2022-06-06

Relationships Among AA-Genome Chenopodium Diploids and a Whole-Genome Assembly of the North American Species, C. watsonii

Lauren Amillicent Young
Brigham Young University

Relationships Among AA-Genome *Chenopodium* Diploids and a Whole-Genome Assembly of the North American Species, *C. watsonii*

Lauren Amillicent Young

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

Relationships Among AA-Genome *Chenopodium* Diploids and a Whole-Genome Assembly of the North American Species, *C. watsonii*

Lauren Amillicent Young
Department of Plant and Wildlife Sciences, BYU
Master of Science

Chenopodium quinoa Willd., an ancient Andean pseudocereal almost exclusively consumed in South America, jumped onto the global stage when Western cultures noted quinoa's advantageous nutritional profile. Quinoa seed's high protein content, nutritionally balanced amino acid profile, low glycemic index, and high fiber, vitamin, and mineral content, make it a highly sought-after 'superfood'. Pitseed goosefoot (*C. berlandieri* Moq.), a closely related North American species sharing quinoa's genome composition (AABB), grows across the North American continent, inhabiting diverse environments including the saline coastal soils of the Gulf of Texas and the drought-prone regions of the Southwest. Quinoa and pitseed goosefoot, along with South American avian goosefoot (*C. hircinum* Schrad.), make up the Allotetraploid Goosefoot Complex (ATGC). We hypothesize that an ancient hybridization event between A- and B-genome diploids, with a subsequent whole-genome duplication, gave rise to the common ancestor of the ATGC. Prior data indicate that allopolyploidization most likely occurred within North America, with long-range dispersal of the ATGC to South America. We have sequenced the genome of the North American AA-genome diploid *C.* watsonii and identified via DNA marker analyses the closest extant species to the AA-genome diploid ancestor of the ATGC from among a panel of 41 AA-genome diploid resequenced accessions, encompassing 30 putative AA-genome diploid species, from North and South America. We also present evidence for reciprocal long-range dispersal of *Chenopodium* diploids between North and South America.

Keywords: *Chenopodium berlandieri, Chenopodium quinoa*, *Chenopodium watsonii*, AA-genome diploid species, whole-genome assembly, phylogenomics

ACKNOWLEDGEMENTS

The time, effort, and support I have received from not only my advisors and mentors, but family, friends and fellow coworkers has no doubt gotten me to this point of completing my thesis and project. Countless hours of inspired and dedicated work by Drs. Rick Jellen, Jeff Maughan, and David Jarvis have made this project possible, teaching me the ropes of plant genetics and genomics. I could not have come this far without the flexibility of my advisor, Dr. Rick Jellen, encouraging me to continue my education and inspiring me though the potential global impact of orphaned crops. I would like to thank the undergraduates who bore several of my burdens, allowing me to spend more time with my son. Lastly, I would not be here without the unfaltering support from my family to make my hours of lab work pass without worry, thank you for all that you do

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Literature Review on New World *Chenopodium* Species

Lauren A. Young

Department of Plant and Wildlife Sciences, BYU
Master of Science

Introduction to the Allotetraploid Goosefoot Complex

The Allotetraploid Goosefoot Complex (ATGC) is comprised of the AABB-genome tetraploid *Chenopodium* species ($2n = 4x = 36$), *C. berlandieri* Moq., *C. hircinum* Schrad., and *C. quinoa* Willd (Jarvis et al., 2017). Often referred to goosefoots due to their shared leaf morphology, *Chenopodium berlandieri* Moq. is found across North America with native varieties ranging from Canada to Central America (Jarvis et al., 2017), while the Andean pseudocereal quinoa (*C. quinoa* Willd.) and weedy avian goosefoot (*C. hircinum* Schrad.) are both native to South America. Through a dispersal event, it is hypothesized that *C. berlandieri* was brought to South America, underwent speciation pressures, and was given the alternate taxonomic designation of *C. hircinum*. The taxon was then domesticated and became known as the species *C. quinoa*, having large, primarily white seeds. Genome analyses of *C. berlandieri* and *C. hircinum* support the hypothesis that the two taxa are indeed the same biological species inhabiting different hemispheres, with quinoa and Mesoamerican *huauzontle* or *chia roja* (*C. berlandieri* Moq. subsp. *nuttaliae* (Saff.) H.D. Wilson & Heiser) and the principal extant domesticated forms (Jellen et al., 2019; Jarvis et al., 2017; Wilson & Heiser, 1979; Wilson, 1980; Wilson, 1981; Wilson, 1988).

As noted above, domestication of chenopods has not been limited to quinoa. Though not a commercial food crop like quinoa, domesticated varieties of pitseed goosefoot exist, notably *huauzontle* (*C. berlandieri* ssp. *nuttalliae*) which is cultivated across central and southern Mexico as a locally consumed inflorescence-vegetable that is steamed or boiled like broccoli (Wilson & Heiser, 1979). Additionally, the Andean diploid *cañhua* or *kañiwa* (*C. pallidicaule* Aell.) is also cultivated as a high-altitude pseudocereal. Though domestication of chenopods is attributed mainly to South America and Mesoamerica, there is evidence that the extinct *C. berlandieri* subsp. *jonesianum* Smith & Funk was domesticated in eastern North America, independent of the domestication events of *C. hircinum* and *C. quinoa* that occurred in South America and Mesoamerica (Kistler & Shapiro, 2011, Smith & Yarnell, 2009). As such, the *jonesianum* cultigen of pitseed goosefoot was a principal component of the Eastern Agricultural Complex (Smith, 2006; Smith & Yarnell, 2009).

As with other domesticated crops, understanding the ancestry, specifically the ancient hybridization event that ultimately led to the evolution of wild-weedy *C. berlandieri* and *C. hircinum* as the free-living ancestors of quinoa and huauzontle, can potentially reveal strategies for utilizing these in quinoa and huauzontle improvement through crossing and selective breeding (Khoury et al., 2020). Recent analyses of the organellar genomes (Maughan et al., 2019) and nuclear sub-genomes (Jarvis et al., 2017) have revealed that the A-genome donor was the maternal parent in the cross that gave rise to the original allotetraploid of the complex. Whole-genome sequencing data of three A-genome diploids, *C. watsonii* A. Nels., *C. sonorense* Benet-Pierce & Simpson, and *C. pallidicaule,* showed *C. watsonii* as being the closest A-genome relative of the three to *C. berlandieri,* though a larger panel of potential A-genome donors is needed to identify the closest extant relative of the maternal ancestor (Jellen et al., 2019). An

initial investigation of relatedness based on analysis of 3,600 microsatellite markers detected by mapping 10x Illumina short-reads of 27 AA- and BB-genome diploids back to the sub-genomes of quinoa var. 'QQ74' whole-genome reference (Jarvis et al., 2017) revealed that *C. desiccatum* A. Nels.*, C. papulosum* Moq.*,* and *C. leptophyllum* (Moq.) Nutt. ex S. Wats. are more closely related to the AA sub-genome of the ATGC than other previously sequenced species, despite the drastic phenotypic differences among the three species and *C. berlandieri* (Maughan, personal communication).

Chenopodium taxonomy

The *sensu stricto* genus *Chenopodium* was largely defined by Fuentes-Bazan et al. (2012) when they broke apart the existing genus, which had included over 150 taxonomic entities but was observed to be polyphyletic, using cpDNA and nuclear internal transcribed spacer (ITS) sequence markers. Mosyakin and Clemants (1996) published the most recent comprehensive taxonomy since Aellen and Just (1943). The described putative *Chenopodium* taxa from Mosyakin and Clemants (1996), 45 in total, are listed in Appendix 1 and described below. Chenopods are found across the Americas, Asia, Europe, and Australia. Although a thorough understanding of species relationships is generally lacking, genetic studies like those of the notoriously confusing *C. album* complex by Mandak et al. (2018) and whole-genome sequencing as in Jarvis et al. (2017) are beginning to shed light on *Chenopodium* systematics. With tens of currently described species native to North America, the continual collection of germplasm not only allows for the conservation of diverse species but also provides potential breeding resources for improving abiotic and biotic stress tolerance in cultivated quinoa varieties specifically that of pitseed goosefoot due to the matching chromosome numbers and sub-genome composition.

Within the *Chenopodium* genus, eight sub-genomes (A-H) have been identified, three of which (C, F, & G) originate from extinct diploid ancestors but can be found in extant polyploids (Walsh et al., 2015; Štorchová et al., 2015; Mandák et al., 2018). Quinoa is the most well-known of the chenopods due to its use as a food crop. The nutritional profile of quinoa seeds has contributed to its consumption spreading from its origins in the Andean regions of South America to becoming a global food crop. Quinoa seeds are high in protein with a diverse amino acid profile, low on the glycemic index, and well-rounded in fiber, vitamin, and mineral content making quinoa seeds a highly sought after 'superfood' (Bhargava, 2005; James, 2009; Wright, 2002).

In addition to quinoa, another notable chenopod used as a food crop is *C. pallidicaule*. Cultivated by the indigenous peoples of Peru and Bolivia, *C. pallidicaule* is commonly referred to as *cañahua* (Peru) or *kañiwa* (Bolivia) and consumed as a pseudocereal, whole seeds being added to soups or toasted and ground into flour (*pito*) and used for baking. Similar to *C. berlandieri*, *C. pallidicaule* grows on marginal lands in arid climates where traditional crops fail (Rastrelli et al., 1996). The nutrition profile of *C. pallidicaule* is comparable to that of quinoa, presenting with high protein and lipid content along with notable quantities of antioxidants (Gross et al., 1989).

The ATGC members are the most common representatives of *Chenopodium* Subsection *Favosa* in both North and South America and have been widespread across temperate and subtropical regions of both continents due to human disturbance. Within North America, *C. berlandieri* includes at least five ecotypes of *C. berlandieri* subsp. *berlandieri*: 1) variety *berlandieri* in far southern Texas; 2) variety *boscianum* (Moq.) Wahl along the Gulf of Mexico Coast; 3) variety *macrocalycium* (Aell.) Cronq. along the New England and Canadian Maritime

seacoasts; 4) variety *sinuatum* (Murr) Wahl in the Southwest; and 5) variety *zschackei* (Murr) Murr ex Asch. throughout the bulk of the continent and possibly into the northern Andes. North American representatives of subsp. *nuttaliae* include the cultigens *huauzontle* in highland South-Central Mexico, *chia roja* in Michoacan, and possibly semi-wild forms of *quelites* throughout Central Mexico. Within South America, Andean quinoa (*C. quinoa*) is usually accompanied by apparently feral forms classified as subsp. milleanum (Aell.) Aell. in the northern Andes and subsp. *melanospermum* Hunziker in the central Andes (Wilson, 1988). Also in South America, weedy *C. hircinum* is found in lowlands and river valleys on both sides of the Andes from central Peru on the Pacific Slope and southeastern Bolivia on the Atlantic slope southward into Patagonia. In addition, hexaploid *C. bushianum* (Aell.) Cronq., which was formerly treated as a variety of *C. berlandieri*, should be considered a separate species due to its hexaploid chromosome number; this taxon is found in the eastern United States as a weed associated with agriculture.

The taxonomic classification of *Chenopodium sensu lato* was extensively reviewed by Jellen et al. (2011) and is in need of revision due to the recent discovery of multiple new taxa by Benet-Pierce. Current taxonomic designations of the *Chenopodium* genus are based on plant and seed morphology rather than molecular data, as seen in the most recent and comprehensive taxonomy as described by Mosyakin and Clemants (1996; Fuentes-Bazan et al., 2012; Appendix 1).

Subsection Polysperma

Subsection *Polysperma* houses a single species, *Chenopodium polyspermum* Kowal ex Mosy. & Clem., found surrounding the Great Lakes and northeastern United States and eastern

Canadian provinces (Clemants & Mosyakin, 2004). The leaves vary from oblong to ovate and the plants produce smooth seeds with non-adhering pericarps.

Subsection Urbica

Only containing a single species, *Chenopodium urbicum* L., Subsection *Urbica* is found in similar regions to *C. polyspermum*, across the northeastern United States and eastern Canada (Clemants & Mosyakin, 2004). Based on the morphological characteristics of *C. urbicum*, discussion of whether reclassification to *Blitem* rather than *Chenopodium* is more appropriate. *C. urbicum* has reddish brown seeds with a papillose to smooth pericarp and triangular leaves which grow from the simple, rarely branched stem.

Subsection Undata

Classified by the rugose to smooth seed coats and acute seed margins, the subsection includes one described species, *Chenopodium murale* L. (Clemants & Mosyakin, 2004). Serrated ovate or triangular leaves are typical for this lesser described subsection. Though found across North America, *C. murale* is native to Eurasia and found worldwide, particularly in the warm to temperate regions of the subtropics.

Subsection Leptophylla

Found growing in the sandy soils of western North America eastward to the Midwest, the taxa within Subsection *Leptophylla* have narrow to linear leaves that range from non-fleshy to fleshy and varying testa textures from smooth to rugose. The taxa have utriculate pericarps that can present as alveolate but are always non-adhering and usually flake easily off the seed when disturbed. Unfortunately, this group is very poorly characterized except for *Chenopodium leptophyllum* subsp. *oblongifolium* (S. Wats.) Wahl and *C. desiccatum*. Other species belonging

to this subsection include *C. albescens* Small, *C. cycloides* A. Nels., *C. foggii* Walh*, C. hians* Stand., *C. pallescens* Stand., *C. pratericola* Rydb.*,* and *C. subglabrum* (S. Wats.) A. Nels.

C. albescens has a black to dark brown, finely warty, non-adhering pericarp, with acute-tipped sepals that spread apart from the fruit at maturity (Figure 2A). *C. cycloides* produces black, irregularly maturing seeds with prominent margins and rugose testa (Nelson, 1902). *C. foggii* grows in the rocky forest of New England and eastern Canada. The plants produce farinose leaves that are ovate-lanceolate in shape and finely rugose seeds with rounded margins (Clemants & Mosyakin, 2004). With similar seeds and foliage to those of *C. foggii*, *C. hians* has a more branched growth habit and farinose stems (Standley, 1916). *C. hians* inhabits the open prairies and pastures of Western North America. Found in the Midwest, *C. pallescens* is a branched species with linear leaves and black rugose seeds with rounded margins like those of other species in the subsection (Standley, 1916). *C. pratericola* is found in the alkaline and saline soils across North America, often near pinyon pines and sagebrush (Rydberg, 1912). Lastly, *C. subglabrum* is found on the sandy riverbanks of the upper Midwest and produces shiny black seeds and a non-adhering pericarp (Nelson, 1902).

Subsection Fremontiana

Chenopodium Subsection *Fremontiana* includes mostly taxa with deltoid to campanulate leaves and seeds having smooth testas and non-adhering pericarps. Found in western North and lowland-temperate South America, the subsection includes the following species: *C. atrovirens* Rydb.*, C. fremontii* S. Wats.*, C. incanum* (S. Wats.) Heller (North America), *C. cordobense* Aell., and *C. ruiz-lealii* Aell. (South America) (Mosyakin & Clemants, 1996).

North American *Fremontiana* species are concentrated west of the Rocky Mountains *C. atrovirens* is a mostly montane species from the Rockies to the Sierra Nevada and intervening

ranges of the Great Basin. *C. fremontii* is sympatric with *C. atrovirens*, though the former generally occupies lower elevations in hills and foothills, often in the shade of pinyon pines, junipers, mountain maples, and scrub oaks. *C. incanum* is a short (<5 dm tall), bushy plant and mostly a desert and semiarid shortgrass prairie species, very scattered in its distribution though often encountered in arroyos or following fires in the Great Basin, and usually highly farinose, giving the plant its characteristic silvery or mealy appearance.

The South American species of *Fremontiana* - *C. cordobense* and *C. ruiz-lealii* - are found in the northwestern Argentine provinces of Cordoba, La Rioja, and San Luis, and closely resemble the Texas endemic *C. albescens*, being upright plants with a somewhat yellow-green appearance, distinct paniculate spikes, and seeds with very pronounced radicle points (Giusti, 1997). The sepals are very fleshy, having obtuse tips, and distinct to the base of the receptacle in *C. cordobense*, with a semi-adhering warty pericarp. *C. ruiz-lealii* has a warty, black pericarp that is non-adhering with non-fleshy, slightly keeled sepals having acute tips.

Subsection Favosa

The *Chenopodium* Subsection *Favosa* includes the ATGC along with 11 diploid species and one hexaploid, *C. bushianum*. The subsection is found across North and South America, with a concentration in the Chihuahuan and Sonoran Deserts of the former. Generally, *Favosa* taxa are characterized by their deltoid, rhomboid-ovate, to campanulate leaves that often produce a foul-smelling odor due to trimethylamine (TMA). The fruits have a pitted testa and an alveolate or papillate, mostly adhering pericarp (Clemants & Mosyakin, 2004).

Diploid species of this subsection occupy an assortment of mostly montane and plateau habitats of the southwestern United States and northern Mexico. One North American exception to this rule is *C. flabellifolium* Stand., an extreme endemic found only on the tiny island of San

Martin off the northern coast of Baja California. *C. arizonicum* Stand. inhabits mostly montane and high-desert habitats of Arizona and New Mexico. *C. neomexicanum* Stand. is most abundant on igneous pine-forest soils of the Mogollon Rim of Arizona. *Chenopodium lenticulare* Aell. is found in the Davis, Guadalupe, and Sierra Blanca Mountains of West Texas and southeastern New Mexico. *C. palmeri* Stand. and *C. sonorense* mostly inhabit arroyos and disturbed roadsides of the Sonoran Desert. *C. parryi* Stand. is found in mountainous and plateau regions of Northeast Mexico, almost up to the Rio Grande.

C. watsonii, commonly known as stinking goosefoot due to its characteristically strong TMA odor, is most abundant on grazing-disturbed sites on plateaus of the Four Corners region and extending northward in the High Plains along the eastern slope of the Rocky Mountains. Stinking goosefoot was chosen as our whole-genome assembly because of its clear characterization and species designation, as well as sharing the same subsection, Subsection *Favosa*, as *C. berlandieri* and *C. quinoa*.

Subsection *Favosa* includes a singular South American diploid, *C. philippianum* Aell. *C. philippianum* is fairly common on disturbed sites in the Cordillera Occidental and upper Atacama Desert of southwestern Bolivia and northeastern Chile, where it can grow as a perennial on year-round soil moisture.

Subsection Cicatriosa

As described by Mosyakin and Clemants (1996), Subsection *Cicatriosa* encompasses several diploid and polyploid species with Eurasian origins, including the BBDD tetraploid, *Chenopodium acerifolium* Andr., B-genome diploid, *C. suecicum* Murr, and two BBEE tetraploid species, *C. karoi* (Murr) Aell. and *C. jenissejense* Aell. & Iljin.

Subsection Standleyana

Currently housing several species, *Chenopodium badachschanicum* Tzvelev, *C. bryoniifolium* Bunge, *C. gracilispicum* Kung, *C. missouriense* Aell., and *C. standleyanum* Aell., Subsection *Standleyana* is characterized as having large and narrow, yet non-linear, and distinctly acute leaves and glomerules of ovate seeds with prominent radicle ends and non-adhering pericarps (Mosyakin & Clemants, 1996). The former three species are found in Eurasia while the latter two are endemic to North America, thus our focus will be on the latter species. *C. standleyanum* is native to the Midwest, generally east of the Missouri River, and is rare to locally common along forest edges and other disturbed areas.

Subsection Chenopodium

Similar to Subsection *Cicatricosa, Chenopodium* subsection *Chenopodium* exclusively houses Eurasian taxa, including di-, tetra-. and hexa- ploids: *C. album* L., *C. strictum* Roth, *C. opulifolium* Schrad. ex A. P. De Cand., *C. vulvaria* L., *C. sosnowskii* Kap., *C. pamiricum* Iljin, *C. nidorosum* Otsch., and *C. iljinii* Gol. Including neither A-genome diploids nor North American species, this subsection is not discussed further due to the scope of our study.

Wild goosefoot species and quinoa breeding

With its ecological diversity and ability to grow in a variety of environments, *Chenopodium berlandieri* offers genetic resources for providing biotic and abiotic stress tolerance in domesticated varieties of *C. quinoa* (Wilson & Manhart, 1993). Varieties of *C. berlandieri* are found growing during the hot summers with temperatures reaching 38°C and in saline coastal soils along the Gulf Coast of Texas, while others are found in some of Canada's northernmost provinces (Clemants & Mosyakin, 2004). The seeds of wild *C. berlandieri* are characterized by their large size, thick, dark seed coats and pitted surface, much like that of a

golf ball. The extinct subspecies of *C. berlandieri*, *jonesianum* was domesticated in prehistoric Eastern North America and had larger seeds with thinner seed coats, closer to the domesticated traits of *C. quinoa* (Gremillion, 1993). When boiled, the seeds remain tough, unlike their thin-coated *C. quinoa* counterpart whose seeds soften to a couscous-like consistency when cooked.

Alternately, *C. quinoa* has been grown and bred for thousands of years in South America, leading to highly specialized ecotypes including Altiplano, Inter-Andean Valley, Salares, Sea-Level and Sub-Tropical (Tapia et al., 1980; Murphy et al., 2019). Many of the cultivated varieties, except for those of the Sea-Level ecotype, are adapted to the high altitudes with different biotic stresses within the Andean region of South America. The susceptibility of quinoa to lowland pests and diseases has led to the slow adoption of the crop on a global scale due to the high yield losses seen when grown in lower elevation environments for which it is not adapted (Murphy et al., 2019). Outside of the primary gene pool of taxa within *C. quinoa*, pitseed goosefoot offers a secondary gene pool from which taxa can easily be crossed with quinoa due to the matching chromosome counts and sub-genome compilations (Jellen et al., 2019).

CHAPTER 2

Relationships Among AA-Genome *Chenopodium* Diploids and a Whole-Genome Assembly of the North American Species, *C. watsonii* A. Nels.

Lauren A. Young

Department of Plant and Wildlife Sciences, BYU
Master of Science

ABSTRACT

Quinoa (*Chenopodium quinoa* Willd.), an Andean pseudocereal, attained global popularity beginning in the early 2000's due to its exceptional amino acid profile, low glycemic index, and high fiber, vitamin, and mineral contents. Pitseed goosefoot (*C. berlandieri* Moq.), quinoa's putative North American wild-weedy ancestor, grows on disturbed and sandy substrates across the North American continent, inhabiting diverse environments including saline coastal sands, southwestern deserts, subtropical highlands, the Great Plains, and boreal forests. Together with South American avian goosefoot (*C. hircinum* Schrad.) the three taxa comprise the Allotetraploid Goosefoot Complex (ATGC; $2n=4x=36$, AABB subgenomes). Superimposed on pitseed goosefoot's range are approximately 35 A-genome diploids, most of which are adapted to a diversity of niche environments, with another nine taxa native to temperate-subtropical South America. We sequenced the genome of the North American AA diploid *C. watsonii* A. Nels., revealing a genome size of 551.6 MB in 1700 scaffolds (N50=55.14, L50=5), with 93.87% single-copy and 3.35% duplicated genes. A high degree of synteny, with minor and mostly telomeric rearrangements, was found when comparing this taxon with the previously reported genome of *C. pallidicaule* Aell. and the A-genome chromosomes of *C. quinoa.* Phylogenetic analysis using 10,588 SNPs on a panel of 41 AA accessions, three AABB, and one HH outgroup encompassing 32 taxa from North and South America indicated that the Rocky Mountains-Great Plains psammophyte *C. subglabrum* A. Nels. was closest to the A-genome ancestor of the ATGC. We also present evidence for reciprocal long-range dispersal of *Chenopodium* diploids between North and South America.

INTRODUCTION

Quinoa (*Chenopodium quinoa* Willd.) is an Andean-origin pseudocereal possessing an appreciable content of high-quality protein for human consumption (Wu, 2015). Aside from this and its other nutritional benefits, such as high fiber, essential mineral content, and fatty acid profile, quinoa is recognized for its tolerance to abiotic stresses including drought, salinity, and cold (Azurita-Silva et al., 2015; Biondi et al., 2015; Bhargava & Srivastava, 2013; Martinez, 2015). On the other hand, quinoa's poor thermal tolerance has presented an impediment to its successful introduction into lowland tropical and subtropical environments (Zurita-Silva et al., 2014). Fortunately, quinoa produces mostly fertile hybrids when cross-pollinated with its free-living, heat-tolerant North American (pitseed goosefoot, *C. berlandieri* Moq.) and South American (avian goosefoot, *C. hircinum* Schreb.) ancestor-species, which together constitute the Allotetraploid Goosefoot Complex (ATGC, $2n=4x=36$, AABB subgenomes; Wilson & Manhart, 1993; Jellen et al., 2019). The ATGC also includes the Mesoamerican domesticated vegetable and pseudocereal forms of huauzontle (*C. berlandieri* Moq. subsp. *nuttaliae* (Saff.) H.D. Wilson & Heiser; Wilson & Heiser, 1979; Cepeda-Cornejo et al., 2016).

Since the ATGC's A- and B- genome ancestors are potential genetic resources for improving quinoa and huauzontle, their characterization is an important step in determining the tertiary gene pool for these cultivated species. The ATGC's subgenome B is only recognized as existing in diploids of Eurasian origin: *C. ficifolium* Sm., *C. suecicum* Murr, and *C. ucrainicum* Mosyakin & Mandák (Mosyakin & Mandák, 2020; Mandák et al., 2018; Walsh et al., 2015). Recently, Subedi et al. (2021) reported on the potential of *C. ficifolium* as a model system for studying quinoa's molecular biology and physiology. In contrast, subgenome A is found throughout the New World in a wide array of diploids ≥ 40 and continually increasing – that are

adapted to mostly disturbed environments including highlands and subtropical to temperate steppes, deserts, alkaline basins, seashores, and forests (Table 1; Figure 1; Aellen & Just, 1943; Aellen, 1960; Giusti, 1970; Mosyakin & Clemants, 1996; Benet-Pierce & Simpson, 2010, 2014, 2017; 2019; WCVP, 2022). This pattern of adaptive species radiation (Gavrilets & Losos, 2009; Schluter, 1996) for AA diploids is unique within the genus *Chenopodium* and elucidation of the mechanism responsible for this variation invites further study.

Whole-genome assemblies can serve as powerful resources for assessing phylogenetic relationships (Eisen & Fraser, 2003), allelic diversity (The 100 Tomato Genome Sequencing Consortium et al., 2014), domestication pathways (Xie et al., 2019), and evolution of structural variation (Filiault et al., 2018; Kamal et al., 2022). Mangelson et al. (2019) reported a short read-based, Hi-C scaffolded whole-genome assembly for the domesticated Andean A-genome diploid *kañiw*a or *cañahua* (*C. pallidicaule* Aell.). However, various studies including the *C. quinoa* whole-genome sequence paper of Jarvis et al. (2017) and the single-gene phylogenies of Brown et al. (2015), Walsh et al. (2015), and Storchova et al. (2015) indicated that North American AA diploids were most likely closer to the ancestral donor of AA to the ATGC. Consequently, we embarked on an effort to construct a reference-quality whole-genome sequence of the well-characterized southwestern North American AA diploid *C. watsonii* A. Nels. (Jellen et al., 2019). The two AA *Chenopodium* whole-genome assemblies were then used to determine phylogenetic relationships among an extensive species panel of mostly North American taxa, with some South American representatives. Unfortunately, taxonomic characterization of some AA diploids is a work in progress (Benet-Pierce & Simpson, 2010, 2014, 2017, 2019) so compilation of a complete species set for resequencing and phylogenetic analysis is not yet possible.

METHODS

Tissue collection and long-read sequencing

The BYU *C. watsonii* accession BYU 873 (Table 2), collected in Humboldt, Arizona, was grown hydroponically in a growth chamber at BYU set to a photoperiod of 11 hours with broad-spectrum lighting. Temperature controls were set between 18ºC and 20 ºC. The hydroponics solution was made using 27 g of MaxiGrow® Hydroponics Plant Food (General Hydroponics, Sevastopol, California) dissolved in 16 L of deionized water. The hydroponics solution was replaced every two weeks.

Prior to extraction, the *C. watsonii* plant was dark treated for 72 hours. Young leaf tissue was harvested, and DNA was extracted using a modified protocol from Oxford Nanopore Technologies (Oxford, United Kingdom), "High molecular weight gDNA extraction from spinach leaves", using the QIAGEN® (Hilden, Germany) Genomic-top 500/G kit (Supplement 1). Following the protocol, DNA quality was analyzed using the Thermo Scientific™ (Waltham, MA) NanoDrop™ One Microvolume UV-Vis Spectrophotometer to check 260/280 and 260/230 absorbance ratios and Invitrogen™ (Waltham, MA) Qubit™ 3 Fluorometer to estimate DNA concentration. Non-fragmented samples were prepared using the DNA Clean & Concentrator-5 kit (ZYMO Research, Irvine, CA). Protocol from the ZYMO kit was followed to produce the fragmented samples. Long-read library preparation was done using the SQK-LSK109 kit from Oxford Nanopore Technologies with Quick T4 DNA Ligase (NEB, M2200L) and 1D Genomic DNA by Ligation MinION protocol (Oxford Nanopore Technologies). Long-read sequencing of the *C. watsonii* genome was completed using R9 flow cells from Oxford Nanopore Technologies on the MinION™ sequencing machine. Short-read sequencing was generated using the same

DNA on a Illumina (San Diego, California) HiSeq platform with a library preparation of 180-bp insert sizes.

Whole-genome assembly of C. watsonii

Long-read sequence data quality was checked using MinIONQC (Lanfear et al., 2018). Nanopore reads were trimmed and filtered using NanoFilt (De Coster et al., 2018) using the following options: -q=8, headcrop=25, and -l=2000. Adaptor sequences were trimmed using Porechop v.0.2.3 (Wick, 2017) with the verbosity option set to 2.

Illumina short-read data was trimmed, removing remnant adapter sequences, using the ILLUMINACLIP option from Trimmomatic v0.39 (Bolger et al., 2014). The following options were used within the pipeline: leading and trailing set to 20 bp, a sliding window of 4:20, and a minimum length of 75 bp.

A preliminary genome was assembled using CANU v.1.8 (Koren et al., 2017). CANU parameters were set with normal corMhapSensitivity, 40 corOutCoverage, and parallel ovsMethod. The CANU-assembled genome was polished using two rounds of RACON (Vaser et al., 2017); the first round using the ONT long reads and the second round using the trimmed Illumina reads. Phase Genomics (Seattle, Washington) produced the scaffolded assembly using the polished CANU assembly and Hi-C data from dark-treated, liquid nitrogen flash frozen leaf tissue. Contaminant reads were identified and removed using BlobTools (Laetsch and Blaxter, 2017). Chloroplast and mitochondrial DNA was identified and removed using NCBI BLAST against the quinoa chloroplast and mitochondrial genomes (Maughan et al., 2019).

Transcriptome assembly

Leaf, root, and stem tissues from the hydroponically grown *C. watsonii* plant in addition to a whole seeding were used for RNA extraction using Trizol (Invitrogen™) and QIAGEN® RNEasy spin column per the manufacturer's instructions. RNA from each tissue and whole seedling was combined in equal parts to create a single bulk sample. Library preparation and transcriptome sequencing was completed using the PacBio (Menlo Park, California) Iso-Seq platform on the Sequel II instrument at the BYU DNA Sequencing Center (Provo, Utah).

The transcriptome was assembled using the Iso-Seq reads and the IsoSeq v3 pipeline from the PacBio SMRT® Tools software. The Iso-Seq reads were aligned to the Hi-C scaffolded assembly using the pbmm2 pipeline, another tool from SMRT® Tools. Lastly, the transcripts were collapsed using IsoSeq v3.

Repeat analysis and gene annotation

RepeatModeler2 v.2.0.1 (Flynn et al., 2020) identified novel repeats in the assembled genome and RepeatMasker v.4.1.2 (Smit et al., 2013) classified the identified repeats using the RepBase/RepeatMasker database. MAKER v.2.31.0 (Holt & Yandell, 2011; Bowman et al., 2017) was used to annotate the final assembly in conjunction with AUGUSTUS (Stanke & Morgenstern, 2005) *ab initio* gene predications, the uniprot_sprot database from UniProtKB, sugar beet (Dohm et al., 2012) and quinoa protein sequences from Jarvis et al. (2017) for expressed sequence tags (EST) and protein homology. Genome completeness of the Hi-C scaffold assembly was estimated using BUSCO v5 (Simão et al., 2015) and two orthologous gene sets, the Embryophyte (embryophyta_obd10) and Viridiplantae (viridiplantae_obd10). A circos plot of the assembled genome was created using Circa (https://omgenomics.com/circa) including chromosome sizes, gene density, GC content, and repeat distributions.

Genome comparison

Synteny plots between the coding sequences of *C. watsonii* and *C. pallidicaule*, a South American AA-genome diploid species, as well as quinoa were generated using CoGe SynMap (https://genomevolution.org/coge). DAGchainer (Haas et al., 2004) output file from the *C. watsonii* vs. *C. pallidicaule* SynMap with the MCScanX toolkit (Wang et al., 2012) generated a collinearity file which was visualized using SynVisio (Bandi & Gutwin, 2020).

Resequencing

Forty-one AA-genome *Chenopodium* diploid accessions, one HH-genome diploid, and three AABB tetraploid accessions from the germplasm collection at BYU (Table 2) were sterilized with 10% bleach, manually scarified, and germinated on filter paper in 9 cm petri plates. Samples were treated with 1ml of 30**µ**M potassium nitrate, 1ml of 100 ppm gibberellic acid, and sprayed with Hi-Yield Captan 50W Fungicide. Young leaf tissue from the established plants was collected, freeze-dried and DNA was extracted using a modified mini-salts extraction protocol (Supplement 2) (Todd & Vodkin, 1996; Dellaporta et al., 1983). Quality control parameters for concentration (<300 **µ**g/ml) and contamination (260/280 and 260/230 ≅ 2.0) were followed before sequencing. All DNA samples were sent to Novogene Corporation, Inc. (San Diego, California) for Illuming NovaSeq 6000 whole-genome sequencing with 10x coverage of 150-bp paired-end reads from a 500-bp insert library. The three tetraploid samples were previously sequenced by Jarvis et al. (2017).

Variant detection between taxa and phylogenetic relationship inferences

Raw data in FASTQ files, approximately 500 megabases per accession, were trimmed using the ILLUMINACLIP option from Trimmomatic v0.39 (Bolger et al., 2014) with the same parameters previously described. The trimmed reads were subsequently mapped to the *C. watsonii* reference genome using Minimap2 v2.17 (Li, 2018) with a minimum read coverage depth of two and a minimum allele frequency of 51%. The output SAM files were sorted, duplicate reads were removed using fixmate and markdup and filtered for quality (MAPQ > 45) using SAMtools v1.9 (Li, 2009). The filtered SAM files were subsequently converted to BAM files using the view tool from SAMtools.

Single nucleotide polymorphisms (SNPs) were identified using the BAM files and InterSnp, a program within BamBam v1.4 pipeline (Page et al., 2014) which produced a SimpleSNP file containing the nucleotides at each location of the genome for the proposed accessions in comparison to the reference genome. The SNPhylo v20160204 pipeline (Lee et al., 2014) removed low-quality data and filter representative SNPs. SNP sites with > 10% missing data, a minor allele frequency > 15%, and linkage disequilibrium > 30% were removed from the dataset. IQ-TREE (Nguyen et al., 2015) in conjunction with the PHYLIP SNP data set (Felsenstein, 1989) produced by SNPhylo was used to generate a phylogenetic tree based on maximum likelihood (ML) with a bootstrap of n=1000 and correcting for ascertainment bias using the +ASC option. SplitsTree5 (Huson, 1998) and FigTree v1.4.4 (Rambaut, 2010) were used for tree visualization.

Gene-based tree analysis

A phylogenetic gene tree of the separated sub-genomes from whole-genome assemblies of six species, three *Chenopodium* diploids (*C. pallidicaule, C. suecicum*, *C. watsonii*), two

Chenopodium polyploids (*C. formosanum* Koidz. (BBCCDD genome composition; Jarvis et al., 2022), *C. quinoa* (AABB genome composition; Jarvis et al., 2017)) and one *Atriplex* diploid (*Atriplex hortensis* L.; Hunt et al., 2020), generated using 1,600 single copy orthologous genes identified with BUSCO and aligned with MAFFT v7.490 (Katoh et al., 2002). ALISCORE (Misof & Misof, 2009) and AliCUT (Kueck, 2017) were used to remove regions of the alignments that were indistinguishable from random noise. The alignments were concatenated using the FASconCAT-G v1.11 software (Kueck & Longo, 2014), a tree was generated using IQ-TREE v2.1.3 (Nguyen et al., 2015) with n=1,000 bootstrap support and visualized with FigTree v1.4.4 (Rambaut, 2010).

RESULTS AND DISCUSSION

Whole-genome assembly of C. watsonii

Oxford Nanopore® long-read sequencing data produced 45.5 Gb included in 3.67 million reads. The N50 from the sequencing reads was 13,761 bp with an average read length of 12,383 bp with a range from 2 -190 kb, read coverage of 82x and average quality score of 13. The primary contig assembly contained 3,517 contigs with a total assembly length of 551.37 Mb. The contig N50 was 553.04 Mb with an L50 of 231. The contig assembly contained 3,520 gaps and 3,533 Ns (Table 3).

Chromatin-contact mapping using Hi-C data yielded nine chromosome-length pseudomolecule scaffolds, with a total of 1,700 scaffolds. The Hi-C-based data produced 48.9 Gb within 163.01 million read pairs (89x coverage). These pseudochromosomes correspond to the nine haploid (n=9) chromosomes of *C. watsonii* (Figure 2). The Hi-C chromosome-scale assembly produced a total genome length of 551.56 Mb and an N50 of 55.14 Mb with an L50 of

5 (Table 3). Each chromosome-scale scaffold contains a range of clustered contigs from 171 to 240, containing a total of 1,844 contigs (52.16%) in the nine largest scaffolds. Chromosome lengths vary from 64.46 Mbp to 54.81 Mbp, with an average length of 57.14 Mbp. The N count was 188 kb with 5,364 assembly gaps. Five contaminant unscaffolded reads of insect DNA from the Thysanoptera and Coleoptera orders were identified using Blobtools and subsequently removed; one of these reads was 29,766 bp while the others were less than 7,000 bp. The Blobplot produced by Blobtools shows an average GC content of 37% with read coverage averaging around 80x (Figure 3).

Repeat analysis and gene annotation

RepeatModeler identified 942,183 repetitive sequences, comprising 60.39% (330 Mb) of the assembled genome. RepeatMasker categorized the repetitive sequences as follows: DNA transposons made up 6.69% of the repetitive sequences; of these, 1.7% were classified as long interspersed elements (LINEs) and 32.66% were classified as long terminal repeats (LTRs), specifically *Copia* (14.93%) and *Gypsy* (17.55%) retrotransposons. In contrast, the *C. pallidicaule* genome was only 8.53% *Copia* elements, a significant decrease that is possibly due to incomplete assembly of the *C. pallidicaule* genome, which was based on Illumina short-read technology which is known to collapse across long repetitive, transposon-rich heterochromatic region of the genome, and is also likely reflected in the substantially smaller size of the *C. pallidicaule* genome (452 Mb) relative to the *C. watsonii* genome assembly (551 Mb). This hypothesis is supported by flow cytometric analyses, wherein Mandak et al. (2018) reported four AA diploids as having haploid genome sizes ranging from 597-637 Mb, along with the calculated size of the A-subgenome of quinoa (524 Mb; Jarvis et al., 2017). Another salient

component of the *C. watsonii* genome is the remaining interspersed 16.87% consisting of unclassified motifs. The high percentage of unclassified motifs is expected in a new species with little representation in the RepeatMasker database. Low-complexity elements, including simple sequence repeats (SSRs), microsatellites, and rRNA, comprised an additional 2.24% of the genome (Table 4).

Figure 4 provides a spatial distribution of key genetic elements along the nine chromosomes of *C. watsonii*. As expected, GC content (37.3%), *Gypsy* and *Copia* retroelement concentrations, and 12-13P centromeric repeats (Kolano et al., 2011) are elevated in the repeat-rich, gene-poor pericentromeric regions and are less abundant distally. Chromosomes 2, 3, 4, 6, 7, and 9 show clear peaks of telomeric sequence distributed at the ends of one or both arms. However, the telomeric sub-repeat track apparently shows redistribution of these sequences interstitially on chromosomes 1, 5, 8, and possibly 9, with chromosome 8 having two telomeric interstitial peaks: one close to the centromere and the other farther out in the chromosome arm. The 5S rDNA sequence located using BLASTn is found on Cw8 and is consistent with the location on Cp8 in *C. pallidicaule* (Kolano et al., 2011; Mangelson et al., 2019)

The MAKER program identified 30,725 gene models and 2,254 tRNA genes. The average gene length was 3,653 bp. Completeness was assessed using BUSCO with the Embryophyta and Viridiplantae BUSCO gene sets. The final assembly contained 1,569 (97.2%) complete clusters of orthologous genes (COGs), which included 1,515 (93.9%) single-copy and 54 (3.3%) duplicated COGs with the embryophyta_obd10 set. Similarly, 419 (98.6%) complete COGs, including 396 (93.2%) single-copy and 23 (5.4%) duplicated COGs were identified with the viridiplantae_odb10 gene set (Figure 2). The low duplication rate is expected for a diploid species, while the high detection rate of complete single copy COGs is indicative of a high-

quality and complete genome. Annotation quality was assessed using annotation edit distance (AED) which considers specificity, sensitivity, and accuracy of the annotation. Eighty-nine percent of the annotated genes had AED values < 0.50 with an overall mean AED value of 0.23, suggesting a high-quality annotation (Holt & Yandell, 2011).

Genome comparison and features

Synteny between *C. watsonii* and *C. pallidicaule* using the DAGChainer output generated by the SynMap feature from CoGe showed 16,521 syntenic coding sequences within 583 syntenic blocks, averaging 28 genes per block. A comparison of these potential internal telomeric sequences on the synteny and ribbon plots with the A-genome diploid *C. pallidicaule* (Figure 5) identified a potential subtelomeric paracentric inversion on Cw1 with telomeric inversions on Cw5 and Cw8 and a potential telomeric inversion on Cp3 that does not show up as an internal telomeric sequence on Cw3.

C. watsonii and *C. quinoa* shared a total of 32,334 syntenic coding sequences, 16,539 and 15,708 from the A- and B-subgenomes, respectively, with the remaining from unscaffolded contigs. The gene count averages 35 and 32 per syntenic block with 473 and 477 blocks per A- and B-subgenomes, respectively. The comparison of *C. watsonii* with *C. quinoa* subgenome A (Figure 6) identified a potential chromosome 4A telomeric inversion in *C. quinoa* in addition to the 1, 5, and 8 inversions of *C. watsonii*. Whether one or all of these internal telomere peaks represent terminal inversions on 1, 5, and 8, with an additional whole-arm inversion on chromosome 8, as opposed to other rearrangements or scaffolding errors, remains to be seen. Mechanisms besides inversion that lead to interstitial migration of telomeric sequences – a relatively common phenomenon in plants – include translocation, transposition, gene

amplification, etc. (Maravilla et al., 2021). Interstitial telomeric inversions have previously been ascribed to chromosome instability in microsatellite-enriched regions in yeast (Aksenova et al., 2013).

Phylogenetic analysis of Chenopodium A-genome diploids

Two phylogenetic trees of 41 New World AA-genome *Chenopodium* diploid accessions and the Eurasian HH-genome outgroup (*C. vulvaria* L.), one including the AA-subgenomes separated out from three AABB allotetraploids, were generated using the same SNP calls and IQ-Tree followed by visualized with two different software packages. SNP-based trees allow for inference of phylogenetic relationships and take into account linkage disequilibrium, though do not provide insight into evolutionary pressures that can be derived from gene-based trees (Boussau & Scornavacca, 2020; Heath et al., 2008). InterSNP called 1,010,399 SNPs across the mapped reads with >10% missing data and <51% heterozygosity due to the primarily autogamous reproductive system of *Chenopodium* species. SNPs were further filtered using SNPhylo with parameters of an LD threshold (0.3), minimum allele frequency <0.15 and a sliding window of 500,000 bp. The final data set fed into IQ-Tree included 10,588 SNPs, with an average of 1,176 SNPs per chromosome.

This analysis yielded robust bootstrap values with 90% of nodes having values >95% and resolved the set of AA diploids into eight arbitrarily assigned monophyletic subgroups with an additional clade housing the AA sub-genomes of the three accessions belonging to the ATGC. Below we note differences between the nine species groups and the accessions they house (Figure 8). While the goal was to survey most or all of the North American AA taxa, we were unable to obtain or include samples of *C. foggii* Wahl, *C. incanum* (S.Wats.) Heller, *C. lineatum*

Benet-Pierce, *C. luteum* Benet-Pierce, and *C. simpsonii* Benet-Pierce. We also included two samples of A-genome *C. pallidicaule* and five other putative AA diploids of South American origin to help determine geographic insularity of the North American species. Each AA subgroup is described systematically based on its order in the phylogenetic tree, from top to bottom (Figure 8).

South American Group (Group I)

Beginning at the top of the tree, all but one of the samples from South America formed a unified clade that was supported by a very high bootstrap value (>95%). As expected, domesticated Andean *C. pallidicaule* (PALL) samples grouped together and were closely related to a Pacific-slope sample of *C. carnosolum* Moq. (CARN 562) from over 3100 meters elevation on the Andean Cordillera Occidental of Tarapacá, Chile. Also grouping together were samples of *C. cordobense* Aell. (CORD 1748) and *C. petiolare* Kunth (PETI 1723) from Huascha, Córdoba Province, and Agua de las Palomas, Catamarca Province, Argentina, respectively. The samples in this group were BYU 1816-2, a sample of *C. albescens* Small (ALBE) from Laguna Salada in Brooks Co., Texas, which turned out to be very similar genetically and morphologically to *C. ruiz-lealii* Aell. (RUIZ 1749) from Chañar, La Rioja Province, Argentina.

Lejosperma-Leptophylla *Group 1 (Group II)*

The next group consisted of five narrow-leaved samples from North America and one from South America, designated for Aellen's Section *Chenopodia* Subsection *Lejosperma* that grouped strict-sense *Chenopodium* taxa having narrow leaves, mostly smooth testas, and mostly non-adhering pericarps (Aellen & Just, 1943) and Mosyakin & Clemants' (1996) designation of

these species as subsection *Leptophylla*. All of the samples falling in this genetic group except for BYU 1959 (*C. howellii* Benet-Pierce, HOWE, from Adel, Oregon) have non-adhering pericarps and are morphologically similar to *C. leptophyllum* (Moq.) Nutt. ex S. Wats. In the case of *C. howellii* there is an adhering pericarp (achene) and rugose testa. Sample NADH 835 (*C. desiccatum* A. Nels. from Elko, Nevada) was closest to HOWE. Accession NADH 20123 (*C. leptophyllum* from Colorado Springs) formed a group with NADH 1816-1 (*C. pratericola* Rydb. from Brooks Co., Texas), *C. papulosum* Moq. (PAPU 1755 from Matagusanos, San Juan Province, Argentina), and NADH 2073 (*C. pratericola* from Palo Pinto Co., Texas).

Atrovirens and "California Hians Aggregate" Group (Group III)

Containing several species recently reclassified by Benet-Pierce & Simpson (2019), this group of accessions included six taxa, all from California: *C. atrovirens* Rydb. (ATRO 1989 from Monitor Pass, Alpine Co.); *C. littoreum* Benet-Pierce & Simpson (LITT 1902, a prostrate psammophyte from coastal dunes in San Luis Obispo Co.); *C. aureum* Benet-Pierce (AURE 19111, 19136, 19140, all from the Sierra Nevada Mountains); *C. twisselmannii* Benet-Pierce (TWIS 19112 from the Kern River Plateau); *C. sandersii* Benet-Pierce (SAND 19291 from the San Gabriel Mountains); and *C. wahlii* Benet-Pierce (WAHL 19269, 19274, 19280, all from the Peninsular Ranges of Riverside Co. and San Diego Co.). The fruits of *C. atrovirens* are the only ones of this group having utricles rather than achenes.

Fremontii Group (Group IV)

The two samples *of C. fremontii* S. Wats., FREM 408 (San Gabriel Mts., California) and FREM 410 (Sierra Nevada Mts., California) grouped together into a single clade. This taxon has

large, warty to smooth seeds with non-adhering pericarps and broad triangular to ovate leaves with a distinct, earthy odor.

Lejosperma-Leptophylla *Group 2 (Group V)*

This group encompasses five taxa, four of which were classified previously in subsections *Lejosperma* or *Leptophylla*: *C. cycloides* A. Nels.*, C. nitens* Benet-Pierce & Simpson*, C. pallescens* Stand.*, C. standleyanum* Aell., and *C. subglabrum* (S. Wats) A. Nels. Our samples of *C. cycloides* (CYCL 2064 and 2067) were collected in the gypsiferous sand hills and along disturbed roadsides of the Permian Basin of West Texas. The taxon *C. pallescens* (PALE 2072, Eastland Co., Texas) is an episodic and apparently declining species that used to be widespread on disturbed, sandy tallgrass prairie (typical vegetation, *Andropogon gerardii*) soils from Northeast Texas and through Oklahoma, eastern Kansas, Missouri, eastern Nebraska, Iowa, and southern Illinois. The psammophytic species *C. subglabrum* (SUBG 2127, Seminoe Sand Dunes, Wyoming) is characterized by having very narrow leaves which range from fleshy to non-fleshy and are minimally farinose. The testa of the characteristically large seeds (~1.5mm) ranges from rugose to pitted with an adhering pericarp. Our sample of *C. standleyanum* (STAN 1310, Scott Co., Missouri) is from sandy oak-hickory woodlands of central North America. In contrast to these Great Plains species, *C. nitens* (NITE 20156, Mogollon Plateau, Arizona) characteristically grows on dry volcanic lake beds in *Pinus ponderosa* forests of western North America.

Cellulata-Favosa *Group (Group VI)*

Aellen & Just (1943) assigned alveolate, honeycombed, achene-fruited species to Sect. *Chenopodia* Subsect. *Cellulata* while Mosyakin & Clemants (1996) designated these as subsect. *Favosa.* This group of taxa similar morphologically to *C. neomexicanum* Stand. was expanded by Benet-Pierce & Simpson (2017) included seven diploid species: *C. arizonicum* Stand. (ARIZ 17238 from Arivaca, Arizona)*; C. lenticulare* Aell. (LENT 17152 from the Davis Mts. in West Texas)*; C. neomexicanum* (NEOM 869 from Coconino Co., Arizona)*, C. palmeri* Stand. (PALM 17231 from Arivaca, Arizona)*, C. sonorense* Benet-Pierce & Simpson (SONO 17220 from Tubac, Arizona)*,* and *C. watsonii* (WATS 873 from Yavapai Co., Arizona)*.* All of the samples in this group have the characteristic alveolate fruit with adhering pericarps (achenes) and leaves ranging from broadly elliptic to campanulate.

Hians Group (Group VII)

This group consisted of two samples of *C. hians* Stand.: HIAN 872 (Yavapai Co., Arizona) and HIAN 877 (Catron Co., New Mexico). This species is found mostly in mountainous terrain of the southwestern United States in and around the Colorado Plateau and is characterized by narrow, farinose, fleshy leaves. The fruits vary in appearance, having smooth testa and adhering to semi-adhering pericarps that are alveolate.

Nevadense Group (Group VIII)

Chenopodium nevadense Stand. grouped by itself in our tree. Found mainly in the sodic clay pans of the western Great Basin and valleys of the eastern Sierra Nevada Mountains, NEVA is a highly episodic taxon having fleshy, farinose leaves that are rhombate to ovate in shape. The

adherent pericarp is papillate and typically a pale white color (Standley, 1916). The sample included here, NEVA 816, was collected on the Soda Lakes Playa in Churchill Co., Nevada.

The midpoint-rooted tree (Figure 9) illustrates the relationships between WATS, PALL, and the ATGC. While PALL is somewhat closer genetically than WATS to the AABB group, both are more distant in comparison with the members of *Lejosperma-Leptophylla* Group II, particularly SUBG 2127.

Homoplasy in AA Chenopodium species

It is interesting to note that several key morphological characters appear in multiple clades, presumably due to convergent evolution. One obvious trait that apparently evolved in at least two lineages is narrow vs. broad leaf blades, presumably in response to hydric stress and/or as an adaptation to sandy soils. While all taxa in Groups II, III, V, VII, and VIII have narrow leaves, all the others in Groups I, IV, and VI plus VULV have broad leaf blades. The fact that all the taxa in Group V, which is most closely allied to the ATGC, have narrow leaves while all of the AABB tetraploids are broad-leaved suggests that this character might have been contributed by the B-genome ancestor, a rational assumption given that all three extant BB species – *C. ficifolium*, *C. suecicum*, and *C. ucrainicum* – also have broad leaves.

Chenopodium taxonomists have long considered the pericarp (fruit wall) as a paramount morphological trait, with species delineated into adhering (achene), semi-adhering, and non-adhering (utricle) forms (Benet-Pierce & Simpson, 2014; Mosyakin & Clemants, 1996). The *Lejosperma* and *Cellulata* subsections proposed by Aellen & Just (1943) divded species based on seed coat texture and pericarp adherence, with *Lejosperma* housing taxa wtih smooth to wavy

seed coats and non-adhering pericarps and *Cellulata* housing taxa with pitted seeds and adhering pericarps. Mosyakin & Clemants (1996) further divided Subsection *Lejosperma* into *Leptophylla, Chenopodium, Fremontiana,* and *Standleyana* based on additional morphological characteristics including leaf, seed and plant morphoogy. In our phylogenetic analysis, however, pericarp morphology was homoplastic. In Group I, all samples except CARN 562 had non-adhering pericarps. In Group II, the same was true for all samples except HOWE 1959. In Group III, ATRO 1989 and LITT 1902 were the only samples with non-adhering pericarps, although SAND 19291 and possibly TWIS 19112 have semi-adhering pericarps (Benet-Pierce & Simpson, 2019). Group V was a mixture of adhering (CYCL, PALE 2072) and non-adhering (NITE 20156, STAN 1310, SUBG 2127) samples. The Group IV FREM samples were both non-adhering while all the samples in Groups VI, VII, and VIII had adhering pericarps. Wentland (1965) described the adhering pericarp trait in *C. album* L. as being associated with enhanced seed dormancy. Based on our analysis, this is a trait that has been under strong selective pressure and its consideration as a key species-delineation trait should be reconsidered.

North-South reciprocal long-range dispersal

The grouping of Argentine Pampa sample PAPU with Group II from North America indicates the potential for an ancient north-south intercontinental dispersal event. Similarly, the placement of the Texas endemic species *C. albescens* squarely amid the South American Group containing CORD, PETI, PALL, and RUIZ suggests a reciprocal south-north dispersal to Texas. On an April, 2018 collection expedition to South Texas, our group collected seed from ten populations of ALBE spread across Brooks, Dimmit, Duval, Jim Hogg, Karnes, La Salle, and Webb Counties, indicating this is a well-established species between San Antonio and the Rio

Grande Valley (Jellen et al., 2019). Cruden (1966) provided an overview of seed dispersal via avian migration, postulating that bird populations carry seeds, stuck to the mud on wings and feet, by "mountain-hopping" to and from South American via Central America. These data suggest that migrating birds following the Central Flyway could have carried *Chenopodium* seeds back and forth between the temperate climates of North and South America at some point, or repeatedly, in antiquity.

Gene-based tree analysis

Using all current whole-genome *Chenopodium* assemblies, the COG-based tree showed distinct groupings of the four different *Chenopodium* sub-genomes with *Atriplex hortensis* L. (Hunt et al., 2020) as the outgroup. IQ-Tree and an input matrix of 618,448 sites, including 28,912 parsimony-informative and 73, 787 singleton sites from within 1,600 single copy orthologous genes identified with BUSCO from the embryphyta_obd10 gene set, generated a high-quality tree backed up by 100% bootstrap support across all nodes. A COG-based anaylsis allows for the inference of relationships based on evolutionary time, data that cannot be inferred from a SNP-based phylogeny. Based on the assumption that all genes evolve similarly, gene-based trees do not consider hybridization, gene conversion or gene transfer (Boussau & Scornavacca, 2020; Heath et al., 2008). Within the three BB-genome accessions, the BB sub-genome of the Taiwanese species *C. formosanum* Koidz. falls closest to the Eurasian BB diploid, *C. suecicum*, with the BB sub-genome of quinoa being the root of the BB genome group. The CC sub-genome of *C. formosanum* is the closest relative to the BB genome group, followed by the DD sub-genome of *C. formosanum*. The AA-genome accessions form a separate group from the other sub-genomes, with AA-diploid *C. pallidicaule* neighboring the AA sub-genome of quinoa

and *C. watsonii* rooting the AA-genome group (Figure 9). This contradicts the initial analyses based on read-mapping from Jellen et al. (2019) that showed *C. watsonii* as a closer relative to *C. quinoa* than *C. pallidicaule*, perhaps due to the use of read-mapping percentages rather than SNPs. Additional high-quality assembled genomes from North and South American AA-genome diploid *Chenopodium* species are needed to provide more evidence regarding which AA-genome diploid is the closest extant relative to the ancestor that gave rise to the ATGC.

CONCLUSIONS

We present a chromosome scale whole-genome assembly of *C. watsonii* and new phylogenetic evidence of *Chenopodium* AA-genome diploid relationships, producing eight distinct clades housing thirty AA-genome diploid species. The *C. watsonii* reference genome provides a new genetic resource for understanding the North American *Chenopodium* AA-genome species. We also find evidence to support a north-south reciprocal dispersal of *Chenopodium* germplasm between the continents of the western hemisphere. We propose adjustments to the current taxonomic subsections and the continuation of assembling whole genomes of *Chenopodium* based on our results, allowing for greater understanding of the evolutionary development of *Chenopodium* species, particularly those carrying the AA genome.

LITERATURE CITED

Aellen, P. (1960). Chenopodiaceae. (In German). In Rechinger, K.H. Hegis (Ed.) *Illustrierte Flora von Mitteleuropa, 3(2)*, 533–762.

Aellen, P. & Just, T. (1943). Key and synopsis of the American species of *Chenopodium* L. *The American Midland Naturalist, 30*, 47-76.

Aksenova, A. Y., Greenwell, P. W., Dominska, M., Shishkin, A. A., Kim, J. C., Petes, T. D., & Mirkin, S. M. (2013). Genome rearrangements caused by interstitial telomeric sequences in yeast. *Biological Sciences, 110(49),* 19866-19871. https://doi.org/10.1073/pnas.1319313110

Azurita-Silva, A., Jacobsen, S., Razzaghi, F., Alvarez-Flores, R., Ruiz, K. B., Morales, A., & Silva, H. (2015). Quinoa drought responses and adaptation. Chapter 2.4 In: FAO and CIRAD. *State of the art report of Quinoa in the world in 2013* (pp. 157-171). https://doi.org/10.13140/RG.2.1.4294.2565

Bandi, V. K. & Gutwin, K. (2020). Interactive exploration of genomic conservation. In *Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20)*. Canadian Human-Computer Communications Society, Waterloo, CA. https://synvisio.github.io/

Benet-Pierce, N. & Simpson, M. G. (2010). *Chenopodium littoreum* (Chenopodiaceae), a new goosefoot from dunes of South-Central coastal California. *Madrono, 57(1)*, 64-72. https://doi.org/10.3120/0024-9637-57.1.64

Benet-Pierce, N. & Simpson, M. G. (2014). The taxonomy of *Chenopodium desiccatum* and *C. nitens*, sp. nov. *Journal of the Torrey Botanical Society, 141(2)*, 161-172. https://doi.org/10.3159/TORREY-D-13-00046.1

Benet-Pierce, N. & Simpson, M. G. (2017). Taxonomic recovery of the species in the Chenopodium neomexicanum (Chenopodiaceae) complex and description of Chenopodium sonorense sp. nov. *The Journal of the Torrey Botanical Society, 144(3)*, 339-356. https://doi.org10.3159/TORREY-D-16-00013.1

Benet-Pierce, N. & Simpson, M. G. (2019). The taxonomy of *Chenopodium hians*, *C. incognitum*, and ten new taxa within the narrow-leaved *Chenopodium* group in western North America, with special attention to California. *Madrono, 66(2)*, 56-75. https://doi.org/10.3120/0024-9637-66.2.56

Bhargava, A., Rana, T. S., Shukla, S., & Ohri, D. (2005). Seed protein electrophoresis of some cultivated and wild species of *Chenopodium*. *Biologia Plantarum, 49(4)*, 505-511. https://doi.org/10.1007/s10535-005-0042-5

Bhargava, A. & Srivastava, S. (2013). Quinoa: Botany, production, and uses. Boston, MA: CABI. https://search.lib.byu.edu/byu/record/lee.6913980

Biondi, S., Ruiz, K. B., Martínez, E. A., Zurita-Silva, A., Orsini, F., Antognoni, F., Dinelli, G., Marotti, I., Gianquinto, G., Maldonado, S., Burrieza, H., Bazile, D., Adolf, V. I., & Jacobsen, F. (2015). Tolerance to saline conditions. Chapter 2.3 In FAO and CIRAD. *State of the art report of Quinoa in the world in 2013* (pp. 143-156). https://doi.org/10.13140/RG.2.1.4294.2565

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics, 30(15),* 2114-20. https://doi.org/10.1093/bioinformatics/btu170

Boussau, B. & Scornavacca, C. (2020). Reconciling gene trees with species trees. In

Scornavacca, C., Delsuc, F., & Galtier, N. (Eds.) *Phylogenetics in the Genomic Era* (pp.3.2:1-3.2:23. https://hal.inira.fr/PGE

Bowman, M. J., Pulman, J. A., Liu, T. L., & Childs, K. L. (2017). A modified GC-specific MAKER gene annotation method revels improved and novel gene predictions of high and low GC content in *Orzya sativa*. *BMC Bioinformatics, 18,* 522. https://doi.org/10.1186/s12859-017-1942-z

Cain, M. L., Milligan, B. G., & Strand, A. E. (2000). Long-distance seed dispersal in plant populations. *American Journal of Botany, 87(9)*, 1217-1227. https://doi.org/10.2307/2656714

Cepeda-Cornejo, V., Brown, D. C., Palomino, G., de la Cruz, E., Fogarty, M., Maughan, P. J., & Jellen, E. N. (2016). Genetic variation of the granule-bound starch synthase I (*GBSSI*) genes in *waxy* and non-*waxy* accessions of *Chenopodium berlandieri* ssp. *nuttaliae* from Central Mexico. *Plant Genetic Resources: Characterization and Utilization, 14(1),* 57-66.

Clemants, S. E. & Mosyakin, S. L. (2004). *Chenopodium.* In Flora of North America Editorial Committee (Eds.). *Flora of North America North of Mexico* (Vol. 4, pp. 275-299). Oxford Univ. Press, New York & Oxford.

Crawford, D. J. & Julian, E. A. (1976). Seed protein profiles in the narrow-leaved species of *Chenopodium* of the western United States: Taxonomic value and comparison with distribution of flavonoid compounds. *American Journal of Botany, 63(3),* 302-308. https://doi.org/10.1002/j.1537-2197.1976.tb11815.x

De Coster, W., D'Hert, S., Schultz, D. T., Crutz, M., & Van Broeckhoven, C. (2018).

NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics, 34(15),* 2666-2669. https://doi.org/10.1093/bioinformatics/bty149

Dellaporta, S. L., Wood, J., & Hicks, J. B. (1983). A rapid DNA minipreparation: version II. *Plant Molecular Biology Reporter, 1*, 19-21. https://doi.org/10.1007/BF02712670

Dohm, J. C., Lange, C., Holtgräwe, D., Sörensen, T. R., Borchardt, D., Schulz, B., Lehrach, H., Weisshaar, B., & Himmelbauer, H. (2012). Palaeohexaploid Ancestry for Caryophyllales Inferred from Extensive Gene-Based Physical and Genetic Mapping of the Sugar Beet Genome (*Beta vulgaris*). *The Plant Journal, 70(3),* 528-540. https://doi.org/10.1111/j.1365-313X.2011.04898.x

Eisen, J. A. & Fraser, C. M. (2003). Phylogenomics: Intersection of evolution and genomics. *Science, 300(5626),* 1706-1707. https://doi.org/10.1126/science.1086292

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS, 117(17),* 9451-9457. https://doi.org/10.1073/pnas.1921046117

Felsenstein, J. (1995). PHYLIP- Phylogeny Inference Package (Version 3.57c). *Cladistics, 5(2),* 163-166.

Fuentes-Bazan, S., Mansion, G., & Borsch, T. (2012). Towards a species level tree of the globally diverse genus *Chenopodium* (Chenopodiacae). *Molecular Phylogenetics and Evolution, 62*, 359-374. https://doi.org/10.1016/j.ympev.2011.10.006

Gavrilets, S. & Losos, J. B. (2009). Adaptive radiation: contrasting theory with data. *Science, 323(5915),* 732-737. https://doi.org/10.1126/science.1157966

Giusti, L. (1970). El género *Chenopodium* (In Spanish) In Argentina: I. Números de cromosomas. *Darwiniana 16(1/2)*, 98-105.

Giusti, L. (1997). Fasc. 40, Chenopodiaceae Vent. (In Spanish) In *Flora Fanerogamica Argentina*. Programa PROFLORA (CONICET). Cordoba, Argentina.

Gremillion, K. J. (1993). Crop and weed in prehistoric Eastern North America: the *Chenopodium* example. *American Antiquity, 58(3)*, 496-509. https://doi.org/10.2307/282109

Gross, R., Koch, F., Malaga, I., de Miranda, A. F., Schoeneberger, H., & Trugo, L. C. (1989). Chemical composition and protein quality of some local Andean food sources. *Food Chemistry, 34(1),* 25-34. https://doi.org/10.1016/0308-8146(89)90030-7

Haas, B. J., Delcher, A. L., Wortman, J. R., & Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics, 20(18),* 3643-3646. https://doi.org/10.1093/bioinformatics/bth397

Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics & Evolution, 46(3),* 239-257. https://doi.org/10.3724/SP.J.1002.2008.08016

Holt, C. & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics, 12*, 491. https://doi.org/10.1186/1471-2105-12-491

Hunt, S. P., Jarvis, D. E., Larsen, D. J., Mosyakin, S. L., Kolano, B. A., Jackson, E. W., Martin, S. L., Jellen, E. N., & Maughan, P. J. (2020). A chromosome-scale assembly of the garden orach (*Atriplex hortensis* L.) genome using Oxford Nanopore sequencing. *Frontiers in Plant Science, 11*. https://doi.org/10.3389/fpls.2020.00624

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics, 14(1),* 68-73. https://doi.org/10.1093/bioinformatics/14.1.68

James, L. E. A. (2009). Quinoa (*Chenopodium quinoa* Willd.): composition, chemistry,

nutritional, and functional properties. *Advances in Food and Nutrition Research, 58*, 1-25. https://doi.org/10.1016/S1043-4526(09)58001-1

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., Ohyanagi, H., Mineta, K., Michell, C. T., Saber, N., Kharbatia, N. M., Rupper, R. R., Sharp, A. R., Dally, N. (14), Boughton, B. A., Woo, Y. H., Gao, G., Schijlen, E. G. W. M., Guo, X., ... Tester, M. (2017). The genome of *Chenopodium quinoa*. *Nature, 542*, 307-312. https://doi.org/10.1038/nature21370

Jarvis, D. E., Sproul, J. S., Navarro-Dominguez, B., Krak, K., Jaggi, K., Huang, Y. F., Huang, T. Y., Lin, T. C., Jellen, E. N., & Maughan, P. J. (2022). Chromosome-scale assembly of the hexaploidy Taiwanese goosefoot 'djulis' (*Chenopodium formosanum*). *Genome Biology and Evolution*.

Jellen, E. N., Jarvis, D. E., Hunt, S. P., Mangelson, H. H., & Maughan, P. J. (2019). New seed collections of North American pitseed goosefoot (*Chenopodium berlandieri*) and efforts to identify its diploid ancestors through whole-genome sequencing. *Ciencia e Investigación Agraria, 46(2)*, 187-196. https://doi.org/10.7764/rcia.v46i2.2150

Jellen, E. N., Kolano, B. A., Sederberg, M. C., Bonifacio, A., & Maughan, P. J. (2011). *Chenopodium*. In Kole, C. (Ed.) *Wild Crop Relatives: Genomic and Breeding Resources: Legume Crops and Forages* (pp 35-61). Springer-Verlag. https://doi.org/10.1007/978-3-642-14387-8_3

Kamal, N., Renhuldt, N. T., Bentzer, J., Gundlach, H., Haberer, G., Juhasz, A., Lux, T., Bose, U., Tye-Din, J. A., Lang, D., van Gessel, N., Reski, R., Fu, Y.-B., Spegel, P., Ceplitis, A., Himmelbach, A., Waters, A. J., Bekele, W. A., Colgrave, M. L., ... Sirijovski, N. (2022).

The mosaic oat genome gives insights into a uniquely healthy cereal crop. *Nature.* https://doi.org/10.1038/s41586-022-04732-y

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transformation. *Nucleic Acids Research, 30(14)*, 3059-3066. https://doi.org/10.1093/nar/gkf426

Khoury, C. K., Carver, D., Greene, S. L., Williams, K. A., Achicanoy, H. A., Schori, M., León, B., Wiersema, J. H., & Frances, A. (2020). Crop wild relatives of the United States require urgent conservation action. *Proceedings of the National Academy of Sciences of the United States of America, 117(52),* 33351-33357. https://doi.org/10.1073/pnas.2007029117

Kistler, L. & Shapiro, B. (2011). Ancient DNA confirms a local origin of domesticated chenopod in eastern North America. *Journal of Archeological Science, 38(12),* 3549-3554. https://doi.org/10.1016/j.jas.2011.08.023

Kolano, B., Gardunia, B. W., Michalska, M., Bonifacio, A., Fairbanks, D., Maughan, P. J., Coleman, C. E., Stevens, M. R., Jellen, E. N., & Maluszynska, J. (2011). Chromosomal localization of two novel repetitive sequences isolated from the *Chenopodium quinoa* Willd. genome. *Genome, 54,* 710-717. https://doi.org/10.1139/G11-035

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

Kueck, P. (2017). AliCUT v2.31 https://github.com/PatrickKueck/AliCUT

Kueck, P. & Longo, G. C. (2014). FASconCAT-G: extensive functions for multiple sequence

alignment preparations concerning phylogenetic studies. *Frontiers in Zoology, 11,* 81. https://doi.org/10.1186/s12983-014-0081-x

La Duke, J. C. & Crawford, D. J. (1979). Character compatibility and phyletic relationships in several closely related species of *Chenopodium* of the western United States. *Taxon, 28(4),* 307-314. https://doi.org/10.2307/1219738

Laetsch, D. R. & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research, 6,* 1287. https://doi.org/10.12688/f1000research.12232.1

Lanfear, R., Schalamun, M., Kainer, D., Wang, W., & Schwessinger, B. (2018). MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics, 35(3),* 523-525. https://doi.org/10.1093/bioinformatics/bty654

Lee, T., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics, 15*, 162. https://doi.org/10.1186/1471-2164-15-162

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25(15),* 2078-2079. https://doi.org/10.1093/bioinformatics/bpt352

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34(18),* 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Mandák, B., Krak, K., Vít, P., Lomonosova, M. N., Belyayev, A., Habibi, F., Wang, L., Douda, J., & Štorhová, H. (2018). Hybridization and polyploidization within the *Chenopodium album* aggregate analysed by means of cytological and molecular markers. *Molecular Phylogenetics and Evolution, 129,* 189-201. https://doi.org/10.1016/j.ympev.2018.08.016

Mangelson, H., Jarvis, D. E., Mollinedo, P., Rollano-Penaloza, O. M., Palma-Encinas, V. D., Gomez-Pando, L. R., Jellen, E. N., & Maughan, P. J. (2019). The genome of *Chenopodium pallidicaule*: An emerging Andean super grain. *Applications in Plant Sciences, 7(11),* e11300. https://doi.org/10.1002/aps3.11300

Maravilla, A. J., Rosato, M., & Rosselló, J. A. (2021). Interstitial Telomeric-like Repeats (ITR) in seed plants as assessed by molecular cytogenetic techniques: a review. *Plants, 10,* 2541. https://doi.org/10.3390/plants10112541

Martínez, E. A. (2015). Quinoa: Nutritional aspects of the rice of the Incas. Chapter 3.4 In FAO and CIRAD *State of the art report of Quinoa in the world in 2013* (pp. 278-285). https://doi.org/10.13140/RG.2.1.4294.2565

Maughan, P. J., Chaney, L., Lightfoot, D. J., Cox, B. J., Tester, M., Jellen, E. N., & Jarvis, D. E. (2019). Mitochondrial and chloroplast genomes provide insights into the evolutionary origins of quinoa (*Chenopodium quinoa* Willd.). *Scientific Reports, 9*, 185. https://doi.org/10.1038/s41598-018-36693-6

Misof, B. & Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology, 58(1),* 21-34. https://doi.org/10.1093/sysbio/syp006

Mosyakin, S. L. & Mandák, B. (2020). *Chenopodium ucrainicum* (*Chenopodiaceae/ Amaranthaceae* sensu APG), a new diploid species: a morphological description and pictoral guide. *Plant Taxonomy, Geography and Floristics, 77(4),* 237-248.

Mosyakin, S. L. & Clemants, S. E. (1996). New infrageneric taxa and combinations in *Chenopodium* L. (Chenopodiaceae). *Novon: A Journal for Botanical Nomenclature, 6(4)*, 398-403. https://doi.org/10.2307/3392049

Murphy, K. M., Matanguihan, J. B., Fuentes, F. F., Gómez-Pando, L. R., Jellen, E. N., Maughan, P. J., & Jarvis, D. E. (2019). Quinoa breeding and genomics. *Plant Breeding Reviews, 42*, 257-320. https://doi.org/10.1002/9781119521358.ch7

Nelson, A. (1902). Contributions from the Rocky Mountain Herbarium. IV. *Botanical Gazette, 34(5),* 363. https://doi.org/10.1086/328298

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution, 32,* 268–274. https://doi.org/10.1093/molbev/msu300

Page, J. T., Liechty, Z. S., Huynh, M. D., & Udall, J. A. (2014). BamBam: genome sequence analysis tools for biologists. *BMC Research Notes 7*, 829. https://doi.org/10.1186/1756-0500-7-829

Plucknett, D. L., Smith, N. J. H, Williams, J. T., & Murthi Anishetty, N. (1983). Crop germplasm conservation and developing countries. *Science 220(4593),* 163-169. https://doi.org/10.1126/science.220.4593.163

Rambaut, A. (2010). FigTree v1.3.1. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. http://tree.bio.ed.ac.uk/software/figtree/

Rastrelli, L., De Simone, F., Schettino, O., & Dini, A. (1996). Constituents of *Chenopodium pallidicaule* (cañihua) seeds: Isolation and characterization of new triterpene saponins. *Journal of Agriculture and Food Chemistry, 44(11),* 3528-3533. https://doi.org/10.1021/jf950253p

Rydberg, P. A. (1912). *Chenopodium pratericola. Bulletin of the Torrey Botany Club, 39(7)*, 310-311.

Schluter, D. (1996). Ecological causes of adaptive radiation. *The American Naturalist, 148*, S40-S64. https://doi.org/10.1086/285901

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

Smit, A. F. A. & Hubley, R. (2008). RepeatModeler Open-1.0. http://www.repeatmasker.org

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

Smith, B. D. (2006). Eastern North America as an independent center of plant domestication. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 12223-12228. https://doi.org/10.1073/pnas.0604335103

Smith, B. D. & Yarnell, R. A. (2009). Initial formation of an indigenous crop complex in eastern North America at 3800 B.P. *Proceedings of the National Academy of Sciences of the United States of America, 106(16),* 6561-6566. https://doi.org/10.1073/pnas.0901846106

Standley, P. C. (1916). *Chenopodiaceae.* In *North American Flora,* Vol. 21, part 1 (pp. 1-93). The New York Botanical Garden.

Stanke, M. & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research, 33,* W465-W467. https://doi.org/10.1093/nar/gki458

Štorchová, H., Drabešová, J., Cháb, D., Kolář, J., & Jellen, E. N. (2015). The introns in

FLOWERING LOCUS T-LIKE (*FTL*) genes are useful markers for tracking paternity in tetraploid *Chenopodium quinoa* Willd.. *Genetic Resources and Crop Evolution, 62,* 913-925. https://doi.org/10.1007/s10722-014-0200-8

Subedi, M., Neff, E., & Davis, T. M. (2021). Developing *Chenopodium ficifolium* as a potential B genome diploid model system for genetic characterization and improvement of allotetraploid quinoa (*Chenopodium quinoa*). *BMC Plant Biology, 21,* 490. https://doi.org/10.1186/s12870-021-03270-5

Tapia, M.E., Mujica, S., & Canahua, A. (1980). Origen distribución geográfica y sistemas de producción en quinua. (In Spanish). *Primera Reunion sobre Genética y Fitomejoramiento de la Quinua, A1–A8*. Puno, Peru: Proyecto PISCA/UNTA/IBTA/IICA/CIID.

Todd, J. J. & Vodkin, L. O. (1996). Duplications that express and deletions that restore expression from a chalcone synthase multigene family. *The Plant Cell, 8(4),* 687-699. https://doi.org/10.1105/tpc.8.4.687

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

Walsh, B. M., Adhikary, D., Maughan, P. J., Emshwiller, E., & Jellen, E. N. (2015). *Chenopodium* polyploidy inferences from *Salt Overly Sensitive 1* (*SOS1*) data. *American Journal of Botany, 102(4),* 533-543. https://doi.org/10.3732/ajb.1400344

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research, 40(7),* e49. https://doi.org/10.1093/nar/gkr1293

WCVP (2022). *Chenopodium ficifolium* Sm. World Checklist of Vascular Plants, version 2.0. https://wcvp.science.kew.org

Wentland, M. J. (1965). The effect of photoperiod on the seed dormancy of *Chenopodium album.* Available from ProQuest Dissertations & Theses Global. (302183605). https://www.proquest.com/dissertations-theses/effect-photoperiod-on-seed-dormancy-chenopodium/docview/302183605/se-2?accountid=4488

Wick, R. R. (2018). Porechop v0.2.4 https://github.com/rrwick/Porechop

Wilson, H. D. (1980). Artificial hybridization among species of *Chenopodium* sect. *Chenopodium*. *Systematic Botany, 5*, 253-268. https://doi.org/10.2307/2418372

Wilson, H. D. (1981). Genetic variation among South American populations of tetraploid *Chenopodium* sect. *Chenopodium* subsect. *Cellulata*. *Systematic Botany, 6*, 380-398. https://doi.org/10.2307/2418450

Wilson, H. D. (1988). Allozyme variation and morphological relationships of *Chenopodium hircinum* (s.l.). *Systematic Botany, 13,* 215-228. https://doi.org/10.2307/2419100

Wilson, H. D. & Heiser, C. B. (1979). The origin and evolutionary relationships of 'Huauzontle' (Chenopodium nuttalliae Safford), domesticated chenopod of Mexico. *American Journal of Botany, 66(2),* 198-206. https://doi.org/10.1002/j.1537-2197.1979.tb06215.x

Wilson, H. & Manhart, J. (1993). Crop/weed gene flow: *Chenopodium quinoa* Willd. and *C. berlandieri* Moq. *Theoretical and Applied Genetics, 86*, 642-648. https://doi.org/10.1007/BF00838721

Wongsriphuek, C., Dugger, B. D., & Bartuszevige, A. M. (2008). Dispersal of wetland plant

seeds by mallards: influence of gut passage on recovery, retention, and germination. *Wetlands, 28(2),* 290-299. https://doi.org/10.1672/07-101.1

Wright, K. H., Huber, K. C., Fairbanks, D. J., & Huber, C. S. (2002). Isolation and characterization of *Atriplex hortensis* and sweet *Chenopodium quinoa* starches. *Cereal Chemistry, 79(5),* 715-719. https://doi.org/10.1094/CCHEM.2002.79.5.715

Wu, G. (2015). Nutritional properties of quinoa. In Murphy, K. & Matanguihan, J. (Eds.), *Quinoa: Improvement and sustainable production* (pp. 193-210). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118628041.ch11

Xie, M., Chung, C. Y., Li, M., Wong, F., Wang, X., Liu, A., Wang, Z., Leung, A. K., Wong, T., Tong, S., Xiao, Z., Fan, K., Ng, M., Qi, X., Yang, L., Deng, T., He, L., Chen, L., Fu, A. ... Lam, H. T. (2019). A reference-grade wild soybean genome. *Nature Communications, 10,* 1216. https://doi.org/10.1038/s41467-019-09142-9

Zurita-Silva, A., Fuentes, F., Zamora, P., Jacobsen, S., & Schwember, A. R. (2014). Breeding quinoa (*Chenopodium quinoa* Willd.): potential and perspectives. *Molecular Breeding, 34*, 13-30. https://doi.org/10.1007/s11032-014-0023-5

The 100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., Bakker, F., Dirks, R., Breit, T., Gravendeel, B., Huits, H., Struss, D., Swanson-Wagner, R., van Leeuwen, H. ... Peters, S. (2014). Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *The Plant Journal, 80(1),* 136-148. https://doi.org/10.1111/tpj.12616

FIGURES

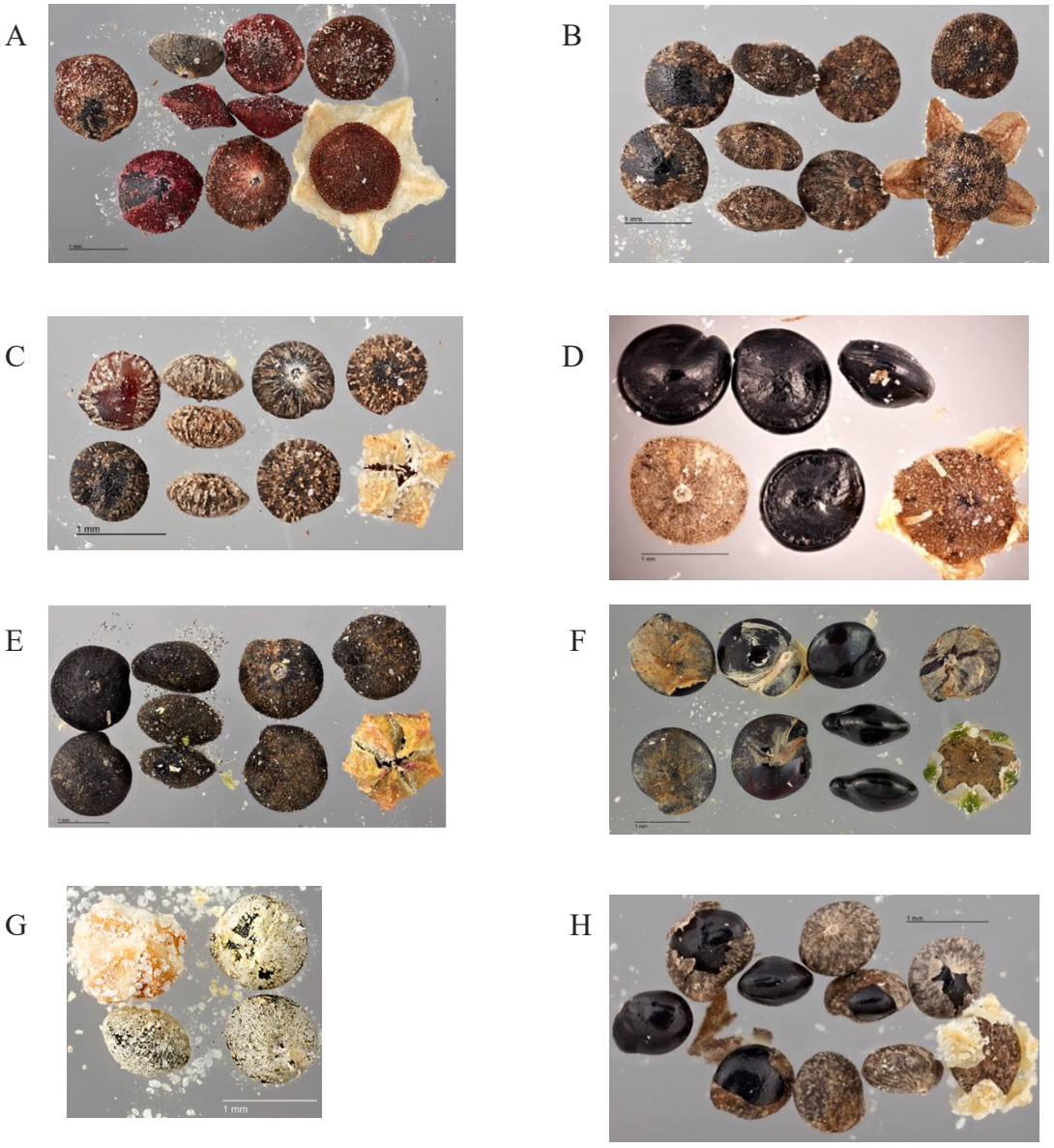


Figure 1: Morphological diversity of *Chenopodium* AA diploid fruits. A. *C. cycloides* BYU 2069; B. *C. neomexicanum* BYU 17178; C. *C. sonorense* BYU 17220; D. *C. fremontii* BYU 17245; E. *C. pallescens* BYU 2072; F. *C. subglabrum* BYU 2127; G. *C. watsonii* BYU 873; and H. *C. albescens* BYU 1811.

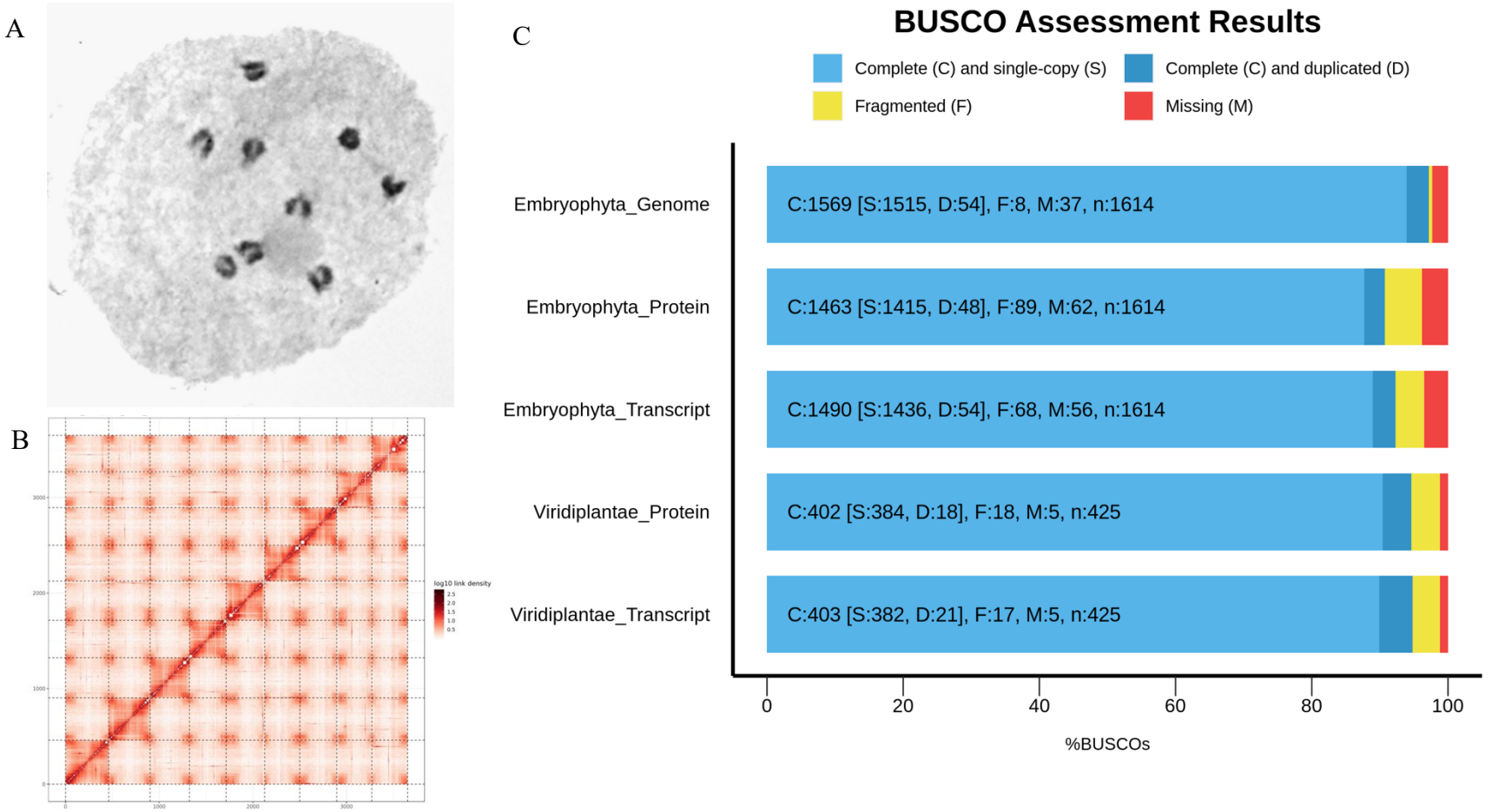


Figure 2: (A) *Chenopodium watsonii* chromosome pairs. Nine chromosome pairs forming bivalent rings in the diakinesis stage of prophase. (B) Hi-C linkage density heat map with nine distinct scaffolds. (C) BUSCO assembly statistics against the embryophyte and viridiplantae orthologous gene sets for the assembled genome, transcriptome, and protein annotation.

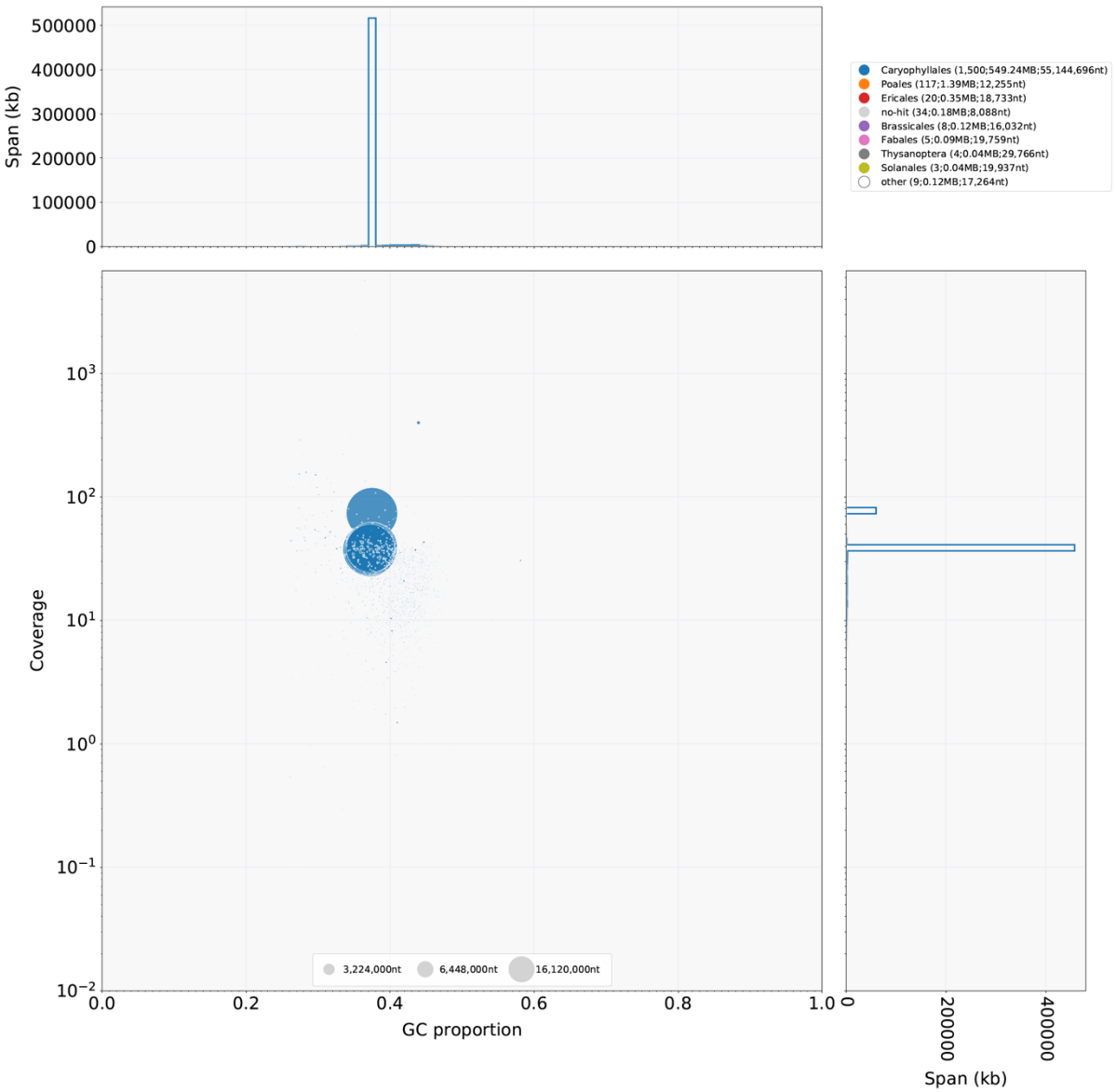


Figure 3: Bloblpot of read statistics in terms of coverage and GC content. Circles and dots represent chromosome-scale scaffolds and unscaffolded contigs with diameter scaled to sequence length and colored based on BLASTn taxonomic annotation.

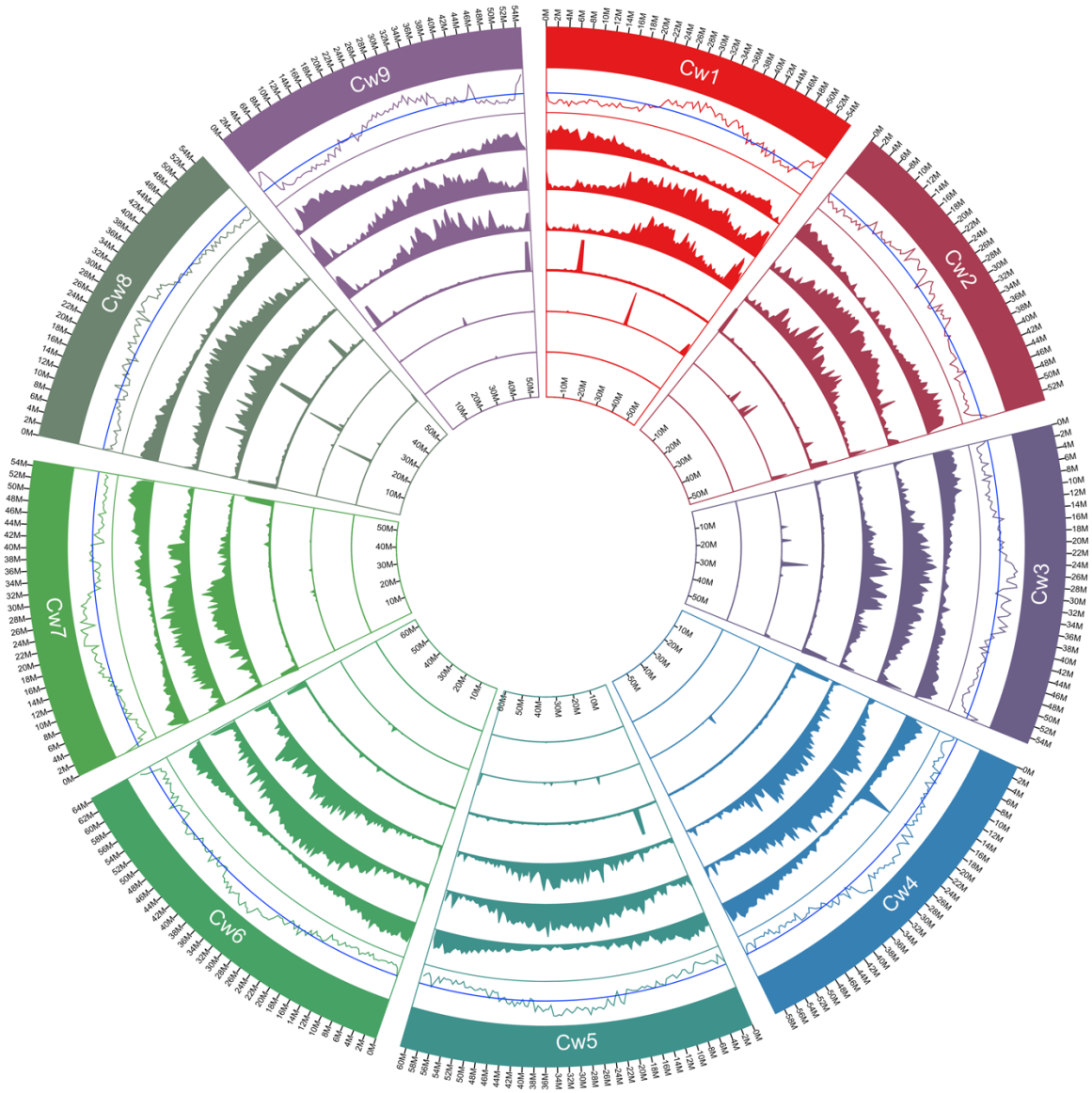


Figure 4: Genome overview of *C. watsonii* in 500 kb windows. Track 1 (outside): Chromosome and sizes; Track 2: GC content with mean (blue line = 37.3%; scale 33 – 43%); Track 3: Annotated gene density; Track 4: LTR-*Gypsy* distribution; Track 5: LTR-*Copia* distribution; Track 6: Telomeric sub-repeat distribution; Track 7: Centromere specific repeat (p12-13; reference) density; Track 8: 5S rRNA gene distribution.

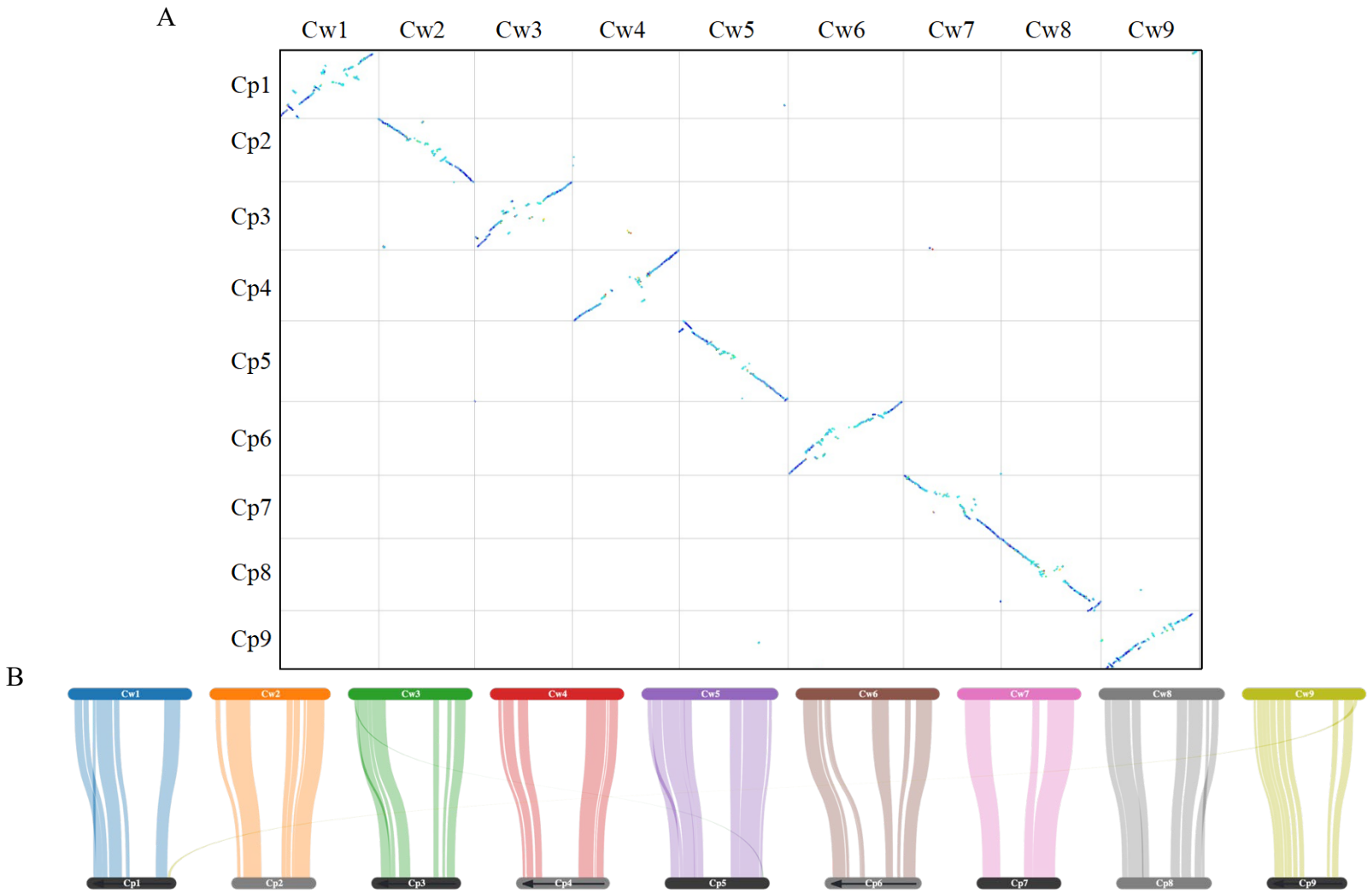
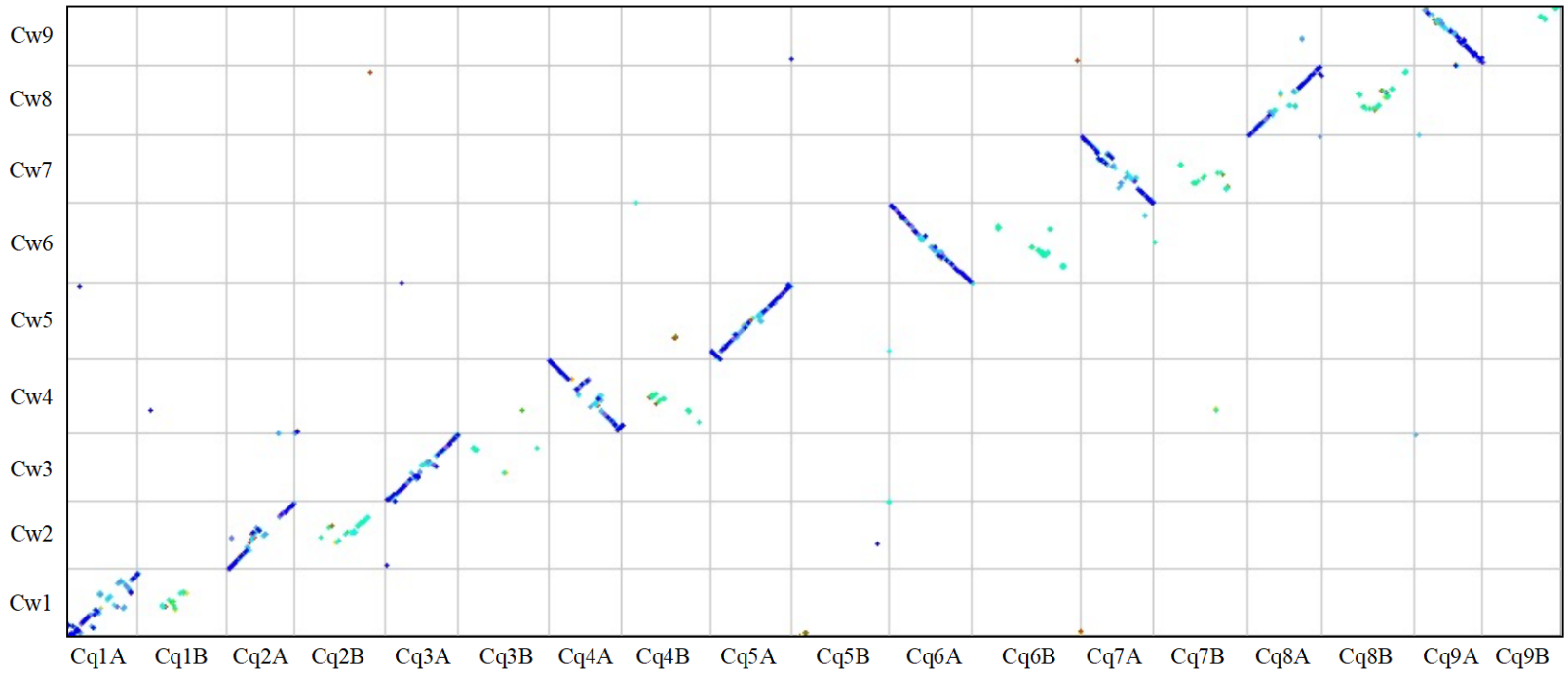


Figure 5: Genomic comparison with *C. pallidicaule*. (A) Synteny dot plot between *C. pallidicaule* (y-axis, Cp) and *C. watsonii* (x-axis, Cw); darker colors reflect high homology. (B) Ribbon plot between *C. pallidicaule* (bottom row, Cp) and *C. watsonii* (top row, Cw) pseudochromosomes; synteny between Cp1-9 and Cw1-9, respectively.

A



B

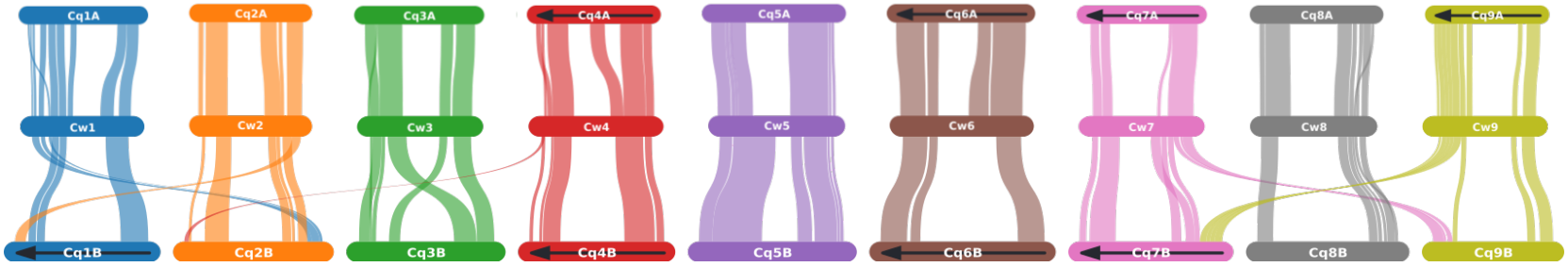


Figure 6: Genomic comparison with quinoa. (A) Synteny dot plot between quinoa (x-axis, Cq) and *C. watsonii* (y-axis, Cw); darker colors reflect high homology. (B) Ribbon plot between quinoa (Cq) and *C. watsonii* (Cw) pseudochromosomes; synteny between Cq 1A-9A (top), Cw1-9 (middle), Cq 1B-9B (bottom).

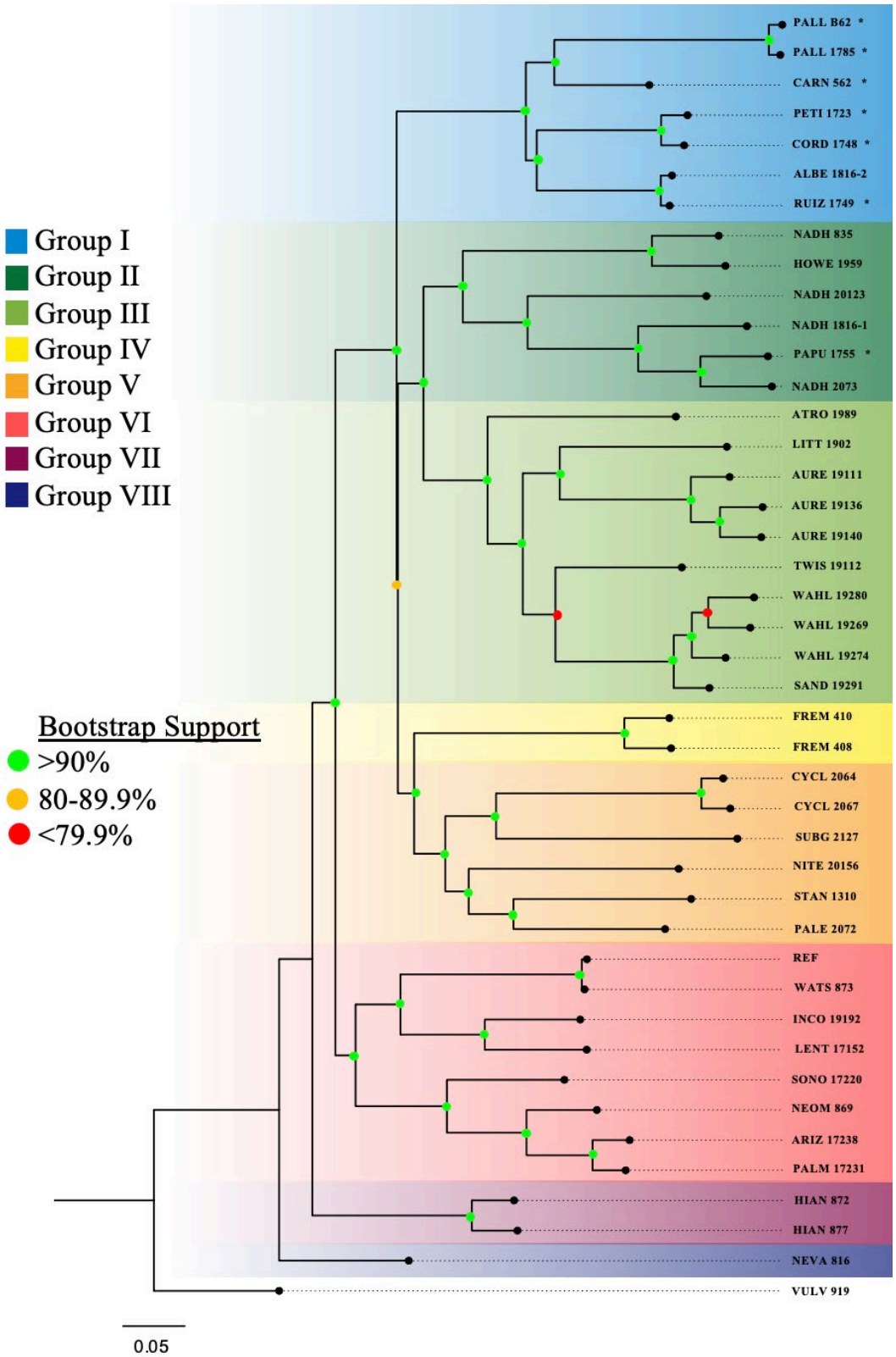


Figure 7: *C. vulvaria* (VULV 919) rooted tree visualized using FigTree. Bootstrap values by IQ-Tree are indicated by colored nodes based on SH-aLRT support values. Passport information is in Table 2.

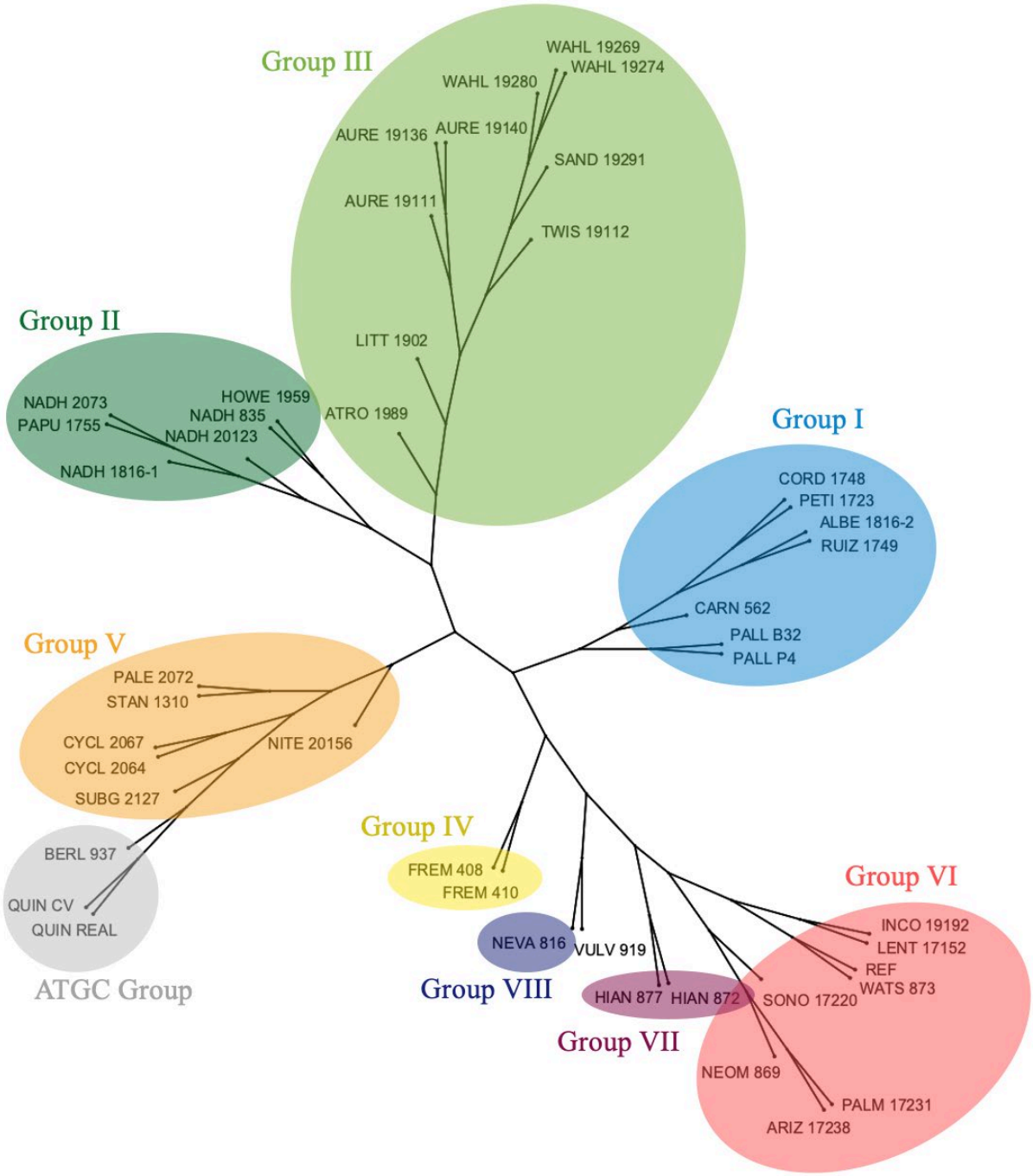


Figure 8: Midpoint rooted tree with colored clades and *C. vulvaria* (VULV 919) outgroup. Generated using SplitsTree and 10,588 SNPs after filtering using the following parameters: <10% missing data, minor allele frequency <15%, and linkage disequilibrium <30%.

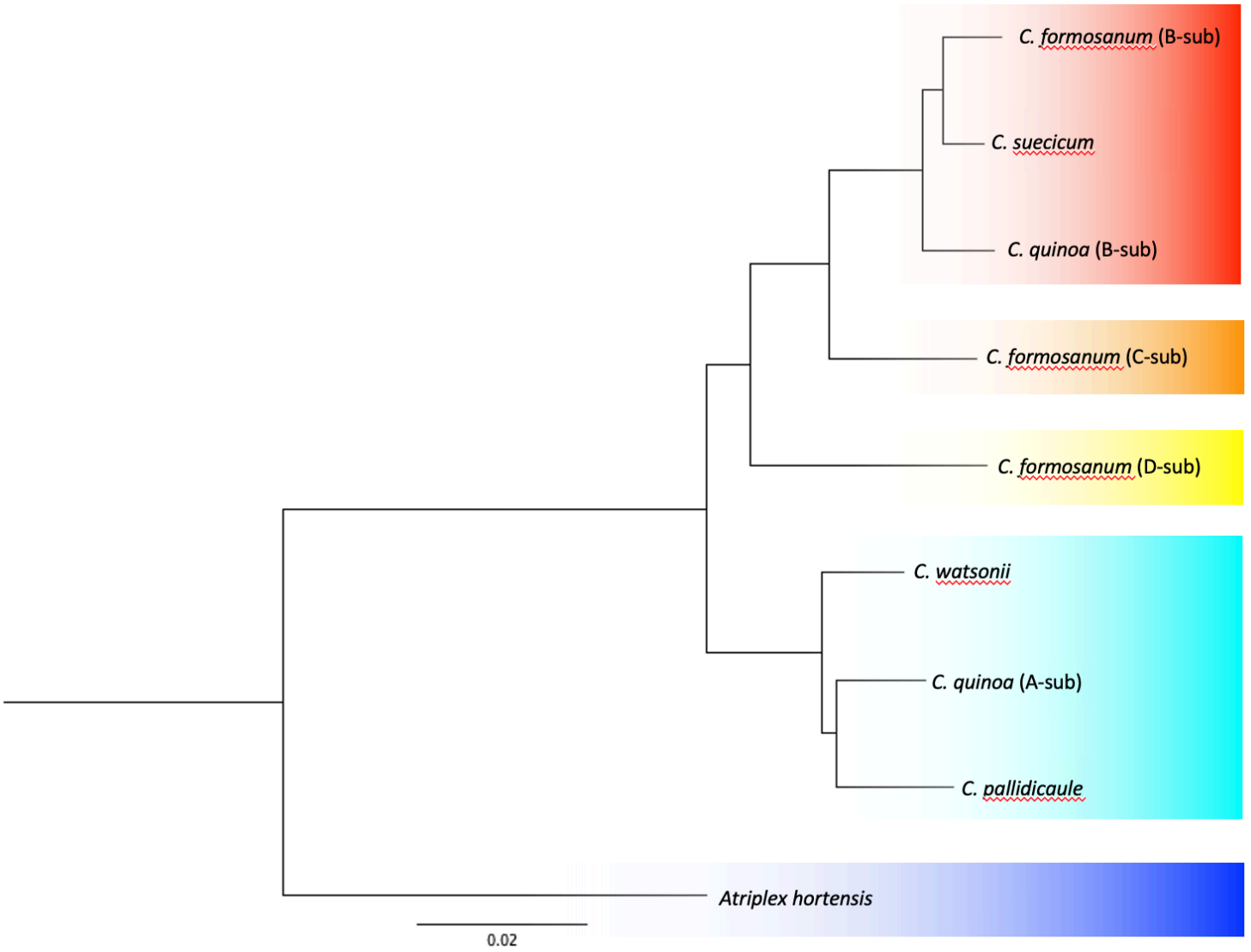


Figure 9: Gene-based tree of four *Chenopodium* sub-genomes (A-D) from five *Chenopodium* species and *A. hortensis* generated using 1,600 COGs.

TABLES

Table 1: Current *Chenopodium* AA-diploid taxonomy

Species	**Origin**	**Species**	**Origin**
C. albescens Small	Texas	*C. luteum* Benet-Pierce	California
C. arizonicum Stand.	North America	*C. neomexicanum* Stand.	North America
C. atrovirens Rydb.	North America	*C. nevadense* Stand.	North America
C. aureum Benet-Pierce	North America	*C. nitens* Benet-Pierce & Simpson	North America
C. brandegeeae Benet-Pierce	North America	*C. pallescens* Stand.	North America
C. bryoniifolium Bunge	Eurasia	*C. pallidicaule* Aell.	Peru-Boliva
C. carnosolum Moq.	Chile	*C. palmeri* Stand.	North America
C. cordobense Aell.	Argentina	*C. papulosum* Moq.	Argentina
C. cycloides A. Nels.	Texas	*C. parryi* Stand.	Mexico
C. desiccatum A. Nels.	California	*C. petiolare* Kunth	Argentina
C. eastwoodiae Benet-Pierce	California	*C. philippianum* Aell.	South America
C. flabellifolium Stand.	San Martin Island, Mexico	*C. pilcomayense* Aell.	Argentina
C. foggii Wahl.	Eastern North America	*C. pratericola* Rydb.	North America

C. fremontii S. Wats.	North America	*C. ruiz-lealii* Aell.	Argentina
C. hians Stand.	North America	*C. sandersii* Benet-Pierce	California
C. howellii Benet-Pierce	California	*C. scabricale* Speg.	Argentina
C. incanum (S. Wats.) Heller	North America	*C. sonorense* Benet-Pierce & Simpson	North America
C. incognitum Wahl	North America	*C. standleyanum* Aell.	North America
C. lenticulare Aell.	Texas	*C. subglabrum* A. Nels.	North America
C. leptophyllum (Moq.) Nutt. ex S. Wats.	North America	*C. twisselmannii* Benet-Pierce	California
C. lineatum Benet-Pierce	California	*C. wahlli* Benet-Pierce	California
C. littorium Benet-Pierce & Simpson	California	*C. watsonii* A. Nels.	North America

Table 2: Resequencing panel

Name	Accession	Species	Origin	Collection Location
ALBE 1816-2	BYU 1816-2	*C. albescens* Small	Texas	27.1569, -98.07140
ARIZ 17238	BYU 17238	*C. arizonicum* Stand.	North America	31.5897, -111.32100
ATRO 1989	BYU 1989	*C. atrovirens* Rydb.	North America	38.6717, -119.58686
AURE 19111	BYU 19111	*C. aureum* Benet-Pierce	North America	36.047496, -118.19806
AURE 19136	BYU 19136	*C. aureum* Benet-Pierce	North America	37.56043, -118.67159
AURE 19140	BYU 19140	*C. aureum* Benet-Pierce	North America	37.53527, -118.70678
BERL 937	BYU 937	*C. berlandieri* Moq.	North America	29.30524, -94.90580
CARN 562	BYU 562	*C. carnosulum* Moq.	Chile	-19.74031, -69.24114
CORD 1748	BYU 1748	*C. cordobense* Aell.	Argentina	-30.5828, -64.73270
CYCL 2064	BYU 2064	*C. cycloides* A. Nels.	Texas	31.778813, -103.33882
CYCL 2067	BYU 2067	*C. cycloides* A. Nels.	Texas	31.687546, -103.02716
FREM 408	BYU 408	*C. fremontii* S. Wats.	North America	34.37963, -117.70712
FREM 410	BYU 410	*C. fremontii* S. Wats.	North America	37.13243, -118.42768
HIAN 872	BYU 872	*C. hians* Stand.	North America	34.51477, -112.00698
HIAN 877	BYU 877	*C. hians* Stand.	North America	34.19945, -108.93878
HOWE 1959	BYU 1959	*C. howellii* Benet-Pierce	California	42.205015, -120.01561
INCO 19192	BYU 19192	*C. incognitum* Wahl	North America	38.806676, -104.85103
LENT 17152	BYU 17152	*C. lenticulare* Aell.	Texas	30.6911, -103.78910
LITT 1902	BYU 1902	*C. littoreum* Benet-Pierce & Simpson	California	35.0559, -120.60330
NADH 1816-1	BYU 1816-1	*C.* sp. NADH	North America	27.1569, -98.07140

NADH 20123	BYU 20123	*C.* sp. NADH	North America	38.799099, -104.73271
NADH 2073	BYU 2073	*C.* sp. NADH	North America	32.4987205, -98.52292
NADH 835	BYU 835	*C.* sp. NADH	North America	40.99821, -115.87011
NEOM 869	BYU 869	*C. neomexicanum* Stand.	North America	34.95468, -111.43585
NEVA 816	BYU 816	*C. nevadense* Stand.	North America	39.51815, -118.88073
NITE 20156	BYU 20156	*C. nitens* Benet-Pierce & Simpson	North America	35.476333, -112.01179
PALE 2072	BYU 2072	*C. pallescens* Stand.	North America	32.4987205, -98.52292
PALL B32	Bol-6.2	*C. pallidicaule* Aell.	Peru-Bolivia	-16.67403, -68.31833
PALL P4	BYU 1785	*C. pallidicaule* Aell.	Peru-Bolivia	-15.7693, -70.27050
PALM 17231	BYU 17231	*C. palmerii* Stand.	North America	31.7733, -111.46600
PAPU 1755	BYU 1755	*C. papulosum* Moq.	Argentina	-31.3342, -68.60660
PETI 1723	BYU 1723	*C. petiolare* Kunth	Argentina	-27.6202, -66.12180
QUIN CV	BYU 1439	*C. quinoa* Willd.	Central Chile	NA
QUIN REAL	BYU 1633	*C. quinoa* Willd.	Bolivia	NA
RUIZ 1749	BYU 1749	*C. ruiz-lealii* Aell.	Argentina	-30.544, -65.95670
SAND 19291	BYU 19291	*C. sandersii* Benet-Pierce	California	34.35966, -118.01100
SONO 17220	BYU 17220	*C. sonorense* Benet-Pierce & Simpson	North America	31.6104, -111.05120
STAN 1310	BYU 1310	*C. standleyanum* Aell.	North America	37.0103, -89.61000
SUBG 2127	BYU 2127	*C. subglabrum* A. Nels.	North America	41.101168, -106.93817
TWIS 19112	BYU 19112	*C. twisselmannii* Benet-Pierce	California	36.047496, -118.19806
VULV 919	BYU 919	*C. vulvaria* L.	Eurasia	NA
WAHL 19269	BYU 19269	*C. wahlii* Benet-Pierce	California	33.736208, -116.71421

WAHL 19274	BYU 19274	*C. wahlii* Benet-Pierce	California	33.31221, -116.86220
WAHL 19280	BYU 19280	*C. wahlii* Benet-Pierce	California	32.924204, -116.48198
WATS 873, REF	BYU 873	*C. watsonii* A. Nels.	North America	34.51477, -112.00698

NA indicates missing values

Table 3: Assembly statistics of *C. watsonii* primary contig and Hi-C scaffold assemblies.

Assembly Statistics		
	Primary	Hi-C
Assembly size (Mb)	551.37	551.56
Number of contigs/scaffolds, resp.	3,517	1,700
N50 (Mb)	553.04	55.14
L50	231	5
Longest (Mb)	5.17	64.48
N count	3,533	187,907
Gaps	3,520	5,338
N90 (Mb)	61.59	53.65
L90	1,399	9
Assembly % in Scaffold N90	--	93.3

Table 4: Repetitive element analysis of the scaffold assembly using RepeatMasker.

Repeat Class	Repeat Name	Count	Bases Masked	% Masked
DNA		2,258	524,299	0.10%
	CMC-EnSpm	21,692	11,690,656	2.13%
	CMC-Transib	1,253	1,466,009	0.27%
	MULE-MuDR	6,942	5,512,320	1.01%
	MuLE-MuDR	8,149	5,957,136	1.09%
	PIF-Harbinger	5,373	2,784,698	0.51%
	TcMar-Fot1	219	78,239	0.01%
	TcMar-Stoaway	21,494	4,132,608	0.75%
	TcMar-Tc1	895	80,610	0.01%
	TcMar-Tigger	47	9,894	0.00%
	hAT	153	57,918	0.01%
	hAT-Ac	10,643	3,530,341	0.64%
	hAT-Tag1	2,935	607,568	0.11%
	hAT-Tip100	757	246,607	0.05%
LINE		--	--	--
	CRE-II	745	1,033,284	0.19%
	Jockey	132	158,397	0.03%
	L1	8,829	7,282,308	1.33%
	L1-Tx1	505	82,039	0.01%
	RTE-BovB	4,246	772,162	0.14%
LTR		2,878	394,345	0.07%
	Caulimovirus	273	428,861	0.08%
	Copia	43,357	81,779,794	14.93%
	ERV1	277	144,991	0.03%
	Gypsy	74,329	96,125,885	17.55%
	Pao	53	4,042	0.00%
RC		--	--	--
	Helitron	1,749	1,028,213	0.19%
SINE		--	--	--
	tRNA	416	198,560	0.04%
Unknown		510,485	92,429,542	16.87%
Low-complexity		26,879	1,402,537	0.26%
Satellite		1,925	443,835	0.08%
	5S	2,645	431,306	0.08%
Simple repeat		179,290	9,745,492	1.78%
rRNA		360	234,795	0.04%
Total		942,183	330,799,291	60.39%

SUPPLEMENTAL MATERIAL

Supplement 1: Modified high molecular weight gDNA extraction protocol

Materials

1-1.5 g leaves
QIAGEN Genomic-tip 500/G
Carlson buffer, pre-warmed to 65°C:
- 100 mM Tris-HCl, pH 9.5
- 2% CTAB
- 1.4 M NaCl
- 1% PEG 8000
- 20 mM EDTA
- *To ensure all CTAB is dissolved, stir the Carlson buffer overnight

β-mercaptoethanol
Chloroform
Isopropanol
Rnase A (100mg/ml)
AMPure XP beads
Liquid nitrogen
Mortar and pestle
Vortex mixer
50 ml Falcon tubes
Centrifuge capable of taking 50 ml tubes
Water baths at 65°C, 55°C, and 50°C
QC buffer
QF buffer
G2 buffer
QBT buffer

Directions:

1. Transfer 20 ml of Carlson buffer to a 50 ml Falcon tube. In a fume hood, add 50 µl β-mercaptoethanol to the Carlson buffer, mix by vortexing and pre-warm to 65°C in a water bath.
2. Pre-cool the mortar and pestle with liquid nitrogen until both are at -80°C. This keeps the sample as a fine powder during grinding and will prevent re-activation of intracellular DNases.
3. Pour ~30 ml of liquid nitrogen into the mortar and add 1-1.5 g of leaves. When the liquid nitrogen has evaporated, grind the tissue for approximately 30 seconds, to a flour-like consistency. Keep the sample at the bottom of the mortar as much as possible. Add another ~30 ml of liquid nitrogen and repeat griding for approximately 30 seconds. Perform three cycles of grinding total.
4. Transfer the frozen powdered leaf tissue to the tube with the pre-warmed Carlson buffer. Add 40 µl of Rnase A and vortex the tube for 5 seconds. Immediately transfer the tube to a 65°C water bath and incubate for 1 hour, mixing the sample by inversion halfway through.

5. Let the sample cool to room temperature and add 1 volume of chloroform. Vortex the sample for two pulses of 5 seconds each.
6. Centrifuge the sample at 5500 g for 10 minutes at 4°C. Carefully transfer the top aqueous phase to a new 50 ml Falcon tube, without disturbing the interphase. Next, add 0.7 volumes of isopropanol to the top phase (e.g., for a sample volume of 18 ml, you will need 12.6 ml of isopropanol), and mix thoroughly by inverting the tube 10 times. Place the tube at -80°C for 15 minutes.
7. Centrifuge the sample at 5500 g for 30 minutes at 4°C. Carefully discard the supernatant, without disturbing the pellet. You can remove any remaining liquid by pressing the rim of the tube with a clean paper towel.
8. Carefully dissolve DNA pellet at 19 ml of G2 buffer. Do not vortex as it can fragment your DNA. Place the sample in a 50°C water bath for 15 minutes, mixing occasionally until the pellet dissolves. The protocol can be paused at this point and the sample kept at 4°C overnight.
9. Equilibrate a QIAGEN Genomic-tip 500/G column with 10 ml of QBT buffer.
   Apply your fully dissolved DNA in G2 buffer to the equilibrated QIAGEN Genomic-tip 500/G column. Allow the DNA to enter the resin by gravity flow.
10. Wash the QIAGEN Genomic-tip 500/G with 20 ml of QC buffer. Wait until all the buffer flows through the resin and repeat the wash with another 20 ml of QC buffer.
11. Place the QIAGEN Genomic-tip 500/G over a clean 50 ml Falcon tube and elute the genomic DNA with 15 ml of QF buffer, pre-warmed to 55°C.
    Allow the eluate to cool down to room temperature.
12. Precipitate the DNA by adding 0.7 volumes of room temperature isopropanol to the eluted DNA and mix by inverting the tube several times. Incubate at room temperature for 15 minutes.
13. Centrifuge at 5500 g for 30 minutes at 4°C and carefully remove the supernatant. Collect the remaining liquid by mopping the rim of the inverted tube with a clean paper towel.
14. Wash the centrifuged DNA pellet with 4 ml of cold 70% ethanol. Shake the tube several times to disturb the pellet and centrifuge at 5500 g for 10 minutes at 4°C.
15. Carefully remove the supernatant without disturbing the pellet. Collect all remaining ethanol by mopping the rim of the inverted tube with a clean paper towel. If small liquid droplets are still visible on the sides of the tube, we recommend carefully removing them with a clean, soft tissue, avoiding the area with the DNA pellet.
16. Air dry the pellet for 10 minutes and resuspend the DNA in 125 µl of TE buffer pH 7.5. Measure the concentration using the DNA BR Qubit assay and 260/280, 260/230 absorbance ratios with a spectrophotometer.
    At this point, the sample can be stored for up to a week at 4°C. For longer storage, freeze the DNA at -20°C.

Supplement 2: Modified mini-salts extraction protocol

Working Solution (10 samples)
Complete DNA extraction buffer (made prior to use):

Phenanthroline	16 mg
100% Ethanol (EtOH)	80 µl
Salts Buffer	8 ml
Sodium Lauryl Sulfate (SDS)	80 mg
2(β)-Mercaptoethanol	5.6 µl
Total volume	~10 ml

Directions:

1. Weigh phenanthroline (Sigma 9375) in a microcentrifuge tube and dissolve in the ethanol.
2. When dissolved, add it all to the appropriate amount of salts buffer (in a 50 ml centrifuge tube).
3. Add the SDS to the tube, cover and heat to 65°C in a water bath until dissolved.
4. Immediately before use, add the 2(β)-mercaptoethanol to the buffer under the fume hood.

Extraction Protocol:
(Turn on the water bath to 65°C for ~1 hour. Make the working extraction buffer (above). Don't attempt to do too many samples at once: 10 samples is a good number).

1. Samples should be freeze-dried and ground into powder before beginning the protocol. To grind the samples, use the shaker set to 5.0 speed for 20 seconds.
2. Add 600 µl of complete, warmed extraction buffer (make sure you added the 2(β)-mercaptoethanol) to each tube. Cap the tube.
3. Shake in the beat mixer for 4 seconds at 5.0 speed.
4. Immediately place the tube in the 65°C water bath for 12 minutes. Invert tube at the 4-minute and 8-minute marks to mix the sample.
5. Add ⅓ volume (~200 µl) of 5 M KOAc to the sample. Invert to mix well and place on ice for 20 minutes.
6. Centrifuge the samples for 10 minutes at >1400 rpm.
7. Carefully transfer the supernatant to the new/labeled tube (this can often be done by simply pouring the supernatant into the new tube).
8. Add an equal amount (600 µl) of pre-saturated phenol:chloroform solution. (In the phenol:chloroform there are two fluids - be sure to take from the lower fluid. The top fluid is a buffer protecting the phenol. Do not take this fluid.)
9. Invert the tube vigorously by hand two times each for 5 seconds and centrifuge at >1400 rpm for 5 minutes.
10. Carefully transfer the upper aqueous phase containing the DNA into a 1.5 ml tube labeled with the sample name (set the pipette to 500 µl).
11. Perform a Sevag extraction by adding an equal volume of Sevag solution (~500 µl of 24:1 chloroform:isoamyl alcohol) to the DNA solution.

12. Invert the tube vigorously by hand two times each for 5 seconds and centrifuge at >1400 rpm for 5 minutes.
13. Very carefully (using a pipette set to 400 **µ**l) transfer the upper aqueous phase to a new labeled tube. Avoid the interface!
14. Add equal volume of isopropanol (~400 **µ**l) into the tube. Cap the tube, wait ~1 minute and swirl the tube 10 times and then invert the tube 5 times to precipitate out the DNA.
15. Put the tubes in the -20°C freezer for ~2 hours. (This is a good stopping point as the samples can be left in the -20°C overnight.) It'll take 1 ½ - 2 hours depending on the number of samples to get to this point.
16. Centrifuge tube for 5 minutes at >1400 rpm. Gently pour off the supernatant in the aqueous waste and place the tube upside-down on a paper towel to wick off the remaining supernatant.
17. Wash the pellet with 500 **µ**l cold 70% EtOH. Finger-flick the tube to dislodge the pellet from the bottom of the tube and centrifuge at >1400 rpm for 5 minutes.
18. Remove the 70% EtOH by pouring the supernatant out and placing the tubes upside-down on a paper towel.
19. Dry the DNA pellet for 7 minutes in the Speed-Vac at 5.1 pressure (no heat).
20. Add 150-200 **µ**l of TeR, depending on how much DNA is present.
21. Finger-flick the sample to dislodge the pellet and place in a 37°C incubator for 30 minutes. *It is very important that the sample is fully resuspended in the TeR which may take some time.
22. Make sure the sample is labeled properly (can easily be read). Seal the sample with parafilm and store in a properly labeled box in the -80°C freezer.

APPENDIX

Appendix 1: Current Chenopodium taxonomy from Mosyakin and Clemants (1996)

Subsection	**Species**
Polysperma Stand.	*polyspermum* Kowal ex Mosy. & Clem.
Urbica (Stand.) Mosy. & Clem.	*urbicum* L.
Undata Aell. & Iljin ex Mosy. & Clem.	*murale* L.
Leptophyllum (Stand.) Clem. & Mosy.	*albescens* Small
	cycloides A. Nels.
	desiccatum A. Nels.
	foggii Wahl
	hians Stand.
	leptophyllum (Moq.) Nutt. ex S. Wats.
	pallescens Stand.
	pratericola Rydb.
	subglabrum (S. Wats.) A. Nels.
Fremontiana (Stand.) Clem. & Mosy.	*atrovirens* Rydb.
	fremontii S. Wats.
	incanum (S. Wats.) Heller
	cordobense Aell.
	ruiz-lealii Aell.

Favosa (Aell.) Mosy. & Clem.	*arizonicum* Stand.
	berlandieri Moq.
	bushianum Aell.
	ficifolium Smith
	hircinum Schrad.
	macrocalycium Aell.
	neomexicanum Stand.
	palmeri Stand.
	philippianum Aell.
	quinoa Willd.
	watsonii A. Nels.
Cicatricosa (Aell.) Mosy. & Clem.	*acerifolium* Andr.
	suecicum Murr
	karoi (Murr) Aell.
	jenissejense Aell. & Iljin
Standleyana Mosy. & Clem.	*badachschanicum* Tzvelev
	bryoniifolium Bunge
	gracilispicum Kung
	missouriense Aell.

	standleyanum Aell.
Chenopodium	*C. album* L.
	C. iljinii Gol.
	C. nidorosum Otsch.
	C. opulifolium Schrad. ex A. P. De Cand.
	C. pamiricum Iljin
	C. sosnowskii Kap.
	C. strictum Roth
	C. vulvaria L.