB without NaAc’ method had 260/280 ratio of 1.88. Thus ‘general CTAB without NaAc’ was the best one.

The 260/230 ratio of DNA extracted using QIAGEN ‘mini kit’ protocol was higher than 2.2, while the ratios of DNA produced by ‘rapid CTAB’, ‘nuclei method’, ‘general CTAB’ and ‘general CTAB without NaAc’ were smaller than 1.5, indicating all the DNA extracts were contaminated by salt. ‘General CTAB without NaAc’ method produced DNA with 260/230 ratio of 1.99, which is close to the 2.0 – 2.2 range required by the Nanopore library.

To summarize, the ‘general CTAB without NaAc’ was the best method among the six protocols to extract high quality carrot DNA, since it produced a larger concentration of DNA with less protein and salt contaminations. The QIAGEN Mini kit produced DNA in the shortest amount of time, but it did not yield as much DNA as the ‘general CTAB’ protocol, and there were salt mixed in the samples. The ‘rapid CTAB’ method produced a relatively large amount of DNA, but the DNA was highly contaminated with protein and salt. The ‘nuclei method’ was the least effective method in terms of concentration of DNA as well as contamination of protein and salt.

Bead binding step

The results of the bead binding tests are in Table 2.2. Comparing samples with similar concentrations but using a different protocol, DNA samples with less salt had a higher recycling rate. For example, using a middle-level concentration of 40 – 42 ng/μL of DNA, 25 % DNA extracted with the ‘general CTAB’ method could be recycled from beads while 32.94 % DNA extracted by ‘CTAB without NaAc’ was recycled. Regarding samples with a low-level concentration of 25 – 27 ng/μL of DNA, 52.8 % DNA extracted by ‘general CTAB’ could be

recycled while 67.18 % DNA extracted with 'CTAB without NaAc' was recycled. DNA samples extracted with the 'general CTAB' protocol, the DNA concentration of the supernatant was higher, indicating that the salt would influence the bead binding, preventing DNA bound to the beads. Since 'CTAB without NaAc' method also performed the best, more concentrations using this method were studied. There was a peak recycling rate around sample concentration of 21 ng/µL, at which more than 70% of the DNA could be recycled from beads (Table 2.2). DNA concentrations between 20.7 ng/µL to 26.2 ng/µL, the recycling rates still kept high. Considering that there are two bead binding steps in the Nanopore library preparation and each expects to maintain 70% of the DNA input, input DNA with approximate concentration of 26 ng/µL was selected, therefore, the amount of input DNA was around 1,200 ng.

Nanopore library preparation and sequence data analysis

The extracted DNA was intact, larger than 10,000 bp with no smears indicating high molecular weight DNA (Fig 2.3.). Then, the DNA was sheared using Covaris g-tube by centrifuging in Eppendorf 5424 centrifuge at 6,000 rpm for 1 min twice. The TapeStation results indicated that the size of the DNA sample after shearing met the requirements of the Nanopore library input (Fig 2.2.).

The Nanopore library was prepared four times using four different DNA samples. The quality, concentration and the average DNA recycling rate for all of the samples in the two bead binding steps were recorded (Table 2.3). Three of the four libraries satisfied the requirement of more than a 70 % recycling rate.

134,751 Nanopore reads were produced with a total length of over 717 Mb, with a maximum read length of 1.38 Mb (Table 2.4). However, the coverage (total read length divided

by genome size) of Nanopore sequences was only 1.5X, indicating that it was not possible to cover most of the carrot genome and the chance of correcting errors in long read sequencing with low coverage is very low. Hence, the Nanopore reads were not used for the de novo genome assembly. On the other hand, the PacBio reads totaled 1,177,331, with a total length of more than 17,472 Mb and was a better material for the genome assembly, due to its high 37 X coverage. Although the maximal read size of Nanopore is more than two times greater than that of PacBio, the mean read size of PacBio is almost three times larger than that of Nanopore.

Construction of v3 carrot genome

In *de novo* assembly, scaffolding and gap filling step, statistics of every step constructing the v3 carrot genome are summarized in Table 2.5. The PacBio reads *de novo* assembly produced 817 contigs with a total length of 436.3 Mb and a contig N50 of 3.37 Mb. This assembled genome account for 92.24 % of the carrot estimated genome size. After scaffolding and gap filling, the number of sequences, including scaffolds and contigs, reduced to 521. The N50 of all the sequences increased to 7.94 Mb.

In order to identify the mis-assemblies, Pacbio reads, Hi-C reads, Illumina PE reads and BAC PE reads were mapped on the v3 assembly. The mapping rates of 40 k, 20 k, 10 k Illumina reads are 98.97 %, 98.11 % and 98.48 %, respectively. Chimeric sequences were identified as those with markers mapped to different linkage groups (Fig. 2.4). V3 assembly had only 3 scaffolds (Contig 35, 348 and 14) that contained mis-assemblies. Two of them (Contig 35 and 348) were broken into 2 sequences respectively while one (contig 14) was broken into 3 sequences at the mis-assembly points using a custom Python script. Then the PE BACs enabled the connection of 82 contigs and scaffolds into 22 super-scaffolds.

According to markers, 86 contigs and scaffolds (21 super-scaffolds) with a total length of 385.9 Mb were anchored to chromosomes in v3 genome. All these super-scaffolds with a total length of 383.1 Mb containing more than one bin marker were oriented (Table 2.7).

Comparison of v2 and v3 carrot genome

In the step of *de novo* assembly, scaffolding and gap filling, the new v3 assembly was compared with the v2 carrot genome assembly (Table 2.6). In comparison the N50 of v3 contigs increased from 31 kb to 4.9 Mb (159-fold increase). Meanwhile, owing to this increase of sequence continuity, the number and total length of gaps decreased. In assembly v2, there are 25,255 gaps with a total size of 33.7 Mb, accounting for almost 8 % of the assembled sequences. But in assembly v3, there are only 171 gaps with a total length of 48.5 kb, accounting for 0.01 % of the assembled sequences. Hence, since the sequences are longer and the number of unknown regions reduces, the fraction of known sequences and the continuity of the carrot genome increase significantly. The fraction of assembled genome as compared with the estimated genome size (473 Mb) is also increased slightly, from 82.58 % to 92.81 %.

As for mis-assembly, compared to the v2 genome which had 135 scaffolds that needed to be corrected, the v3 assembly constructed with long reads only had 3 scaffolds (Contig 35, 348 and 14) that contained mis-assemblies. Besides, due to the benefit of longer sequencing reads, number of super-scaffolds in v3 carrot genome, 22, is much less than that of v2 carrot genome, which is 89. It would reduce the difficulty of anchoring in the following step and increase the continuity of the genome.

According to Table 2.7, v3 genome contained 86 contigs and scaffolds (21 super-scaffolds) with a total length of 385.9 Mb anchored to chromosomes. The anchored super-

scaffolds had a total length of 383.1 Mb. In carrot genome v2, 914 contigs and scaffolds (60 super-scaffolds) with total length of 362 Mb were anchored to nine chromosomes. Among them, 52 super-scaffolds were anchored by at least three markers. 353 Mb in total was anchored and oriented (Iorizzo *et al*. 2016). Table 2.8 compares the chromosome scale assembly of v2 and v3. In addition, compared to v2 assembly, the length of unknown regions in the v3 assembly decreased by 170 times.

Discussion and conclusion

In summary, DNA of DH1 carrot was extracted and sequenced using Nanopore sequencing. In order to get high quality DNA to satisfy the requirements of Nanopore sequencing, the 'general CTAB' protocol was modified. The crucial modification step was not adding sodium acetate during DNA precipitation. Therefore, less salt was introduced into the sample. By doing the bead binding test, the crucial factors, DNA concentration and salt contaminations, that should be paid attention to during sample preparation were figured out. Hence, most of our DNA samples would not be lost during library preparation.

Nanopore sequencing can sequence long reads, but Nanopore library preparation step is challenging, resulting from its strict quality requirements of the input DNA. Beads binding step in the library preparation protocol aims to clean and select the DNA of specific length, which is quite sensitive and easy to be influenced by the intactness and purity of DNA. If the input DNA is too long, the beads-DNA mixture will become thick and hard to separate; while the DNA is over-fragmented, most DNA will be lost during beads washing. In addition, even if there is little amount of salt contamination, more DNA than expected will lose connection with beads and be

washed away. All these situations reduce DNA recycling rate, resulting in insufficient DNA input into sequencing machine.

The DH1 carrot assembly v3 was constructed mainly with PacBio read, Hi-C reads and Nanopore reads. Compared with the DH1 carrot assembly v2, which was constructed using Illumina short reads, the number of contigs in the v3 assembly decreased from 33,351 to 692. The N50 of contigs increases by 159-fold and the total length increased by 48.3 Mb, resulting in the number and total length of gaps in the v3 assembly decreased significantly. In assembly v3, gaps only account for 0.01% of the assembled sequences. After anchoring, mis-assembly and orienting, the number of anchored sequences and super-scaffolds in v3 genome becomes smaller and the length of unknown regions reduced by 170 times compared to the v2 genome. Hence, the workload and difficulty of connecting contigs into scaffolds, super-scaffolds and chromosomes are reduced, but the accuracy and continuity of the genome increases.

To date, numerous genomes have been sequenced with third generation sequencing technologies, mainly PacBio sequencing. In 2019, the genome assembly of a tropical maize (*Zea mays*) small-kernel inbred line was published with a predicted genome length of 2.16 Gb, accounting for 93.1 % of estimated genome size, in 708 scaffolds with an N50 of 73.24 Mb and a contig N50 of 15.78 Mb (Yang, N. *et al*. 2019). A high-quality peanut (*Arachis hypogaea*) genome sequence was published at the same year with a predicted genome length of 2.54 Gb, accounting for 94 % of estimated genome size, in 7,232 contigs with N50 of 1.5 Mb or 20 pseudomolecules with an N50 of 135 Mb (Zhuang *et al*. 2019). The genome of upland cotton (*Gossypium hirsutum* L.) was also sequenced with PacBio and assembled with the help of Hi-C technology, 2.23 Gb of the assembly was anchored, representing 97.4% of the assembly, which has a large improvement comparing to the former assembly (Yang, Z. E. *et al*. 2019). Apart from

plants, the genomes of spotted lanternfly (*Lycorma delicatula*) (Kingan *et al*. 2019), goldfish (*Carassius auratus*) (Chen, Z. L. *et al*. 2019), human (Wenger *et al*. 2019) and so on were also assembled by long read sequencing and demonstrated significant improvements in term of contiguity and genome coverage.

Indeed, with long reads sequencing technologies, genome coverage of assemblies stated above are all more than 90 %. Our DH1 v3 assembly has genome coverage of 92.81 %, which is not the highest but also in this range. N50 of contigs varies a lot, depending on species and ranging from 1.5 Mb to more than 15 Mb. Our assembly has contig N50 of around 5 Mb, also in this range. As for the percent of anchored sequences in assembled sequences, our genome is 92.67 %, which is lower than those of upland cotton genome (97.4 %) and aphid (99.2 %). Most importantly the fraction of the genome anchored was 24 Mb higher than the v2 genome assembly. The gap rate of our assembly (0.01 %), calculated by total length of gaps / total length of assembled sequences, is lower compared to that of small-kernel maize genome, which is 0.34 %.

Chapter 3: R gene prediction with v3 carrot genome

Abstract

Carrots (*Daucus carota* L.) are a popular nutritious vegetable consumed worldwide, however, it is susceptible to several disease-causing pathogens such as bacteria, fungi, viruses and nematodes, which can significantly reduce yields. Use of natural or innate resistance in terms of resistance genes or R genes, is an economical and sustainable method of preventing and managing disease. High quality sequenced genomes with accurate annotations can facilitate the research process. The aim of this study was to compare R genes predicted from the v2 carrot genome with the improved long read v3 genome to determine improvements using third generation sequencing technologies over previous short read technologies.

The results found that in v3 genome, over 300 more R genes and over 3,500 more R gene domains were predicted than in v2 genome. Percent of R genes that locate on the nine chromosomes of carrot also has a slight increase, from 98.3 % to 99.4 %. In addition, in the v3 genome, there are more short R genes predicted as well as less intron regions. R genes in the v3 genome also contain more domains on average. Pairwise comparison shows that R genes in the v3 genome has better continuity and many are unique to the v3 genome.

Introduction

Carrot (*Daucus carota* L.) is a popular crop around the world, which is grown on more than one million hectares in temperate climate regions (Grzebelus *et al*. 2014). The world area harvested and world production area of carrot also increased during the last 50 years, according

to FAO. However, carrot is affected by various plant pathogens. For example, Alternaria leaf blight (ALB) is a major foliar disease caused by the fungus *Alternaria dauci* (Le Clerc *et al*. 2019), which caused yellow leaves and brown spots on leaves, decreasing plants' ability of photosynthesis and nutrient accumulation significantly. Other important diseases which affect carrot include soft-rot, bacterial leaf blight, root-knot and rot are caused by *Pectobacterium carotovorum*, *Xanthomonas campestris* pv. *carotae*, *Meloidogyne* spp. and *Fusarium solani*, respectively (Siddiqui *et al*. 2019). If these pathogens infect carrots during flowering and seed development process, seed-borne diseases will increase. These diseases caused huge lose to carrot production. For example, it is recorded crater rot led to losses as large as 50-70 % in Denmark (Simon et al. 2019); *Fusarium* spp. resulted in up to 80 % losses in China (Zhang, X. Y. *et al*. 2014).

Resistance genes (R genes) help plants fight against pathogens when plants are exposed to pathogen attacks (Staskawicz *et al*. 1995). As indicated by the 'gene-for-gene model', the inheritance of both resistance in the host and the parasites' ability to cause disease is controlled by pairs of matching genes -- one is R gene of plants; the other is a gene in the parasite called the avirulence (Avr) gene. Plants producing a specific R gene product are resistant towards a pathogen that produces the corresponding Avr gene product (Flor. 1971), by direct or indirect interaction (Staskawicz *et al*. 1995).

An important goal of breeders is to develop varieties highly resistance to multiple diseases (Le Clerc *et al*. 2019). Molecular methods can deliver a targeted approach for integrating multiple disease resistance loci into improved cultivars. Segregating populations and the corresponding genetic linkage maps with phenotypic data for disease resistance allow for identification of QTLs and molecular markers associated with resistance. Markers associated

with these regions can then be used by breeders to ensure that selected lines retain resistance genes by having the correct 'allele' at the corresponding QTL or region in the genome. Furthermore, the use of integrated genetic maps with sequenced genomes enables the quick identification of candidate genes underlying QTL, quickening the research process over older methods of genome walking to identify candidate genes. This is all dependent on a high-quality reference genome, as the sequence of interest must be included in the genome in order to be located.

There have been improvements to the carrot v3 over the v2 genome, which include an increase in sequence length and continuity. To infer improvements of the v3 over v2 genome, predicted R genes from the two carrot genomes will be used as a point of reference to compare the two genomes. A high-quality genome with accurate annotations of genes and gene models will enable the identification of candidate resistance genes which can be utilized to manage disease in carrot.

Materials and Methods

R genes prediction

DRAGO2 software (Osuna-Cruz *et al*. 2018) was used to predict R gene domains and the corresponding R gene families. To compare the number and the structure of R genes between the v2 and v3 genomes, predicted protein sequences from these two genomes were used as input files for DRAGO2. Multiple custom scripts were developed and used to analyze the R gene prediction results from DRAGO2 and obtain multiple summary statistic parameters that are reported in Table 3.1 and 3.2.

Pairwise comparison of R gene structures in v2 and v3 genome

To identify the genomic regions/loci in the v3 genome that span the sequences of the R genes predicted in the v2 genome and vice versa, R genes predicted in the v2 genome and v3 genomes were mapped onto the v3 and v2 genomes, respectively. Bedtools multiinter was used with default parameters to compare the structures of the predicted R genes. In addition, Iso-seqs, or the full-length transcript isoform sequences, were checked in Jbrowse manually. If multiple genes shared the same Iso-seq, they can be assumed to have actually originated from one gene.

This pairwise comparison led to identify nine structural categories: 'completely identical', 'identical positions but different domains', 'mutually extend', 'v2 inside v3', 'v3 inside v2', 'interlaced', 'only in v2', 'only in v3' and 'split gene' (Fig.3.1). 'Completely identical' means that the R genes span the same region/locus and contain identical domains in both v2 and v3 genomes. 'Identical positions but different domains' means that the R genes span the same region/locus in the two genomes, but the domains they contain are different. 'Mutually extend' means that the R genes span different start points and end points but they share a genomic region in v2 and v3 genomes; only the start point in one genome and the end point in the other genome are extended. 'v2 inside v3' means that the R gene is predicted in both genomes, but the start and end points in the v2 genome are both extended compared to the start and end points in v3. 'v3 inside v2' means that the R gene is predicted in both genomes, but the start and end points spanning the gene locus in the v3 genome are both extended compared to the start and end points in v2. 'Interlaced' is a combination of mutually extended and the inside categories. 'Only in v2' refers to R gene predicted only in the v2 genome but not in v3; and vice versa for 'only in v3'. In the former eight structural categories, not every R gene has the

corresponding Iso-seqs, but two different R genes will not share the same Iso-seqs. In ‘split gene’ structure, however, two different R genes share the same Iso-seq.

Results

R gene prediction

In total 1,646 R genes were predicted in the v2 carrot genome, containing 10,494 domains (Table.3.1). Among them, 1,618 R genes were anchored to the nine chromosomes, accounting for 98.3 % of all the R genes predicted in the v2 genome. In the v3 genome, 1,982 R genes containing 14,075 domains were predicted, with 1,971 R genes anchored to the nine chromosomes, accounting for 99.4 % of all the R genes predicted in the v3 genome. This indicates that there were over 300 more R genes with 3,500 more domains predicted in the v3 genome than in the v2 genome. The percent of R genes anchored to chromosomes also slightly increased in the v3 genome. Hence, researchers will benefit more from the v3 genome since it can provide additional, varied, and more complete R genes, most of which are anchored to chromosomes.

A comparative analysis of the R gene structure revealed the average and median length of the genomic loci (intron plus exons) spanning the predicted R genes were less (Table.3.2) in the v3 genome compared to v2 genome. Apart from this observation, the longest R gene in v3 genome is approximately 10 kb longer than that in the v2 genome.

As for R gene coding sequences (CDS), the average and median lengths in the v3 are larger than those in the v2, but the number of CDS per R gene in v3 is a little less than that in v2, which implies the average length of every CDS in the v3 genome is larger. Genes are composed

of CDS, introns, and untranslated region (UTR), considering this we found that the average and median length of R genes in the v3 genome is less but the length of CDS in v3 is larger, it can be concluded that introns and UTR in v3 R genes are less than those in v2 genome. An example is in Fig.3.2, which shows less intronic regions in the v3 genome, indicating better continuity of v3 genome. Besides, the total domain number has an obvious increase in the v3 genome. According to t-test, it is 99 % confident that the average number of domains per R gene of v3 genome is higher than that of v2 genome.

Pairwise comparison of R gene structures in the v2 and v3 genomes

Examples of all nine situations of R gene comparisons between the v2 and v3 genomes are shown in Fig.3.3. 585 R genes were predicted in both genomes spanning the same region/locus and containing the same domains (Table.3.3). More than 200 R genes were predicted in both genomes spanning the same start and end points, but within these genes around 200 more domains were detected in the v3 genome.

There were 69 original R genes that were predicted in the v2 genome; 58 had corresponding predicted genes in the v3 genome but were not predicted as R genes by DRAGO2, and 11 of the genes did not have corresponding genes in the v3 genome. In contrast, 382 R genes were predicted in the v3 genome but not in the v2 genome. Among these, 278 genes had corresponding genes in the v2 genome but those genes were not predicted as R genes by DRAGO2; 104 R genes had no corresponding genes in the v2 genome.

By checking R genes and Iso-seq in JBrowse manually, 2 pairs of R genes were identified in the v2 carrot genome sharing the same Iso-seq, which means they were split from three genes (Fig.3.4). However, there were no R genes sharing the same Iso-seq found in the v3 genome.

This indicates that in the improved v3 carrot genome, more than 100 R genes were detected that were not predicted in the v2 genome. The v3 genome had more continuity and accuracy of R genes compared to the v2, and additionally, there were no split R genes in the v3 genome.

Discussion and conclusion

In summary, approximately 1/3 of the predicted R genes in the v2 and v3 genome were the same, including the spanning regions and the domains. The majority of the R genes could be predicted in both genomes, although their spanning regions and domains contained differed. Additionally, there were situations in which one R gene in a genome had multiple corresponding R genes in the other genome.

Overall, more than 300 additional R genes and 3,500 more R gene domains were predicted in the v3 carrot genome. R genes predicted in the v3 genome typically contained more domains than in the v2 genome. Many of the R genes identified in the v3 genome were unique, which is more informative for studying disease resistance. In the v3 genome, the average R gene coding sequences were longer, indicating that there were less introns and UTR regions compared to the v2. Additionally, there was no split R gene in the v3 carrot genome, indicating an improvement of genome continuity.

In the pairwise comparison of R genes predicted in these two genomes, there are 58 R genes in v2 genome that have corresponding genes in v3 genome; however, these genes in v3 are not predicted as R genes. In order to figure out the possible reason, the consistency of DNA sequences of these 58 R genes in the two genomes were checked. The results showed about 20 out of 58 R genes had the same DNA sequences in the two genomes. Hence, one of the reasons

why some genes were predicted as R genes in v2 but not in v3 is the DNA base pair differences in the two genomes.

A drawback of R gene comparison in this chapter is that the previous gene prediction steps in v2 and v3 genome are not exactly the same. Owing to the technology development, Iso-seq provided a large amount of information for gene prediction in v3 genome. However, the gene prediction in the v2 genome was not guided by Iso-seq. It indicates the advantages of the v3 genome in gene prediction might not only result from its better continuity, but also from the information provided by Iso-seq. The solution to this problem is also including Iso-seq information during gene prediction in the v2 genome.

Although there is still room to improve this experiment, the v3 genome has better continuity and is superior to v. 2. An improved reference genome with a larger, more varied and complete R gene collection with more domains and higher continuity will benefit the study and development of resistant carrot cultivars.

REFERENCES

Alasalvar, C., M. Al-Farsi, P. C. Quantick, F. Shahidi and R. Wiktorowicz, 2005 Effect of chill storage and modified atmosphere packaging (MAP) on antioxidant activity, anthocyanins, carotenoids, phenolics and sensory quality of ready-to-eat shredded orange and purple carrots. Food Chem. **89:** 69-76.

Alessandro, M. S., C. R. Galmarini, M. Iorizzo and P. W. Simon, 2013a Molecular mapping of vernalization requirement and fertility restoration genes in carrot. Theor. Appl. Genet. **126:** 415-423.

Alessandro, M. S., C. R. Galmarini, M. Iorizzo and P. W. Simon, 2013b Molecular mapping of vernalization requirement and fertility restoration genes in carrot. Theor. Appl. Genet. **126:** 415-423.

Ali, A., W. C. Matthews, P. F. Cavagnaro, M. Iorizzo, P. A. Roberts *et al*, 2014 Inheritance and mapping of mj-2, a new source of root-knot nematode (*meloidogyne javanica*) resistance in carrot. J. Hered. **105:** 288-291.

Allen, G. C., S. Kumar, S. Krasynanski, W. F. Thompson and M. A. Flores-Vergara, 2006 A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nature Protocols **1:** 2320-2325.

Arumuganathan, K., and E. D. Earle, 1991 Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9:** 415.

Ayoib, A., U. Hashim, S. C. B. Gopinath and M. K. Md Arshad, 2017 DNA extraction on bio-chip: History and preeminence over conventional and solid-phase extraction methods. Appl. Microbiol. Biotechnol. **101:** 8077-8088.

Battulin, N., V. S. Fishman, A. M. Mazur, M. Pomaznoy, A. A. Khabarova *et al*, 2015 Comparison of the three-dimensional organization of sperm and fibroblast genomes using the hi-C approach. Genome Biol. **16:** 77.

Beckman Coulter, 2016 *Instructions for use: Agencourt AMPure XP, PCR Purification*.

Beckman Coulter, 2012 *SPRIselect User Guide: SPRI Based Size Selection*.

Bennett, M. D., and L. J. Leitch, 2011 Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. Annals of Botany **107:** 467-590.

Ben-Noon, E., D. Shtienberg, E. Shlevin and A. Dinoor, 2003 Joint action of disease control measures: A case study of alternaria leaf blight of carrot. Phytopathology **93:** 1320-1328.

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

Boiteux, L. S., J. G. Belter, P. A. Roberts and P. W. Simon, 2000 RAPD linkage map of the genomic region encompassing the root-knot nematode (*meloidogyne javanica*) resistance locus in carrot. TAG Theoretical and Applied Genetics **100:** 439-446.

Bradeen, J. M., and P. W. Simon, 2007 Carrot, pp. vol.5 in *Vegetables.* Genome Mapping and Molecular Breeding in Plants, edited by C. Kole.

Bradeen, J. M., and P. W. Simon, 1998 Conversion of an AFLP fragment linked to the carrot Y2 locus to a simple, codominant, PCR-based marker form. TAG Theoretical and Applied Genetics **97:** 960-967.

Buckingham, L., and M. L. Flaws, 2007 Molecular Diagnostics; Fundamentals, Methods, & Clinical Applications. Ringgold, Inc, Portland.

Budahn, H., R. Barański, D. Grzebelus, A. Kiełkowska, P. Straka *et al*, 2014 Mapping genes governing flower architecture and pollen development in a double mutant population of carrot. Frontiers in plant science **5:** 504.

Buermans, H. P. J., and J. T. den Dunnen, 2014 Next generation sequencing technology: Advances and applications. BBA - Molecular Basis of Disease **1842:** 1932-1941.

Bushnell, B., BBMap Short Read Aligner, and Other Bioinformatic Tools.

Cassia da Silva Linge, L. Antanaviciute, A. Abdelghafar, P. Arús, D. Bassi *et al*, 2018 High-density multi-population consensus genetic linkage map for peach. PLoS One **13:** e0207724.

Cavagnaro, P. F., S. Chung, S. Manin, M. Yildiz, A. Ali *et al*, 2011a Microsatellite isolation and marker development in carrot - genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across *apiaceae*. BMC Genomics **12:** 386.

Cavagnaro, P. F., M. Iorizzo, M. Yildiz, D. Senalik, J. Parsons *et al*, 2014 A gene-derived SNP-based high resolution linkage map of carrot including the location of QTL conditioning root and leaf anthocyanin pigmentation. BMC Genomics **15:** 1118.

Cavagnaro, P. F., S. Chung, M. Szklarczyk, D. Grzebelus, D. Senalik *et al*, 2009 Characterization of a deep-coverage carrot (*daucus carota* L.) BAC library and initial analysis of BAC-end sequences. Molecular Genetics and Genomics **281:** 273-288.

Cavagnaro, P. F., S. Chung, S. Manin, M. Yildiz, A. Ali *et al*, 2011b Microsatellite isolation and marker development in carrot - genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across apiaceae. BMC Genomics **12:** 386.

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long and J. J. Emerson, 2016 Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. **44:** e147.

Chantaro, P., S. Devahastin and N. Chiewchan, 2008 Production of antioxidant high dietary fiber powder from carrot peels. LWT - Food Science and Technology **41:** 1987-1994.

Chen, W. P., and Z. K. Punja, 2002 Transgenic herbicide- and disease-tolerant carrot (*daucus carota* L.) plants obtained through agrobacterium-mediated transformation. Plant Cell Rep. **20:** 929-935.

Chen, X. e., Z. e. Kang and SpringerLink, 2017 *Stripe Rust*. Springer Netherlands :Imprint: Springer, Dordrecht.

Chen, Z. L., Y. Omori, S. Koren, T. Shirokiya, T. Kuroda *et al*, 2019 De novo assembly of the goldfish (*carassius auratus*) genome and the evolution of genes after whole-genome duplication. SCIENCE ADVANCES **5:** eaav0547.

Clotault, J., E. Geoffriau, E. Lionneton, M. Briard and D. Peltier, 2010 Carotenoid biosynthesis genes provide evidence of geographical subdivision and extensive linkage disequilibrium in the carrot. Theor. Appl. Genet. **121:** 659-672.

Cseke, L. J., P. B. Kaufman, G. K. Podila and C. Tsai, 2005 Handbook of molecular and cellular methods in biology and medicine. American Journal of Clinical Dermatology **6:** 411.

Dahm, R., and R. Dahm, 2008 Discovering DNA: Friedrich miescher and the early years of nucleic acid research. Hum. Genet. **122:** 565-581.

Ellison, S., 2019 Carrot domestication, pp. 77-91 in The Carrot Genome, edited by C. Kole.

Ellison, S., D. Senalik, H. Bostan, M. Iorizzo and P. Simon, 2017 Fine mapping, transcriptome analysis, and marker development for Y 2, the gene that conditions β-carotene accumulation in carrot (daucus carota L.). G3 (Bethesda, Md.) **7:** 2665.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee *et al*, 2012 Mind the gap: Upgrading genomes with pacific biosciences RS long-read sequencing technology. PloS one **7:** e47768.

Ersus, S., and U. Yurdagel, 2007 Microencapsulation of anthocyanin pigments of black carrot (*daucus carota* L.) by spray drier. J. Food Eng. **80:** 805-812.

FAO, 2019 Production/Yield Quantities of Carrots and Turnips in World.

Flor, H. H., 1971 Current status of the gene-for-gene concept. Annu. Rev. Phytopathol. **9:** 275-296.

Fournier, T., J. Gounot, K. Freel, C. Cruaud, A. Lemainque *et al*, 2017 High-quality de novo genome assembly of the *dekkera bruxellensis* yeast isolate using nanopore MinION sequencing. G3 (Bethesda, Md.) .

Ghurye, J., M. Pop, S. Koren, D. Bickhart and C. Chin, 2017 Scaffolding of long read assemblies using long range contact information. BMC Genomics **18:** 527-11.

Glasel, J. A., 1995 Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. BioTechniques **18:** 62.

Grzebelus, D., M. Iorizzo, D. Senalik, S. Ellison, P. Cavagnaro *et al*, 2014 Diversity, genetic mapping, and signatures of domestication in the carrot (*daucus carota* L.) genome, as revealed by diversity arrays technology (DArT) markers. Mol. Breed. **33:** 625-637.

Gugino, B. K., G. S. Abawi and J. W. Ludwig, 2006 Damage and management of meloidogyne hapla using oxamyl on carrot in new york. J. Nematol. **38:** 483-490.

Gulig, P. A., V. de Crecy-Lagard, A. C. Wright, B. Walts, M. Telonis-Scott *et al*, 2010 SOLiD sequencing of four *vibrio vulnificus* genomes enables comparative genomic analysis and identification of candidate clade-specific virulence genes. BMC Genomics **11:** 512.

Gunderson, K. L., J. Fan and M. S. Chee, 2006 Highly parallel genomic assays. Nature Reviews Genetics **7:** 632-644.

Hackl, T., R. Hedrich, J. Schultz and F. Förster, 2014 Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics **30:** 3004-3011.

Hohenlohe, P. A., J. W. Davey, P. D. Etter, J. M. Catchen, J. Q. Boone *et al*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics **12:** 499-510.

Holden, J. M., A. L. Eldridge, G. R. Beecher, I. Marilyn Buzzard, S. Bhagwat *et al*, 1999 Carotenoid content of U.S. foods: An update of the database. Journal of Food Composition and Analysis **12:** 169-196.

Howorka, S., and Z. Siwy, 2009 Nanopore analytics: Sensing of single molecules. Chem. Soc. Rev. **38:** 2360.

Hu, X., J. Yuan, Y. Shi, J. Lu, B. Liu *et al*, 2012 pIRS: Profile-based illumina pair-end reads simulator. Bioinformatics **28:** 1533-1535.

Hulbert, S. H., C. A. Webb, S. M. Smith and Q. Sun, 2001 Resistance gene complexes: Evolution and utilization. Annu. Rev. Phytopathol. **39:** 285-312.

Idrovo Encalada, A. M., M. F. Basanta, E. N. Fissore, M. D. De'Nobili and A. M. Rojas, 2016 Carrot fiber (CF) composite films for antioxidant preservation: Particle size effect. Carbohydr. Polym. **136:** 1041-1051.

Iorizzo, M., S. Ellison, M. Pottorff and P. F. Cavagnaro, 2019 Carrot molecular genetics and mapping, pp. 101-117 in The Carrot Genome.

Iorizzo, M., D. A. Senalik, D. Grzebelus, M. Bowman, P. F. Cavagnaro *et al*, 2011 De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics **12:** 389.

Iorizzo, M., D. A. Senalik, S. L. Ellison, D. Grzebelus, P. F. Cavagnaro *et al*, 2013 Genetic structure and domestication of carrot (*daucus carota* subsp. *sativus*) (*apiaceae*). Am. J. Bot. **100:** 930-938.

Iorizzo, M., S. Ellison, D. Senalik, P. Zeng, P. Satapoomin *et al*, 2016 A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat. Genet. **48:** 657-666.

Iovene, M., P. F. Cavagnaro, D. Senalik, C. R. Buell, J. Jiang *et al*, 2011a Comparative FISH mapping of daucus species (*apiaceae* family). Chromosome Research **19:** 493-506.

Jansen, H. J., M. Liem, S. A. Jong-Raadsen, S. Dufour, F. A. Weltzien *et al*, 2017 Rapid de novo assembly of the European eel genome from nanopore sequencing reads. Scientific Reports **7:** 7213-13.

Jayaraj, J., and Z. K. Punja, 2007 Combined expression of chitinase and lipid transfer protein genes in transgenic carrot plants enhances resistance to foliar fungal pathogens. Plant Cell Rep. **26:** 1539-1546.

Jiang, Y., P. Ninwichian, S. Liu, J. Zhang, H. Kucuktas *et al*, 2013 Generation of physical map contig-specific sequences useful for whole genome sequence scaffolding. PLoS One **8:** e78872.

Jiao, W., and K. Schneeberger, 2017 The impact of third generation genomic technologies on plant genome assembly. Curr. Opin. Plant Biol. **36:** 64-70.

Jiao, W., G. G. Accinelli, B. Hartwig, C. Kiefer, D. Baker *et al*, 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:** 778-786.

Jibran, R., H. Dzierzon, N. Bassil, J. M. Bushakra, P. P. Edger *et al*, 2018 Chromosome-scale scaffolding of the black raspberry (*rubus occidentalis* L.) genome based on chromatin interaction data. Horticulture Research **5:** 1-11.

Just, B. J., C. A. F. Santos, M. E. N. Fonseca, L. S. Boiteux, B. B. Oloizia *et al*, 2007 Carotenoid biosynthesis structural genes in carrot (*daucus carota*): Isolation, sequence-characterization, single nucleotide polymorphism (SNP) markers and genome mapping. Theor. Appl. Genet. **114:** 693-704.

Kaplan, N., and J. Dekker, 2013 High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat. Biotechnol. **31:** 1143-1147.

Kaul, S., H. L. Koo, J. Jenkins, M. Rizzo, T. Rooney *et al*, 2000 Analysis of the genome sequence of the flowering plant *arabidopsis thaliana*. Nature **408:** 796-815.

Kawakami, T., L. Smeds, N. Backström, A. Husby, A. Qvarnström *et al*, 2014 A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol. Ecol. **23:** 4035-4058.

Khalifa, M. E., A. Varsani, A. R. D. Ganley and M. N. Pearson, 2016 Comparison of illumina de novo assembled and sanger sequenced viral genomes: A case study for RNA viruses recovered from the plant pathogenic fungus *sclerotinia sclerotiorum*. Virus Res. **219:** 51-57.

Kilian, A., P. Wenzl, E. Huttner, J. Carling, L. Xia *et al*, 2012 Diversity arrays technology: A generic genome profiling technology on open platforms. Methods Mol. Biol. **888:** 67.

Kingan, S. B., J. Urban, C. C. Lambert, P. Baybayan, A. K. Childers *et al*, 2019 A high-quality genome assembly from a single, field-collected spotted lanternfly (*lycorma delicatula*) using the PacBio sequel II system. GigaScience **8:** .

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

LaFramboise, T., 2009 Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. Nucleic Acids Res. **37:** 4181-4193.

Lam, E. T., A. Hastie, C. Lin, D. Ehrlich, S. K. Das *et al*, 2012 Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. **30:** 771-776.

Laver, T., J. Harrison, P. A. O'Neill, K. Moore, A. Farbos *et al*, 2015 Assessing the performance of the oxford nanopore technologies MinION. Biomolecular Detection and Quantification **3:** 1-8.

Le Clerc, V., C. Aubert, V. Cottet, C. Yovanopoulos, M. Piquet *et al*, 2019 Breeding for carrot resistance to alternaria dauci without compromising taste. Mol. Breed. **39:** 1-15.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Oxford University Press 2013 **00:** .

Liu, D., C. Ma, W. Hong, L. Huang, M. Liu *et al*, 2014 Construction and analysis of high-density linkage map using high-throughput sequencing data. PLoS One **9:** e98855.

Liu, J., L. Shi, J. Han, G. Li, H. Lu *et al*, 2014 Identification of species in the angiosperm family *apiaceae* using DNA barcodes. Molecular Ecology Resources **14:** 1231-1238.

Macko-Podgorni, A., G. Machaj, K. Stelmach, D. Senalik, E. Grzebelus *et al*, 2017 Characterization of a genomic region under selection in cultivated carrot (*daucus carota* subsp *sativus*) reveals a candidate domestication gene. FRONTIERS IN PLANT SCIENCE **8:** 12.

Madoui, M. A., S. Engelen, C. Cruaud, C. Belser, L. Bertrand *et al*, 2015 Genome assembly using nanopore-guided long and error-free DNA reads. BMC Genomics **16:** 327.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al*, 2006 Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005). Nature **441:** 120.

Matsumoto, T., J. Wu, H. Kanamori and Y. Katayose, 2005 The map-based sequence of the rice genome. Nature **436:** 793-800.

Michael, T. P., F. Jupe, F. Bemm, S. T. Motley, J. P. Sandoval *et al*, 2018 High contiguity arabidopsis thaliana genome assembly with a single nanopore flow cell. NATURE COMMUNICATIONS **9:** 541-8.

Mifsud, B., F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder *et al*, 2015 Mapping long-range promoter contacts in human cells with high-resolution capture hi-C. Nat. Genet. **47:** 598-606.

Milos, P. M., and F. Ozsolak, 2011 RNA sequencing: Advances, challenges and opportunities. Nature Reviews Genetics **12:** 87-98.

Nagano, T., Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe *et al*, 2013 Single-cell hi-C reveals cell-to-cell variability in chromosome structure. Nature **502:** 59-64.

NanoDrop Technologies Inc, 2007 260/280 and 260/230 Ratios NanoDrop® ND-1000 and ND-8000 8-Sample Spectrophotometers.

Nathans, D., and H. O. Smith, 1975 Restriction endonucleases in the analysis and restructuring of DNA molecules. Annu. Rev. Biochem. **44:** 273-293.

Niedzicka, M., K. Dudek, A. Fijarczyk, P. Zieliński and W. Babik, 2017 Linkage map of. G3 (Bethesda, Md.) **7:** 2115.

Ninomiya, M., Y. Ueno and T. Shimosegawa, 2014 Application of deep sequence technology in hepatology: Application of deep sequencing in hepatology. Hepatology Research **44:** 141-148.

Nivala, J., D. B. Marks and M. Akeson, 2013 Unfoldase-mediated protein translocation through an alpha -hemolysin nanopore. Nat. Biotechnol. **31:** 247-250.

Osuna-Cruz, C. M., A. Paytuvi-Gallart, A. Di Donato, V. Sundesha, G. Andolfo *et al*, 2018 PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. Nucleic Acids Res. **46:** D1197-D1201.

Oxford Nanopore Technologies, 2017 1D Lambda Control Experiment (SQK-LSK108).

Pacific Biosciences of California Inc., 2019 Pacific Biosciences Launches New Sequel II System, Featuring ~8 Times the DNA Sequencing Data Output.

Pareek, C. S., R. Smoczynski and A. Tretyn, 2011 Sequencing technologies and genome sequencing. J. Appl. Genet. **52:** 413-435.

Parsons, J., W. Matthews, M. Iorizzo, P. Roberts and P. Simon, 2015 Meloidogyne incognita nematode resistance QTL in carrot. Mol. Breed. **35:** 1-11.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood *et al*, 2009 The sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

Paux, E., D. Roger, E. Badaeva, G. Gay, M. Bernard *et al*, 2006 Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. The Plant Journal **48:** 463-474.

Pfeiffer, F., C. Gröber, M. Blank, K. Händler, M. Beyer *et al*, 2018 Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific reports **8:** 10950-14.

Pinheiro Sant'Ana, H. M., P. C. Stringheta, S. C. Cardoso Brandão and Cordeiro de Azeredo, Raquel Monteiro, 1998 Carotenoid retention and vitamin A value in carrot (*daucus carota* L.) prepared by food service. Food Chem. **61:** 145-151.

Postlethwait, J. H., I. G. Woods, P. Ngo-Hazelett, Y. L. Yan, P. D. Kelly *et al*, 2000 Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res. **10:** 1890-1902.

Pryor, B. M., J. O. Strandberg, R. M. Davis, J. J. Nunez and R. L. Gilbertson, 2002 Survival and persistence of alternaria dauci in carrot cropping systems. Plant Dis. **86:** 1115-1122.

Punja, Z. K., and S. Raharjo, 1996 Response of transgenic cucumber and carrot plants expressing different chitinase enzymes to inoculation with fungal pathogens. Plant Dis. **80:** 999-1005.

Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette *et al*, 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. **26:** 342-350.

Qi, Z., L. Huang, R. Zhu, D. Xin, C. Liu *et al*, 2014 A high-density genetic map for soybean based on specific length amplified fragment sequencing. PLoS One **9:** e104871.

QIAGEN, 2016 DNeasy® Plant Mini Kit Quick-Start Protocol.

Ramani, V., X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers *et al*, 2017 Massively multiplex single-cell Hi-C. Nature Methods **14:** 263-266.

Rao, M. S., M. Kamalnath, R. Umamaheswari, R. Rajinikanth, P. Prabu *et al*, 2017 *Bacillus subtilis* IIHR BS-2 enriched vermicompost controls root knot nematode and soft rot disease complex in carrot. Scientia Horticulturae **218:** 56-62.

Reuter, J., D. V. Spacek and M. Snyder, 2015 High-throughput sequencing technologies. Mol. Cell **58:** 586-597.

Rhee, M.,Mark A., 2007 Nanopore sequencing technology: Nanopore preparations. Trends Biotechnol. **25:** 174-181.

Rhoads, A., and K. F. Au, 2015 PacBio sequencing and its applications. Genomics, proteomics & bioinformatics **13:** 278.

Rieseberg, L. H., S. J. E. Baird and K. A. Gardner, 2000 Hybridization, introgression, and linkage evolution. Plant Mol. Biol. **42:** 205-224.

Roach, J. C., C. Boysen, K. Wang and L. Hood, 1995 Pairwise end sequencing: A unified approach to genomic mapping and sequencing. Genomics **26:** 345-353.

Rock, C. L., 1997 Carotenoids: Biology and treatment. Pharmacology and Therapeutics **75:** 185-197.

Rothberg, J. M., and J. H. Leamon, 2008 The development and impact of 454 sequencing. Nat. Biotechnol. **26:** 1117-1124.

Royaert, S., J. Jansen, d. Silva Daniela Viana, d. Jesus Branco Samuel Martins, D. S. Livingstone *et al*, 2016 Identification of candidate genes involved in witches' broom disease resistance in a segregating mapping population of *theobroma cacao* L. in brazil. BMC Genomics **17:** 107.

Rubatzky, V. E., C. F. Quiros and P. W. Simon, 1999 Carrots and Related Vegetable *mbelliferae*. Wallingford, Oxon, UK : New York, NY, USA : CABI Publishing.

S. Ellison, M. Iorizzo, D. Senalik and P.W. Simon, 2017 The next generation of carotenoid studies in carrot (*daucus carota* L.). ISHS Acta Horticulturae **1153:** 93-100.

Saha, S., and S. Rajasekaran, 2014 Efficient and scalable scaffolding using optical restriction maps. BMC Genomics **15:** S5.

Samad, A., E. F. Huff, W. Cai and D. C. Schwartz, 1995 Optical mapping: A novel, single-molecule approach to genomic analysis. Genome Res. **5:** 1-4.

Sanger, F., and A. R. Coulson, 1975 A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. **94:** 441,IN19,447-446,IN20,448.

Santos, C., and P. Simon, 2002 QTL analyses reveal clustered loci for accumulation of major provitamin A carotenes and lycopene in carrot roots. Molecular Genetics and Genomics **268:** 122-129.

Santos, C. A. F., and P. W. Simon, 2004a Merging carrot linkage groups based on conserved dominant AFLP markers in F2 populations. J. Am. Soc. Hort. Sci. **129:** 211.

Santos, C. A. F., and P. W. Simon, 2004b Merging carrot linkage groups based on conserved dominant AFLP markers in F2 populations. J. Am. Soc. Hort. Sci. **129:** 211-217.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al*, 2009 The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112-1115.

Schulz, B., L. Westphal and G. Wricke, 1994 Linkage groups of isozymes, RFLP and RAPD markers in carrot (*daucus carota* L. *sativus*). Euphytica **74:** 67-76.

Shao, W., V. F. Boltz, J. E. Spindler, M. F. Kearney, F. Maldarelli *et al*, 2013 Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. Retrovirology **10:** 18.

Sharma, K. D., S. Karki, N. S. Thakur and S. Attri, 2012 Chemical composition, functional properties and processing of carrot—a review. Journal of Food Science and Technology **49:** 22-32.

Shirasawa, K., H. Fukuoka, H. Matsunaga, Y. Kobayashi, I. Kobayashi *et al*, 2013 Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA RESEARCH **20:** 593-603.

Siddiqui, Z. A., A. Parveen, L. Ahmad and A. Hashem, 2019 Effects of graphene oxide and zinc oxide nanoparticles on growth, chlorophyll, carotenoids, proline contents and diseases of carrot. Scientia Horticulturae **249:** 374-382.

Sim, S., G. Durstewitz, J. Plieske, R. Wieseke, M. W. Ganal *et al*, 2012 Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One **7:** e40563.

Simon, P. W., R. E. Freeman and J. V. Vieira, 2008 Carrot, pp. 327-357 in Handbook of Crop Breeding, edited by J. Prohens, M. J. Carena and F. Nuez.

Simon, P., M. Iorizzo, D. Grzebelus and R. Baranski, 2019 The Carrot Genome. Springer, Cham.

Simon, P. W., 2019a Classical and molecular carrot breeding in The Carrot Genome.

Simon, P. W., 2019b Economic and academic importance, pp. 1-8 in The Carrot Genome.

Simon, P. W., 2010 Domestication, historical development, and modern breeding of carrot, pp. 157-190 in . John Wiley & Sons, Inc, Oxford, UK.

Soufflet-Freslon, V., M. Jourdan, J. Clotault, S. Huet, M. Briard *et al*, 2013 Functional gene polymorphism to reveal species history: The case of the CRTISO gene in cultivated carrots. PLoS One **8:** e70801.

Soylu, S., S. Kurt, E. M. Soylu and F. M. Tok, 2005 First report of alternaria leaf blight caused by alternaria dauci on carrot in turkey. Plant Pathol. **54:** 252.

St. Pierre, M. D., R. J. Bayer and I. M. Weis, 1990 An isozyme-based assessment of the genetic variability within the *daucus carota* complex (*apiaceae*: *caucalideae*). Canadian Journal of Botany **68:** 2449-2457.

Staskawicz, B. J., F. M. Ausubel, B. J. Baker, J. G. Ellis and Jonathan D. G. Jones, 1995 Molecular genetics of plant disease resistance. Science **268:** 661-667.

Tan, S. C., and B. C. Yiap, 2009 DNA, RNA, and protein extraction: The past and the present. Journal of biomedicine & biotechnology **2009:** 574398-10.

Thompson, J. F., and P. M. Milos, 2011 The properties and applications of single-molecule DNA sequencing. Genome Biol. **12:** 217.

Thorvaldsdóttir, H., E. S. Lander, M. Guttman, W. Winckler, J. P. Mesirov *et al*, 2011 Integrative genomics viewer. Nat. Biotechnol. **29:** 24-26.

Tsiatis, A. C., A. Norris-Kirby, R. G. Rich, M. J. Hafez, C. D. Gocke *et al*, 2010 Comparison of sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: Diagnostic and clinical implications. The Journal of molecular diagnostics: JMD **12:** 425.

van Berkum, N. L., E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke *et al*, 2010 Hi-C: A method to study the three-dimensional architecture of genomes. Journal of Visualized Experiments .

van Dijk, E. L., H. Auger, Y. Jaszczyszyn and C. Thermes, 2014 Ten years of next-generation sequencing technology. Trends in Genetics **30:** 418-426.

van Heesch, S., W. P. Kloosterman, N. Lansu, F. P. Ruzius, E. Levandowsky *et al*, 2013 Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC Genomics **14:** 257.

Veale, A. J., L. Williams, P. Tsai, V. Thakur and S. Lavery, 2016 The complete mitochondrial genomes of two chiton species (*sypharochiton pelliserpentis* and *sypharochiton sinclairi*) obtained using illumina next generation sequencing. MITOCHONDRIAL DNA PART A **27:** 537-538.

Venkatesan, B. M., and R. Bashir, 2011 Nanopore sensors for nucleic acid analysis. Nature Nanotechnology **6:** 615-624.

Vieira, M. L. C., L. Santini, A. L. Diniz and C. d. F. Munhoz, 2016 Microsatellite markers: What they mean and why they are so useful. Genetics and Molecular Biology **39:** 312-328.

Vivek, B. S., and P. W. Simon, 1999 Linkage relationships among molecular markers and storage root traits of carrot (*daucus carota* L. ssp. *sativus*). TAG Theoretical and Applied Genetics **99:** 58-64.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van De Lee *et al*, 1995 AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407-4414.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al*, 2014 Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9:** e112963.

Wally, O., J. Jayaraj and Z. Punja, 2009a Comparative resistance to foliar fungal pathogens in transgenic carrot plants expressing genes encoding for chitinase, β-1,3-glucanase and peroxidise. Eur. J. Plant Pathol. **123:** 331-342.

Wally, O., J. Jayaraj and Z. K. Punja, 2009b Broad-spectrum disease resistance to necrotrophic and biotrophic pathogens in transgenic carrots (*daucus carota* L.) expressing an arabidopsis NPR1 gene. Planta **231:** 131-141.

Wang, W. Q., and J. Messing, 2011 High-throughput sequencing of three *lemnoideae* (duckweeds) chloroplast genomes from total DNA. PLOS ONE **6:** e24670.

Wang, X. V., N. Blades, J. Ding, R. Sultana and G. Parmigiani, 2012 Estimation of sequencing error rates in short reads. BMC Bioinformatics **13:** 185.

Wang, Z., N. Hobson, L. Galindo, S. Zhu, D. Shi *et al*, 2012 The genome of flax (*linum usitatissimum*) assembled de novo from short shotgun sequence reads. The Plant Journal **72:** 461-473.

Warman, P. R., and K. A. Havard, 1997 Yield, vitamin and mineral contents of organically and conventionally grown carrots and cabbage. Agriculture, Ecosystems and Environment **61:** 155-162.

Waterston, R. H., K. Lindblad-Toh and E. Birney, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

Wenger, A. M., P. Peluso, W. J. Rowell, P. Chang, R. J. Hall *et al*, 2019 Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. **37:** 1155-1162.

Westphal, L., and G. Wricke, 1991 Genetic and linkage analysis of isozyme loci in *daucus carota* L. Euphytica **56:** 259-267.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen *et al*, 2008 The complete genome of an individual by massively parallel DNA sequencing. Nature **452:** 872-U5.

Wicker, T., E. Schlagenhauf, A. Graner, T. J. Close, B. Keller *et al*, 2006 454 sequencing put to the test using the complex genome of barley. BMC Genomics **7:** 275.

Wiggin, M., Y. V. Pershin, T. Butler, M. Wanunu, R. Riehn *et al*, 2008 The potential and challenges of nanopore sequencing. Nat. Biotechnol. **26:** 1146-1153.

Wrzodak, A., J. Szwejda-Grzybowska, K. Elkner and I. Babik, 2012 Comparison of the nutritional value and storage life of carrot roots from organic and conventional cultivation. Vegetable Crops Research Bulletin **76:** 137-150.

Xu, L., S. Wang, X. Fan, D. Xu, X. Zhang *et al*, 2016 Sequencing of complete mitochondrial genome of brown algal *saccharina* sp. ye-C6. Mitochondrial DNA Part A **27:** 3733-3734.

Xue, W., J. Li, Y. Zhu, G. Hou, X. Kong *et al*, 2013 L_RNA_scaffolder: Scaffolding genomes with transcripts. BMC Genomics **14:** 604.

Yang, N., J. Liu, Q. Gao, S. Gui, L. Chen *et al*, 2019 Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat. Genet. **51:** 1052-1059.

Yang, Z. E., X. Y. Ge, Z. R. Yang, W. Q. Qin, G. F. Sun *et al*, 2019 Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. NATURE COMMUNICATIONS **10:** 2989-13.

Yildiz, M., D. K. Willis, P. F. Cavagnaro, M. Iorizzo, K. Abak *et al*, 2013a Expression and mapping of anthocyanin biosynthesis genes in carrot. Theor. Appl. Genet. **126:** 1689-1702.

Zhang, M., Y. Zhang, C. F. Scheuring, C. Wu, J. J. Dong *et al*, 2012 Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. Nature Protocols **7:** 467-478.

Zhang, X. Y., J. Hu, H. Y. Zhou, J. J. Hao, Y. F. Xue *et al*, 2014 First report of fusarium oxysporum and F. solani causing fusarium dry rot of carrot in china. Plant Dis. **98:** 1273.

Zhuang, W. J., H. Chen, M. Yang, J. P. Wang, M. K. Pandey *et al*, 2019 The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nat. Genet. **51:** 865.

Table 1.1. Summary of carrot genetic linkage maps

Mapping population	Generation	Type of markers	Dominant markers mapped	Codominant markers mapped	References
B9304×YC7262	F2	AFLP, SNP	6	1	(Bradeen and Simon. 1998)
B9304×YC7262	F2	RFLP, RAPD, AFLP, SSR		10	(Vivek and Simon. 1999)
Brasilia-1252×B6274	F2	RAPD	4		(Boiteux *et al*. 2000)
Brasilia×HCM	F2	AFLP	164		(Santos, C. and Simon. 2002; Santos, Carlos A. F. and Simon. 2004b)
B493×QAL	F2	AFLP	141		(Santos, C. and Simon. 2002; Santos, Carlos A. F. and Simon. 2004b)
B493×QAL	F2	FISH			(Iovene, Cavagnaro *et al*. 2011a)
Biennial×Criolla	F2	AFLP, SSR, RAPD, SCAR	355	23	(Alessandro *et al*. 2013b)
B1896×B7261	F2	AFLP, SSR, SNP	279	38	(Yildiz *et al*. 2013a)
70349	F2, F3, F4	SNP, SSR		894	(Cavagnaro *et al*. 2014)
2569	F4	SNP		811	(S. Ellison *et al*. 2017)

Table 1.1 (continued).

Br1091×HM1	F2	SNP		389	(Parsons *et al*. 2015)
SFF×HM2	F2	AFLP, SNP, SSR	20	138	(Parsons *et al*. 2015)
HM3	F5	SSR, SNP		70	(Parsons *et al*. 2015)
74146	F4	SNP		2999	(Ellison *et al*. 2017)
PI652188×B7262	F3	RAPD, SSR		8	(Ali *et al*. 2014)

Table 2.1. Comparison of carrot DNA extraction methods.

QC	DNA extraction methods					
	Rapid CTAB	Nuclei method	QIAGEN mini kit	General CTAB	CTAB with ethanol	CTAB without NaAc
DNA (ng/g)[1]	3,494.28	8.58	2,013.33	4750.00	5,383.00	16,579.00
260/280	1.67	1.32	2.13	2.08	2.00	1.88
260/230	0.67	0.44	2.52	1.35	1.49	1.99

[1]: ng/g indicates the amount of DNA quantified (ng) divided by the amount of tissue used for the extraction (g).

Table 2.2. Comparison of DNA recycling rate using AMPure XP beads on DNA extracted using different protocols and with different concentrations.

DNA extraction method	Input DNA conc. (ng/µL)	Eluted DNA conc. (ng/µL)	DNA recycling rate[1]
General CTAB	54.60	5.58	10.22 %
	40.00	5.00	25.00 %
	25.00	6.60	52.80 %
CTAB without NaAc	42.00	27.80	32.94 %
	37.20	41.20	55.38 %
	31.60	10.70	33.86 %
	26.20	17.60	67.18 %
	22.00	15.00	68.18 %
	21.10	15.90	75.36 %
	16.60	10.30	62.00 %

[1]: DNA recycling rate was calculated using: (Eluted DNA amount (ng)/ Input DNA amount (ng)) *100.

Table 2.3. Quality and concentration of DNA samples used as input to prepare four Nanopore libraries and run four PacificBiosciences SMRT flow cells.

DNA Samples	260/280	260/230	DNA conc.[2] (ng/μL)	Volume (μL)	Total DNA (ng)	Average recycling rate[1]
1	1.90	1.90	26.7	45	~1200	48.15%
2	1.88	1.99	26.7	45	~1200	73.90%
3	1.88	1.99	26.7	45	~1200	74.60%
4	1.88	1.99	26.7	45	~1200	84.40%

[1]: The average recycling rate is calculated as average percentage of DNA recovered after two bead binding steps.

Table 2.4. Statistics of DH1 carrot sequence reads obtained from different sequencing technologies.

	PacBio	Nanopore	Illumina_40k	Illumina_20k	Illumina_10k	BACs
Number of reads	1,177,331	134,751	65,674,576	88,465,135	199,475,162	2,9875
Total reads size (bp)	17,472,208,758	717,429,905	3,218,054,224	4,334,791,615	9,774,282,938	19027,398
Depth coverage[1]	37X	1.5X	6.8X	9.2X	20.7X	0.04X
Min read size (bp)	1,001	68	49	49	49	100
Median read size (bp)	13,898	5,284	49	49	49	669
Mean read size (bp)	14,841	5,324	49	49	49	637
Max read size (bp)	63,432	1,380,954	49	49	49	998

[1]: Depth coverage was calculated as: total length of reads/estimated genome size (473 MB).

Table 2.5. Summary statistics of carrot DH1 v3 genome assembly.

		Assembly steps					
		De-novo assembly	Polishing1	Polishing2	Scaffolding	Gap filling	Final assembly[2]
Number of sequences	All	817	817	817	532	521	443
	>= 10kb	817	817	817	532	521	443
	>= 25kb	769	769	768	488	483	407
	>= 50kb	508	510	508	338	337	261
	>= 100kb	317	317	345	215	216	142
	>= 250kb	196	196	196	115	112	51
	>= 500kb	156	156	156	77	73	21
	>= 1kb	99	99	99	63	61	15
	>= 2.5Mb	46	46	46	40	38	9
	>= 5Mb	18	18	18	24	22	9
	>= 10Mb	3	3	3	8	9	9
	>= 25Mb	0	0	0	2	2	9
	>= 50Mb	0	0	0	0	0	1
Sequence length (bp)	All	436,301,639	436,360,025	436,156,097	436,314,597	438,970,282	439,079,848
	>= 10kb	436,301,639	436,360,025	436,156,097	436,314,597	438,970,282	439,079,848
	>= 25kb	435,365,120	435,423,226	435,202,331	435,477,100	438,248,651	438,404,179
	>= 50kb	425,524,924	425,679,353	425,437,558	430,019,580	432,832,586	432,988,114
	>= 100kb	412,322,673	412,368,904	412,087,019	421,510,408	424,430,018	424,698,336
	>= 250kb	394,359,015	394,392,593	394,381,077	406,121,029	408,371,827	410,857,323
	>= 500kb	379,841,218	379,869,839	379,867,229	393,311,895	394,936,713	400,744,879
	>= 1Mb	339,986,519	340,010,326	340,010,430	383,558,349	386,476,311	396,015,367
	>= 2.5Mb	259,238,712	259,257,386	259,259,485	344,476,886	347,105,140	385,960,123
	>= 5Mb	151,716,401	151,731,245	151,733,691	289,371,224	290,750,271	385,960,123
	>= 10Mb	42,133,621	42,145,211	42,146,329	181,331,611	202,565,211	385,960,123
	>= 25Mb	0	0	0	72,977,930	73,086,915	385,960,123
	>= 50Mb	0	0	0	0	0	54,951,991

Table 2.5 (continued).

Sequence size (bp)	Min	10,322	10,384	10,383	10,383	11,048	11,049
	Median	67,313	67,413	67,368	70,702	74,109	59,113
	Mean	534,029	534,100	533,853	820,140	842,553	991,152
	Max	19,887,722	19,897,079	19,897,977	41,609,876	41,680,558	54,951,992
N50 (Mb)		3.37	3.369	3.369	7.4	7.944	44.038
N90 (kb)		273.182	273.279	273.245	530.019	489.748	1,054
L50 (#)		35	35	35	13	11	5
L90 (#)		190	190	190	76	74	15
Genome Coverage (%)[1]		92.24	92.25	92.21	92.24	92.81	92.83

[1] Genome coverage was calculated using the following formula: (total length (bp)/estimated genome size (473 MB))*100.
[2] This column is the statistics of the final fasta file.

Table 2.6. Comparison of v2 and v3 genome assemblies.

	Genome v2			Genome v3		
	Contigs	Scaffolds	Total	Contigs	Scaffolds	Total
Total number of sequences	33,351	4,182	8,096	692	100	521
Total size (bp)	390,582,546	418,822,324	424,246,664	438,921,773	292,699,918	438,970,282
N50 (bp)	31,102	807,326	787,280	4,945,074	19,238,000	7,944,388
L50 (#)	3,638	137		20	6	11
N90 (bp)	12,974	2,663,518	2,663,518	331,432	2,166,000	489,748
L90 (#)	8,584	13		128	24	74
Longest (bp)	258,656	5,109,390	5,109,390	28,632,103	41,680,558	41,680,558
Genome coverage	82.58%	88.55%	89.69%	92.80%	61.88%	92.81%
Number of gaps			25,255			171
Gaps size (bp)			33,685,185			43,897
Gap rate[1]			7.94%			0.01%

[1] Gap rate was calculated as total length of gaps/total length of assembled sequences.

Table 2.7. Genome anchoring summary statistics.

	Number	Length (bp)
Anchored super-scaffolds	21	290,639,203
Anchored sequences[1]	42	385,854,123
Anchored and oriented super-scaffolds	21	290,639,203
Anchored and oriented sequences[2]	38	383,110,696

[1,2] Sequences include contigs, scaffolds and super-scaffolds.

Table 2.8. Comparison of chromosome scale assembly of DH1 genome v2 and genome v3.

	Genome v2				Genome v3			
Chromo-some	Number of sequences	Total Length (bp)	Gap Length (bp)	Number of super-scaffolds	Number of sequences	Total length (bp)	Gap length (bp)	Number of super-scaffolds
Chr1	117	51,465,339	3,094,207	8	5	48,188,873	7,579	1
Chr2	96	43,913,520	2,550,549	8	14	47,282,954	23,532	3
Chr3	137	50,312,079	2,927,773	8	5	54,951,991	8,876	2
Chr4	74	35,924,511	1,790,350	3	6	44,959,275	6,150	1
Chr5	84	41,956,025	2,279,819	4	11	44,038,122	15,995	4
Chr6	92	36,610,139	1,988,013	6	8	37,338,840	13,255	2
Chr7	105	36,358,036	2,304,027	7	10	37,758,459	15,612	4
Chr8	102	31,745,509	2,345,311	6	10	32,272,103	18,025	2
Chr9	107	33,682,890	2,352,050	10	13	39,168,506	18,088	2
Total	914	361,968,048	21,632,099	60	82	385,959,123	127,112	21

Table 3.1. Number of genes and domains in each chromosomes of carrot genome v2 and v3.

	Carrot genome v2								Carrot genome v3							
	# of genes	# of domains	CC	Kinase	LRR	NBS	TIR	TM	# of genes	# of domains	CC	Kinase	LRR	NBS	TIR	TM
Chr1	244	1,439	48	508	244	48	6	585	303	2,170	48	802	336	118	34	832
Chr2	217	1,261	19	440	152	177	13	460	256	1,757	23	584	225	239	81	605
Chr3	237	1,614	43	440	288	206	0	637	310	2,335	38	658	415	322	30	872
Chr4	171	1,064	22	418	165	47	1	411	199	1,405	23	578	196	93	12	503
Chr5	139	920	20	296	212	53	1	338	166	1,158	20	425	237	56	3	417
Chr6	147	920	21	286	165	56	2	390	171	1,132	23	427	207	63	8	404
Chr7	180	1,340	41	425	134	231	0	509	213	1,516	31	518	179	224	8	556
Chr8	143	862	11	332	122	62	0	335	176	1,205	8	471	163	107	18	438
Chr9	140	922	13	236	263	37	6	367	177	1,329	20	362	274	131	57	485
Anchored total	1,618	10,342	238	3,381	1,745	917	29	4,032	1,971	14,007	234	4,825	2,232	1,353	251	5,112
Not anchored	28	152	4	60	28	0	0	60	11	68	2	39	6	0	0	21
Total	1,646	10,494	242	3,441	1,773	917	29	4,092	1,982	14,075	236	4,864	2,238	1,353	251	5,133

*Abbreviations: LRR: Leucine-rich repeats, NBS: nucleotide-binding site, TIR: toll-interleukin region, CC: coiled-coil, TM: transmembrane, Kinase: kinase domain.

Table 3.2. Statistics of R genes, coding sequences, and domains in carrot genomes v2 and v3.

	Carrot genome v2	Carrot genome v3
Total R gene number	1,646	1,982
Total R gene length (bp)	7,165,333	7,588,227
Average R gene length (bp)	4,353	3,829
Median R gene length (bp)	3,239	2,964
Smallest R gene (bp)	222	183
Longest R gene (bp)	54,629	63,679
Total CDS number	9,305	10,133
Total CDS length (bp)	3,218,805	3,824,280
Average CDS length (bp)	346	377
Median CDS length (bp)	147	156
Smallest CDS (bp)	3	3
Longest CDS (bp)	4,770	4,956
Average CDS per R gene	5.56	5.11
Total domain number	10,494	14,075
Average domain per R gene	6.38	7

*Abbreviations: R genes: resistance genes, CDS: coding sequence.

Table 3.3. Comparison of pair-to-pair R genes in carrot genomes v2 and v3.

	Number of R genes		Number of domains	
	v2	v3	v2	v3
Identical	585	585	3,675	3,675
Identical position/different domains[1]	221	214	1,559	1,756
Mutually extend	44	45	209	353
V2 inside v3	313	303	1677	2264
V3 inside v2	361	376	2635	2470
Interlaced[2]	49	77	422	637
Only in v2	69	0	312	0
Only in v3	0	382	0	2920
Split gene	4	0	5	0
Total	1,646	1,982	10,494	14,075

*Abbreviations: R genes: resistance genes

[1] Identical position / different domains means that R genes have the same start points and end points on v2 and v3 genome but the number or types of domains are different.

[2] Interlaced includes both extended and inside.

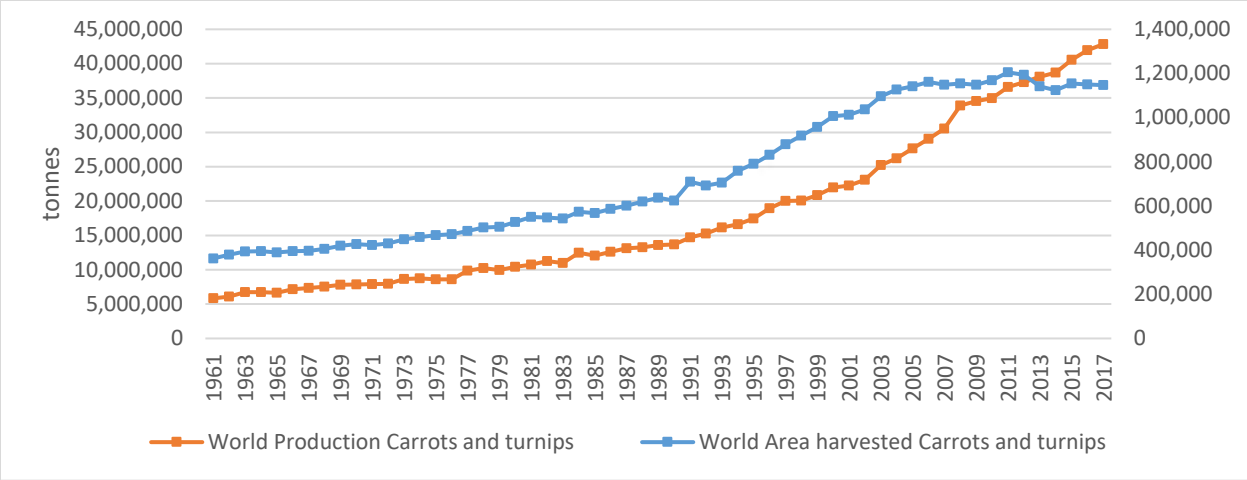


Figure 1.1. FAO world harvest and production of carrots and turnips.

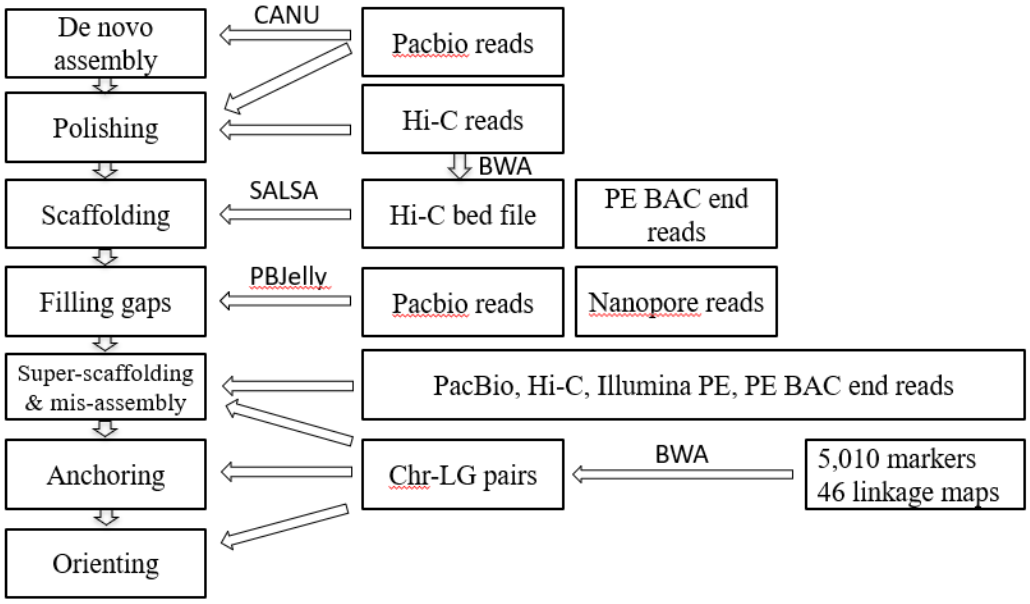
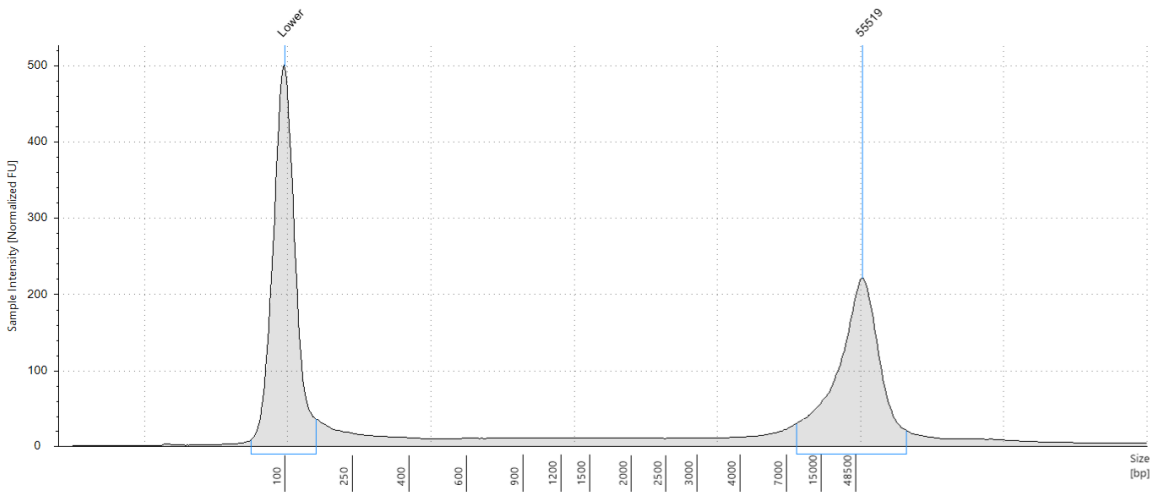
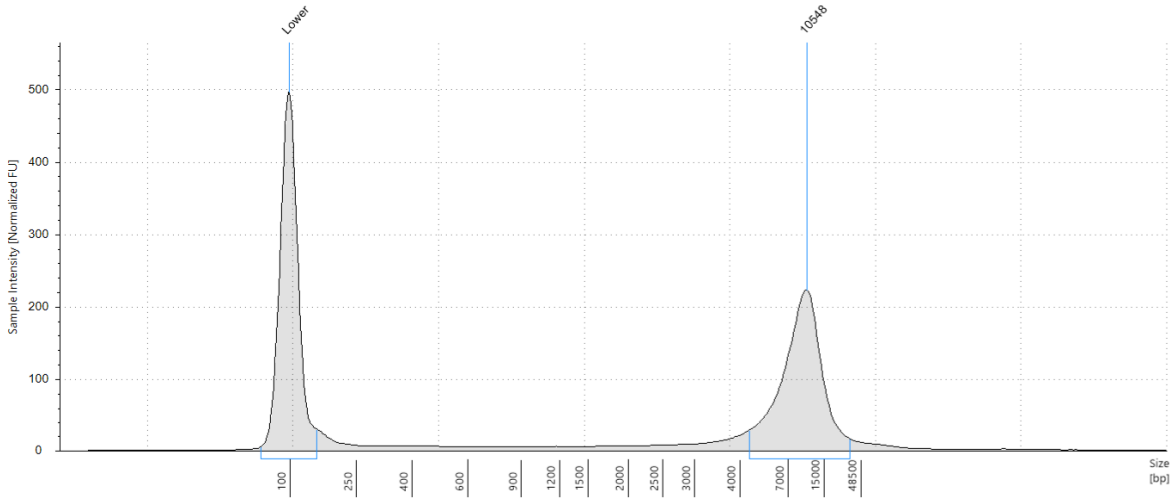


Figure 2.1. Workflow of the DH1 v3 genome construction.



(a)



(b)

Figure 2.2. Tape station results of DH1 carrot DNA size distribution (a) before and (b) after shearing.

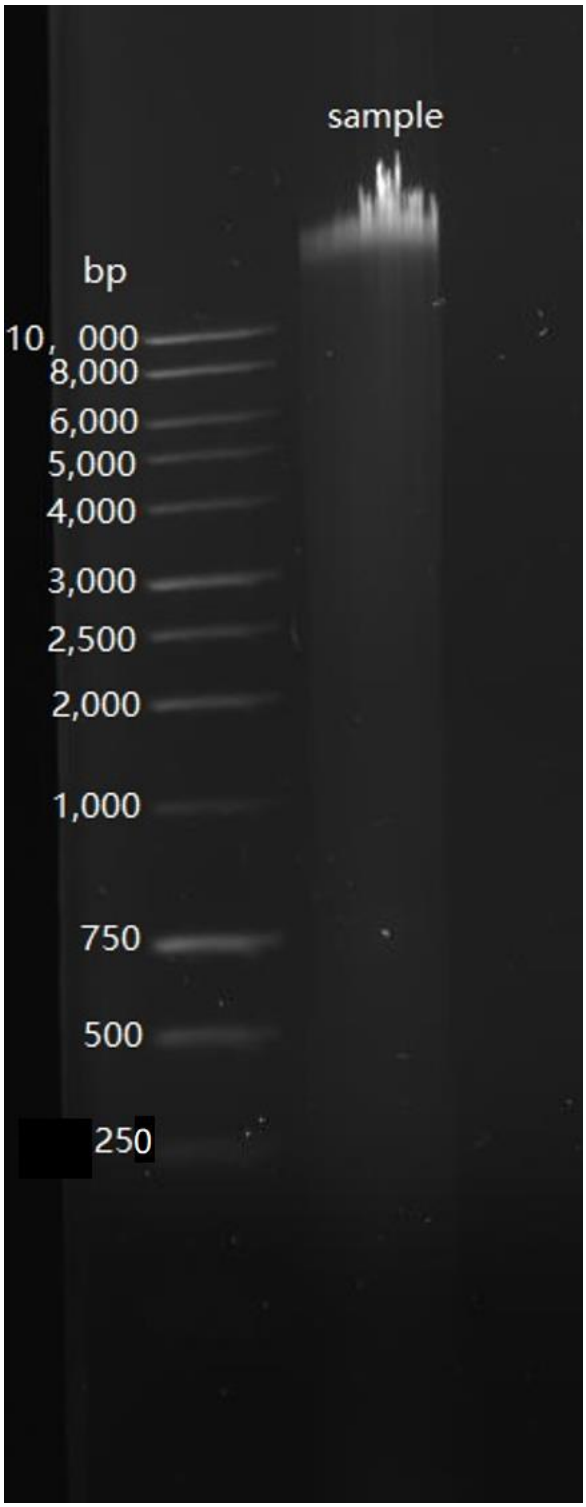
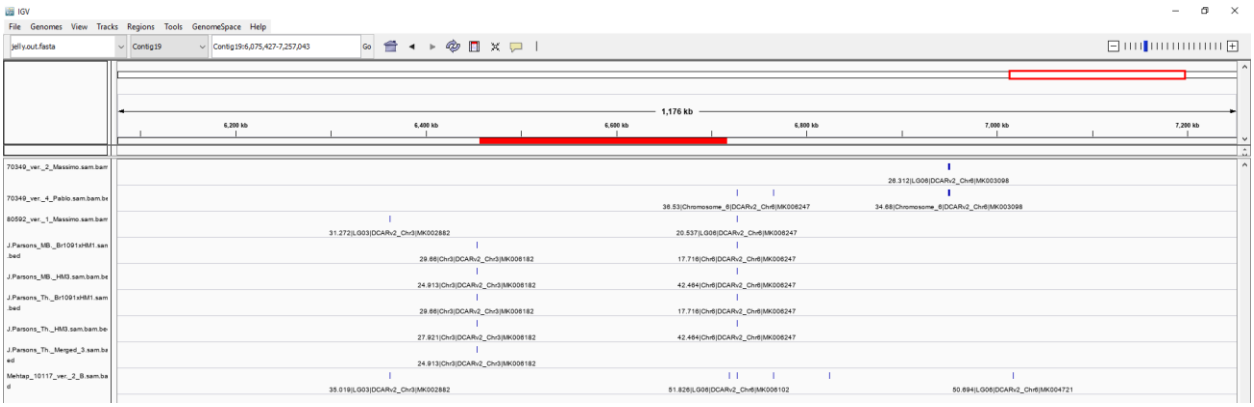
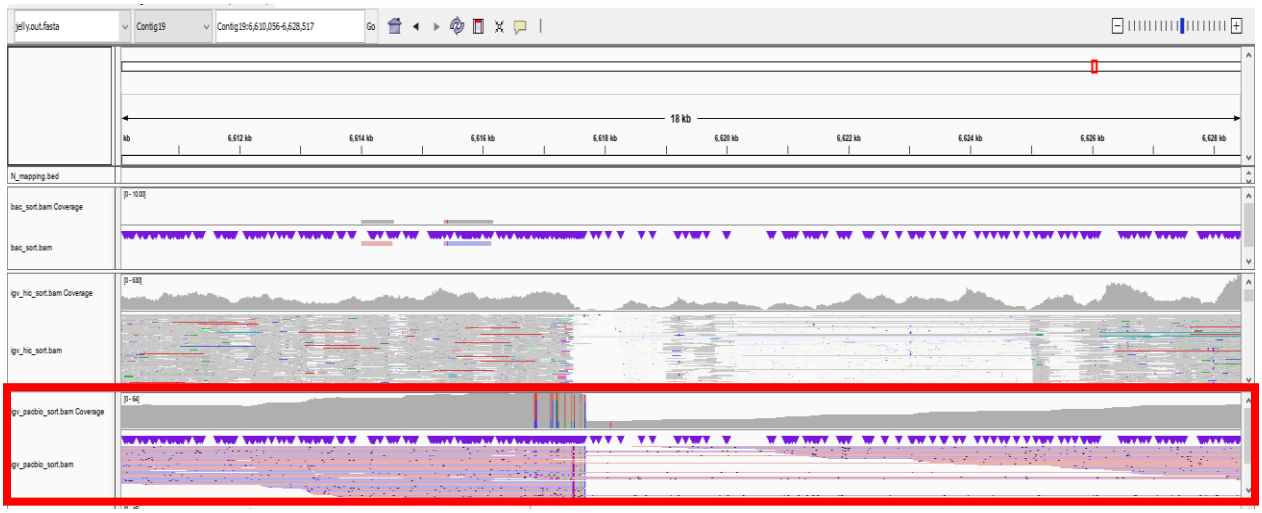


Figure 2.3. 0.8% gel image of DH1 carrot DNA extracted before shearing. (The ladder represents the size of DNA.)

Figure 2.4. Mis-assembly example of v3 genome. (a) Mapped markers on assemblies and uploaded into IGV_2.5.0. On contig19, markers clustered on the left were from chromosome3 while markers clustered on the right were from chromosome 6, between these two groups of markers (marked in red on track), it was highly possible existing a mis-assembly point. (b) Zooming in the marked area, a sudden drop of PacBio read coverage existed at 6,617,697 bp. Hi-C and other PE reads lacked coverage around this point. (c) Checked the BAC PE reads on the right of this point. Many reads were connected to contig293, which belong to chromosome6. (d) On the left of this point, many BAC PE reads were connected to contig508, which belong to chromosome3. Due to the abnormal sudden drop of PacBio read coverage at 6,617,697 bp, this point was designated the mis-assembly point and contig19 would break there.


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
jelly.out.fasta
Contig19
Contig19:6,075,427-7,257,043
Go
1,176 kb
6,200 kb
6,400 kb
6,600 kb
6,800 kb
7,000 kb
7,200 kb

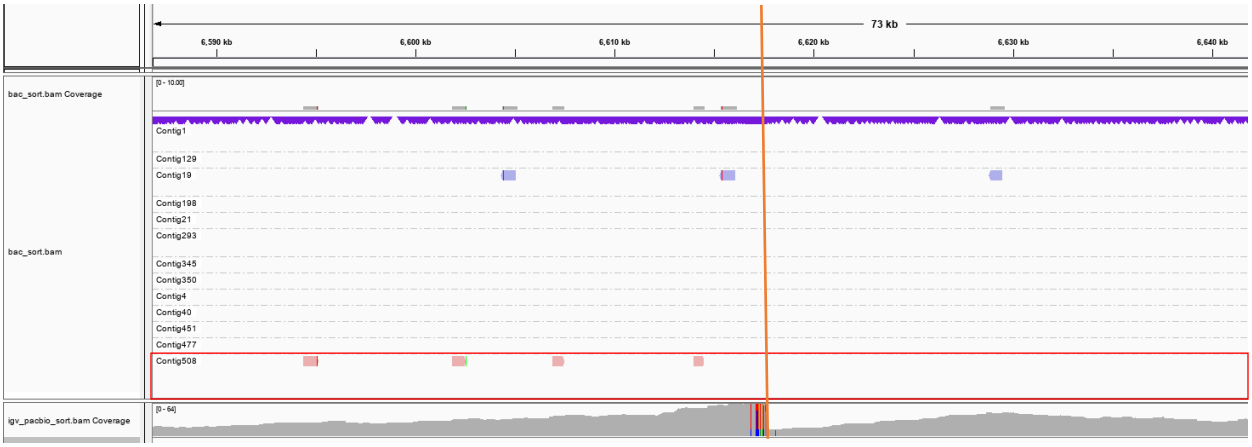
(a)


jelly.out.fasta
Contig19
Contig19:6,610,056-6,628,517
Go
18 kb
6,612 kb
6,614 kb
6,616 kb
6,618 kb
6,620 kb
6,622 kb
6,624 kb
6,626 kb
6,628 kb
N_mapping.bed
bac_sort.bam Coverage
bac_sort.bam
igv_hic_sort.bam Coverage
igv_hic_sort.bam
igv_pacbio_sort.bam Coverage
igv_pacbio_sort.bam

(b)


73 kb
6,700 kb
6,710 kb
6,720 kb
6,730 kb
6,740 kb
Contig1
Contig129
Contig19
Contig198
Contig21
Contig293
Contig345
Contig350
Contig4
Contig40
Contig451
Contig477
Contig508

(c)


73 kb
6,590 kb
6,600 kb
6,610 kb
6,620 kb
6,630 kb
6,640 kb
bac_sort.bam Coverage
[0 - 10.00]
Contig1
Contig129
Contig19
Contig198
Contig21
Contig293
bac_sort.bam
Contig345
Contig350
Contig4
Contig40
Contig451
Contig477
Contig508
igv_pacbio_sort.bam Coverage
[0 - 64]

(d)

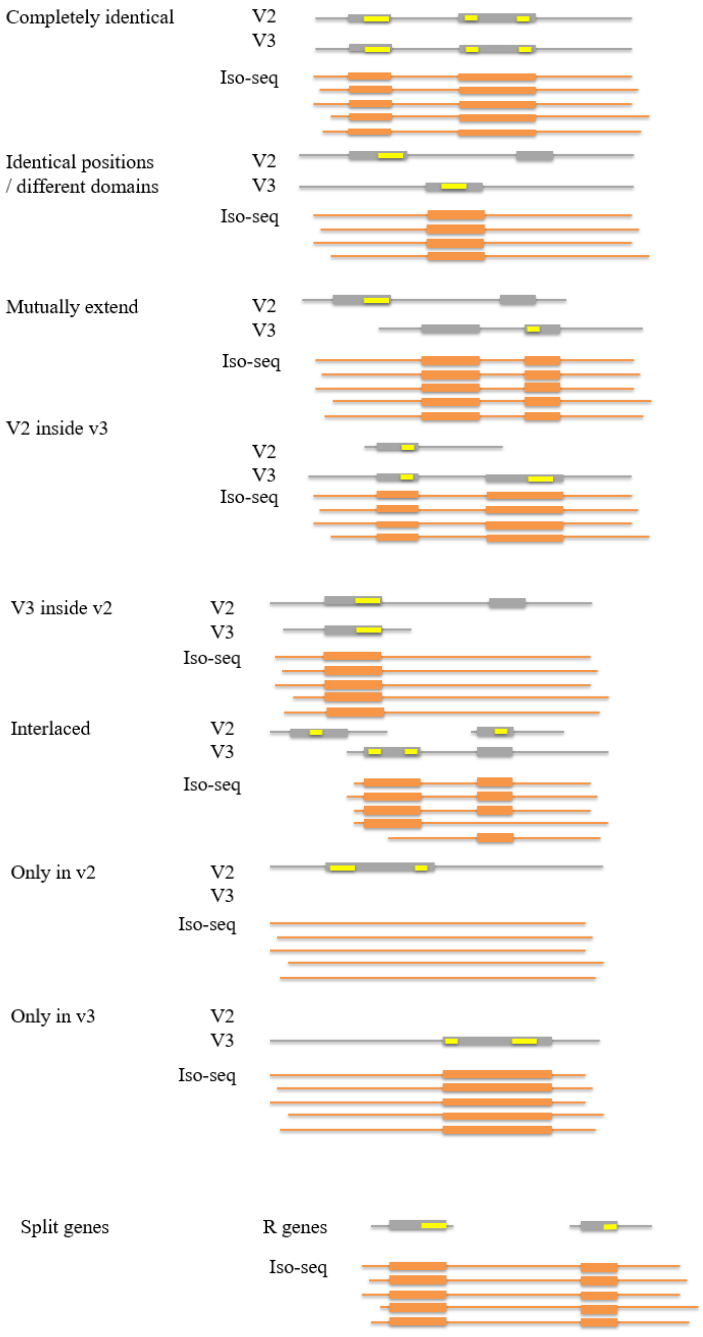


Figure 3.1. Nine categories of R gene comparison in v2 and v3 genomes. (Grey line segments represent R genes. The grey squares on it are exons. Yellow strips represent R gene domains. The orange lines are Iso-seqs.)

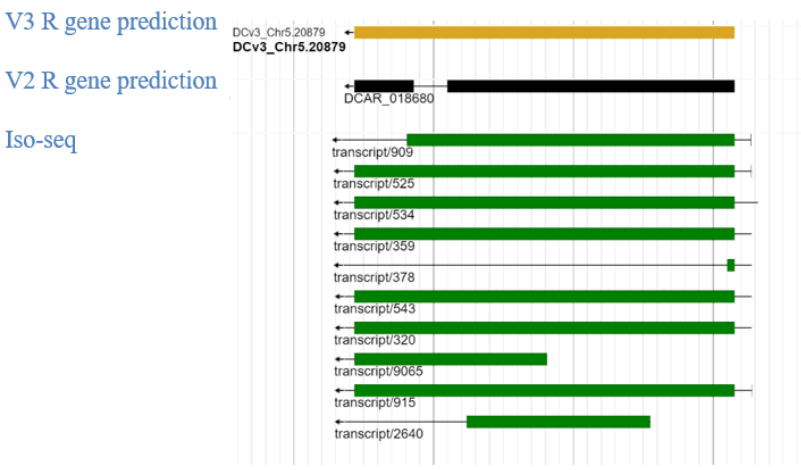


Figure 3.2. An example of less intron regions in v3 genome compared to v2 genome, shown in JBrowse. (Yellow strip represents R gene in v3 genome; black strip represents R gene in v2 genome. Bolder regions are coding regions)

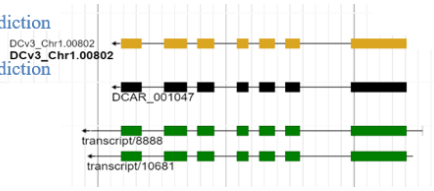
Fig.3.3. Examples of nine categories of pairwise comparison of R genes in the v2 and v3 genomes. (Yellow ones are R genes in the v3 genome. Black ones represent R genes in the v2 genome. Green ones are Iso-seq.)

Completely identical

V3 R gene prediction

V2 R gene prediction

Iso-seq

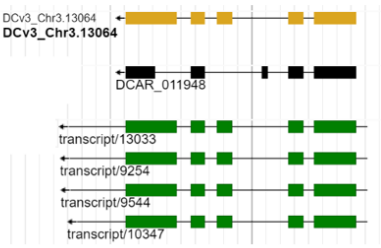


Identical positions / different domains

V3 R gene prediction

V2 R gene prediction

Iso-seq

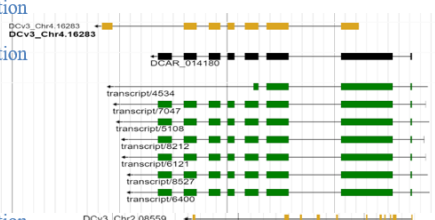


Mutually extend

V3 R gene prediction

V2 R gene prediction

Iso-seq



V2 inside v3

V3 R gene prediction

V2 R gene prediction

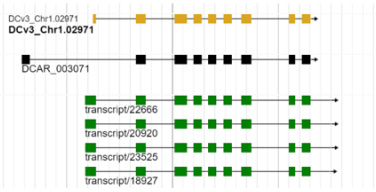
Iso-seq

V3 inside v2

V3 R gene prediction

V2 R gene prediction

Iso-seq

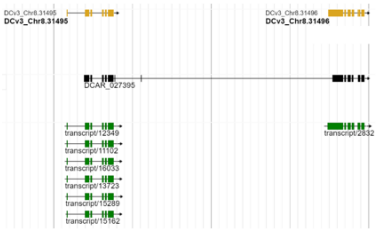


Interlaced

V3 R gene prediction

V2 R gene prediction

Iso-seq

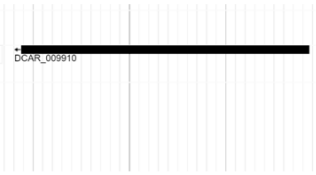


Only in v2

V3 R gene prediction

V2 R gene prediction

Iso-seq

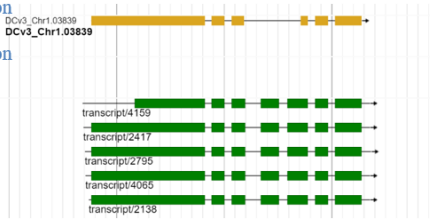


Only in v3

V3 R gene prediction

V2 R gene prediction

Iso-seq



Split gene

V3 R gene prediction

V2 R gene prediction

Iso-seq