2019-07-01

Whole-Genome Assembly of Atriplex hortensis L. Using OxfordNanopore Technology with Chromatin-Contact Mapping

Spencer Philip Hunt
Brigham Young University

Whole-Genome Assembly of *Atriplex hortensis* L. Using Oxford Nanopore Technology with Chromatin-Contact Mapping

Spencer Philip Hunt

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Eric N. Jellen, Chair
Peter J. Maughan
David E. Jarvis

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

Whole-Genome Assembly of *Atriplex hortensis* L. Using Oxford Nanopore Technology with Chromatin-Contact Mapping

Spencer Philip Hunt
Department of Plant and Wildlife Sciences, BYU
Master of Science

Atriplex hortensis ($2n = 2x = 18$, 1C genome size ~1.1 gigabases), also known as garden orach, is a highly nutritious, broadleaf annual of the Amaranthaceae-Chenopodiaceae family that has spread from its native Eurasia to other temperate and subtropical environments worldwide. *Atriplex* is a highly complex and polyphyletic genus of generally halophytic and/or xerophytic plants, some of which have been used as food sources for humans and animals alike. Although there is some literature describing the taxonomy and ecology of orach, there is a lack of genetic and genomic data that would otherwise help elucidate the genetic variation, phylogenetic position, and future potential of this species. Here, we report the assembly of the first high-quality, chromosome-scale reference genome for orach cv. 'Golden'. Sequence data was produced using Oxford Nanopore's MinION sequencing technology in conjunction with Illumina short-reads and chromatin-contact mapping. Genome assembly was accomplished using the high-noise, single-molecule sequencing assembler, Canu. The genome is enriched for highly repetitive DNA (68%). The Canu assembly combined with the Hi-C chromatin-proximity data yielded a final assembly containing 1,325 scaffolds with a contig N_{50} of 98.9 Mb and with 94.7% of the assembly represented in the nine largest, chromosome-scale scaffolds. Sixty-eight percent of the genome was classified as highly repetitive DNA, with the most common repetitive elements being Gypsy and Copia-like LTRs. The annotation was completed using MAKER which identified 31,010 gene models and 2,555 tRNA genes. Completeness of the genome was assessed using the Benchmarking Universal Single Copy Orthologs (BUSCO) platform, which quantifies functional gene content using a large core set of highly conserved orthologous genes (COGs). Of the 1,375 plant-specific COGs in the *Embryophyta* database, 1,330 (96.7%) were identified in the Atriplex assembly. We also report the results of a resequencing panel consisting of 21 accessions which illustrates a high degree of genetic similarity among cultivars and wild material from various locations in North America and Europe. These genome resources provide vital information to better understand orach and facilitate future study and comparison.

Keywords: *Atriplex hortensis*, orach, Oxford Nanopore, DNA sequencing, proximity-guided assembly, genome assembly

ACKNOWLEDGEMENTS

I would like to thank my mentor, friend and committee chair, Dr. Eric N. Jellen for his advice, support and guidance thus far and for allowing me to be his graduate student. I would also like to thank Dr. Peter J. Maughan for his great patience and assistance in helping with many of the bioinformatic aspects of my project. I would also like to thank Dr. David Jarvis for his advice and help in expanding my vision and generating figures. Thank you to Sarah Martin from the department of Agriculture and Agri-Food Canada (AAFC) for helping with genome size estimation work. I would also like to thank Hayley Hansen Mangelson from Brigham Young University for sharing her sequence data with us for this study. Lastly, I would like to thank the other members of my laboratory, department, friends, family and most importantly my wonderful wife Shelby for their support, assistance and patience throughout this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Whole-Genome Assembly of *Atriplex hortensis* L. Using Oxford Nanopore Technology with Chromatin-Contact Mapping

Spencer P. Hunt[1], Peter J. Maughan[1], David E. Jarvis[1], Dallas J. Larsen[1], Eric W. Jackson[2], Bozena A. Kolano[3], and Eric N. Jellen[1]
[1]Department of Plant & Wildlife Sciences, Brigham Young University, Provo, UT
[2]25-2 Solutions LLC, Rockford, MN, 55373, USA
[3]Department of Plant Anatomy & Cytology, University of Silesia, Jagiellonska 28, 40-032, Katowice, Poland

LITERATURE REVIEW

Garden orach (*Atriplex hortensis* L., $2n = 9x = 18$), also known as mountain spinach, is a highly nutritious, C_3 leafy annual plant that has adapted to several harsh ecosystems. Orach is just one of nearly 200 species within the genus that are xero-halophytic, making it resistant to some of the most extreme biotic and abiotic stressors; among them, highly saline soils, wide temperature ranges and drought conditions. Orach is a member of the Amaranthaceae-Chenopodiaceae family (previously known as solely the Chenopodiaceae family and sometimes just as Amaranthaceae) of the flowering Dicotyledonae (Hernández-Ledesma et al., 2015). It is in the same family as some economically important crops such as spinach (*Spinacea*), amaranth (*Amaranthus*), quinoa and goosefoot (*Chenopodium*), and sugar beets (*Beta*) (Dohm et al., 2012).

Morphology and Physiology

Orach demonstrates incredible phenotypic plasticity in pigmentation, height and seed production. Orach plants may be lanky to shrubby and grow between 4-8 feet tall. Garden orach finds itself amidst a compendium of Caryophyllales that are known for their rich variation in color. This variation has led to a common varietal classification system used to separate common orach into four distinct categories corresponding to their coloration. The first category is white

orach which is the most common variety. Its leaves are typically a very pale green, almost yellow color. The second category is red orach which has dark-red stems and leaves. Red orach is the variety typically harvested for human consumption even though species in each variety type are edible. Red orach is also more drought tolerant than varieties in other categories (Sai Kachout et al., 2011). The third category is green orach. Green orach is vigorous, with stout, angular, branching stems (Stephens, 1994). The leaves tend to be reminiscent of spinach leaves as they are rounder, less toothed, and darker green than those of the white variety. The fourth variety is a copper-colored variety that is rarely grown (Stephens, 1994).

This rich diversity in coloration is the result of the production of betalains which are a class of red and yellow indole-derived pigments (Tanaka et al., 2008). Betalain production makes orach unique as betalains are only produced by plants in the order Caryophyllales and some fungi, unlike most other plants that derive their pigmentation from the production of anthocyanins and/or carotenoids (Stafford, 1994; Rohrer et al., 1997). Genotypic as well as environmental variation can result in varied production of specific subcategories of betalains that control coloration. These subcategories include betacyanins, which are reddish to violet pigments, as well yellow to orange betaxanthins. The accumulation of these pigments tends to increase as plant tissues mature, especially in leaves and axils. While orach is mostly used as either a vegetable or ornamental today, it has been documented that the betalains produced in orach have been used to create a variety of dyes derived from both seeds and leaves (Frankton and Bassett, 1968).

Orach plants produce impressive panicles with hundreds of seeds localized at the top of the plant's stalk. Orach has three main fruit/seed types, all achenes, differing in shape (round to pointed), size (small to medium to large) and color (tan to black). These different fruit types vary

in their germination rates and salt tolerance, all of which contribute toward the fitness of the species. The large fruits are brown, thin, flat and vertical, surrounded by a light-yellow leafy membrane. They have the fastest and highest germination rates of all the fruits as they lack a thick testa which would inhibit penetration and imbibition of water. These large fruits are non-dormant and germinate as soon as conditions are favorable, thus ensuring species survival in the short term (Wertis et al., 1986). They are also the least affected by saline pressures (Kahn and Ungar, 1984).

The medium and small fruits are dark brown and black, respectively. They are usually shiny and pitted. Many of these have significantly lower germination rates due to their thicker testas that are recalcitrant to imbibition (Frankton and Bassett, 1968; Stephens, 1994). Their germination rates are higher, however, than the larger fruits after prolonged inactive periods, thus contributing to long-term reproduction and species survival (Venable and Levin, 1985). Scarification of the seed coat of medium and small fruits is usually necessary for successful germination.

Often, various species of *Atriplex* can be among the most frequent and important shrubs on saline, fine-textured substrates. On occasion, *Atriplex* species, especially *A. hortensis*, are among a small number of species inhabiting salty ecosystems, which is why they have been used extensively in land rehabilitation and roadside plantings because of their ability to establish well, grow rapidly, reduce soil erosion, provide excellent wildlife and livestock forage, resist road salt used to melt ice, and grow well with other native plants (Mcarthur et al., 1983; Simon et al., 1994; Wright et al., 2002). Despite its affinity for saline areas where it has little competition (except from other halophytes), orach can also grow where total soluble salts are low – making it broadly adaptable (Welsh and Crompton, 1995).

Orach maintains ionic balance in the high-saline environments that it occupies by depositing salt onto the surface of its leaves via bladders and trichomes. These bladders burst when they contain too much salt and the saline solution dries and crystallizes on the surface of the leaf (Karimi and Ungar, 1989). This removes salt from the surrounding soil and maintains cellular pH so that vital cellular processes can proceed uninhibited. The crystallization of salt on the leaf surface also functions as a UV screen to filter out potentially damaging short-wave radiation and thereby reduce the risk of reactive chemical species such as free radicals from being generated (Grašič et al., 2017; Karimi and Ungar, 1989) as well as deterring insects and herbivores from inflicting foliar damage (LoPresti, 2014).

Systematics

Considering recent phylogenetic and taxonomic developments among Caryophyllales, several distinctions have been made, resulting in more accurate positioning and repositioning of different species, especially within the Chenopodiaceae family. For decades, species characterization and ordering has been based on differentiating characters including free central placentation, perisperm and embryo shape (Kubitzki et al., 1993). Since the advent of DNA sequencing, however, phylogenetic analyses are now opening this characterization up to produce a more comprehensive and accurate order of species relationships. As a result, *Atriplex* circumscription and placement, especially as it relates to other species within its own tribe and family, has changed several times.

For a time, the official description of *Atriplex* as monophyletic or paraphyletic was contested. It was Flores and Davis (2001) who used a morphology-based cladistic analysis to show that *Atriplex* was indeed paraphyletic. Flores and Davis supported this statement by claiming that

genera of both *Atripliceae* and *Chenopodieae* tribes were interrelated, rendering neither to be truly monophyletic (Flores and Davis, 2001). Additionally, Kadereit et al. (2010) and Zacharias & Baldwin et al. (2010) generated molecular data demonstrating that *Atriplex* is not monophyletic, which in turn led to the grouping of several satellite genera that had been incorrectly separated in the past (Hernández-Ledesma et al., 2015).

Fuentes-Bazan has since redefined and extended the circumscription of *Atripliceae* to include the genus *Chenopodium* with this newest definition including all *Atripliceae* as monophyletic (Fuentes-Bazan et al., 2012). Consequentially, the tribes *Atripliceae* and *Chenopodieae* have been merged under the new name of *Atripliceae* (Hernández-Ledesma et al., 2015) to accommodate the new monophyletic definition. This change was made as *Chenopodieae* is paraphyletic to *Atripliceae*. Additionally, Hernandez-Ledesma et al. (2010) placed *Atriplex* within the order Caryophyllales, family Chenopodiaceae, while acknowledging the monophyletic nature of Chenopodiaceae and Amaranthaceae (Hernández-Ledesma et al., 2015). This new positioning was supported most recently by Sukhorukov et al. (Sukhorukov et al., 2018) who used molecular phylogenetic data as well as seed coat anatomy to resolve the disputed position of some *Amaranthaceae-Chenopodiaceae* species.

History

Garden orach is part of the Caryophyllales with a unique tolerance for aridity and salinity. Originating in Eurasia (especially in Siberia), it is widely believed that garden orach is one of the oldest wild, edible cultivated plants in existence (Mcarthur et al., 1983; Stephens, 1994; Wright et al., 2002). This is evident as many different tribes who settled in the Trans-Himalayan region of Tibet and India still depend largely on wild edibles such as orach for their livelihood. Orach is

a food source of particular significance to the peoples of the Ladakh and Nubra Valley regions of India and Pakistan as it is the first leafy vegetable to appear after the prolonged winters that are characteristic to the area, giving people much-needed nutrients that cannot be found in high abundance in milk or meat (Rinchen and Singh, 2015; Rinchen et al., 2017).

Although never utilized as a large-scale crop, orach was commonly used during the Middle Ages as a leafy garden vegetable throughout the Mediterranean and other parts of greater Eurasia (Harvey, 1984). Its popularity decreased as spinach was consumed with greater frequency throughout the world. Orach is still used as a garden vegetable throughout parts of Europe, especially in France and Italy where it is commonly incorporated into local cuisine. Orach has since become naturalized throughout the Americas. It can be found as a free-living weed in the cold, temperate areas of northern Alberta all the way down to the much warmer climates of northern and central Mexico.

Medicine and Agriculture

Garden orach has been known for its medicinal properties. Many of orach's remedial characteristics are still utilized in traditional eastern medicinal practices today. While the list of health claims and potential benefits for orach are long, proven benefits include better digestion, increased circulation and a boosted immune system among others (Rinchen et al., 2017). Additionally, orach leaves are diuretic making them useful in treating vomiting and efficacious in the treatment of gout (Simon et al., 1994).

As the world clamors for new ways to feed its ever-growing population, novel food sources have gained popularity that have helped provide diversity to diets while capitalizing on less

desirable, underutilized or even fallow landscapes for agriculture. One such species of interest is quinoa (*Chenopodium quinoa* Willd.). However, some quinoa varieties require processing by washing to remove saponins – a bitter compound that resides in the seed coat. Orach seeds also contain saponins in their seed coat, although they are thought to reside deeper within, potentially making them more difficult to remove for safe consumption in large quantities. Despite this, orach seeds could be a potential substitute for quinoa as they have a similar nutritional profile (Wright et al., 2002). Amino acid profiles for garden orach seeds are very similar to sweet and bitter quinoa. Garden orach seeds have essential amino acid levels equal to or exceeding recommended adult levels by the World Health Organization, the Food and Agriculture Organization of the United Nations, and the United Nations University (FAO/WHO/UNU) (Wright et al., 2002).

Where orach and quinoa begin to differ is in protein content. Orach seeds have a high protein content (26% dry weight) which is comparable to that of some legumes (Wright et al., 2002), whereas quinoa seeds only contain between 15% - 17% protein (Ranhotra et al., 1992). Garden orach seeds have higher fat, ash, and fiber contents as well as substantially higher lysine contents than most cereal grains (Wright et al., 2002). Orach leaves also have a very high protein content of 35% (per dry weight tissue). The high protein content and well-balanced amino acid profile of garden orach also make it a very attractive, novel protein source (Wright et al., 2002). Because both its leaves and seeds are edible, orach is a doubly productive crop making it an extremely attractive food plant in comparison with related pseudocereals like quinoa and amaranth.

DNA Sequencing Methods – Advantages and Disadvantages

DNA sequencing technologies have rapidly evolved over the past 30 years. Today, there are many available technologies, each with its own advantages and disadvantages that make it useful depending on the scope and budget of the study. Different combinations of data generated from these technologies are frequently used to capitalize on individual strengths while masking the weaknesses of each method. This results in highly accurate polished genome assemblies that can potentially generate chromosome-sized scaffolds.

Sanger sequencing was the first widely used DNA sequencing technology. As a result, this technique was instrumental in paving the way for future technologies and studies. Data generated by this technology is still in use today as Sanger sequencing yields very precise reads that are extremely beneficial when doing targeted sequencing, for example of individual genes, using flanking PCR primers. Unfortunately, Sanger sequencing is extremely expensive, time-consuming, and yields read lengths approaching 800 bases albeit with low throughput. As a result, the popularity of this technology for whole-genome sequencing rapidly diminished as cheaper, high-throughput next-generation sequencing platforms became accessible to scientists beginning in the early 2000's.

Roughly 10 years ago, Illumina sequencing emerged with a new platform and chemistry that produced high-throughput, high-fidelity short-read sequences. Many researchers currently rely on Illumina sequencing to achieve greater read depth due to Illumina's ultra-high throughput. This technology led to a significant increase in published plant genome assemblies. Illumina's biggest disadvantage is linked to its greatest strength. While Illumina's short-read lengths allow for high throughput, they make it difficult to correctly assemble certain areas of the genome; this is especially true in plant species having highly repetitive genomes that require longer reads to

span problematic regions. Additionally, since Illumina library preparation is relatively expensive, it can be cost-prohibitive to sequence large numbers of genomes in parallel using this technology.

In contrast to Illumina, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) sequencing have developed distinct methods that are capable of sequencing long DNA fragments in the kilobase to megabase range with high throughput. As a result, many high-quality assemblies in terms of contiguity and completeness of repetitive regions have been produced. The main disadvantage to both PacBio and ONT sequencing is their low precision due to high sequencing error rates of 10%-15%. The ONT platform is distinct from PacBio in that it is portable and can be easily integrated into studies at remote locations unlike all other sequencing machines. It is also user friendly and generates longer sequence data at a very low cost. To date, ONT sequencing has only produced a handful of complete plant genomes.

Often, short and long reads produced by the aforementioned technologies alone are still not enough to resolve some areas of the genome that are difficult to correctly assemble. Chromatin-contact sequencing is another method that is frequently used to solve this problem. Chromatin-contact sequencing relies on the proximity of either endogenous or reconstituted DNA molecules to determine spatial relationships that are useful in understanding which sequences exist closest to and farthest from each other (Lieberman-Aiden et al., 2009; Sati and Cavalli, 2017). Together, these technologies each contribute towards producing a more complete genome.

Objectives of the Present Study

To better understand the underlying differences that grant orach its xero-halophytic, nutritive and unique pigmentation characteristics, not to mention the development of more

accurate phylogenetic relationships within its family and genus, we present the diploid orach genome. We show that ONT technology-based reads with short-read polishing is an effective way to generate a high-quality genome assembly. We demonstrate the quality and utility of the chromosome-level genome assembly by using a widely accepted assessment to prove genome completeness and by genomic comparison to other Caryophyllales, more specifically, species within the *Amaranthaceae-Chenopodiaceae* family. Additionally, we sequenced the genomes of other diploid orach accessions to characterize genetic diversity and to better understand genome evolution in orach. Together, these resources provide the foundation for a deeper understanding of orach and how it may be used to contribute to improve global food security and the potential genetic improvement of this fascinating crop.

INTRODUCTION

Atriplex hortensis L. ($2n = 9x = 18$), also known as garden orach or mountain spinach, is a highly nutritious, leafy annual plant. Orach is a xero-halophytic Caryophyllale that is resistant to salinity, a wide range of temperatures, and drought. Orach is a member of the Amaranthaceae-Chenopodiaceae family (previously known as solely the Chenopodiaceae family) of the flowering Dicotyledonae (Hernández-Ledesma et al., 2015). Originating in Eurasia, orach has been an important local food source for natives of certain areas of the Trans-Himalayan region. Orach has since become naturalized throughout the Americas. Orach exhibits incredible variation in pigmentation as a result of betalains as well as substantial differences in height and seed production (Tanaka et al., 2008).

Orach is a broadly adaptable species that has many uses. Orach has been known for its medicinal properties which have been shown to improve digestion, increase circulation and boost

the immune system (Rinchen et al., 2017). Additionally, orach has been used in land rehabilitation projects because of its ability to establish well, grow rapidly, reduce soil erosion and grow well with other native plants (Mcarthur et al., 1983; Simon et al., 1994; Wright et al., 2002). As a result, orach is important for both domestic and wild browsing animals where other forage crops are lacking (Simon et al., 1994). Despite its affinity for saline areas where it has little competition (except from other halophytes), orach can also grow where total soluble salts are low, making it well suited to a multitude of different environments (Welsh and Crompton, 1995).

As the world continues to search for new ways to feed its ever-growing population, new food sources have gained popularity that have helped provide diversity to diets while capitalizing on less desirable, underutilized or even fallow landscapes for agriculture. Given its xero-halophytic characteristics, orach is an interesting candidate for contributing to the solution of the food security, especially in saline soils. Orach is a doubly productive crop as both its leaves and seeds are edible. In comparison to other leafy vegetable, grain and pseudocereal crops, orach seeds and leaves have a high protein content with 26% (dry weight) in seeds, which is comparable to some legumes (Wright et al., 2002), and 35% (dry weight) in leaves, which is higher than spinach, a close relative of orach with similar nutritional characteristics. Orach seeds contain antinutritional saponins that must be removed by washing. This problem is not unique to orach seeds and is frequently seen in the emerging super grain quinoa as well. Orach seeds are known to have higher fat, ash, and fiber and substantially higher lysine contents than most cereal grains (Wright et al., 2002). Its high protein content, which includes an essential amino acid profile that meets the WHO and UN-FAO recommended adult levels, also make orach very attractive as a novel protein source (Wright et al., 2002).

Few studies have been focused on orach in recent years. Some of the most notable developments surrounding orach involve molecular studies that have led to phylogenetic and taxonomic improvements among Caryophyllales (Kadereit et al., 2010; Flores and Davis, 2001). As a result, there have been several adjustments in the positioning and circumscription of *A. hortensis* with current consensus branding it as a paraphyletic member of the new Amaranthaceae-Chenopodiaceae family (Flores and Davis, 2001). There have also been studies conducted that test the limits of salt-tolerance of orach (Sai Kachout et al., 2011; Vickerman et al., 2002). Unfortunately, there has been little to no research conducted to develop genetic tools that are necessary to accelerate the improvement of orach.

To better understand the underlying genetic basis of orach's xero-halophytic, nutritive and unique pigmentation characteristics, not to mention the development of more accurate phylogenetic relationships within its family and genus, we present the orach genome. We show that ultra-long reads produced by the portable, real time Oxford Nanopore Technology (Oxford, UK) ONT sequencing system (Lu et al., 2016) with short-read polishing and chromatin-contact mapping is an effective approach to generate a high-quality genome assembly in a moderately large-genome diploid plant species. Additionally, we annotated the genome with a deeply sequenced transcriptome from various orach plant tissues. Lastly, we demonstrate the quality of the chromosome-level genome assembly by using a widely accepted tool to assess genome completeness and by genomic comparison to other Caryophyllales within the Amaranthaceae-Chenopodiaceae family. Together, these resources provide an initial, but important foundation for future accelerated genetic improvement of this potentially valuable crop needed to advance global food security.

MATERIALS AND METHODS

Plant Material

Atriplex hortensis, cv. 'Golden' seed was used in estimating genome size and for cytogenetic analysis and was obtained from Wild Garden Seed (Philomath, Oregon). *Atriplex hortensis,* cv. 'Triple Purple' seed used in other cytogenetic analysis was obtained from the same vendor. The resequencing panel consisted of 21 *A. hortensis* accessions: 15 from the United States Department of Agriculture collection (USDA; Ames, Iowa, USA; https://npgsweb.ars-grin.gov/), five from the seed vendors Baker Creek Heirloom Seed Company (Mansfield, Missouri) and Wild Garden Seed (Philomath, Oregon). One accession was collected in the wild in Utah and is reposited at Brigham Young University (BYU 1317; Provo, Utah). Plants used in the resequencing panel were originally collected from across Europe (France, Poland, former Soviet Union, former Serbia/Montenegro and Norway) and North America (Oregon, USA, Utah, USA, and Alberta, Canada). A complete list of all plant materials including accession information is provided in Table 1.

Genome Size Estimation

Atriplex hortensis, cv. 'Golden' seed was grown hydroponically in a growth chamber at BYU. An 11-hour photoperiod was maintained using broad-spectrum light sources. Growing temperatures ranged from 18° C to 20° C. Hydroponic growth solution was made from MaxiBloom® Hydroponics Plant Food (General Hydroponics, Sevastopol, CA, USA) at 27 g/16 L. Hydroponic solution was changed every five days. Plants were grown for 25 days.

Genome-size estimation was conducted using flow cytometry by Agriculture and Agri-Food Canada (AAFC) and at BYU using the CytoFLEX flow cytometer (Beckman Coulter, 405/488

nm lasers). Leaf tissue was prepared using standardized protocols (Galbraith et al., 1983) and DNA was stained using propidium iodide. *Solanum lycopersicum* (tomato) was used as a standard to measure orach genome size as it has a known size of 900 megabases (Mb) (Consortium, 2012).

DNA Extraction, Library Preparation and Oxford Nanopore Sequencing

The 'Golden' variety of orach was grown hydroponically in a growth chamber at BYU as previously described. Plants were dark-treated for 72 hours at which point young leaf tissue was harvested. A Qiagen-Nanopore high molecular weight (HMW) genomic DNA (gDNA) extraction protocol produced by Oxford Nanopore Technologies (Oxford Nanopore Technologies, 2018) was used to extract DNA. DNA quality was checked to ensure that 260/280 and 260/230 absorbance ratios were within acceptable ranges using Thermo Scientific's NanoDrop™ (ThermoFisher Scientific/Nanodrop, Wilmington, DE). DNA concentration was then checked by using the dsDNA High Sensitivity DNA Assay (ThermoFisher Scientific) on the Qubit® 2.0 Fluorimeter (Invitrogen, Merelbeke, Belgium).

Samples for DNA sequencing were prepared without fragmentation and with fragmentation using Covaris g-TUBEs (Covaris, Woburn, MA, USA, 520079) and a ZYMO DNA Clean & Concentrator-5 column (ZYMO Research, Cat. No. D4010). Samples fragmented using the ZYMO kit were prepared following the manufacturer's instructions. Samples prepared using the Covaris g-TUBEs were centrifuged at 3,800, 4,000 and 4,200 RPM depending on desired read lengths also following the manufacturer's instructions. In total, nine libraries from the same DNA stock were prepared for sequencing using ONT's 1D Genomic DNA by Ligation MinION

library preparation protocol. ONT's SQK-LSK109 kit was used for library preparation with Quick T4 DNA Ligase (NEB, M2200L).

Oxford Nanopore Technology R9 flow cells were used for sequencing on the MinION™ (Oxford Nanopore Technologies). Samples were run for 48 hours using MinKNOW 2.0 software with the following settings: DNA, PCR-free, no multiplexing, SQK-LSK109 kit. No alterations were made to voltage or time. Base calling was done in parallel using MinKNOW 2.0 software for the first three samples. The remaining samples were base-called on the supercomputer cluster at the Fulton Supercomputing Laboratory, BYU using Albacore v2.3.1 with options K=SQK-LSK109 and F=FLO-MIN106.

Read Cleaning, Draft Genome Assembly and Polishing

All mux scans from Nanopore runs were omitted from the assembly as they are often truncated reads. MinIONQC (Lanfear et al., 2018) was used with default settings to summarize sequence data. NanoFilt (De Coster et al., 2018) was then used to trim and filter reads using the following options: -q = 8, headcrop = 25, -l = 2000. Porechop v.0.2.3 (Wick, 2017) was used to trim adaptors from sequence data with the following options: -t (threads) 24 -v (verbosity) 2. Draft genomes were assembled using CANU v.1.7.1 (Released June 2018) (Koren et al., 2017) with the following options: corMhapSensitivity=normal, corOutCoverage=40 and ovsMethod=parallel, MaSuRCA v.3.2.8 (Released August 2018) (Zimin et al., 2013), Flye v.2.3.6 (Released September 2018) (Kolmogorov et al., 2018) and wtdbg (Ruan, 2018). Illumina reads were used to polish sequence data using Nanopolish (Loman et al., 2015), Pilon v.1.22 (Walker et al., 2014) and RACON (Vaser et al., 2017). Three rounds of polishing were conducted with different combinations of the previously mentioned polishing programs to

determine which polishing iterations and how much polishing was necessary for optimal assembly accuracy.

Proximity-based Sequencing and Scaffolding

Orach tissue was dark-treated for 72 hours and flash-frozen in liquid nitrogen before being shipped to Dovetail Genomics™ for Chicago and Hi-C proximity ligation sequencing (Dovetail Genomics LLC, Santa Cruz, CA, USA). Dovetail Chicago libraries are similar to Hi-C libraries but differ in that they rely on library preparation from *in vitro* rather than *in vivo* reconstituted chromatin that has been cross-linked and subsequently sheared (Moll et al., 2017). Assembly was completed for both Chicago and Hi-C sequence data using the HiRise™ assembler (Dovetail Genomics LLC, Santa Cruz, CA, USA).

Illumina Sequencing and Transcriptome Assembly

The 'Golden' variety of orach was grown hydroponically in a growth chamber at BYU as previously described. Plants were either grown in normal hydroponic solution or in hydroponic solution with an augmented concentration of NaCl. Once plants were one week old, NaCl was added incrementally to the hydroponic solution, 50 mM at a time on a daily (24 hour) basis until a concentration of ~350 mM NaCl was reached. Tissue for RNA extraction was harvested 24 hours after ~350 mM NaCl concentration was achieved. Root, stem and leaf tissue was harvested from plants in both treatments. One-week old whole plantlet and inflorescence (tissue and immature seed) tissues from untreated plants were also collected.

In total, seven libraries were prepared with 180 bp inserts. Sequencing was conducted using the Illumina HiSeq platform at the Beijing Genomics Institute (BGI, Shenzhen, China). Reads

were trimmed using the program Trimmomatic-0.35 (Bolger et al., 2014). The ILLUMINACLIP option was used to remove adapters from reads. SLIDINGWINDOW option was set to 4:20. LEADING and TRAILING options were set to 20. The MINLEN option was set to 75. RNA-seq data were aligned to the Hi-C assembly using HiSat v2.2.1 with the max intron length set to 50,000 bp and the number of threads set to 32 (Kim et al., 2015). Data was then assembled into potential transcripts using StringTie (Pertea et al., 2015) with default parameters.

Repeat Analysis and Annotation

Repeat motif analysis was conducted using RepeatModeler v.1.0.11 (Smit and Hubley, 2008) and RepeatMasker v.4.0.7 (Hubley et al., 2018). RepeatModeler consists of two subprograms: RECON v1.08 and RepeatScout v1.0.5 that work to find novel repeats in the input genome. RepeatMasker was run with rmblastn version 2.2.27+. The query was compared to classified sequences in the consensi file using RepBase/RepeatMasker database version 20160829 which was then used to quantify and classify the RepeatModeler output.

The MAKER2 v2.31.10 pipeline (May, 2018) (Holt and Yandell, 2011) was used to annotate the polished *A. hortensis* genome with *ab initio* gene predictions from AUGUSTUS. Evidence for expressed sequence tags (EST) and protein homology included the *C. quinoa* and *C. pallidicaule* transcriptomes provided by Hayley Hansen Mangelson from BYU (Hansen Mangelson et al., 2019) as well as RNA-seq data from previously described stem, leaf, root, floral and whole plantlet tissues. *Chenopodium quinoa* coding sequence (CDS) gene models were obtained from the previously reported genome assembly (Jarvis et al., 2017). Protein evidence included the uniprot_sprot database (ftp://ftp.uniprot.org/pub/ databases/uniport /current_release/knowledgebase/complete/uniprot_sprot.fasta.gz/; downloaded 11/13/2018).

Resequencing

Seeds were grown in BYU's greenhouse in Provo, UT. After approximately three weeks, leaf tissue was collected and lyophilized. Genomic DNA was extracted using the mini-salts protocol reported by Todd and Vodkin (Todd and Vodkin, 1996). DNA was resuspended in TE buffer. DNA concentration was checked using the dsDNA BR Assay from Qubit® 2.0 Fluorimeter. Libraries were sent to Novogene (San Diego, CA) for whole-genome Illumina HiSeq X Ten sequencing (150-bp paired-end) where approximately 13x coverage was achieved. Trimmomatic was used to trim paired-end Illumina reads (Bolger et al., 2014) using the same options as stated previously. Reads from each accession were then aligned to the orach genome using BWA-MEM v0.7.17 (Li, 2013) to produce SAM files which were then converted to BAM format, sorted and indexed using SAMtools v1.9 (Li et al., 2009). The InterSnp tool from the program BamBam v1.4 (Herold et al., 2009) was used to identify SNPs. Hapmap output files were analyzed by SNPhylo v20160204 (Lee et al., 2014) with the bootstrapping parameter set to 1000. Samples with missing data, a minor allele frequency lower than 10%, or linkage disequilibrium greater than 10% were removed. Following filtering, SNPs were used to generate a phylogeny which was visualized with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree).

Genome Quality and Comparison

Each newly polished genome was run through the BUSCO v3.0.2 (Stanke et al., 2004) pipeline using the flowering plant (embryophyte_odb10) orthologous gene data set. The BUSCO pipeline tests for conserved orthologous genes (COGs) expected to be found in all flowering plants and is a widely accepted assessment for genome completeness (Simão et al., 2015). BUSCO scores were used in tandem with AUGUSTUS v.3.3 (Stanke et al., 2004) to make an

AUGUSTUS training set specific to orach which was later used for downstream annotation of the assembled genome. BUSCO scores were generated from assemblies of *C. quinoa, C. suecicum, C. pallidicaule*, *A. hortensis, A. hypochondriacus* and *B. vulgaris* which were used to generate a phylogeny using the multiple sequence alignment function from Clustal Omega (Sievers and Higgins, 2014) with the PHYLIP output format selected.

Cytogenetics

Atriplex hortensis cv. "Golden" seeds were germinated on petri dishes for 36 hours. Root meristems were severed, collected and immersed in ice water for 24 hours. Root meristems were then treated for another 24 hours in a 3:1 mixture of ethanol (95%) – glacial acetic acid. Root tips were prepared under a dissecting microscope where they were placed on slides, treated with iron-acetocarmine, warmed on an alcohol burner, and squashed. Chromosomes were examined using a Zeiss Axioplan 2 phase-contrast microscope and images were captured on an Axiocam (Carl Zeiss, Jena, Germany) CCD camera.

Fluorescent in-situ hybridization (FISH) rDNA images of mitotic chromosome preparations of *A. hortensis* cv. 'Triple Purple' were taken using yellow-green fluorescing digoxygenin to highlight the NOR-35S region and red fluorescing rhodamine to highlight the 5S region using the protocol described by Maughan et al. (2006). Tissue for squashes and probes were prepared using the protocol described in Kolano et al. (2012).

RESULTS

Genome Size and Cytogenetics

Flow cytometry performed at Agriculture and Agri-Food Canada as well as at the BYU Research Instrumentation Core facility (RIC) verified that the orach genome is approximately 1.2 Gb (Table 2).

Cytogenetic analysis showed that garden orach contains nine pairs of chromosomes ($2n = 9x = 18$). Chromosomes were metacentric to slightly submetacentric (Figure 1), and similar in length. A FISH analysis of mitotic chromosome preparations for *A. hortensis* cv. 'Triple Purple' conducted prior to the start of this project revealed the physical positions of the single NOR-35S (green) locus and 5S (red) rRNA tandem repeat-array locus (Figure 1). A Basic Local Alignment Search Tool (BLAST) search (Altschul et al., 1990) of the complete rRNA gene sequence found in *C. quinoa* (DQ187960.1) was conducted to identify the 25S rRNA gene (NOR) location in the *A. hortensis* pseudochromosome assembly using the *C. quinoa* sequence as query. The 25S rRNA locus was located on chromosome Ah06. Another BLAST search was conducted to identify matches for the 5S rRNA gene locus in *A. hortensis*, again using the complimentary 5S repeat sequence in *C. quinoa* (DQ187967.1) as the query. The 5S sequences mapped primarily to chromosome 4 and to several other smaller scaffolds that did not assemble into specific chromosomes. The appearance of these smaller scaffolds in the BLAST search results is not surprising as this is a low-complexity, highly repeated region. These 5s rRNA and 25s rRNA features give unique identities to two of the nine chromosome pairs.

Sequencing, Assembly and Hybrid Scaffolding

Because ONT sequencing is still relatively new, we tested the relationship between fragmentation strategy, read length and total sequence output to discover the optimal sample preparation method. To achieve sufficient coverage, we developed nine different libraries that were each sequenced independently on different flow cells. In total, the nine libraries yielded 65.4 Gb of data from 4,969,313 reads. Means were generated for each library and total means were generated to describe the pooled sequence data. Sequence data had a mean N50 of 22,087 bp, a mean read length of 13,487 bp and a mean quality score of 9 (Supplemental Figure 1, Table 3). DNA samples prepared with fragmentation (Covaris g-TUBEs and ZYMO DNA Clean & Concentrator-5 column kit) and without fragmentation yielded varied results. Not unexpectedly, the sample prepared without fragmentation produced long read lengths but low overall throughput. Samples prepared using Covaris g-TUBEs were centrifuged at variable speeds (3,800, 4,000 and 4,200 rpm). Higher centrifugation speeds producing shorter fragments yielding greater throughput, whereas slower centrifugation speeds produced longer fragments yielding less throughput (Supplemental Figure 1, Table 3). Covaris g-TUBEs yielded an average throughput of 9.38 G of data per run compared to 3.93 G of data per run when the ZYMO DNA kit was used.

Canu (Koren et al., 2017), Masurca (Zimin et al., 2013), Flye (Kolmogorov et al.) and wtdbg (Ruan, 2018) assemblers were used to determine which would most optimally assemble the DNA sequencing data. The high scaffold number from the wtdbg assembly, the low scaffold N50 from the Flye assembly and the low scaffold size from both the wtdbg and Flye assemblies led us to look more closely at the Masurca and Canu assemblies as better options (Figure 2). Although the Masurca assembly was appealing, the larger assembly size of the Canu assembly was closest to

the actual genome size of *A. hortensis* which ultimately informed our final decision to move forward with the Canu assembly. The Canu assembly had 3,183 scaffolds with BUSCO identifying only 694 (50.5%) COGs. We then conducted three rounds of polishing with different combinations of Nanopolish, RACON and Pilon polishing programs (Figure 3). The combination of Nanopolish-Pilon-Pilon yielded the highest BUSCO score 97.5% (1,340) (Figure 3). The final Canu assembly after polishing resulted in 2,191 scaffolds, a contig N_{50} of 816.58 kb with the longest scaffold being 9.6 Mb in size. After running BUSCO on this assembly, BUSCO identified 1,340 (97.5%) complete COGs from the assembly resulting in an overall reduction of 992 scaffolds and a 47% increase in the BUSCO score is an improvement from the pre-polished assembly that had 3,183 scaffolds with a BUSCO score of 694 (50.5%).

The input assembly for Hi-C proximity sequencing had 3,183 scaffolds, an N50 1,114.7 kb and the longest scaffold spanning 9,632,068 bp. First, *in vivo* chromatin structures were used to produce Chicago reads. These sequences were aligned to the draft assembly and a likelihood model was produced that describes features pertaining to genomic distance was created. From this model, putative breaks, joins and gap closures were identified which were used to align and scaffold the Chicago data. This same process was performed with the Hi-C data. The Chicago scaffolding made 429 breaks and 1,421 joins via the HiRise assembler producing an assembly with 2,191 scaffolds, an N_{50} of 816.58 kb and the longest scaffold spanning 15,147,297 bp. The i*n vivo* chromatin Hi-C scaffolding process made 868 joins and 0 breaks producing a final assembly containing 1,325 scaffolds with an N_{50} of 98.9 Mb with the longest scaffold spanning 113.5 Mb in size. Nine chromosome-scale scaffolds were assembled representing 94.7% of the total sequence length (Figure 4A). The chromosome-scale scaffolds ranged in size from 93.6 Mb to 113.5 Mb which corresponds to the similarly sized chromosomes visualized in the previously

described cytogenetic analysis. Chromosomes were numbered one through nine based on scaffold length (e.g., Ah01 – Ah09). After running BUSCO on this new assembly, 1,330 (96.7%) of the 1,375 COGs in the *Embryophyta* database were identified demonstrating a high level of completeness (Complete: 96.7% [Single: 95.0%, Duplicated: 1.7%], Fragmented: 0.8%, Missing: 2.5%). This supports the overall quality of the hybrid assembly. Compared to the input assembly, 10 fewer BUSCOs were identified from the Hi-C data (Table 4). This slight decrease, while not too worrisome, is the result of fewer single copy and duplicated orthologs being identified. This could be due to the differences in assembly methods between Canu and HiRise as HiRise relies on proximity ligation data to create scaffolds whereas Canu is a hierarchical assembly pipeline that relies solely on overlap detection data to assemble genomes.

The transcriptome analysis of root, stem, leaf, floral and whole plantlet tissues resulted in 30 to 40 million reads per tissue library, totaling approximately 4 Gb of data. The StringTie assembler (Pertea et al., 2015) produced a transcriptome with 302,037 transcripts with an N_{50} of 3,815 bp and a mean length of 2,136 bp.

Repeat Modeling and Genome Annotation

The RepeatModeler and RepeatMasker pipelines revealed that the genome of orach is highly repetitive with 68.23% (657.8 Mb) of the total assembly being marked as repetitive. Repeat Masker also identified 1.95% (18.8 Mb) as low-complexity elements (simple repeats, satellites and small RNA species). The most common elements were long-terminal repeat (LTR) retrotransposons with LTRs and transposable elements constituting 49.28% (480.6 Mb) of the genome with the two most frequent types being Gypsy (33.25%) and Copia (10.85%) elements.

An additional 16.08% (155 Mb) of the genome was characterized as unclassified repetitive elements (Table 5). Additionally, there were a total of 3,300 microsatellites identified.

The MAKER pipeline identified a total of identified 31,010 gene models and 2,555 tRNA genes. The average length of genes identified by MAKER was 1,177 bp with the longest gene (without introns) spanning 21,489 bp. The completeness of the annotation was assessed by BUSCO which identified 1,331 (96.8%) complete COGs from the annotation (Complete: 96.8% [Single Copy: 95.1%, Duplicated: 1.7%], Fragmented: 0.7%, Missing: 2.5%). To assess the quality of the assembly, we used the mean Annotation Edit Distance (AED) which is calculated by combining annotation values corresponding to specificity and sensitivity. AED values of 0.5 and below are reasonable annotations and values of 0.25 and below are high quality annotations (Holt and Yandell, 2011). The AED score of 0.5 coupled with the BUSCO assessment provides evidence for a reasonable genome annotation (Figure 5). The majority (58.5%) of genes identified in the annotation had AED values below 0.25 (Figure 5).

Genomic Comparison and Features

Synteny plots were generated showing relationships between homoeologous chromosomes in *B. vulgaris* (Dohm et al., 2014) (n = 9), *C. quinoa* (Jarvis et al., 2017) (n = 18) and *A. hypochondriacus* (Lightfoot et al., 2017) (n = 16) (Figure 6). Previous research suggests that *A. hortensis* is more closely related to *C. quinoa* than *A. hypochondriacus* and *B. vulgaris*. To verify this, we quantified the synteny results. The *A. hortensis* and *C. quinoa* plot had a combined total of 31,229 syntenic gene pairs. Both species have 78,341 gene models amounting to 40% of annotated gene models being conserved between the two species (Table 6). The *A. hortensis* and *A. hypochondriacus* plot had a combined total of 17,793 syntenic gene pairs. Of the combined

57,412 annotated gene models, 31% of are found in syntenic gene pairs between orach and amaranth (Table 6). The *A. hortensis* and *B. vulgaris* plot had 18,553 syntenic gene pairs. Combined, there are 60,986 annotated genes models demonstrating 30% of genes are conserved between the two species (Table 6). Based on these results, we see that indeed *C. quinoa* is more closely related to *A. hortensis* than *A. hypochondriacus* and *B. vulgaris.* It should be noted that the allotetraploid species *A. hypochondriacus* and *C. quinoa* may inflate the quantified results as a gene model in *A. hortensis* may result in more than one match in the comparator genomes as genes are doubled. This closer relationship to *C. quinoa* is also reflected in the decreased amount of chromosomal rearrangements present in comparison compared to the other two species which reinforces the phylogeny in Figure 7. Synteny can also be seen between homologous orach and beet for chromosomes Ah01-Bv03, Ah02-Bv02, Ah03-Bv01/09, Ah04-Bv05, Ah05-Bv06, Ah06-Bv01/04, Ah07-Bv06, Ah08-Bv07, and Ah09-Bv04/09 in the circular synteny plot in Figure 8. Quantitative support for these chromosome assignments between *A. hortensis* and *B. vulgaris* further highlights these homoeologous relationships and is provided by the number of syntenic blocks and syntenic gene pairs that were found (Table 7). In total, there were 557 syntenic blocks and 9,721 syntenic gene pairs identified from the chromosomes.

The sequence for telomeric repeats in plants has been identified as TTTAGGG (Richards and Ausubel, 1988). A BLAST search of this sequence against the nine chromosome-sized scaffolds identified tandemly repeated telomeric sequences on every chromosome with 13 repetitive regions identified in total (Figure 9). All chromosomes had the telomeric repeat sequence present at one or both ends of the pseudochromosome as expected.

Resequencing

A diversity panel, consisting of 21 diverse varieties of orach (Table 8), underwent whole-genome, paired-end Illumina sequencing resulting in an average of 13x coverage. Following alignment of the sequencing reads to the orach reference, InterSnp identified 327,645 SNPs from the diversity panel. These were then filtered based on minor allele frequency, missing data and linkage disequilibrium resulting in 1,708 SNPs that were used to develop the phylogeny. There are an average of 190 SNPs per chromosome contributing to the phylogeny (Table 8). When visualized using FigTree, three distinct nodes appeared in the phylogeny (Figure 10) with two accessions clustering in a North America-specific group on the left, five accessions clustering in a Europe-specific group on the right and the remaining 14 clustering in a North America/Europe group in the middle.

DISCUSSION

Library Preparation Findings

Libraries were prepared for ONT sequencing with fragmentation using Covaris g-TUBEs and the ZYMO DNA Clean & Concentrator-5 column kit and without fragmentation to ascertain if fragmentation influenced sequencing output and which kits and centrifugation speeds produced the most desirable results. We found that Covaris g-TUBEs were the most effective fragmentation technique for orach library preparation based on improved throughput. Samples prepared using Covaris g-TUBEs were centrifuged at variable speeds (3,800, 4,000 and 4,200 rpm) which yielded variation in throughput. We also found that sample fragmentation improved the overall output of the MinION flowcell as pore activity did not decrease as fast when

compared to the library we ran without fragmentation. Thus, higher fragmentation speeds produced shorter fragments and yielded greater throughput while slower speeds produced longer fragments yielding less throughput (Supplemental Figure 1, Table 3).

This notion is supported by Kubota et al. (2019) who demonstrated a correlation between DNA length and nanopore clogging with clogging increasing exponentially in relation to increasing DNA size. One possible reason for this occurrence could be that longer read lengths correlate with an increased presence of secondary and/or tertiary structures. Nanopores are restricted to the width of one DNA molecule at a time. If these structures are present in DNA reads, they could quickly clog nanopores rendering them useless (Nivala et al., 2013). The combination of libraries prepared with higher and lower centrifugation speeds resulted in a total dataset with enough throughput to provide ample coverage to compensate for the high error rate of ONT sequencing while still yielding long reads to span repetitive or otherwise problematic regions.

Sequencing, Whole Genome Assembly and Hybrid Scaffolding

Canu (Koren et al., 2017), Masurca (Zimin et al., 2013), Flye (Kolmogorov et al.) and wtdbg (Ruan, 2018) assemblers were used to assemble sequence data to ascertain which assembly program would perform best with ONT sequence data. The Flye and wtdbg assemblies were inferior when assembly statistics for number of scaffolds, scaffold N_{50} and assembly size were compared with the Masurca and Canu assemblies (Figure 2). Prior to polishing, the Masurca assembly had a BUSCO score 90% (1,238) compared to the Canu assembly which had a score of 50.5% (694). Initially, this made the Masurca assembly the more attractive option. Three rounds of polishing were conducted utilizing Illumina reads with different combinations of Nanopolish, RACON and Pilon polishing programs. Nanopolish works by creating an index which is used in

detecting misassembles based on sequencing-generated signal levels that correspond to likelihood ratios (Simpson, 2016). Racon corrects raw contigs by using mapping information to construct a partial-order alignment graph (Vaser et al., 2017). Pilon uses evidence from read alignments to identify specific differences from the input genome supported by the sequencing data which it then applies to the draft genome to produce an improved assembly (Walker et al., 2014).

The Nanopolish+Pilon+Pilon polished assembly yielded the highest BUSCO score of C:97.5% (1,340) which demonstrates a high degree of assembly completeness. The Nanopolish+Racon+Racon and Racon+Racon+Racon polishing combinations yielded similar results after two to three rounds of polishing with a slight degree of assembly degradation based on BUSCO scores after the third round of polishing (Figure 3). This result suggests that too much polishing can negatively affect genome assembly. This observation has been noted in other publications that have shown how too many repeat polishing iterations can have a negative impact on the overall quality of an assembly (Miller et al., 2018).

Comparing the BUSCO scores of the Nanopolish+Pilon+Pilon polished assembly to the Masurca assembly made it easier to choose the Canu assembly moving forward despite nominal differences in scaffold number and N_{50}. Our decision was only reinforced when assembly size was considered as the Canu assembly has the closest assembly size compared to the actual genome size of *A. hortensis.* The decreased assembly size generated from the Masurca assembly could potentially reflect collapsed repeats. This notion was supported by Kolmogorov et al. (2018) who demonstrated the difficulty the MaSuRCA assembler has in assembling telomeric and centromeric chromosome regions. To avoid this, we chose to move forward with the Canu

assembler based on its demonstrated abilities in producing assemblies with high contig continuity (Lu et al., 2016).

The genome of orach is highly repetitive with 68% of the sequence containing repetitive motifs. By comparison, the genome of quinoa is 64.5% repetitive (Jarvis et al., 2017). Genomes that contain substantial repetition can be difficult to correctly assemble. To overcome this challenge, chromosome-contact maps were used for genome scaffolding using Hi-C technology which significantly decreased the number of scaffolds and produced nine chromosome-sized scaffolds, reflecting the actual chromosome number of orach. Additionally, Hi-C was able to generate a more accurate overall assembly because of the technology's ability to leverage the spatial orientation of the chromatin; something that is not possible with the other sequencing technologies that were used. This data complements the Illumina and Nanopore data by more accurately recreating the order and orientation of the DNA sequence.

Phylogeny, Synteny and Comparative Genomics

If *A. hortensis* is more closely related to *C. quinoa* than *A. hypochondriacus* (Kadereit et al., 2010; Hernández-Ledesma et al., 2015) and has the same base chromosome number as *C. quinoa*, then we might expect that the chromosomal rearrangements that altered the chromosome number in *A. hypochondriacus* would be absent in *A. hortensis*, as they are in *C. quinoa*. From the phylogeny in Figure 7, we see that *A. hortensis* is most closely related to *C. quinoa.* Based on this information, we would expect the synteny plots in Figure 6 to reflect this pattern as well with fewer rearrangements between *A. hortensis* and *C. quinoa* compared to *A. hortensis* and *B.* vulgaris and *A. hypochondriacus.* Visually, we observe that this is supported as there are indeed fewer rearrangements and inversions between *A. hortensis* and *C. quinoa* with the majority

appearing in the centromeric regions. We do not see the same pattern with *A. hypochondriacus* however, with portions of several different chromosomes aligning to *A. hortensis.* For example, *A. hortensis* chromosome 1 contains components of *A. hypochondriacus* chromosomes 1-4 and 10-12. This demonstrates how *A. hypochondriacus* is more distantly related to *A. hortensis* than *C. quinoa* and *B. vulgaris.*

If *A. hortensis* is more closely related to *C. quinoa* than *A. hypochondriacus* and has the same base chromosome number as *C. quinoa*, then we might expect that the chromosomal rearrangements that altered the chromosome number in *A. hypochondriacus* would be absent in *A. hortensis*, as they are in *C. quinoa*. By visual inspection of the synteny blocks, we see this pattern is also consistent, further confirming the relationships seen in Figure 7. The circular synteny plot in Figure 8 revealed homeologous gene pairs between orach and beet for chromosomes Ah01-Bv03, Ah02-Bv02, Ah03-Bv01/09, Ah04-Bv05, Ah05-Bv06, Ah06-Bv01/04, Ah07-Bv06, Ah08-Bv07, and Ah09-Bv04/09. There is a high degree of synteny between these two genomes These plots which could provide future insight into large-scale rearrangements which have led to chromosome evolution in the *Amaranthaceae-Chenopodiaceae*.

Genomic Features

The nine chromosome pairs in garden orach are metacentric to slightly submetacentric (Figure 1). Due to the difficulty in assembling highly conserved and repetitive sequence regions within telomeres, the identification of 13 of the possible 18 telomeric ends is indicative of a highly complete chromosome-scale genome assembly (Figure 9). We acknowledge that the unexpected location of telomeric sequences in the subtelomeric region of one of the arms of

chromosome 5 could reflect a mis-assembly. This could also be the product of an inversion event or some other chromosomal rearrangement. Tek and Jiang et al. demonstrated that major paracentric inversions can occur that result in telomere-specific tandem repeats being present in abnormal locations in plant chromosomes (Tek and Jiang, 2004). This occurrence could explain why there is a peak for corresponding to tandem repeat telomere sequence appearing outside of the traditional telomeric regions on chromosome 5. Upon investigation of the zoomed-in portion of chromosome 5 in Figure 4B, there is no indication of an assembly problem with the Hi-C data. Additionally, there is no evidence of any significant rearrangement events between homologous chromosomes *A. hortensis* 5 and *B. vulgaris* 8 in Figure 8. This indicates that there has not been a mis-assembly and that a potential rearrangement may have occurred in a common ancestor that has since been conserved by both species.

Resequencing

The analysis of the consensus phylogeny of the orach diversity panel shows three clusters among the 21 accessions (Figure 10). These clusters correspond according to location with the leftmost grouping containing accessions from North America, the middle grouping containing a mix of European and American accessions and the rightmost grouping consisting of solely European accessions. The genetic similarity among accessions seen in the middle grouping suggests that several accessions have been transported between North America and Eurasia over the past several hundred years. This corresponds to the literature as it is commonly believed that several accessions were brought to the Americas in colonial times (Ruas, 2012). A future diversity panel consisting of accessions from other continents would aid in creating a more

complete story of location-specific variation seen in orach including further insight into the origin and spread of the species.

CONCLUSIONS

The *A. hortensis* genome assembly described here is the first reported reference assembly for this species. The final assembly was composed of nine scaffolds, with a N50 of 98.9 Mb. Pseudo-chromosome scaffold sizes were achieved with the incorporation of Hi-C data. The analysis of the genome assembly demonstrates that 68% of the sequence is comprised of repetitive DNA. The BUSCO analysis of the annotation of this assembly demonstrates a high level of completeness, as 96.8% of conserved orthologs were present and complete. The annotation successfully identified 31,010 gene models and 2,555 tRNA genes. When compared with close relatives such as quinoa and beet, strong syntenic patterns contribute to the quality and completeness of the assembly. As this is the first attempt to generate genomic data for this species and genus, this assembly, transcriptome and resequencing information will serve as important resources for the identification of salt and heat-resistant genes as well as other genes and biochemical pathways of interest in agriculture. Additionally, these resources provide an important foundation that contributes to a deeper understanding of orach which may help initiate and accelerate breeding strategies to improve the potential genetic improvement of this fascinating crop.

LITERATURE CITED

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.

Claudete Ruas, Paulo Ruas, Howard Stutz, D.F. (2012). Cytogenetic studies in the genus Atriplex (Chenopodiaceae). Caryologia: 129–145.

Consortium, T.T.G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635–641.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics **34**: 2666–2669.

Dohm, J.C. et al. (2014). The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature **505**: 546–549.

Dohm, J.C., Lange, C., Holtgräwe, D., Sörensen, T.R., Borchardt, D., Schulz, B., Lehrach, H., Weisshaar, B., and Himmelbauer, H. (2012). Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (Beta vulgaris). Plant J. **70**: 528–540.

Flores, H. and Davis, J.I. (2001). A Cladistic Analysis of Atripliceae (Chenopodiaceae) Based on Morphological Data. Torrey Bot. Soc. **128**: 297–319.

Frankton, C. and Bassett, I.J. (1968). The genus Atriplex (Chenopodiaceae) in Canada. I. Three introduced species: A. heterosperma, A. oblongifolia, and A. hortensis. Can. J. Bot. **46**: 1309–1313.

Fuentes-Bazan, S., Uotila, P., and Borsch, T. (2012). A novel phylogeny-based generic classification for Chenopodium sensu lato, and a tribal rearrangement of Chenopodioideae (Chenopodiaceae). Willdenowia **42**: 5–24.

Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P., and Firoozabady, E. (1983). Rapid flow cytometric analysis of the cell cycle in intact plant tissues. (American Association for the Advancement of Science).

Grašič, M., Budak, V., Klančnik, K., and Gaberščik, A. (2017). Optical properties of halophyte leaves are affected by the presence of salt on the leaf surface. Biologia (Bratisl). **72**.

Hansen Mangelson, H., Maughan, P.J., Jellen, E., Jarvis, D.E., and Geary, B. (2019). The Genome of Cañahua: an Emerging Andean Super Grain. Brigham Young Univ.

Harvey, J.H. (1984). Vegetables in the Middle Ages - Garden History (The Gardens Trust).

Hernández-Ledesma, P. et al. (2015). A taxonomic backbone for the global synthesis of species diversity in the angiosperm order Caryophyllales. Willdenowia **45**: 281.

Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P., and Becker, T. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics **25**: 3275–3281.

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

Hubley, R., Smit, A., and Green, P. (2018). RepeatMasker.: 1.

Jarvis, D.E. et al. (2017). The genome of Chenopodium quinoa. Nature **542**: 307.

Kadereit, G., Mavrodiev, E. V., Zacharias, E.H., and Sukhorukov, A.P. (2010). Molecular phylogeny of Atripliceae (Chenopodioideae, Chenopodiaceae): Implications for systematics, biogeography, flower and fruit evolution, and the origin of C4 photosynthesis. Am. J. Bot. **97**: 1664–1687.

Kahn, M.A. and Ungar, I.A. (1984). The Effect of Salinity and Temperature on the Germination of Polymorphic Seeds and Growth of Atriplex triangularis Willd. Am. J. Bot. **71**: 481–489.

Karimi, S.H. and Ungar, I.A. (1989). Development of Epidermal Salt Hairs in Atriplex triangularis Willd. in Response to Salinity, Light Intensity, and Aeration. Source Bot. Gaz. **150**: 68–71.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**: 357–60.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2018). Assembly of Long Error-Prone Reads Using Repeat Graphs. bioRxiv.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. Assembly of Long Error-Prone Reads Using Repeat Graphs.

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

Kubitzki, K., Rohwer, J., and Bittrich, V. (1993). Flowering Plants · Dicotyledons: Magnoliid, Hamamelid and Caryophyllid Families 2nd ed. (Springer, Berlin, Heidelberg: Berlin, Heidelberg).

Kubota, T., Lloyd, K., Sakashita, N., Minato, S., Ishida, K., and Mitsui, T. (2019). Clog and Release, and Reverse Motions of DNA in a Nanopore. Polymers (Basel). **11**.

Lanfear, R., Schalamun, M., Kainer, D., Wang, W., and Schwessinger, B. (2018). MinIONQC: fast and simple quality control for MinION sequencing data. Bioinformatics **35**: 523–525.

Lee, T.-H., Guo, H., Wang, X., Kim, C., and Paterson, A.H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics **15**: 162.

Li, H. (2013). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**: 3094–3100.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., and Telling, A. (2009). Comprehensive mapping of long-range interactions revelas folding principles of the human genome. Science (80-.). **326**: 289–293.

Lightfoot, D.J., Jarvis, D.E., Ramaraj, T., Lee, R., Jellen, E.N., and Maughan, P.J. (2017). Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. BMC Biol. **15**: 74.

Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods **12**: 733–735.

LoPresti, E.F. (2014). Chenopod salt bladders deter insect herbivores. Oecologia **174**: 921–930.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics. Proteomics Bioinformatics **14**: 265–279.

Mcarthur, E.D., Stevens, R., and Blauer, A.C. (1983). Management Growth Performance Comparisons among 18 Accessions of Fourwing Saltbush [Atriplex canescens] at Two Sites in Central Utah. Source J. Range Manag. **36**: 78–81.

Miller, D.E., Staber, C., Zeitlinger, J., and Hawley, R.S. (2018). Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3 (Bethesda). **8**: 3131–3141.

Moll, K.M., Zhou, P., Ramaraj, T., Fajardo, D., Devitt, N.P., Sadowsky, M.J., Stupar, R.M., Tiffin, P., Miller, J.R., Young, N.D., Silverstein, K.A.T., and Mudge, J. (2017). Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. BMC Genomics **18**: 578.

Nivala, J., Marks, D.B., and Akeson, M. (2013). Unfoldase-mediated protein translocation through an α-hemolysin nanopore. Nat. Biotechnol. **31**: 247–50.

Oxford Nanopore Technologies (2018). High molecular weight gDNA extraction from plant leaves. community.nanoporetech.com: 1–9.

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

Ranhotra, G.S., Gelroth, J.A., Glaser, B.K., Lorenz, K.J., and Johnson, D.L. (1992). Composition and Protein Nutritional Quality of Quinoa. Cereal Chem. **70**: 303–305.

Richards, E.J. and Ausubel, F.M. (1988). Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. Cell **53**: 127–136.

Rinchen, T. and Singh, N. (2015). Exploring nutritional potential of Atriplex hortensis. Indian Hort. **60**: 16–17.

Rinchen, T., Singh, N., Maurya, S.B., Soni, V., Phour, M., and Kumar, B. (2017). Morphological characterization of indigenous vegetable (Atriplex hortensis L.) from trans-Himalayan region of Ladakh (Jammu and Kashmir), India. AJCS **11**: 258–263.

Rohrer, W.L., Porter, D.M., Shirley, B.W., and Turner, B.J. (1997). A biosystematic study of the rare plant Paronychia virginica Spreng. (Caryophyllaceae) employing morphometric and allozyme analyses.

Ruan, J. (2018). wtdbg. github.com: 1.

Sai Kachout, S., Ben Mansoura, A., Jaffel Hamza, K., Leclerc, J.C., Rejeb, M.N., and Ouerghi, Z. (2011). Leaf-water relations and ion concentrations of the halophyte Atriplex hortensis in response to salinity and water stress. Acta Physiol. Plant. **33**: 335–342.

Sati, S. and Cavalli, G. (2017). Chromosome conformation capture technologies and their impact in understanding genome function. Chromosoma **126**: 33–44.

Sievers, F. and Higgins, D.G. (2014). Clustal Omega. Curr. Protoc. Bioinforma. **48**: 3.13.1-3.13.16.

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

Simon, R.D., Abeliovich, A., and Belkin, S. (1994). A novel terrestrial halophilic environment: The phylloplane of Atriplex halimus, a salt-excreting plant. FEMS Microbiol. Ecol. **14**: 99–109.

Simpson, J.T. (2016). Supporting R9 data in nanopolish. Simspon Lab Blog.

Smit, A. and Hubley, R. (2008). RepeatModeler Open 1.0.

Stafford, H.A. (1994). Anthocyanins and betalains: evolution of the mutually exclusive pathways. Plant Sci. **101**: 91–98.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. **32**: W309–W312.

Stephens, J.M. (1994). Orach-Atriplex hortensis L. (Gainesville).

Sukhorukov, A.P., Nilova, M. V., Krinitsina, A.A., Zaika, M.A., Erst, A.S., and Shepherd, K.A. (2018). Molecular phylogenetic data and seed coat anatomy resolve the generic position of some critical Chenopodioideae (Chenopodiaceae – Amaranthaceae) with reduced perianth segments. PhytoKeys **109**: 103–128.

Tanaka, Y., Sasaki, N., and Ohmiya, A. (2008). Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. Plant J. **54**: 733–749.

Tek, A. and Jiang, J. (2004). The centromeric regions of potato chromosomes contain megabase-sized tandem arrays of telomere-similar sequence. Chromosoma **113**: 77–83.

Todd, J.J. and Vodkin, L.O. (1996). Duplications That Suppress and Deletions That Restore Expression from a Chalcone Synthase Multigene Family (American Society of Plant Physiologists).

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

Venable, D.L. and Levin, D.A. (1985). Ecology of Achene Dimorphism in Heterotheca Latifolia: I. Achene Structure, Germination and Dispersal. Source J. Ecol. J. Ecol. **73**: 133–145.

Vickerman, D.B., Shannon, M.C., Bañuelos, G.S., Grieve, C.M., and Trumble, J.T. (2002). Evaluation of Atriplex lines for selenium accumulation, salt tolerance and suitability for a key agricultural insect pest. Environ. Pollut.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One **9**: e112963.

Welsh, S.L. and Crompton, C. (1995). Names and types in perennial Atriplex Linnaeus (Chenopodiaceae) in North America selectively exclusive of Mexico. Gt. Basin Nat. **55**: 322–334.

Wertis, B.A., Ungar, I.A., Triangularis, A., Barbara, W., and Wertis, A. (1986). Seed Demography and Seedling Survival in a Population of Atriplex triangularis Willd Seed Demography and Seedling Survival in a Population of. Source Am. Midl. Nat. **116**: 152–162.

Wick, R. (2017). Porechop.: 1.

Wright, K.H., Pike, O.A., Fairbanks, D.J., and Huber, C.S. (2002). Composition of *Atriplex hortensis*, Sweet and Bitter Chenopodium quinoa Seeds. J. Food Sci. **67**: 1383–1385.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics **29**: 2669–2677.

FIGURES

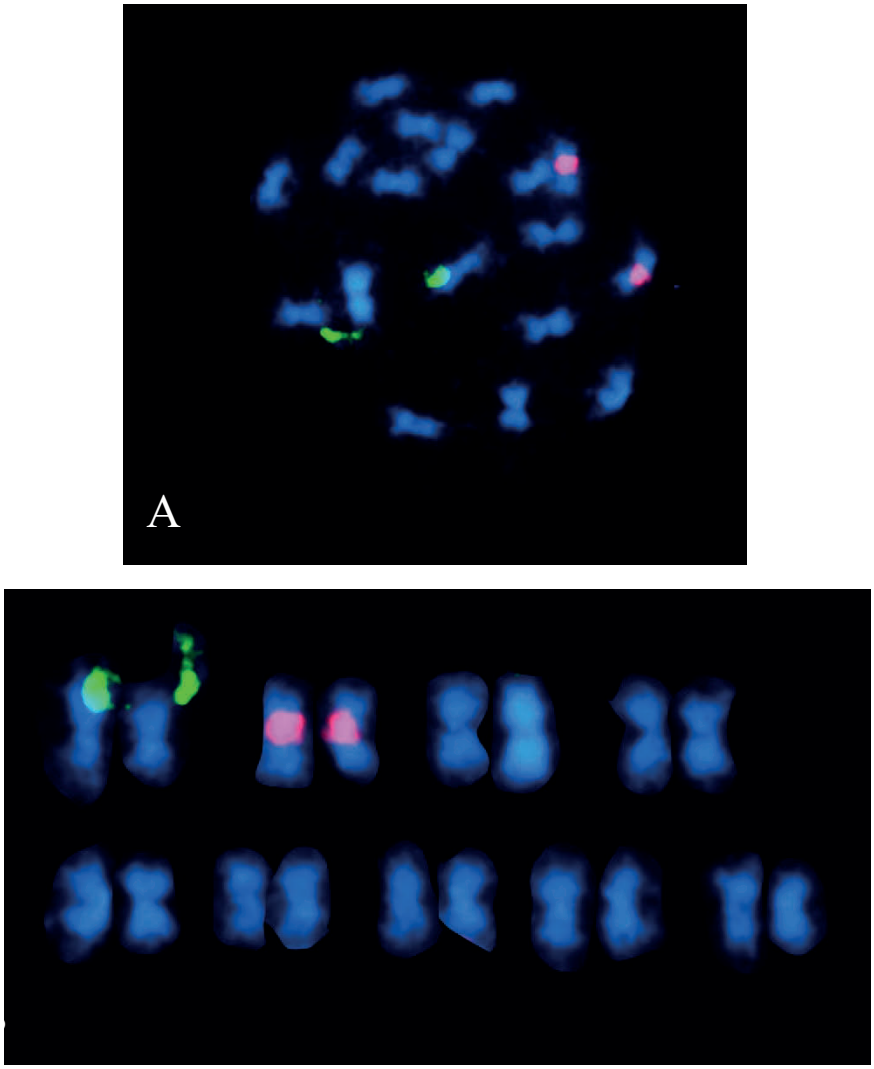


Figure 1. *Atriplex hortensis* chromosome pairs. Nine metacentric chromosome pairs are visible. A) Fluorescent in-situ hybridization (FISH) using NOR-35S (green) and 5S (red) labeled rDNA probes on mitotic chromosome preparations of *Atriplex hortensis* cv. "Triple Purple". B) Chromosomes from (A) arranged as a karyotype. Note the metacentric to submetacentric centromere positions on all nine chromosome pairs.

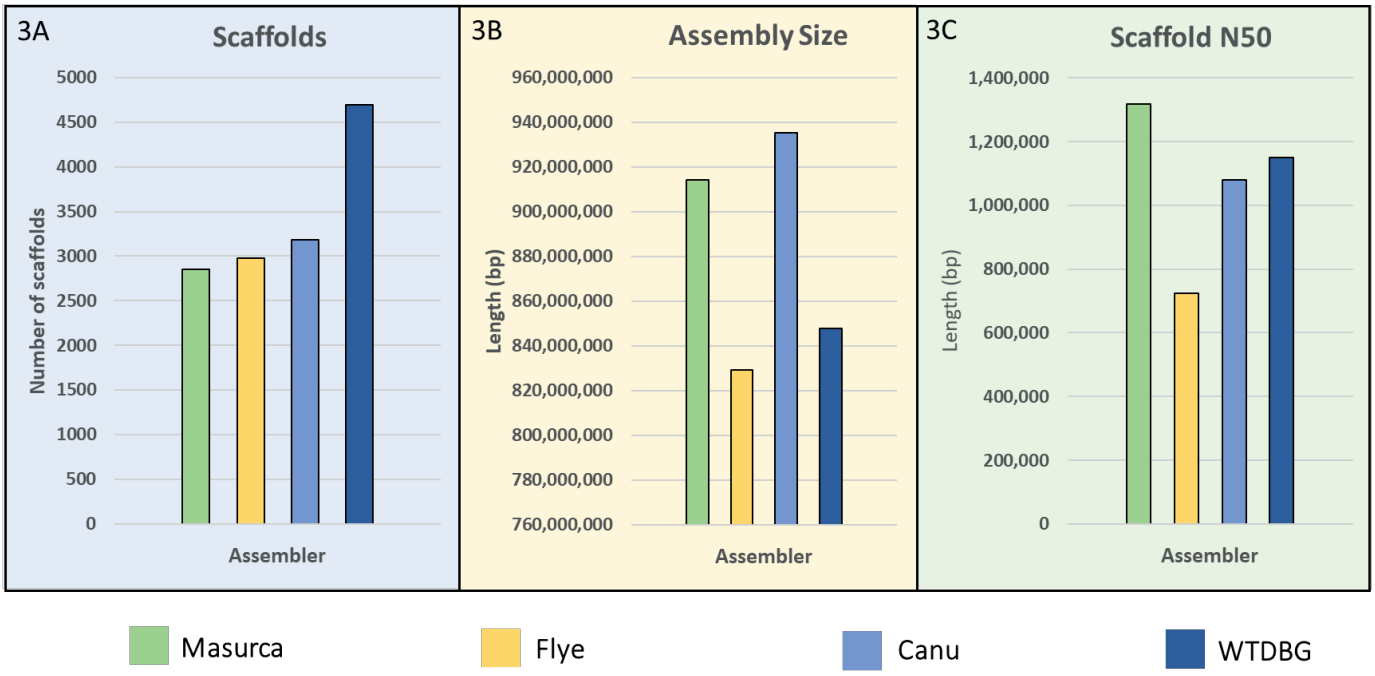


Figure 2. Comparison of genome assembly methods for Oxford Nanopore reads. Assembly metrics including number of scaffolds, scaffold size and scaffold N50 produced from Masurca, Flye, Canu and wtdbg assemblies were compared.

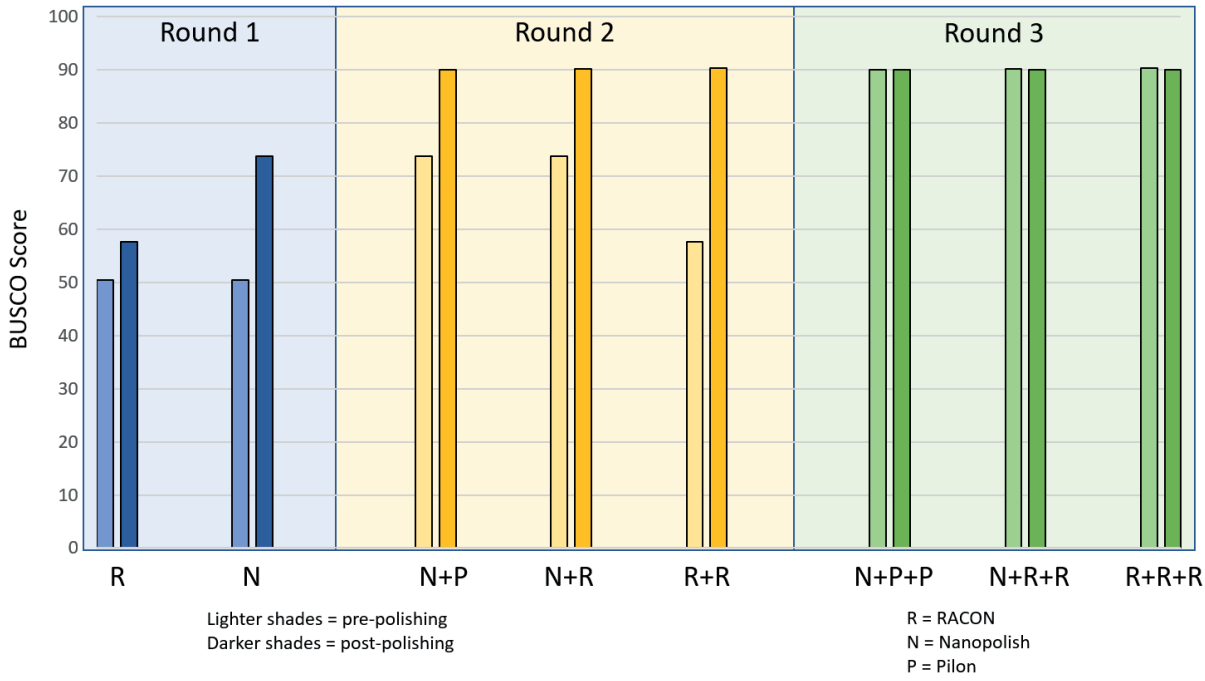
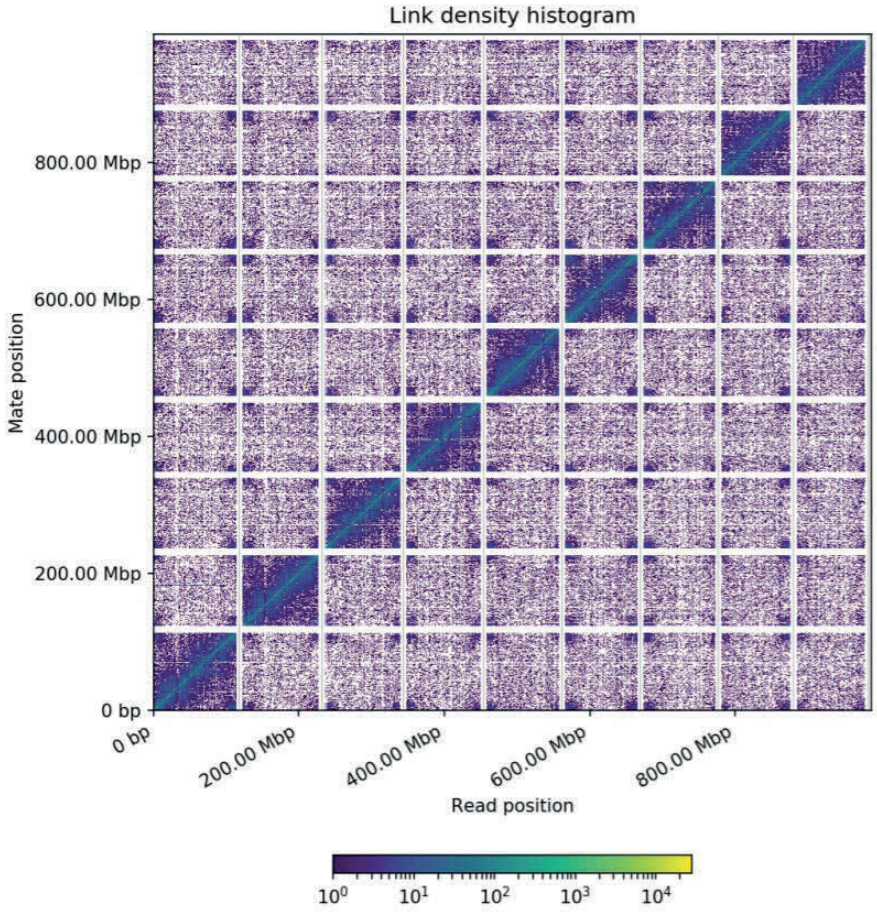


Figure 3. Assembly Polishing. A comparison of polishing methods for assembly improvement over three rounds of polishing using three different polishing programs: RACON, Nanopolish and Pilon. The Nanopolish + Pilon + Pilon combination yielded the assembly with the highest BUSCO score of 97.5% (1,340).

A



B

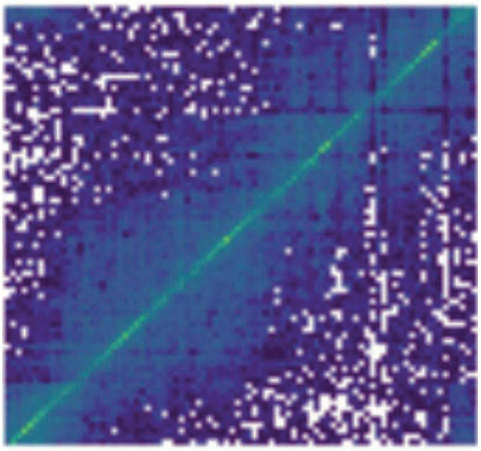


Figure 4. Hi-C link-density histogram. A) The *x* and *y* axes give the mapping positions of the first and second read in the read-pair, respectively, grouped into bins. The color of each square gives the number of read-pairs within that bin. White vertical and gray horizontal lines have been added to show the borders between scaffolds. Scaffolds less than 1 Mb are excluded. B) A zoomed-in image of chromosome five demonstrating that there is no mis-assembly.

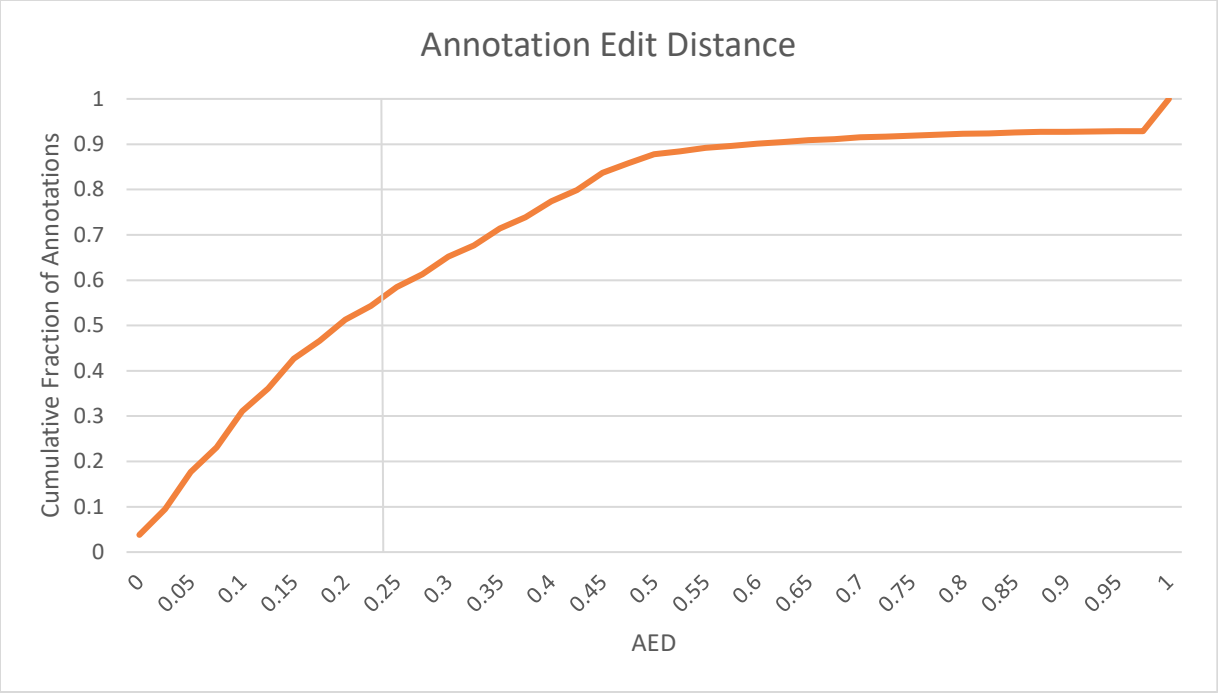


Figure 5. Annotation Edit Distance for MAKER Annotation. Annotation Edit Distance (AED) is used to measure the quality of a genome annotation. This is calculated by combining annotation values corresponding to specificity and sensitivity. AED values of 0.5 and below are reasonable annotations and values of 0.25 and below are high quality annotations (Holt and Yandell, 2011). The AED score of 0.5 coupled with the BUSCO assessment provides evidence for a reasonable genome annotation. The majority (58.5%) of genes identified in the annotation had AED values below 0.25.

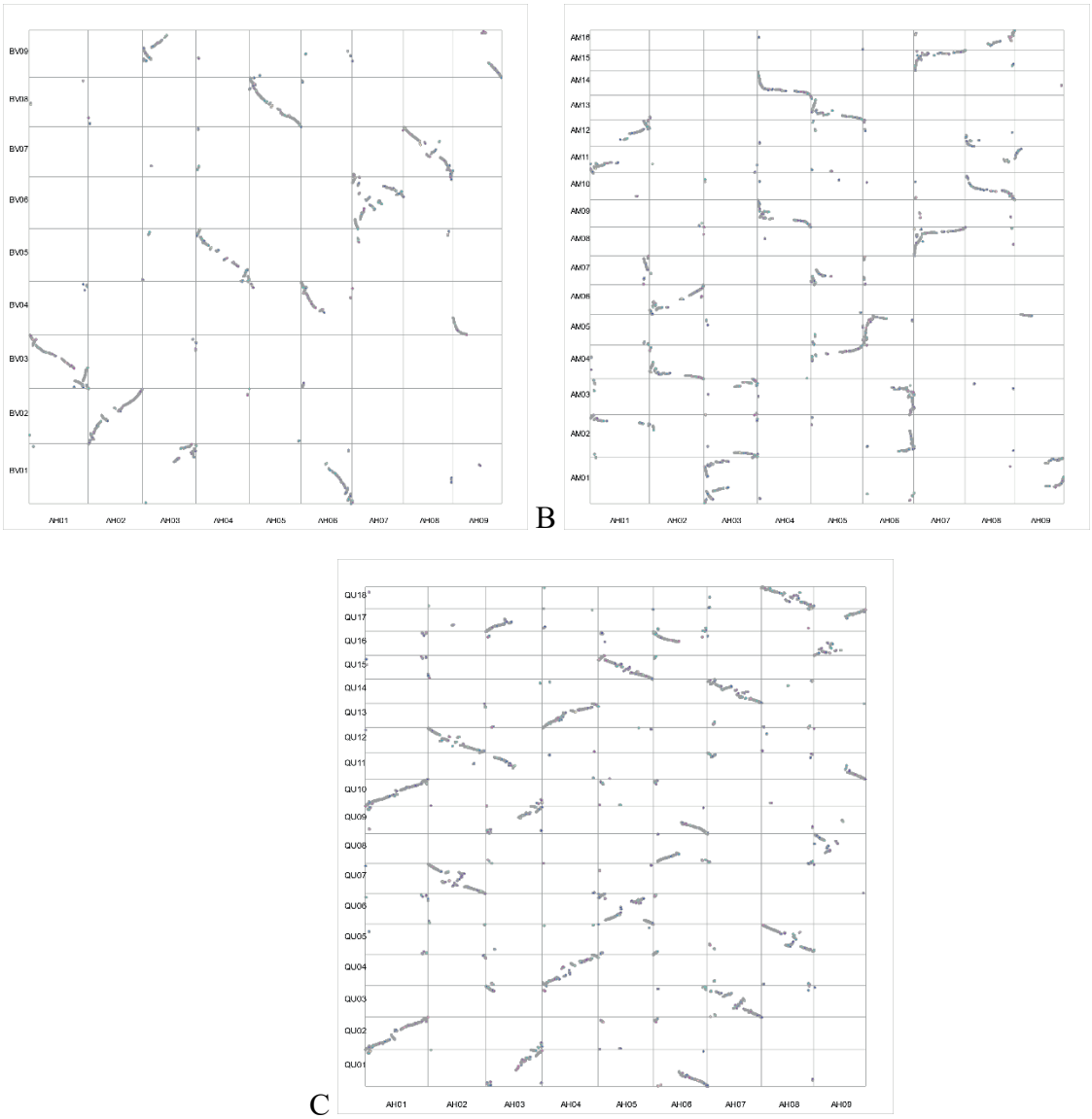


Figure 6. Synteny between related species and orach. Synteny plots showing syntenic relationships between orach (*x*-axis) and A) *B.* vulgaris, B) *A. hypochondriacus* and C) *C. quinoa* (*y*-axis) homologs.

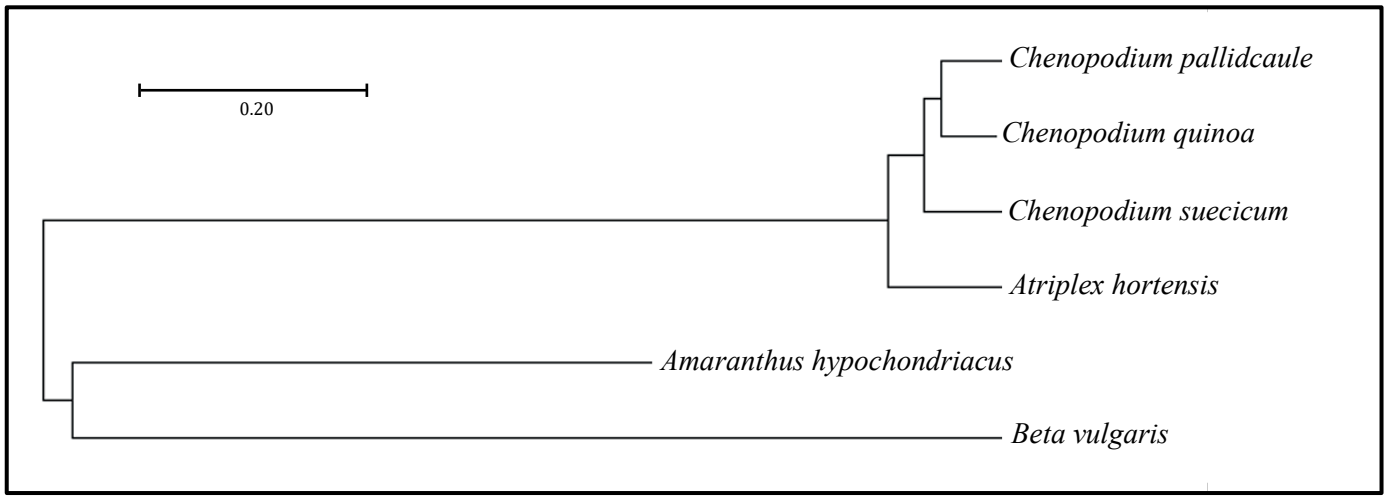


Figure 7. Relationships among *Amaranthaceae*-*Chenopodiaceae* species. COGs were used to generate the phylogeny using the multiple sequence alignment function from Clustal Omega (Sievers and Higgins, 2014) with the PHYLIP output format selected. The scale bar represents residue substitution per site.

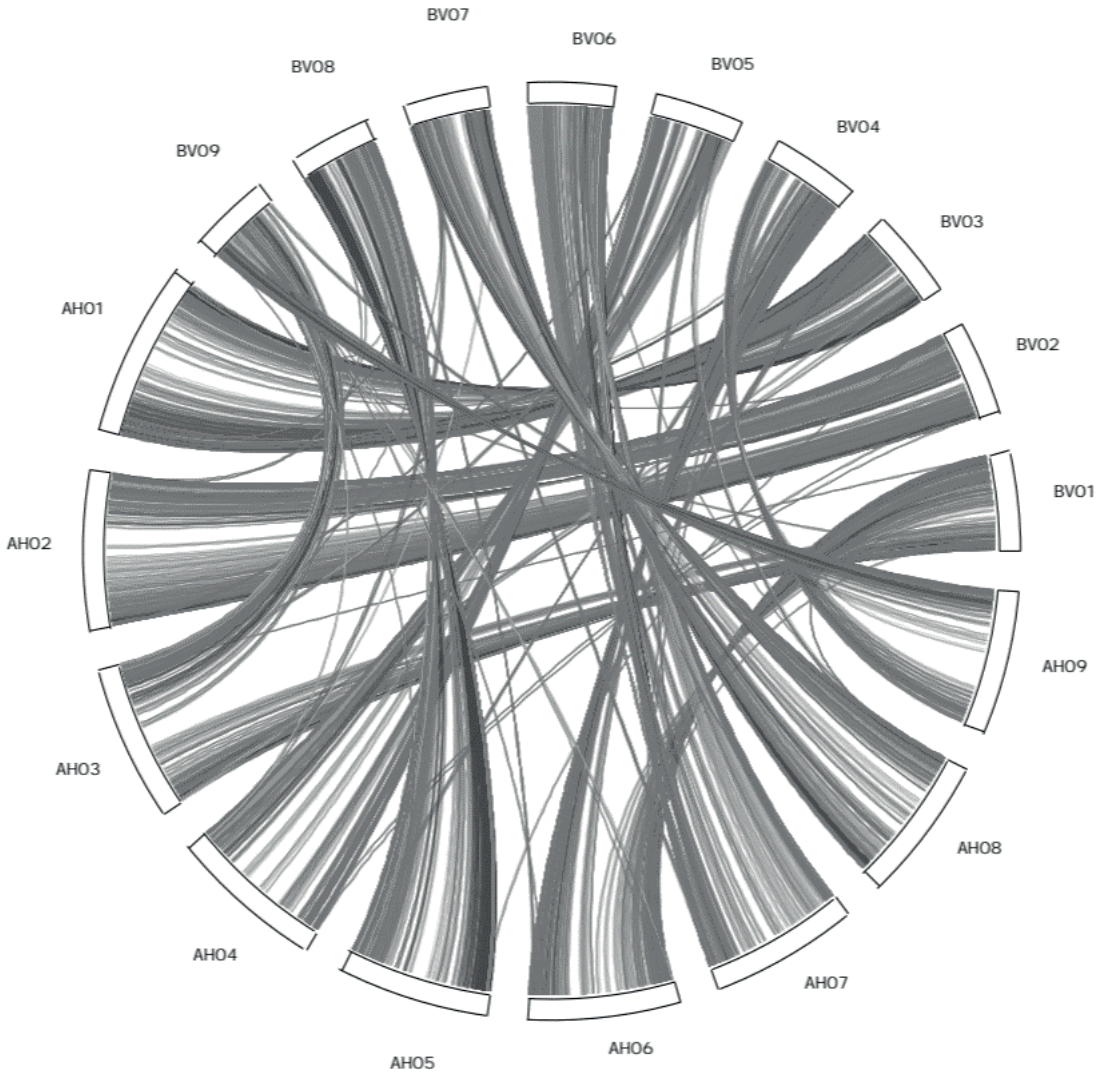


Figure 8. Circular synteny plot illustrating chromosomal synteny between orach (Ah) and beet (Bv, *Beta vulgaris*) pseudochromosomes. Note synteny for Ah01-Bv05, Ah02-Bv04, Ah03-Bv06/09, Ah04-Bv08, Ah05-Bv03, Ah06-Bv01/06, Ah07-Bv07, Ah08-Bv01/09, and Ah09-Bv02.

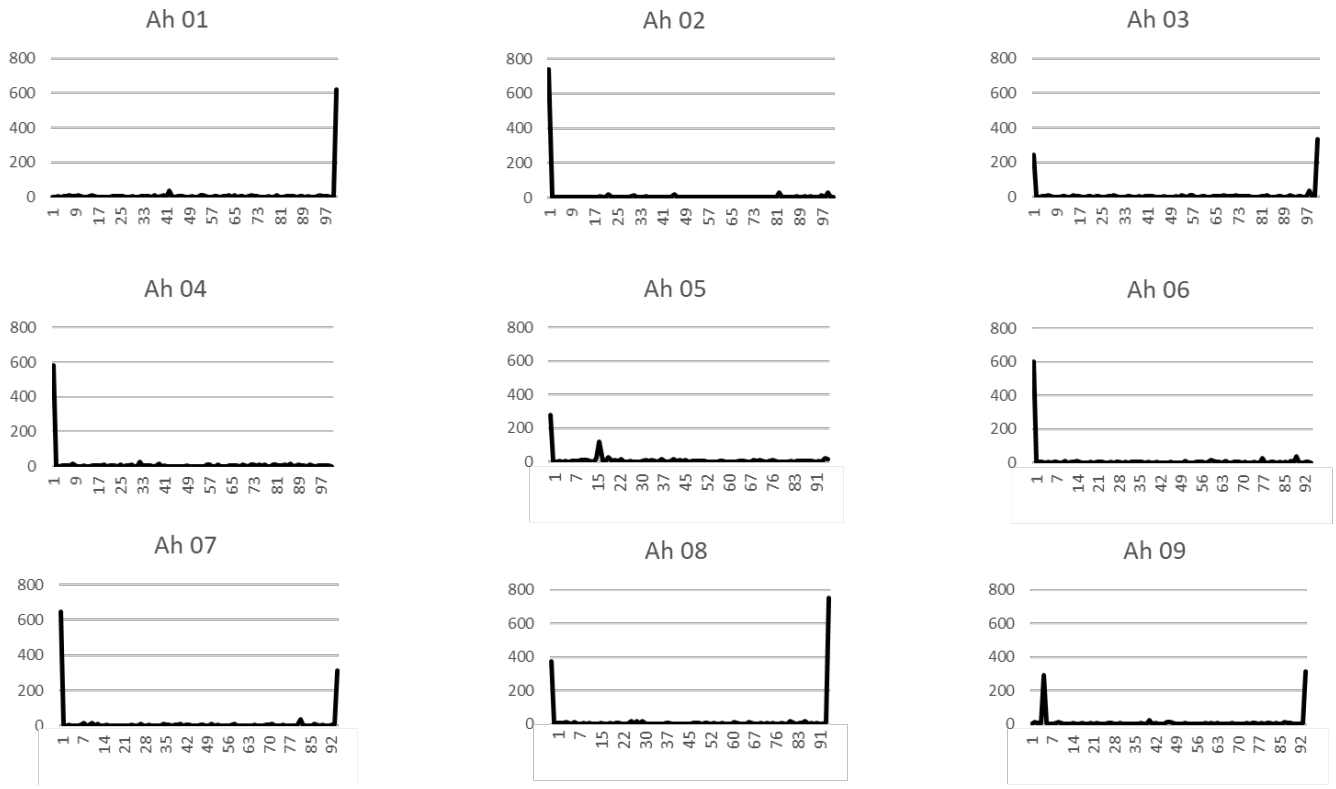


Figure 9. Telomere positioning for *A. hortensis* chromosomes. A conserved telomere repeat sequence in plants was used to locate telomere position in pseudochromosomal scaffolds. A BLAST search of this sequence against the nine chromosome-sized scaffolds identified tandemly repeated telomeric sequences on every chromosome with 13 repetitive regions identified in total. The *x* axis represents each bin consisting of one million bases. The *y* axis represents the number of sequence repeats in each bin.

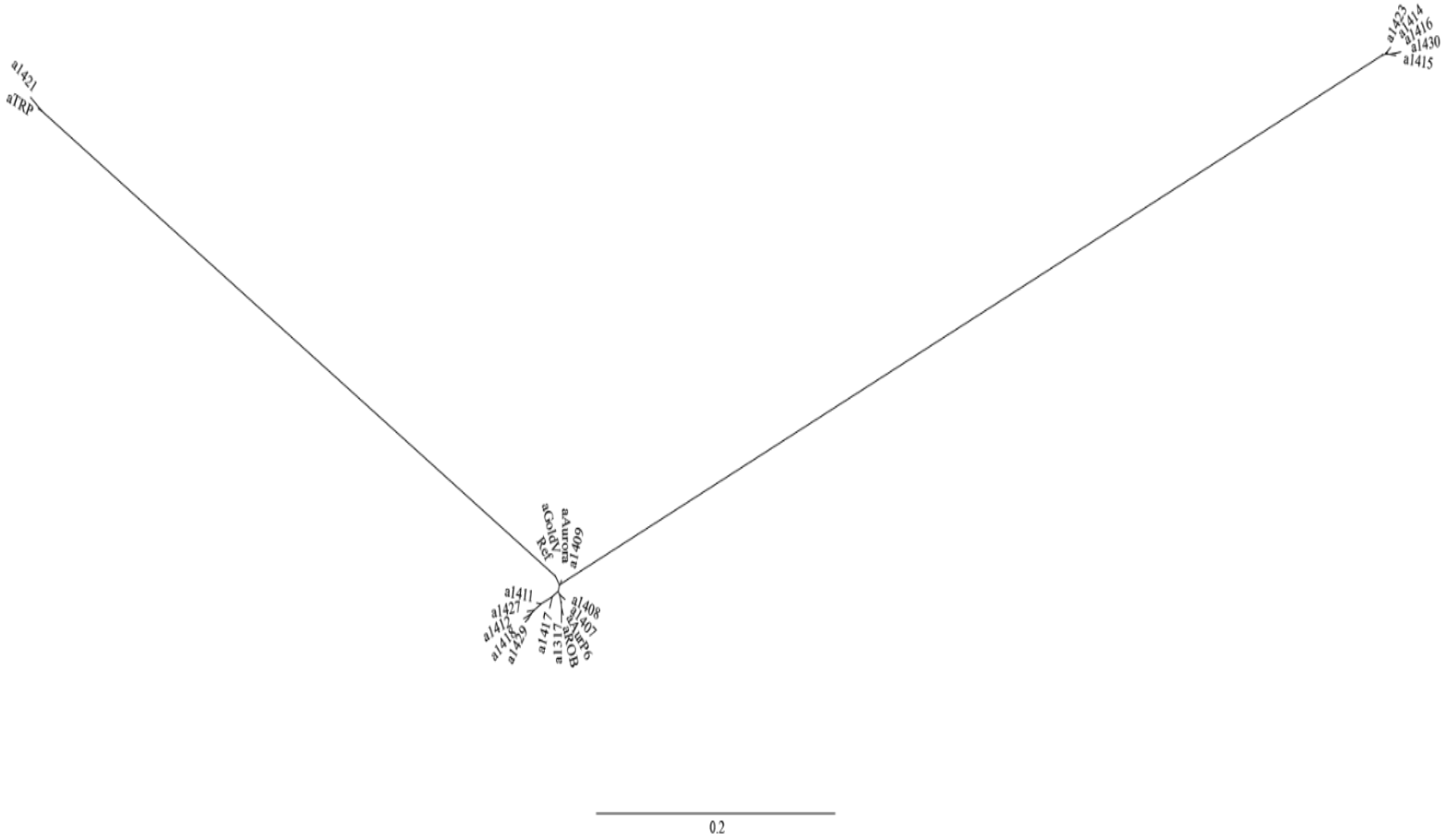


Figure 10. Diversity Panel. This unrooted tree was designed using 1,708 SNPs filtered to remove SNPs with > 10% missing data, minor allele frequency < 5%, and LD < 40%. Accession numbers in the panel correspond with those found in Table 1.

TABLES

Table 1. Passport and ecotype information for plant materials used for the resequencing panel. Accessions for *A. hortensis* were gathered throughout Europe and North America. N/A indicates data. Elevation is reported in meters above sea level.

ID	Accession	Source	Collection Location	Latitude/Longitude	Elevation
BYU 1317	*A. hortensis*	Personal Collection	Park City, UT	N/A	N/A
BYU 1402	Red Orach	Baker Creek Heirloom Seeds	Mansfield, MO	N/A	N/A
BYU 1407	PI 310383	USDA; 2046	Former Soviet Union	N/A	N/A
BYU 1408	PI 323313	USDA	Poland	N/A	N/A
BYU 1409	PI 345962	USDA; 1042	Norway	N/A	N/A
BYU 1411	PI 357340	USDA	Zolta, Former Serbia/Montenegro	41.91667000/22.41667000	350
BYU 1412	PI 357342	USDA	Zolta Prilepska, Former Serbia/Montenegro	41.34640000, 21.55440000	660
BYU 1414	PI 357344	USDA	Lokalna Zolta, Former Serbia/Montenegro	41.81200000, 21.99470000	250
BYU 1415	PI 357346	USDA	Gradinarska, Former Serbia/Montenegro	41.57920000, 21.57190000	460
BYU 1416	PI 357347	USDA	Debarska, Former Serbia/Montenegro	41.52500000, 20.52750000	680
BYU 1417	PI 370353	USDA	Lokalna, Former Serbia/Montenegro	41.89890000, 21.40810000	400
BYU 1418	PI 370354	USDA	Mestna, Former Serbia/Montenegro	41.94140000, 21.41280000	510
BYU 1421	PI 372512	USDA	Alberta, Canada	N/A	N/A
BYU 1423	PI 379088	USDA; 2261	Former Serbia/Montenegro	41.84890000, 21.82030000	500
BYU 1427	PI 379093	USDA; 2475	Former Serbia/Montenegro	41.38250000, 22.28750000	310
BYU 1429	PI 379095	USDA	Skopska, Former Serbia/Montenegro	42.00000000, 21.43330000	240
BYU 1430	PI 420154	USDA; 218	France	N/A	N/A
BYU 1432	Golden	Wild Garden Seed Co.	Philomath, OR	N/A	N/A
BYU 1433	Triple Purple	Wild Garden Seed Co.	Philomath, OR	N/A	N/A
BYU 1434	P1	Wild Garden Seed Co.	Philomath, OR	N/A	N/A
BYU 1434	P6	Wild Garden Seed Co.	Philomath, OR	N/A	N/A

Table 2. Flow Cytometry results. Young *A. hortensis* cv. "Golden" leaf tissue was used. A C-value of 2.4 picograms yielded a genome size estimate of 1.171 G.

Genus	Species	Sample	Tech_reps	pg	Std. Dev	Notes
Atriplex	hortensis	S1	3	2.39	0.0342	
Atriplex	hortensis	S2	3	2.39	0.0184	Strong 4C peak
Atriplex	hortensis	S3	3	2.41	0.0061	
		Average:		2.40		

Table 3. Oxford Nanopore library preparation and sequencing statistics. Non-fragmentation as well as fragmentation techniques were used in sample preparation.

Sample	Fragmentation	Total Gigs	Totals Reads	N_{50}	Mean Length	Median Length	Max Length	Mean q	Median q
1	No Fragmentation	1.26	55,551	40,434	22,617	15,607	194,834	9.1	9.4
2	Zymo	5.61	567,514	23,595	9,877	4,522	153,389	9.4	9.6
3	Zymo	2.24	133,660	33,394	16,770	9,857	199,575	9.2	9.5
4	Covaris, 4,200 RPM	13.04	1,005,270	15,878	11,760	11,017	181,817	8.3	8.8
5	Covaris, 3,800 RPM	10.08	854,994	15,277	11,788	11,104	133,274	9.1	9.3
6	Covaris, 3,800 RPM	6.94	501,526	20,681	13,760	12,431	164,726	8.9	9.2
7	Covaris, 4,200 RPM	8.86	617,385	19,276	14,343	12,932	231,794	9.1	9.4
8	Covaris, 4,200 RPM	10.72	1,221,530	12,664	8,778	8,300	149,453	9.1	9.3
9	Covaris, 4,000 RPM	6.64	568,017	17,580	11,686	11,115	128,743	8.9	9.5
Avg/Total	-	65.4	5,525,447	22,087	13,487	10,765	170,845	9	9

Table 4. Dovetail chromatin proximity-based assembly statistics. A) Comparative assembly statistics showing improvements genome improvements after Dovetail HiRise Assembly B) Other statistics outlining breaks, joins and gaps. C) Summary of orach genes identified in BUSCO odb10 eukaryota gene set. BUSCO statistics show improvement in COGs identified after Dovetail HiRise Assembly.

Comparative Assembly Statistics						
	Input Assembly	**Dovetail HiRise Assembly**				
Longest Scaffold	15,147,297 bp	113,540,706 bp				
Number of scaffolds	2191	1325				
Contig N50	816.58 kb	98.9 Mb				
Number of gaps	1421	2290				
Percent of genome in gaps	0.01%	0.02%				
Other Statistics						
Number of breaks made to input assembly by HiRise	0					
Number of joins made by HiRise	868					
Number of gaps closed after HiRise	0					
Library 1 stats	200M read pairs; 2x150 bp					
BUSCO Statistics						
	Complete	**Single Copy**	**Duplicated**	**Fragmented**	**Missing**	**Total**
Input Assembly	1340	1310	30	8	27	1375
	97.50%	95.30%	2.20%	0.60%	1.90%	
Dovetail HiRise Assembly	1330	1306	24	11	34	1375
	96.70%	95.00%	1.70%	0.80%	2.50%	

Table 5. Summary of repeat element content in the orach genome assembly identified by RepeatMasker relative to the RepBase-derived RepeatMasker libraries. There are 3,183 sequences total. Total length excludes N/X-runs. GC level is 37.06%. Most repeats fragmented by insertions or deletions have been counted as one element. SINE, short interspersed nuclear elements; LINE, long interspersed nuclear elements; LTR, long terminal repeat; RC, Rolling circle.

Repeat Class	Count	Bases Masked	Masked (%)
DNA elements	2,848	419,543	0.04%
CMC- EnSpm	19,647	18,707,200	1.94%
MULE-MuDR	12,625	6,746,714	0.70%
Maverick	229	41,390	0.00%
MuLE-MuDR	4,004	3,329,840	0.35%
PIF-Harbinger	1,334	657,055	0.07%
Sola	545	271,130	0.03%
TcMar-Mogwai	1,391	492,330	0.05%
TcMar-Stowaway	29,143	5,481,810	0.57%
hAT-Ac	9,320	2,590,163	0.27%
hAT-Tag1	1,949	319,829	0.03%
hAT-Tip100	1,478	447,851	0.05%
LINEs	--	--	--
CRE-II	486	275,712	0.03%
Jockey	1,275	305,260	0.03%
L1	7,159	6,706,326	0.70%
L2	11,294	15,454,784	1.60%
Penelope	179	42,364	0.00%
RTE-BovB	5,868	2,014,023	0.21%
LTR	5,141	1,344,555	0.14%
Caulimovirus	625	806,239	0.08%
Copia	52,628	104,589,194	10.85%
DIRS	3,354	1,385,663	0.14%
Gypsy	181,399	320,566,514	33.25%
Pao	7,499	7,885,285	0.82%
RC	--	--	--
Helitron	4,935	2,833,179	0.29%
SINE	274	70,971	0.01%
tRNA	394	46,669	0.00%
Unknown	421,611	155,026,626	16.08%
Total Interspersed	788,644	657,841,458	68.23%
Low complexity	33,619	1,739,656	0.18%
Satellite	3,300	902,683	0.09%
Simple Repeat	197,461	15,870,526	1.65%
rRNA	386	329,627	3.00%
Total	1,023,410	676,683,950	70.18%

Table 6. Comparison of syntenic gene features and gene models in *Amaranthaceae* species based on data generated from RepeatModeler.

	C. quinoa	*B. vulgaris*	*A. hypochondriacus*
Syntenic Features	20,307	9,271	9,438
Syntenic Features with *A. hortensis*	10,922	9,282	8,355
Total Syntenic Features	31,229	18,553	17,793
Gene Models	44,776	27,421	23,847
Gene Models in *A. hortensis*	33,565	33,565	33,565
Total Gene Models between species	78,341	60,986	57,412
Percent of Total Features Conserved	40%	30%	31%

Table 7. Comparison of Gene Synteny between *A. hortensis* and *B. vulgaris* chromosomes.

Atriplex Chromosome	**Syntenic Blocks**	**Total Syntenic Genes**
1	77	1,254
2	74	1,087
3	79	1,302
4	51	911
5	68	1,170
6	73	1,708
7	46	958
8	62	1,059
9	27	272
Total	557	9,721

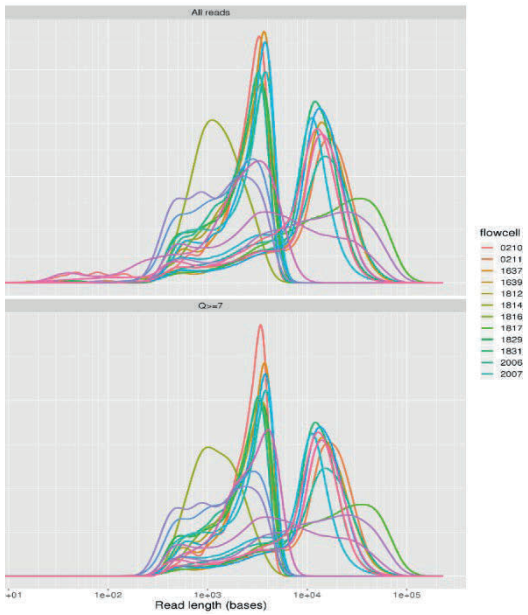
Table 8. Resequencing Panel - SNPs per Chromosome. Chromosomes are ordered based on size. After filtering, 1,708 SNPs were identified and used in making the unrooted tree in Figure 10. Here, the number of SNPs chosen per chromosome are compared to the total number SNPs identified on each chromosome giving a percentage showing each chromosome's SNP contribution.

Orach Chromosome	**Percentage**	**SNPs chosen**
Chromosome 1	0.14	212/148,688
Chromosome 2	0.14	193/140,403
Chromosome 3	0.14	194/143,116
Chromosome 4	0.16	192/119,733
Chromosome 5	0.13	190/141,811
Chromosome 6	0.10	189/186,323
Chromosome 7	0.14	184/128,242
Chromosome 8	0.14	181/129,634
Chromosome 9	0.22	173/79,017

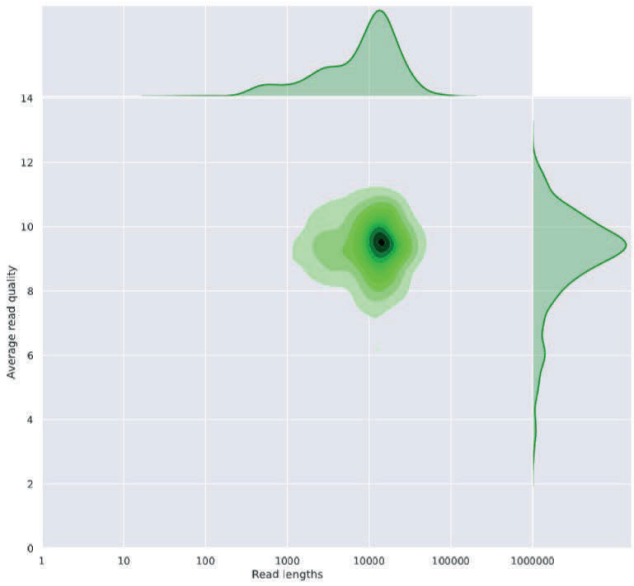
SUPPLEMENTAL MATERIAL

Supplemental Figures

A



B



Supplemental Figure 1. Read Quality vs Length. A) Average read lengths are represented on the x-axis for each individual run. Each run is represented by a unique color with 11 total including one redux and one restart of the flowcell. B) Read length vs read quality is represented for Oxford Nanopore reads.

Supplemental Appendices

Albacore Base-Calling

```
#FULL-PATH-TO-RAW-DATA
I=fast5
#FULL-PATH-TO-SAVE-LOCATION make the Albacore_OUTPUT folder before running.
S=Albacore_OUTPUT
#Kit used during the sequencing run (i.e., SQK-LSK108 or SQK-LSK109)
K=SQK-LSK109
#Flowcell used in the sequencing run (i.e., FLO-MIN106)
F=FLO-MIN106
#Number of threads
T=24
read_fast5_basecaller.py -i ${I} -s ${S} -k ${K} -f ${F} -t ${T} -r --
disable_pings
```

BUSCO

```
module purge
module load conda-pws
module load conda/busco

#Add full path to assembly
assembly=p_19_S_4.ctg.lay.fas

#Add output name
output=wtdgb_p19S4_arrow_long_newbusco

#Add species (rice or arabidopsis)
species=arabidopsis

#Add the mode (genome, protein, transcriptome)
m=genome
```

```
#Add the dataset (plants: embryophyta_odb10, eukaryotes: eukaryota_odb10)

l=/fslhome/pjm43/fsl_groups/fslg_pws_module/software/.conda/envs/busco/datasets/embryophyta_odb10

run_BUSCO.py -i ${assembly} -o ${output}_${species} -l ${l} -m ${m} -c 24 --long -sp ${species} -f
```

BWA

```
#!/bin/sh

BWA_INDEX=path_to_assembly

bwa_scripts=path_to_bwa_scripts_directory

trimmed_reads=path_to_trimmed_reads_directory

alignments=path_to_alignments_directory

T=12 #threads

#Make sure to run script from where the trimmed read files are.

#Make sure you make the bwa_scripts and alignments directories prior to running the shell script

for forward_file in *_1P.fq.gz

do

name=`echo $forward_file | sed 's/_1P.fq.gz//'`

cat > ${bwa_scripts}/${name}.sh <<EOF

#!/bin/bash

#SBATCH -c 12 --mem=64gb --qos=pws -t 72:00:00

module purge

module load conda-pws

module load conda/bwa_0.7.17
```

```
bwa mem -M -t $T $BWA_INDEX ${Trimmed_Reads}/${name}_1P.fq.gz
${Trimmed_Reads}/${name}_2P.fq.gz > ${alignments}/${name}.sam
EOF


sbatch ${bwa_scripts}/${name}.sh
done
```

Canu

```
canu -d canu1_7_atriplex_60 -p canu1_7_atriplex_60 genomeSize=1100m
maxMemory=500g maxThreads=24 corMhapSensitivity=normal corOutCoverage=40 \
merylMemory=500g merylThreads=24 ovsMethod=parallel \
gridOptions="--qos=pws --time=72:00:00"  \
gridOptionsOVS="--mem-per-cpu=64g --time=72:00:00" \
gridOptionsExecutive="--mem-per-cpu=24g --time=72:00:00" \
gridOptionsCORMHAP="--mem-per-cpu=10g --time=72:00:00" \
gridOptionsOBTMHAP="--mem-per-cpu=10g --time=72:00:00" \
gridOptionsUTGMHAP="--mem-per-cpu=10g --time=72:00:00" \
gridOptionsCOROVL="--mem-per-cpu=10g --time=72:00:00" \
gridOptionsOBTOVL="--mem-per-cpu=10g --time=72:00:00" \
gridOptionsUTGOVL="--mem-per-cpu=6g --time=72:00:00" \
gridOptionsRED="--mem-per-cpu=12g --time=72:00:00" \
gridOptionsOEA="--mem-per-cpu=12g --time=72:00:00" \
gridOptionsOVB="--mem-per-cpu=12g --time=71:00:00" \
gridOptionsCNS="--mem-per-cpu=12g --time=70:00:00" \
-nanopore-raw trimmed.q8_l2000.porechop.fastq.gz
```

InterSNP

```
#!/bin/bash


#SBATCH -c 24 --mem=256gb -t 3-0:00:00


module purge
```

```
module load htslib/1.2.1

module load bambam/1.4

interSnp -r path_to_reference -w path_to_hapmap -m 6 -f 0.35 -t 24
~/BWA/alignments/*.bam.sorted.bam > path_to_output.snp
```

MAKER2.0

```
#!/bin/sh


data_dir=/panfs/pan.fsl.byu.edu/scr/grp/fslg_atriplex/annotation/MAKER/Doveta
ilData/complete_reference


scripts_dir=/panfs/pan.fsl.byu.edu/scr/grp/fslg_atriplex/annotation/MAKER/Dov
etailData/complete_reference/scripts


for file in *fasta

do

name=`echo $file | sed 's/.fasta//'`

mkdir ${scripts_dir}/${name}

cat > ${scripts_dir}/${name}/${name}.sh <<EOF

#!/bin/bash


#SBATCH --time=168:00:00   # walltime

#SBATCH --ntasks=8   # number of processor cores (i.e. tasks)

#SBATCH --nodes=1   # number of nodes

#SBATCH --mem-per-cpu=8G   # memory per CPU core


module purge

module load conda-pws

module load conda/maker_v2.31.10


maker -c 8 ${scripts_dir}/${name}/maker_opts.ctl
${scripts_dir}/${name}/maker_bopts.ctl ${scripts_dir}/${name}/maker_exe.ctl

EOF
```

```
cat > ${scripts_dir}/${name}/maker_exe.ctl <<EOF

#-----Location of Executables Used by MAKER/EVALUATOR


makeblastdb=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.3
1.10/bin/makeblastdb #location of NCBI+ makeblastdb executable

blastn=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.10/
bin/blastn #location of NCBI+ blastn executable

blastx=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.10/
bin/blastx #location of NCBI+ blastx executable

tblastx=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.10
/bin/tblastx #location of NCBI+ tblastx executable

formatdb= #location of NCBI formatdb executable

blastall= #location of NCBI blastall executable

xdformat= #location of WUBLAST xdformat executable

blasta= #location of WUBLAST blasta executable


RepeatMasker=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.
31.10/bin/RepeatMasker #location of RepeatMasker executable


exonerate=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.
10/bin/exonerate #location of exonerate executable



#-----Ab-initio Gene Prediction Algorithms

snap=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.10/bi
n/snap #location of snap executable

gmhmme3= #location of eukaryotic genemark executable

gmhmmp= #location of prokaryotic genemark executable


augustus=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.1
0/bin/augustus #location of augustus executable

fgenesh= #location of fgenesh executable

tRNAscan-
SE=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.10/bin/
tRNAscan-SE #location of trnascan executable


snoscan=/fslgroup/fslg_pws_module/compute/software/.conda/envs/maker_v2.31.10
/bin/snoscan #location of snoscan executable
```

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

cat > ${scripts_dir}/${name}/maker_bopts.ctl <<EOF
#-----BLAST and Exonerate Statistics Thresholds
blast_type=ncbi+ #set to 'ncbi+', 'ncbi' or 'wublast'

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

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

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

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

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

en_score_limit=20 #Exonerate nucleotide percent of maximal score threshold

EOF

cat > ${scripts_dir}/${name}/maker_opts.ctl <<EOF

#-----Genome (these are always required)

genome=${data_dir}/${file} #genome sequence (fasta file or fasta embeded in
GFF3 file)

organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic


#-----Re-annotation Using MAKER Derived GFF3

maker_gff= #MAKER derived GFF3 file

est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no

altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no

protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no

rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no

model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no

pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no

other_pass=0 #passthrough anyything else in maker_gff: 1 = yes, 0 = no


#-----EST Evidence (for best results provide a file for at least one)

est=${data_dir}/pallidicaule.transcripts.fa #set of ESTs or assembled mRNA-
seq in fasta format

altest=${data_dir}/quinoa.transcripts.fa #EST/cDNA sequence file in fasta
format from an alternate organism

est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file

altest_gff= #aligned ESTs from a closly relate species in GFF3 format


#-----Protein Homology Evidence (for best results provide a file for at least
one)


protein=${data_dir}/uniprot_sprot.fa,${data_dir}/pallidicaule.protein.fa,${da
ta_dir}/quinoa.protein.fa, #protein sequence file in fasta format (i.e. from

protein_gff=  #aligned protein homology evidence from an external GFF3 file
```

```
#-----Repeat Masking (leave values blank to skip repeat masking)

model_org= #select a model organism for RepBase masking in RepeatMasker

rmlib=${data_dir}/consensi.fa.classified #provide an organism specific repeat
library in fasta format for RepeatMasker

repeat_protein=${data_dir}/te_proteins.fa #provide a fasta file of
transposable element proteins for RepeatRunner

rm_gff= #pre-identified repeat elements from an external GFF3 file

prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this),
1 = yes, 0 = no

softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and
dust filtering)


#-----Gene Prediction


snaphmm=/panfs/pan.fsl.byu.edu/scr/grp/fslg_pws_module/software/.conda/envs/m
aker_v2.31.10/share/snap/HMM/A.thaliana.hmm #SNAP HMM file

gmhmm= #GeneMark HMM file


augustus_species=BUSCO_canu1_7_atriplex_60.contigs_nanopolished_pilon_pilon_a
rabidopsis_2266460549 #Augustus gene prediction species model

fgenesh_par_file= #FGENESH parameter file

pred_gff= #ab-initio predictions from an external GFF3 file

model_gff= #annotated gene models from an external GFF3 file (annotation
pass-through)

est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no

protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no

trna=1 #find tRNAs with tRNAscan, 1 = yes, 0 = no

snoscan_rrna= #rRNA file to have Snoscan find snoRNAs

unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 =
yes, 0 = no


#-----Other Annotation Feature Types (features MAKER doesn't recognize)

other_gff= #extra features to pass-through to final MAKER generated GFF3 file


#-----External Application Behavior Options
```

```
alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST
databases

cpus=8 #max number of cpus to use in BLAST and RepeatMasker (not for MPI,
leave 1 when using MPI)



#-----MAKER Behavior Options

max_dna_len=100000 #length for dividing up contigs into chunks
(increases/decreases memory usage)

min_contig=1000 #skip genome contigs below this length (under 10kb are often
useless)



pred_flank=200 #flank for extending evidence clusters sent to gene predictors

pred_stats=0 #report AED and QI statistics for all predictions as well as
models

AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)

min_protein=0 #require at least this many amino acids in predicted proteins

alt_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes,
0 = no

always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 =
no

map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes,
0 = no

keep_preds=0 #Concordance threshold to add unsupported gene prediction (bound
by 0 and 1)



split_hit=10000 #length for the splitting of hits (expected max intron size
for evidence alignments)

single_exon=1 #consider single exon EST evidence when generating annotations,
1 = yes, 0 = no

single_length=250 #min length required for single exon ESTs if 'single_exon
is enabled'

correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes



tries=2 #number of times to try a contig if there is a failure for some
reason

clean_try=1 #remove all data from previous run before retrying, 1 = yes, 0 =
no
```

```
clean_up=0 #removes theVoid directory with individual analysis files, 1 =
yes, 0 = no

TMP= #specify a directory other than the system default temporary directory
for temporary files

EOF


sbatch ${scripts_dir}/${name}/${name}.sh


done
```

MAKER 2.0 Merge

```
#!/bin/sh


 #mkdir ALLGFFS

 #mkdir ALLGFFS/scripts

 #mkdir ALLFASTAS

 #mkdir ALLFASTAS/scripts


 for file in *.fasta

 do

 name=`echo $file | sed 's/.fasta//'`

 cat > ./ALLGFFS/scripts/${name}.sh <<EOF

 #!/bin/bash

 #SBATCH -c 1 --qos=pws --mem=1gb -t 0:60:00


module purge

module load conda-pws

module load conda/maker_v2.31.10


gff3_merge -d ./${name}.maker.output/${name}_master_datastore_index.log
${name}.all.gff -n

EOF
```

```
sbatch ./ALLGFFS/scripts/${name}.sh


cat > ./ALLFASTAS/scripts/${name}.sh <<EOF

#!/bin/bash


#SBATCH -c 1 --qos=pws --mem=1gb -t 0:10:00


module purge

module load conda-pws

module load conda/maker_v2.31.10


fasta_merge -d ./${name}.maker.output/${name}_master_datastore_index.log
${name}.all.maker.proteins.fasta ${name}.all.maker.transcripts.fasta
${name}.all

EOF


sbatch ./ALLFASTAS/scripts/${name}.sh


done
```

MAKER 2.0 Maker Functional

```
module purge

module load conda-pws

module load conda/maker_v2.31.10



maker_functional_gff uniprot_sprot.fa maker2uni.blasp
canu1_7_atriplex_60.contigs _nanopolished_pilon_pilon_allscaffolds.gff >
canu1_7_atriplex_60.contigs_nanopolished_pilon_pilon_allscaffolds.functional_
blast.gff

maker_functional_fasta uniprot_sprot.fa maker2uni.blasp
canu1_7_atriplex_60.contigs_nanopolished_pilon_pilon_allscaffolds_all_maker_p
roteins.fasta >
canu1_7_atriplex_60.contigs_nanopolished_pilon_pilon_allscaffolds_all_maker_p
roteins_functional_blast.fasta

maker_functional_fasta uniprot_sprot.fa maker2uni.blasp
canu1_7_atriplex_60.contigs_nanopolished_pilon_pilon_allscaffolds_all_maker_t
```

```
ranscripts.fasta >
canu1_7_atriplex_60.contigs_nanopolished_pilon_pilon_allscaffolds_all_maker_t
ranscript_functional_blast.fasta
```

MaSuRCA - Config File

```
DATA

PE= pe 250 20

NANOPORE=/fullpath/nanopore.fa

END


PARAMETERS

EXTEND_JUMP_READS=0

GRAPH_KMER_SIZE = auto

USE_LINKING_MATES = 0

GRID_QUEUE=all.q

GRID_BATCH_SIZE=300000000

LHE_COVERAGE=30

LIMIT_JUMP_COVERAGE = 300

CA_PARAMETERS =  cgwErrorRate=0.15

KMER_COUNT_THRESHOLD = 1

CLOSE_GAPS=1

NUM_THREADS = 32

JF_SIZE = 200000000

SOAP_ASSEMBLY=0

END
```

MinION QC

```
#-i INPUT, --input=INPUT; Input file or directory (required). Either a full
path to a sequence_summary.txt file, or a full path to a directory containing

I=inputdirectory

#-o OUTPUTDIRECTORY, --outputdirectory=OUTPUTDIRECTORY; Output directory
(optional, default is the same as the input directory). If a single
sequencing_s

O=outputdirectory
```

```
#-q QSCORE_CUTOFF, --qscore_cutoff=QSCORE_CUTOFF; The cutoff value for the
mean Q score of a read (default 7).

#-p PROCESSORS, --processors=PROCESSORS; Number of processors to use for the
anlaysis (default 1).

P=4

#-s SMALLFIGURES, --smallfigures=SMALLFIGURES; TRUE or FALSE (the default).
When true, MinIONQC will output smaller figures, e.g. suitable for publicatio

S=FALSE

MinIONQC.R -i ${I} -o ${O} -p ${P} -s ${S}
```

NanoFilt

```
gunzip -c file.fastq.gz | NanoFilt -q 8 --headcrop 25 -l 2000 | gzip >
trimmed.fastq.gz

#USAGE:

#NanoFilt [-h] [-q QUALITY] [-l LENGTH] [--headcrop HEADCROP] [--tailcrop
TAILCROP]
```

Nanopolish Index

```
nanopolish index -d input_directory total.fastq
```

Nanopolish – Makerange

```
module purge

module load conda-pws

module load conda/nanopolish

module load python/3/6

python nanopolish_makerange.py file.fasta | parallel --results

nanopolish.results -P 8 nanopolish variants --consensus -o polished.{1}.vcf -

w {1} -r total.fasta -b reads.sorted.bam -g file.fasta -t 3 --min-candidate-

frequency 0.1
```

Nanopolish - Minimap

```
module purge

module load minimap2/2.12

module load samtools/1.6

minimap2 -ax map-ont -t 24 p_19_S_4.ctg.lay.fa total.fasta | samtools sort -o
reads.sorted.bam -T reads.tmp

samtools index reads.sorted.bam
```

Nanopolish Slurm Makerange

```
module purge

module load conda-pws

module load conda/nanopolish


# Get ranges with nanopolish_makerange.py

 RANGES=$(python `which nanopolish_makerange.py` file.fasta)

 NUM_RANGES=$(wc -w <<< $RANGES)

 RANGES_PER_TASK=$(( ($NUM_RANGES + 999) / 1000 ))


 # Which ranges is this task going to do? (e.g. 1-21, 22-42, 43-63, etc.)

 FIRST=$(( 1 + ($SLURM_ARRAY_TASK_ID - 1) * $RANGES_PER_TASK ))

 LAST=$(( $FIRST + $RANGES_PER_TASK - 1 ))


 # Do all the jobs this task is assigned

 cut -d' ' -f $FIRST-$LAST <<< $RANGES | tr " " "\n" | parallel --results
nanopolish.results -P 6 nanopolish variants --consensus -o polished.{1}.vcf -
w {1} -r total.fasta -b reads.sorted.bam -g file.fasta -t 4 --min-candidate-
frequency 0.1
```

NanoPlot

```
NanoPlot --color green --format pdf –summary sequencing_summary.txt --
loglength -o summary-plots-log-transformed
```

Pilon

```
reference_genome=genome.fasta

sorted_bam1=Q_Pool_1.all.sam.bam.sorted
```

```
sorted_bam2=Q_Pool_2.all.sam.bam.sorted

output_dir=output

output_prefix=output_prefix


module purge

module load conda-pws

module load conda/pilon_1.22


pilon -Xmx512G --genome ${reference_genome} --bam ${sorted_bam1} --bam
${sorted_bam2} --outdir ${output_dir} --output ${output_prefix} --changes --
fix bases --diploid --threads 1
```

Porechop

```
porechop -i trimmed.fastq.gz -t 24 -v 2 -o trimmed.porechop.fastq.gz --
discard_middle > porechopDM.log
```

SNPhylo

```
#!/bin/bash


#SBATCH --time=12:00:00 # walltime

#SBATCH --ntasks=2 # number of processor cores (i.e. tasks)

#SBATCH --nodes=1 # number of nodes

#SBATCH --mem-per-cpu=4096M  # memory per CPU core


#Load R module

#Load miniconda

#Source activate muscle


sh path_to_snphylo.sh -H path_to_hapmap -p 10 -l .4 -m .05 -P snphylo.output
-b B 1000 -a 5000 -A
```

Trimmomatic

```
#!/bin/sh
```

```
for forward_file in *.fq.gz

do

name=`echo $forward_file | sed 's/.fq.gz//'`

cat > ./trim_scripts/${name}.sh <<EOF

#!/bin/bash


module purge

module load conda-pws

module load conda/trimmomatic


trimmomatic PE -threads 4 -summary ${name}_stats_trim.txt ${name}_1.fq.gz
${name}_2.fq.gz -baseout ./Trimmed_Reads/${name}.fq.gz
ILLUMINACLIP:/fslhome/pj


EOF


sbatch ./trim_scripts/${name}.sh

done
```

wtdbg

```
p=19

S=4

reads=Reads.fasta.gz

wtdbg-1.2.8 -t 20 -i ${reads} -fo p_${p}_S_${S} -p ${p} -S ${S} --tidy-reads

5000 --edge-min 2 --rescue-low-cov-edges && wtdbg-cns -t 20 -i

p_${p}_S_${S}.ctg.lay -o p_${p}_S_${S}.ctg.lay.fas && assemblathon_stats_2.pl

p_${p}_S_${S}.ctg.lay.fas > p_${p}_S_${S}.ctg.lay.fas.assembly_stats
```