2019-05-01

The Genome of Cañahua: An Emerging Andean Super Grain

Hayley Jennifer Mangelson
Brigham Young University

The Genome of Cañahua: An Emerging Andean Super Grain

Hayley Jennifer Hansen Mangelson

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

Master of Science

Peter J. Maughan, Chair
Eric Jellen
David Jarvis
Brad Geary

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

The Genome of Cañahua: An Emerging Andean Super Grain

Hayley Jennifer Hansen Mangelson
Department of Plant and Wildlife Sciences, BYU
Master of Science

Chenopodium pallidicaule, known commonly as cañahua, is a semi-domesticated crop grown in high-altitude regions of the Andes. It is an A-genome diploid ($2n = 2x = 18$) relative of the allotetraploid (AABB) *Chenopodium quinoa* and shares many of its nutritional benefits. Both species contain a complete protein, a low glycemic index, and offer a wide variety of nutritionally important vitamins and minerals. Due to its minor crop status, few genomic resources for its improvement have been developed. Here we present a fully annotated, reference-quality assembly of cañahua. The reference assembly was developed using a combination of established techniques, including multiple rounds of Hi-C based proximity-guided assembly. The final assembly consists of 4,633 scaffolds with 96.6% of the assembly contained in nine scaffolds representing the nine haploid chromosomes of the species. Repetitive element analysis classified 52.3% of the assembly as repetitive, with the most common (27.3% of assembly) identified as LTR retrotransposons. MAKER annotation of the assembly yielded 22,832 putative genes with an average length of 4.6 Kb. When compared with quinoa, strong patterns of synteny support the hypothesis that cañahua is a close A-genome diploid relative, and thus potentially a model diploid species for genetic analysis and improvement of quinoa. Resequencing and phylogenetic analysis of a diversity panel of 30 cañahua accessions collected from across the Altiplano suggests that coordinated efforts are needed to enhance genetic diversity conservation within *ex situ* germplasm collections.

Keywords: *Chenopodium pallidicaule*, proximity-guided assembly, *in vivo* Hi-C, Andean crops, genome assembly

ACKNOWLEDGEMENTS

Completing this thesis project has been a dream come true for me, and it wouldn't have been possible without the help of amazing mentors, teachers, family members, and friends. Thanks to Dr. Maughan for teaching me everything I know about genome assembly and working through every step of this project with me, Dr. Jellen for inspiring me and teaching me to love plant genetics, Dr. Jarvis for getting me hooked on figure design and sharing your tips and tricks, and Dr. Geary for encouraging me and making all outdoor adventures much more fun (I am always diagnosing plant diseases). I have also taken incredible courses that have helped me progress in my thesis work. Notably, Dr. Piccolo's bioinformatics class gave me the confidence and the skills to dig deeper into data analysis. PWS administrative staff (particularly Jana, Carolyn, and Kerly) have all been incredibly useful resources during my time as a graduate student. I have also had great support from collaborators, including Dr. Patricia Mollinado and Patty Deza who helped me to see the potential and importance of this work on the lives of farmers and consumers in South America.

As a student and new mother, I could not have accomplished this enormous task without the support and assistance provided by my supportive family members and friends. Finally, I am lucky to have such a bright and joyful son that has happily shared me with graduate school. I hope this makes him proud one day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Introduction to Cañahua

Chenopodium pallidicaule Aellen is a species of goosefoot related to the increasingly popular seed crop, quinoa (*C. quinoa* Willd). Gade (1970) noted that cañahua is a partially domesticated crop that provides food security to many subsistence farmers across the Altiplano, the high plateau situated at 3,500 – 4,200 meters above sea level between the Occidental and Oriental Andean Cordilleras of west-central South America. He also states that its cultivation dates back over 7,000 years when it was a staple crop in ancient Incan and Aztec societies. It has several common names in native languages, including *cañahua* in Quechua and alternatively as *cañigua*, *cañihua*, *cañawa*, and *kañiwa* in other languages (Gade, 1970). Following the Spanish conquest, cultivation was likely discouraged due to its association with Incan society in the minds of European colonists (Ruas et al., 1999), as it was believed that consumption of indigenous foods was inferior (Earle, 2012). While it never regained its former status, subsistence farmers across the Altiplano and other high-altitude parts of the Andes continue to grow cañahua due to its resistance to frost, drought, salinity, and pests in addition to its high nutritional quality (Gade, 1970). It is grown alongside Andean tubers and traditional pseudocereals such as quinoa and kiwicha (*Amaranthus caudatus,* L.). In spite of the increasing popularity of its close relative, quinoa, cañahua remains practically unknown and underutilized as a food resource (Rastrelli et al., 1996).

Cañahua has a unique nutritional profile that is ideal for human consumption in areas where protein is limited. Its seed contains 15-18% protein, with a complete set of essential amino acids, including 5-6% lysine, which is typically limiting in monocotyledonous grain crops (Peñarrieta et al., 2008). Repo-Carrasco et al (2003) state that quinoa and cañahua are principle protein

sources due to the scarcity of available animal protein in many native areas. With a poverty rate of nearly 50% in the rural highlands of the Altiplano, cañahua represents an incredibly important resource in the prevention of poverty-induced malnutrition and in improving food security throughout the region (Repo-Carrasco et al., 2003).

In addition to high quality protein, Peñarrieta et al. (2008) note that cañahua offers a wide variety of antioxidants, phenolic compounds, and flavonoids. The high concentration of antioxidants is thought to be a result of high-altitude cultivation and free radicals that result from intense ultraviolet (UV) light exposure in living cells (Peñarrieta et al., 2008). Appreciable concentrations of antioxidants and phenolic compounds signify that cañahua may have considerable value for human nutrition. Repo-Carrasco-Valencia et al. (2010) compared flavonoid concentrations to berries, which are known to have very high flavonoid content, and found that the amount of flavonoids per 100 g dry matter was comparable (an average of 37 mg in quinoa and 33 mg in cañahua). Quercetin and isorhamnetin in particular were found in exceptionally high concentrations (an average of 60 mg/100 g and 30 mg/100 g, respectively). Traditional cereals contain no flavonoids, thus cañahua may prove an important source of these health-promoting compounds (Repo-Carrasco-Valencia et al., 2010). Cañahua seeds also contain vanillic acid, a phenolic compound which acts as a flavor enhancer and lends a pleasant taste to cañahua, particularly when ground and toasted as a flour called cañihuaco (Peñarrieta et al., 2008; Repo-Carrasco-Valencia et al., 2010).

Genetic Resources for Cañahua

Gade (1970) noted nearly half a century ago that the continued presence of cañahua in the Altiplano will depend on its genetic transformation into a more efficient crop. Agronomic issues

that have prevented more extensive cultivation of cañahua include non-uniform seed ripening and small seed size that make harvesting and processing of the seed difficult (Mujica, 1994). In spite of its unique agronomic and nutritional qualities, very few of the genetic resources needed to accelerate the improvement of cañahua have been reported. Raus et al. (1999) published a phylogenetic study of 19 *Chenopodium* species based on random amplified polymorphic DNAs (RAPDs) to analyze genetic variation within the genus. The analysis included two cañahua accessions that were found to be nearly identical; yet, were only distantly related to quinoa. Vargas et al. (2011) developed the first microsatellite markers for cañahua. From a total of 616 quinoa microsatellite markers, 34 polymorphic cañahua markers were identified, exhibiting a total of 154 different alleles. Nearly 40% of the quinoa-derived markers amplified in cañahua, consistent with shared ancestry between these two species. A phylogeny of 43 cañahua accessions showed clear distinctions between wild and cultivated lines, including a distinct subclade of only erect morphotypes. Other morphotypes were not predictive of genetic distance, nor were there clear associations between geographic origin and genetic distance seen in the data. The authors attributed this to the well documented and extensive trading culture of the native Andean people (Vargas et al., 2011). Kolano et al. (2011) cytologically characterized the genome size and rDNA loci of 23 *Chenopodium* diploid species ($2n = 2x = 18$), including cañahua. Their findings indicated that the New World diploids possess much smaller genomes than the Eurasian diploids. For example, the 2C value for cañahua measured 0.886 ± 0.034 pg (~433 Mb per haploid genome) whereas the 2C value for *C. suecicum* M., an Old World diploid species and the closest known living B-subgenome relative to quinoa, measured 1.763 ± 0.016 pg (~862 Mb). Cañahua was determined to have a single copy of both 35S (subterminal) and 5S (interstitial) rDNA loci.

Quinoa is an allotetraploid ($2n = 4x = 36$), presumably resulting from an ancient polyploidization event between North American and Eurasian diploids representing the A and B subgenomes of modern quinoa, respectively (Štorchová et al., 2015). While cañahua is not believed to be the direct A-genome donor of quinoa, it is a related A-genome diploid. As a part of the genome analysis of quinoa, Jarvis et al. (2017) reported a draft assembly of cañahua (PI 478407). The draft was based solely on Illumina short reads and was thus highly fragmented, consisting of 3,015 scaffolds and spanning a total length of 337 Mb (77.8% of the predicted genome size), with an N50 of 356 Kb.

Here we report the use of PacBio long-reads and Hi-C based proximity-guided assembly to develop a reference-quality, chromosome-scale assembly of cañahua. The genome was fully annotated using a deeply sequenced transcriptome developed from six combinations of tissue types and abiotic stresses. Additionally, genetic diversity within the species was characterized by a diversity panel of 30 accessions of cultivated and wild accessions of cañahua. The reference assembly and annotation reported here should clarify the phylogeny of cañahua within the *Amaranthaceae* family (Brown et al., 2015; Jarvis et al., 2017), facilitate the identification of genes controlling important agronomic traits through traditional bi-parental mapping populations or genome-wide association studies, and subsequently allow the implementation of accelerated breeding programs via genomic selection (Jannink et al., 2010; Brachi et al., 2011).

METHODS

Plant Material

The cañahua accession PI 478407 was used to develop a reference assembly. It was originally collected in 1981 at the Instituto Boliviano de Tenologia, Patacamaya, Bolivia and is

freely available from the United States Department of Agriculture (USDA; Ames, Iowa, USA; https://npgsweb.ars-grin.gov/). For the diversity panel, 30 accessions from three germplasm collections consisting of seven cañahua varieties from the USDA collection, one landrace and two wild accessions from the Universidad Nacional Agraria La Molina (UNALM; Lima, Peru), and 21 accessions from Universidad Major de San Andrés (UMSA; La Paz, Bolivia) were sampled. Two additional *Chenopodium* diploids, *C. watsonii* A. Nels (BYU 873; Yavapai Co., Arizona), and *C. sonorensis* Benet-Pierce & M.G. Simpson (BYU 17220; Santa Cruz Co., Arizona) were collected by BYU personnel and included for read-mapping comparisons. A complete list of all plant materials used is provided in Table 1.

Whole Genome Assembly

In vivo Hi-C and proximity-guided assembly techniques were used to improve the previously published short-read draft assembly reported by Jarvis et al. (2017), referred to hereafter as the ALLPATHS-LG Short-Read Assembly (ASRA). Fresh leaf tissue from a single dark-treated (72 h), 3-week old plant, derived directly from selfing of the original cañahua 'PI 478407' plant used by Jarvis et al. (2017), was sent to Phase Genomics (Seattle, WA, USA) for *in vivo* Hi-C based proximity-guided ligation and 80-bp paired-end sequencing followed by alignment to the ASRA assembly using BWA v0.7 (Li and Durbin, 2010). Only reads that aligned uniquely to the scaffolds were retained. Proximo^{TM}, a proximity-guided assembly method based on the Ligating Adjacent Chromatin Enables Scaffolding *In situ* assembler (LACHESIS; Burton et al., 2013), was used to cluster, order, and orient scaffolds from the ASRA assembly, producing the first Proximity-Guided Assembly (PGA1).

Following the development of PGA1, long-reads were used for gap-filling. High molecular weight DNA was extracted from leaf tissue of a single, 72-h dark-treated cañahua (PI 478407) plant using the Qiagen Genomic-tip 500/G Kit (Hilden, Germany) with a modified protocol (Supplemental Material 1). Single-molecule, real-time sequencing using the PacBio Sequel platform (Menlo Park, CA, USA) was performed at the BYU DNA Sequencing Center (Provo, Utah, USA). The PBJelly2 pipeline from PBSuite v15.8.24 (English et al., 2012) was used to align the long-reads to PGA1 in order to gap-fill the assembly. Arrow v0.22.0 (Chin et al., 2013) and Pilon v1.22 (Walker et al., 2014) were used for genome-polishing with the previously described PacBio long-reads and Illumina paired-end reads, respectively. This gap-filled and polished assembly is henceforth referred to as PGA1.5. To correct for possible errors introduced by low PacBio read coverage and relaxed PBJelly2 parameters, a contig-breaking tool, Polar Star (https://github.com/phasegenomics/polar_star), was employed. Polar Star aligns long-reads to an assembly, then calculates the read depth at each base. Read depth is smoothed in a 100-bp sliding window, then regions of high, low, and normal read depth are merged. These classifications are made based on the read depth distribution. Low read depth outliers are identified, and the assembly is broken at each such location. Following Polar Star, the PGA1.5 underwent a second *de novo,* proximity-guided assembly. Assembly errors (inversions and rearrangements) were identified and adjusted manually using Juicebox v1.9.8 (Durand et al., 2016; https://github.com/aidenlab/Juicebox/). The result was a chromosome-scale, polished assembly referred to as PGA2. (Figure 1).

Transcriptome Assembly

RNA-seq data was generated using the Illumina Hi-Seq platform from cañahua (PI 478407) leaf, root, inflorescence, and apical meristem tissues grown in both non-stressed and salt-stressed conditions, as detailed by Jarvis et al. (2017). The reads were trimmed using Trimmomatic v0.32 (Bolger et al., 2014) to remove Illumina adapters and trailing bases with a quality score below 20, then aligned to the PGA2 reference using HiSat2 v2.0.4 (Kim et al., 2015; Pertea et al., 2016) with default parameters except the max intron length was set to 50,000 bp. Following alignment, the resulting SAM file was sorted and indexed using SAMtools v1.6 (Li et al., 2009) and assembled into putative transcripts using StringTie v1.3.4 (Pertea et al., 2015, 2016).

Repeat Modeling and Gene Annotation

RepeatModeler v1.0.11 (Smit and Hubley, 2008) and RepeatMasker v4.0.7 (Smit et al., 2013) were used to identify and classify repetitive elements in the final (PGA2) assembly relative to Repbase-derived RepeatMasker libraries v20181026 (Bao et al., 2015). Whole-genome annotation of the PGA2 assembly was performed by MAKER v2.31.10 (Cantarel et al.; Holt and Yandell, 2011) using the cañahua transcriptome as expressed sequence tag (EST) evidence, the uniprot_sprot database (downloaded September 25, 2018) and quinoa protein sequences (Jarvis et al., 2017) as protein homology evidence, and the consensi.fa.classified output from RepeatModeler for soft repeat masking. Gene prediction models included an Augustus gene prediction model for cañahua produced by Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (Waterhouse et al.; Simão et al., 2015) and the *Arabidopsis thaliana* SNAP HMM file (Leskovec and Krevl, 2014) for gene prediction. BUSCO v3.0.2 (Simão et al.,

2015) assessed the completeness of the assembly and annotation using the Embryophyta odb10 dataset.

Chloroplast Genome Assembly and Annotation

A reference-guided assembly of the cañahua chloroplast genome was constructed by the Assembly by Reduced Complexity (ARC) assembler (v1.1.4; Hunter et al., 2015) using a subset of six million whole-genome, paired-end Illumina reads with the quinoa chloroplast genome (Maughan et al., 2019) as a target. The ARC algorithm uses Bowtie2 (Langmead and Salzberg, 2012) with relaxed parameters to map reads against targets, extract mapped reads from each target, and assemble mapped reads using the SPAdes assembler (Kulikov et al., 2012). The targets are then replaced with newly assembled contigs and the process is iterated for a predetermined number of cycles or until no additional reads can be incorporated. The ARC pipeline extended the assembled cañahua chloroplast contigs through four (numcycles = 4) successive rounds of mapping and re-assembly. Since chloroplast read depth should be significantly higher than nuclear genome read depth, only assembled contigs with read depth > 50X coverage were selected for further assembly. Pacific Biosciences long-reads (> 15 Kb; n = 246,847) were used to fill gaps between contigs using PBJelly2, a subprogram from PBSuite v15.8,24 (English et al., 2012). A circularized contig representing the complete chloroplast genome was constructed using the circularize tool from Geneious (v11.1.5; https://www.geneious.com/), then the assembly was polished using the same six million paired-end Illumina reads as used in initial assembly.

Annotation of the cañahua chloroplast was performed using GeSeq v1.65 (Tillich et al., 2017; https://chlorobox.mpimp-golm.mpg.de/geseq.html) with the quinoa chloroplast annotation

(Maughan et al., 2019) and the MPI-MP chloroplast database as references. ARAGORN v1.2.3 and HMMER profile search were enabled, the latter using the Embryophyta chloroplast (CDS + rRNA) database. Comparison to the quinoa chloroplast (Maughan et al., 2019) was performed by the nucmer tool from MUMmer v4.0beta (Marçais et al., 2018) followed by MUMmerplot with all default parameters.

Resequencing and SNP Discovery

DNA was extracted from single plants for each of the 30 cañahua accessions using cetyl trimethylammonium bromide extraction method as described by Doyle JJ and Doyle JL, (1987). Samples were sent to Novogene (San Diego, CA) for whole-genome Illumina HiSeq (150-bp paired-end) sequencing from 500-bp insert libraries, for each accession (Table 2). Trimmomatic v0.32 (Bolger et al., 2014) was used to remove Illumina adapters and trailing bases with a quality score below 20 or average per-base quality of 20 over a four-nucleotide sliding window. Reads from each accession were aligned to PGA2 using BWA-MEM v0.7.17 (Li, 2013) to produce SAM files that were converted to BAM format, sorted and indexed using SAMtools v1.9 (Li et al., 2009). The BAM files were used as input for InterSnp, a subprogram of the BamBam v1.4 pipeline (Page et al., 2014), for SNP genotyping. SNPhylo v20160204 (Lee et al., 2014) used the HapMap output files produced by InterSnp to filter and remove SNPs with > 10% missing data and minor allele frequency < 5%. SNPhylo also filters SNP datasets using linkage disequilibrium estimates (SNPs with LD < 40% are removed) prior to building bootstrapped (n = 1000) phylogenies based on MUSCLE (Edgar 2004) sequence alignments. The resulting tree was visualized using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree). Structure v2.3.4 (Novembre et al., 2000) used Bayesian clustering analysis with a range of K = 1 through K = 5 to

assess and visualize population structure from a 1,000 SNP subset of the InterSnp output. The most parsimonious fit occurred at K = 4. ArcMap v10.3.1 (ESRI, 2011) mapping software was used to map the geographic locations of the source materials. The clustering partitions produced by STRUCTURE were used to construct a pie chart representing the allelic composition of each mapped individual.

Genome Comparison

A phylogenetic tree showing relationships between cañahua and four other *Amaranthaceae* species was created by aligning 254 conserved orthologous genes (COGs) using MUSCLE v3.8.31 (Edgar, 2004), then combining the gene alignments with trimAl v1.2 (Capella-Gutiérrez et al., 2009) followed by FASconCAT v1.11 (Kück and Meusemann, 2010). The complete alignment was analyzed and developed into a maximum-likelihood phylogeny (model VT+F+G4) with 1,000 rounds of bootstrapping in IQ-TREE (Nguyen et al., 2015) supported by UFBoot2 (Hoang et al., 2018), then visualized in FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree).

Initial genomic comparisons to quinoa, beet (*Beta vulgaris* L.; Funk et al., 2018) and amaranth (*Amaranthus hypochondriacus* L.; Lightfoot et al., 2017) were developed using the nucmer tool from MUMmer v4.0beta (Marçais et al., 2018) with the minimum number of clusters set to 500 (*c* = 500) to minimize noise. Visualization was done by mummerplot with the layout, filter, and color parameters set to true. Comparisons of coding sequences for each genome were made using the CoGe SynMap tool (https://genomevolution.org/coge/), then DAGchainer (Haas et al., 2004) output files were used as input for the MCScanX toolkit (Wang et al., 2012; https://github.com/wyp1125/MCScanX).

Read-mapping percentages were obtained by first generating paired-end, Illumina Hi-Seq reads for cañahua, *C. watsonii*, and *C. sonorensis*. Read trimming to remove Illumina adapters and trailing bases with a quality score lower than 20 was performed by Trimmomatic v0.32 (Bolger et al., 2014), then the BWA-MEM v0.7.17 (Li and Durbin, 2010) algorithm aligned trimmed reads from each species to the quinoa reference genome. Output SAM files were converted to sorted BAM files by SAMtools v1.9. Picard from GATK v4.0 (McKenna et al., 2010) produced alignment summary statistics.

RESULTS AND DISCUSSION

Whole Genome Assembly

A draft assembly of cañahua, accession PI 478407, was previously reported by Jarvis et al. (2017). This assembly was based solely on Illumina short reads assembled using the ALLPATHS-LG assembler (Gnerre et al., 2011). While an excellent draft assembly, the lack of long-jump libraries (fosmids) resulted in a fragmented assembly. ASRA consisted of 8,982 contigs in 3,013 scaffolds with a contig and scaffold N50 of 84 Kb and 357 Kb, respectively, spanning a total length of 337 Mb (Table 3). To improve ASRA, 179 million Hi-C-based paired-end reads were generated and used to scaffold ASRA using the Proximo™ pipeline (Phase Genomics, Seattle, WA). Seventy-nine percent (2,392) of the ASRA scaffolds were clustered into nine pseudomolecules, presumably corresponding to the nine haploid chromosomes of cañahua ($2n = 2x = 18$; Figure 1), producing a substantially improved assembly (PGA1). The unincorporated scaffolds (621) were small, with an N50 of 97.9 Kb and mean scaffold size of 25.8 Kb, making them much more difficult to incorporate accurately into pseudochromosomes. The unincorporated scaffolds represented < 5% of the total sequence length of ASRA. The

number of scaffolds clustered to specific chromosomes ranged from 203 to 317, and the length of the assembled pseudochromosomes were 31.3 to 40.4 Mb. Thus, the PGA1 scaffolds contained 95.3% of total sequence length (99.7% excluding N gaps) with an N50 and L50 of 35.6 Mb and five, respectively (Table 3). Ns occupied 12.3 Mb (4%) of the assembly, with an average of 1,047 gaps (20 or more contiguous Ns) per scaffold.

The PGA1 was further improved by applying a combination of gap-filling and genome-polishing techniques. To close gaps, 10.21 Gb (1,101,202 reads) of PacBio long reads were generated with a mean read length of 9.3 Kb, providing 23.6X coverage of the cañahua genome (Table 4). PBJelly2 (English et al., 2012) aligned PacBio long reads to PGA1 and closed 75% of existing gaps. Due to potential errors introduced into gaps because of the inherent high error rate of PacBio reads, the assembly quality was improved using two genome-polishing tools: Arrow (Chin et al., 2013), which produces consensus-quality assemblies from PacBio sequences, followed by Pilon (Walker et al., 2014), which performs a similar function but takes advantage of the significantly lower error rate of Illumina reads to improve the consensus assembly. These polishing steps made changes at 593,821 positions, representing < 0.165% of PGA1. The resulting assembly, PGA1.5, had a total assembly size of 363 Mb, an approximate 7.7% increase from the ASRA. The scaffold N50 of PGA1.5 increased slightly to 37.8 Mb, while the number of gaps decreased dramatically from 8,013 to 2,007, which is also reflected in a 10-fold decrease in the number of Ns in the assembly (4% to 0.2%; Table 3).

A second round of proximity-guided assembly using PGA1.5 assessed and improved the chromosome-scale assembly. Polar Star (https://github.com/phasegenomics/polar_star), which aggressively breaks contigs at low-PacBio depth locations based on deviation from mean depth, introduced 5,241 breaks which were then tested for rescaffolding using Hi-C based proximity-

guided assembly. This acts as a check on the error-prone PacBio reads and low coverage depth used in the gap-filling process. The result is a dramatically improved proximity-based assembly, evident by the consistent pattern of Hi-C crosslink density along pseudochromosomes and the resolution of erroneous inversions and rearrangements in three of the scaffolds (2, 5, and 7; Figure 1). The final assembly (PGA2) spans 362.5 Mb, has a scaffold N50 and L50 of 38.1 Mb and five, respectively, with < 0.1% of the assembled sequence found in 3,586 gaps. Eighty-four percent of the estimated genome size is represented; the remaining 16% is likely comprised of repetitive sequence that has collapsed in regions such as centromeres and telomeres due to the use of short-reads for the initial assembly. Nine pseudochromosomes contain 96.7% of the total sequence length (99.9% excluding N gaps), ranging in size from 33.5 Mb to 45.4 Mb (Table 5). The value of incorporating Hi-C data and long-reads into the assembly is clear when comparing ASRA and PGA2 assemblies. The Hi-C data increased contiguity of PGA2 significantly by reducing the assembly from 3,015 scaffolds to nine pseudochromosomes, while the long-read sequence dramatically reduced the number of gaps (by 75%) in the assembly as well as increasing the total assembly size (Table 3).

Repeat Modeling and Gene Annotation

A transcriptome assembly of cañahua was developed by sequencing RNA-seq libraries from six unique tissue and abiotic stress combinations. The resulting RNA-seq libraries generated 66.3 Gb of data from 663,493,956 paired-end reads with an average of 11.05 Gb per library (Table 6). Ninety-eight percent (649,273,284) of the paired RNA-seq reads aligned to the final PGA2 assembly and 255,893 features (214,170 exons and 41,723 transcripts) were identified with a mean transcript length of 2.19 Kb and an average of 28,246 features per pseudochromosome.

One significant disadvantage to developing a genome assembly based on short-reads is the difficulty of properly assembling repetitive elements (Richards, 2018). For example, the telomeric repeat in PGA2 was largely collapsed into a single contig that was not scaffolded to any of the pseudochromosomes. While there are traces of telomere sequence on several of the nine scaffolds (Figure 2A), the integrity of this element was largely lost. In spite of this disadvantage, RepeatModeler and RepeatMasker were still able to obtain some useful information about the assembled repeats in the cañahua genome. Fifty-three percent (191 Mb) was classified as repetitive, with an additional 1.9% (7 Mb) classified as low complexity (satellites, simple repeats, and small RNAs). A total of 129 Mb (35.5%) was identified as retrotransposons or DNA elements, with an additional 61 Mb (16.8%) classified as unknown elements. The most common elements identified were long terminal repeat (LTR) retrotransposons, including 990 Class I endogenous retrovirus (ERV) elements spanning a total of 113 Kb (31.2%). The most common DNA transposon was a hAT-Charlie element, covering 41 Kb (0.01%) of the genome (Table 7). The large fraction of unknown elements was unsurprising given that the only published studies of repetitive elements in the *Chenopodium* genus have been limited to the rDNA sequences (Maughan et al., 2006; Kolano et al., 2011) and two repetitive sequences, 18-24J and 12-13P, that were only recently characterized cytogenetically (Orzechowska et al., 2018). BLASTn was used to identify the 5S rDNA sequence and the two *Chenopodium* repetitive elements. Consistent with the findings of Kolano et al. (2011), the 5S rDNA sequence was found only in a single genomic location in the centromeric region of Cp8. While the 18-24J repeat was present in cañahua*,* it only occupies 55.4 Kb (0.012%) of the genome compared to 1.4 Mb (0.18%) in *C. suecicum*, a B-genome diploid. This supports the findings of Orzechowska et al. (2018) stating that 18-24J is found almost

exclusively in the *Chenopodium* B-genome. The 12-13P repetitive element was twice as common as the 18-24J repeat, occupying 124.6 Kb (0.027%), and localized to the centromeric region on all nine pseudochromosomes (Figure 2A).

The MAKER pipeline was used to annotate PGA2 using as evidence the cañahua transcriptome described previously, cañahua repetitive element features as annotated by RepeatModeler, and quinoa protein sequences as reported by Jarvis et al. (2017) as well as the uniprot_sprot database. A total of 22,832 genes were identified, which is just over half of the 44,776 genes annotated in the tetraploid quinoa (Figure 2A) and an increase of 4,871 genes relative to the annotation of the ASRA annotation. The average length of genes identified was 4.6 Kb, the longest of which spanned 19,183 bp (CP013000) and is predicted to encode the sacsin gene found in many eukaryotes, including other *Amaranthaceae* species such as quinoa, beet, and spinach. The mean Annotation Edit Distance (AED), which is a quality measure combining values for sensitivity, specificity, and accuracy to give evidence of a high-quality annotation, was 0.23 (Figure 2B). AED values < 0.25 are indicative of high-quality annotations (Holt and Yandell, 2011).

Completeness of the gene space was assessed using the Benchmarking Universal Single Copy Orthologs (BUSCO) platform, which quantifies functional gene content using a large core set of highly conserved orthologous genes (COGs). Of the 1,375 plant specific COGs in the *Embryophta* database, 1,341 (97.5%) were identified in the cañahua genome as complete with another nine (0.7%) COGs classified as fragmented (Complete: 97.5% [Single: 95.9%, Duplicated: 1.6%], Fragmented: 0.7%, Missing: 1.8%). Relative to the MAKER *de novo* annotated proteins and transcripts, BUSCO identified 1,260 (91.6%) and 1303 (94.8%) complete COGs, respectively (Figure 2C). The discrepancies between the whole genome, protein, and

transcript BUSCO findings may be attributed to the difference in gene annotation method between BUSCO and MAKER. While BUSCO uses BLAST to identify known genes, MAKER uses an approach that requires sufficient evidence from a combination of protein, EST, and *ab initio* gene prediction inputs. The annotation could potentially be improved by further training of the input gene prediction model (Augustus) and multiple rounds of MAKER annotation.

Chloroplast Genome Reconstruction

The cañahua chloroplast assembly spans 151,799 bp in a single, circular molecule. Annotation reveals the anticipated quadripartite structure, including two copies of an inverted repeat region (IR) separating large and small single-copy regions. One hundred thirty-two genes were identified, including 88 protein-coding genes, 36 tRNA genes, and 8 rRNA genes (Figure 3A). Twenty-one genes occupy each IR, including a pseudogene previously characterized in other *Amaranthaceae* species as *rpl23* (Park et al., 2018; Maughan et al., 2019). Morton et al. (1993) performed an analysis of the *rpl23* gene in seven Poaceae species and hypothesize that gene conversion is preserving the pseudogene as double strand break repair mechanisms use the functional homolog as a template for DNA synthesis.

With a length of 151,799 bp, the cañahua chloroplast is of a similar size to that of quinoa, which has been reported for multiple quinoa accessions ranging in size from 152,079 - 152,282 bp, with an average length of 152,134 bp (Hong et al., 2017; Maughan et al., 2019). Due to lack of recombination of chloroplast genomes and the relatively recent allotetraploidization event creating quinoa (3.3 – 6.3 million years ago; Jarvis et al., 2017), the extreme similarity between the cañahua and quinoa chloroplasts (Figure 3B) supports the existing hypothesis that the maternal parent of quinoa was an A-genome species. It is unlikely that cañahua is the direct ancestor of the A-subgenome in quinoa, but it does suggest that future analyses of the organellar

genomes of the more than 45 putative A-genome diploid *Chenopodium* species should provide important insight into the polyploidization that underlies the evolution and domestication of the New World AABB species complex that includes free-living *C. berlandieri* ssp. *berlandieri* Moq., *C. quinoa* ssp. *melanospermum*, *C. quinoa* ssp. *milleanum* Aellen, and *C. hircinum* Schrad., along with their domesticated forms *C. quinoa* and *C. berlandieri* ssp. nuttaliae (Wilson, 1990).

Resequencing

A diversity panel consisting of 30 varieties of cañahua, including 28 landrace varieties and two wild accessions, underwent whole-genome, paired-end Illumina sequencing resulting in an average of 10.9X coverage (4.7 Gb) per accession. Following BWA alignment to the PGA2 reference, the InterSnp tool from BamBam identified 358,461 SNPs in the diversity panel, which were then filtered to include 16,194 SNPs based on minor allele frequency, missing data and linkage disequilibrium. Analysis of the consensus, 1,000-bootstrap phylogeny of the cañahua diversity panel suggests several major points of interest (Figure 4A). First, the USDA collection of the species is limited to only two of three major nodes with the majority (seven out of eight accessions) on a single node, highlighting the need for international collection efforts to preserve the diversity of its germplasm. Second, the Mantel test suggests that there is no correlation between collection site and genotype ($Z = 11{,}296.22$, $r = -0.12326$, and $p = 0.837$). This is likely due to a lack of true collection site data for many of the accessions. Indeed, four accessions each from the UMSA and USDA collections have as their passport data the latitude and longitude coordinates of the research facilities where they are stored in germplasm collection instead of the coordinates of the original collection site (Figure 4B, Table 1). Vargas et al. (2011) suggest that

another complicating issue is the well-known cultural practice of seed trading among ancient Andean societies that has been an important part of agriculture in the Altiplano region for thousands of years. Lastly, the collection sites of the three UNALM accessions (two wild, one cultivated) are in close proximity, yet they are found on distinct nodes of the phylogeny and have a structure that is distinct from the landraces with little or no admixture occurring (Figure 4C). This finding agrees with those of Vargas et al. (2011) and is further evidence that wild accessions may be useful sources of genetic diversity for improving cañahua.

Genome Comparison

The *Amaranthaceae* family contains approximately 165 genera comprised of over 2,000 species, including food crops like the amaranths (*A. hypochondriacus, A. caudatus, A. cruentus,* and *A. hybridus*), spinach (*Spinacia oleracea*), the foliar and root beet crops (*B. vulgaris*), and quinoa (*C. quinoa)*. A maximum-likelihood tree including these important members of *Amaranthaceae* was developed using 254 conserved orthologous genes and 1,000 rounds of bootstrapping in a VT+F+G4 model (Figure 5A). The relationships reflected therein are somewhat unresolved in that the tree does not definitively show whether amaranth or beet is a closer relative to the *Chenopodium* species. However, synonymous mutation rates (Ks) generated by CoGe (Figure 5B, Table 8) support relationships shown in previous phylogenies (Pratt, 2003; Brown et al., 2015; Jarvis et al., 2017) where amaranth is more significantly diverged than beet with an estimated divergence from the last common ancestor approximately 21.33 - 39.51 compared to16 - 29.63 MYA.

The first species of the family with a reference-quality genome assembly was beet ($2n = 2x = 18$; Dohm et al., 2014), so a genomic comparison with beet was performed and we decided to

maintain the family naming convention by assigning the cañahua chromosomes the same number as the beet homologs (Figure 6B). Comparison of the two species in CoGe identified 13,436 syntenous genes occupying 522 synteny blocks. Interestingly, a comparison to amaranth, a more distant relative to cañahua than beet (Figure 5, Table 8), identified 12.8% more syntenous genes (15,153) than with beet (Table 8). The increase in synteny is likely attributed to the greater number of annotated genes in the paleopolyploid amaranth ($2n = 2x = 32$) and the similarity in assembly methodology of the amaranth and cañahua genomes rather than greater genetic similarity. Lightfoot et al. (2017) noted evidence of chromosomal loss (the homeolog of Ah5) and fusion (Ah1) events in the amaranth genome. This was anticipated because beet and cañahua are diploids that share a base chromosome number of $x = 9$, whereas the base number in *Amaranthus* was reduced to $x = 8$. This is confirmed by the presence of two full-length homologs of Cp9 within Ah1 and a mostly missing second Cp1 homolog that is homologous to Ah5 (Figure 7A). Interestingly, 123 genes syntenous to Cp1 (13% compared to the 931 genes shared by Ah6 and Cp1) have been translocated to another chromosome, Ah11 (Figure 7B). These 123 remaining genes may provide useful insight to the process of chromosome loss and gene function in the *Amaranthaceae* family.

Comparison of cañahua with quinoa confirmed the work of Jarvis et al. (2017) suggesting that cañahua is representative of the A-genome of *Chenopodium*. While both the A and B genomes have maintained similar chromosomal structure, the A-subgenome homologs in quinoa can be clearly identified by visual inspection of the alignment output by MUMmer (Figure 8A). Quantitative support for the A-subgenome chromosome assignments in quinoa is provided by the number of syntenous gene pairs, where 13,574 are found in the A-subgenome and 10,703 in the B-subgenome chromosomes (Table 9). This is even more significant considering that the B-

subgenome of quinoa is much larger than the A-subgenome (531 Mb in the A and 670 Mb in the B-subgenome). All quinoa chromosomes assigned to the A-subgenome have a higher number of syntenous genes than their B-subgenome homeologs, except for Cq4A and Cq4B which show 1,444 and 1,491, respectively. Further inspection of Cq4A and Cq4B in the MUMmer plot, which identifies regions of synteny at the genome level, validates the assignment of Cq4A to the A-subgenome and suggests that gene loss may have occurred on Cq4A that is compensated by Cq4B, albeit with conservation of homoeologous chromosome structure and genetic collinearity. Indeed, Cq4A was annotated with 4,584 genes compared to 5,080 on Cq4B. There is also a notable difference in the estimated time since the A and B subgenomes of quinoa shared a common ancestor with cañahua. While the A-subgenome diverged approximately 0.830 - 1.54 MYA, the B-subgenome has been diverged for nearly twice as long with an approximate age of 1.67 - 3.09 MYA.

Careful evaluation of chromosomes within the *Amaranthaceae* family can shed light on how these genomes evolve over time and what role structural changes have played in biological function. For example, homologs of Cp5 are highly conserved in both the A and B subgenomes of quinoa (Cq5A and Cq5B), but there is clear structural variation in comparison to the homolog in beet, Bv5 (Figure 6A). One of the amaranth homologs of Cp5 is collinear (Ah2), while the second homolog is split between two chromosomes (Ah11 and Ah12) but also reflects a similar order. This may be evidence that a terminal inversion occurred in the evolution of beet after the divergence from a common ancestor. Homologs of Cp9 also show an evolutionarily interesting pattern. While it is very well conserved in the A-subgenome of quinoa (Cq4A), demonstrated both by a CoGe dot plot (Figure 6A) and a high number of syntenous genes (1,323; Table 9), the B-subgenome homolog has a much different structure and less than half the number of syntenous

genes (536). Meanwhile, beet and amaranth both have unique rearrangements of this homolog (Bv9 and Ah1, respectively), suggesting that the order of genes along this molecule may not hold significant biological importance.

Overall, the high level of synteny between cañahua chromosomes and the A-subgenome of quinoa provides strong evidence supporting a New World diploid as the donor of that subgenome in the allopolyploidization of quinoa. However, other A-genome diploid candidates have emerged as the closest known, living A-genome relative to quinoa. Given the closer proximity between the Eurasian landmass (B-subgenome origin) with North America versus South America, a hypothetical North American A-genome diploid donor is more logical than a South American donor. A comparison of read-mapping percentages has revealed that *C. watsonii* and *C. sonorensis,* both wild diploids from southwestern North America, align more closely to the quinoa genome than does cañahua. With a mapping percentage of 98.36%, *C. watsonii* is presently the most likely A-genome donor (Table 10). This discovery does not diminish the importance of understanding the genome of cañahua, as it will act as a model for the structure and contents of New-World diploid *Chenopodium* species and provide tools for improvement of an important Andean food source.

CONCLUSIONS

The reference-quality, chromosome-scale assembly of cañahua presented here has dramatically improved the existing resources for this important subsistence crop. Providing this critical genomic tool to breeding programs may spark new interest in the crop and lead to improved breeding strategies. We also present sequence data for 30 unique varieties that can provide preliminary data and minimize sequencing costs for researchers as they pursue core

breeding lines by identifying and selecting for key agronomic quality and stress resistance genotypes. While another A-genome diploid (*C.watsonii*) has emerged as the closest known, living relative to the A-genome parent in the allotetraploidization of quinoa, cañahua can act as a model for the A-genome of *Chenopodium*, providing phylogenetic context and insight into chromosomal evolution in the genus.

LITERATURE CITED

Bao, W., K.K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6:11. doi:10.1186/s13100-015-0041-9

Bolger, A.M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi:10.1093/bioinformatics/btu170

Brachi, B., G.P. Morris, and J.O. Borevitz. 2011. Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol. 12:232. doi:10.1186/gb-2011-12-10-232

Brown, D.C., V. Cepeda-Cornejo, P.J. Maughan, and E.N. Jellen. 2015. Characterization of the Granule-Bound Starch Synthase I Gene in Chenopodium. Plant Genome 8:0. doi:10.3835/plantgenome2014.09.0051

Burton, J.N., A. Adey, R.P. Patwardhan, R. Qiu, J.O. Kitzman, and J. Shendure. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31:1119–1125. doi:10.1038/nbt.2727

Cantarel, B., I. Korf, S. Robb, G. Parra, E. Ross, B. Moore, C. Holt, J.A. Alvarado, and M. Yandell. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. genome.cshlp.org

Capella-Gutiérrez, S., J.M. Silla-Martínez, and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.. Bioinformatics 25:1972–3. doi:10.1093/bioinformatics/btp348

Chin, C.-S., D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, S.W. Turner, and J. Korlach. 2013. Nonhybrid,

finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10:563–569. doi:10.1038/nmeth.2474

Dohm, J.C., A.E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski, H. Tafer, O. Rupp, T.R. Sörensen, R. Stracke, R. Reinhardt, A. Goesmann, T. Kraft, B. Schulz, P.F. Stadler, T. Schmidt, T. Gabaldón, H. Lehrach, B. Weisshaar, and H. Himmelbauer. 2014. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505:546–549. doi:10.1038/nature12817

Doyle JJ, and Doyle JL. 1987. A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf TIssue. Phytochem. Bull. 19:11–15

Durand, N.C., J.T. Robinson, M.S. Shamim, I. Machol, J.P. Mesirov, E.S. Lander, and E.L. Aiden. 2016. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.. Cell Syst. 3:99–101. doi:10.1016/j.cels.2015.07.012

Earle, R. 2012. Food and the Colonial Experience: Food, Race and the Colonial Experience in Spanish America

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.. Nucleic Acids Res. 32:1792–7. doi:10.1093/nar/gkh340

English, A.C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D.M. Muzny, J.G. Reid, K.C. Worley, and R.A. Gibbs. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One 7:e47768. doi:10.1371/journal.pone.0047768

ESRI. 2011. ArcGIS Desktop:Release 10

Funk, A., P. Galewski, and J.M. McGrath. 2018. Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. Plant J. 95:659–671. doi:10.1111/tpj.13977

Gade, D.W. 1970. Ethnobotany of cañihua (Chenopodium pallidicaule), rustic seed crop of the Altiplano. Econ. Bot. 24:55–61. doi:10.1007/BF02860637

Gnerre, S., I. Maccallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker, T. Sharpe, G. Hall, T.P. Shea, S. Sykes, A.M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E.S. Lander, and D.B. Jaffe. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data.. Proc. Natl. Acad. Sci. U. S. A. 108:1513–8. doi:10.1073/pnas.1017351108

Haas, B.J., A.L. Delcher, J.R. Wortman, and S.L. Salzberg. 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20:3643–3646. doi:10.1093/bioinformatics/bth397

Hoang, D.T., O. Chernomor, A. von Haeseler, B.Q. Minh, and L.S. Vinh. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518–522. doi:10.1093/molbev/msx281

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

Hong, S.-Y., K.-S. Cheon, K.-O. Yoo, H.-O. Lee, K.-S. Cho, J.-T. Suh, S.-J. Kim, J.-H. Nam, H.-B. Sohn, and Y.-H. Kim. 2017. Complete Chloroplast Genome Sequences and Comparative Analysis of Chenopodium quinoa and C. album. Front. Plant Sci. 8:1696.

doi:10.3389/fpls.2017.01696

Hunter, S.S., R.T. Lyon, B.A.J. Sarver, K. Hardwick, L.J. Forney, and M.L. Settles. 2015. Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences.. bioRxiv 014662. doi:10.1101/014662

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9:166–177. doi:10.1093/bfgp/elq001

Jarvis, D.E., Y.S. Ho, D.J. Lightfoot, S.M. Schmöckel, B. Li, T.J.A. Borm, H. Ohyanagi, K. Mineta, C.T. Michell, N. Saber, N.M. Kharbatia, R.R. Rupper, A.R. Sharp, N. Dally, B.A. Boughton, Y.H. Woo, G. Gao, E.G.W.M. Schijlen, X. Guo, A.A. Momin, S. Negrão, S. Al-Babili, C. Gehring, U. Roessner, C. Jung, K. Murphy, S.T. Arold, T. Gojobori, C.G. van der Linden, E.N. van Loo, E.N. Jellen, P.J. Maughan, and M. Tester. 2017. The genome of Chenopodium quinoa. Nature 542:307–312. doi:10.1038/nature21370

Kim, D., B. Langmead, and S.L. Salzberg. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12:357–360. doi:10.1038/nmeth.3317

Kolano, B., B.W. Gardunia, M. Michalska, A. Bonifacio, D. Fairbanks, P.J. Maughan, C.. Coleman, M.R. Stevens, E.N. Jellen, and J. Maluszynska. 2011. Chromosomal localization of two novel repetitive sequences isolated from the *Chenopodium quinoa* Willd. genome. Genome 54:710–717. doi:10.1139/g11-035

Kolano, B., D. Siwinska, J. McCann, and H. Weiss-Schneeweiss. 2015. The evolution of genome size and rDNA in diploid species of Chenopodium s.l. (Amaranthaceae). Bot. J. Linn. Soc. 179:218–235. doi:10.1111/boj.12321

Kück, P., and K. Meusemann. 2010. FASconCAT: Convenient handling of data matrices.. Mol. Phylogenet. Evol. 56:1115–8. doi:10.1016/j.ympev.2010.04.024

Kulikov, A.S., A.D. Prjibelski, G. Tesler, N. Vyahhi, A. V. Sirotkin, S. Pham, M. Dvorkin, P.A. Pevzner, A. Bankevich, S.I. Nikolenko, A. V. Pyshkin, S. Nurk, A.A. Gurevich, D. Antipov, M.A. Alekseyev, and V.M. Lesin. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477. doi:10.1089/cmb.2012.0021

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359. doi:10.1038/nmeth.1923

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

Leskovec, J., and A. Krevl. 2014. SNAP Datasets: Stanford. https://snap.stanford.edu/citing.html (accessed February 11, 2019).

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595. doi:10.1093/bioinformatics/btp698

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Lightfoot, D.J., D.E. Jarvis, T. Ramaraj, R. Lee, E.N. Jellen, and P.J. Maughan. 2017. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. BMC Biol. 15:74. doi:10.1186/s12915-017-0412-4

Marçais, G., A.L. Delcher, A.M. Phillippy, R. Coston, S.L. Salzberg, and A. Zimin. 2018. MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14:e1005944. doi:10.1371/journal.pcbi.1005944

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

Maughan, P.J., B.A. Kolano, J. Maluszynska, N.D. Coles, A. Bonifacio, J. Rojas, C.E. Coleman, M.R. Stevens, D.J. Fairbanks, S.E. Parkinson, and E.N. Jellen. 2006. Molecular and cytological characterization of ribosomal RNA genes in *Chenopodium quinoa* and *Chenopodium berlandieri*. Genome 49:825–839. doi:10.1139/g06-033

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.. Genome Res. 20:1297–303. doi:10.1101/gr.107524.110

Morton, B.R., and M.T. Clegg. 1993. A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the grass family (Poaceae). Curr. Genet. 24:357–365. doi:10.1007/BF00336789

Mujica, A. 1994. Andean Grains and Legumes. FAO, Rome, Italy.

Nguyen, L.-T., H.A. Schmidt, A. von Haeseler, and B.Q. Minh. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32:268–274. doi:10.1093/molbev/msu300

Novembre, J., J.K. Pritchard, M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. doi:10.1534/genetics.116.195164

Orzechowska, M., M. Majka, H. Weiss-Schneeweiss, A. Kovařík, N. Borowska-Zuchowska, and B. Kolano. 2018. Organization and evolution of two repetitive sequences, 18-24J and 12-13P, in the genome of *Chenopodium* (Amaranthaceae). Genome 61:643–652. doi:10.1139/gen-2018-0044

Page, J.T., Z.S. Liechty, M.D. Huynh, and J.A. Udall. 2014. BamBam: genome sequence analysis tools for biologists. BMC Res. Notes 7:829. doi:10.1186/1756-0500-7-829

Park, J.-S., I.-S. Choi, D.-H. Lee, and B.-H. Choi. 2018. The complete plastid genome of *Suaeda malacosperma* (Amaranthaceae/Chenopodiaceae), a vulnerable halophyte in coastal regions of Korea and Japan. Mitochondrial DNA Part B 3:382–383. doi:10.1080/23802359.2018.1437822

Peñarrieta, J.M., J.A. Alvarado, B. Åkesson, and B. Bergenståhl. 2008. Total antioxidant capacity and content of flavonoids and other phenolic compounds in canihua (Chenopodium pallidicaule): An Andean pseudocereal. Mol. Nutr. Food Res. 52:708–717. doi:10.1002/mnfr.200700189

Pertea, M., D. Kim, G.M. Pertea, J.T. Leek, and S.L. Salzberg. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11:1650–1667. doi:10.1038/nprot.2016.095

Pertea, M., G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, and S.L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33:290–295. doi:10.1038/nbt.3122

Pratt, D.B. 2003. Phylogeny and morphological evolution of the Chenopodiaceae-Amaranthaceae alliance. Iowa State University,

Quail, M., M.E. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13:341. doi:10.1186/1471-2164-13-341

Rastrelli, L., F. De Simone, O. Schettino, and A. Dini. 1996. Constituents of Chenopodium pallidicaule (Cañihua) Seeds: Isolation and Characterization of New Triterpene Saponins. J. Agric. Food Chem. 44:3528–3533. doi:10.1021/JF950253P

Repo-Carrasco-Valencia, R., J.K. Hellström, J.-M. Pihlava, and P.H. Mattila. 2010. Flavonoids and other phenolic compounds in Andean indigenous grains: Quinoa (Chenopodium quinoa), kañiwa (Chenopodium pallidicaule) and kiwicha (Amaranthus caudatus). Food Chem. 120:128–133. doi:10.1016/J.FOODCHEM.2009.09.087

Repo-Carrasco, R., C. Espinoza, and S.-E. Jacobsen. 2003. Nutritional Value and Use of the Andean Crops Quinoa (Chenopodium quinoa) and Kañiwa (Chenopodium pallidicaule). Food Rev. Int. 19:179–189. doi:10.1081/FRI-120018884

Richards, S. 2018. Full disclosure: Genome assembly is still hard. PLOS Biol. 16:e2005894. doi:10.1371/journal.pbio.2005894

Ruas, P.M., A. Bonifacio, C.F. Ruas, D.J. Fairbanks, and W.R. Andersen. 1999. Genetic relationship among 19 accessions of six species of Chenopodium L., by Random Amplified Polymorphic DNA fragments (RAPD). Euphytica 105:25–32. doi:10.1023/A:1003480414735

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

Smit, A., and R. Hubley. 2008. RepeatModeler Open-1.0. http://www.repeatmasker.org/ (accessed February 11, 2019).

Smit, A., R. Hubley, and P. Green. 2013. RepeatMasker Open-4.0. http://www.repeatmasker.org/ (accessed February 11, 2019).

Štorchová, H., J. Drabešová, D. Cháb, J. Kolář, and E.N. Jellen. 2015. The introns in flowering locus t-like (FTL) genes are useful markers for tracking paternity in tetraploid Chenopodium quinoa Willd.. Genet. Resour. Crop Evol. 62:913–925. doi:10.1007/s10722-014-0200-8

Tillich, M., P. Lehwark, T. Pellizzer, E.S. Ulbricht-Jones, A. Fischer, R. Bock, and S. Greiner. 2017. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 45:W6–W11. doi:10.1093/nar/gkx391

Vargas, A., D.B. Elzinga, J.A. Rojas-Beltran, A. Bonifacio, B. Geary, M.R. Stevens, E.N. Jellen, and P.J. Maughan. 2011. Development and use of microsatellite markers for genetic diversity analysis of cañahua (Chenopodium pallidicaule Aellen). Genet. Resour. Crop Evol. 58:727–739. doi:10.1007/s10722-010-9615-z

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

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

Waterhouse, R.M., M. Seppey, F.A. Sim~, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V Kriventseva, E.M. Zdobnov, and M. Rosenberg. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. doi:10.1093/molbev/msx319

Wilson, H.D. 1990. Quinua and Relatives (Chenopodium sect.Chenopodium subsect.Celluloid). Econ. Bot. 44:92–110. doi:10.1007/BF02860478

FIGURES

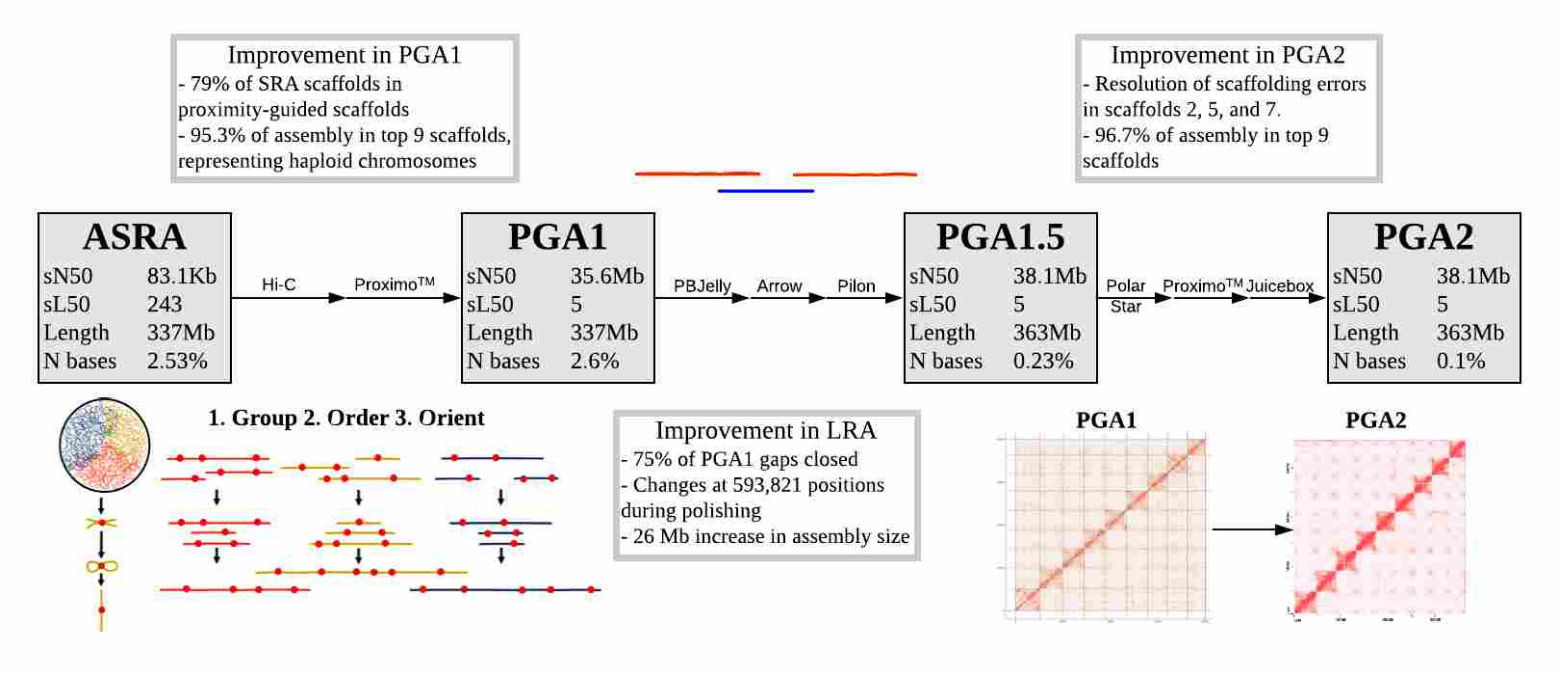

Figure 1. Outline of the Genome Assembly Process. An initial assembly was developed from Illumina short-reads using the ALLPATHS-LG assembler (ASRA). A first proximity-guided assembly was performed using Hi-C data and the Proximo^{TM} pipeline (PGA1), diagrammed in the bottom left of the flowchart. Overlapping chromatin was formalin-fixed, the genome was fragmented, then fixed fragments were selected and circularized. Illumina reads were generated and forward and reverse reads were aligned to the ASRA scaffolds. Crosslink frequency was used to first group, then order, then orient the scaffolds along pseudochromosomes. Proximity-guided assembly was followed by gapfilling with PacBio long-reads, as demonstrated in the top center, and genome-polishing by Arrow and Pilon (PGA1.5). PGA1.5 was broken at all N-gaps and areas of low PacBio read coverage (PolarStar), then underwent a second round of proximity-guided assembly (PGA2). A comparison of PGA1 and PGA2 is shown in the bottom right of the diagram, where increasing frequency of cross-linking is illustrated by increasing color intensity.

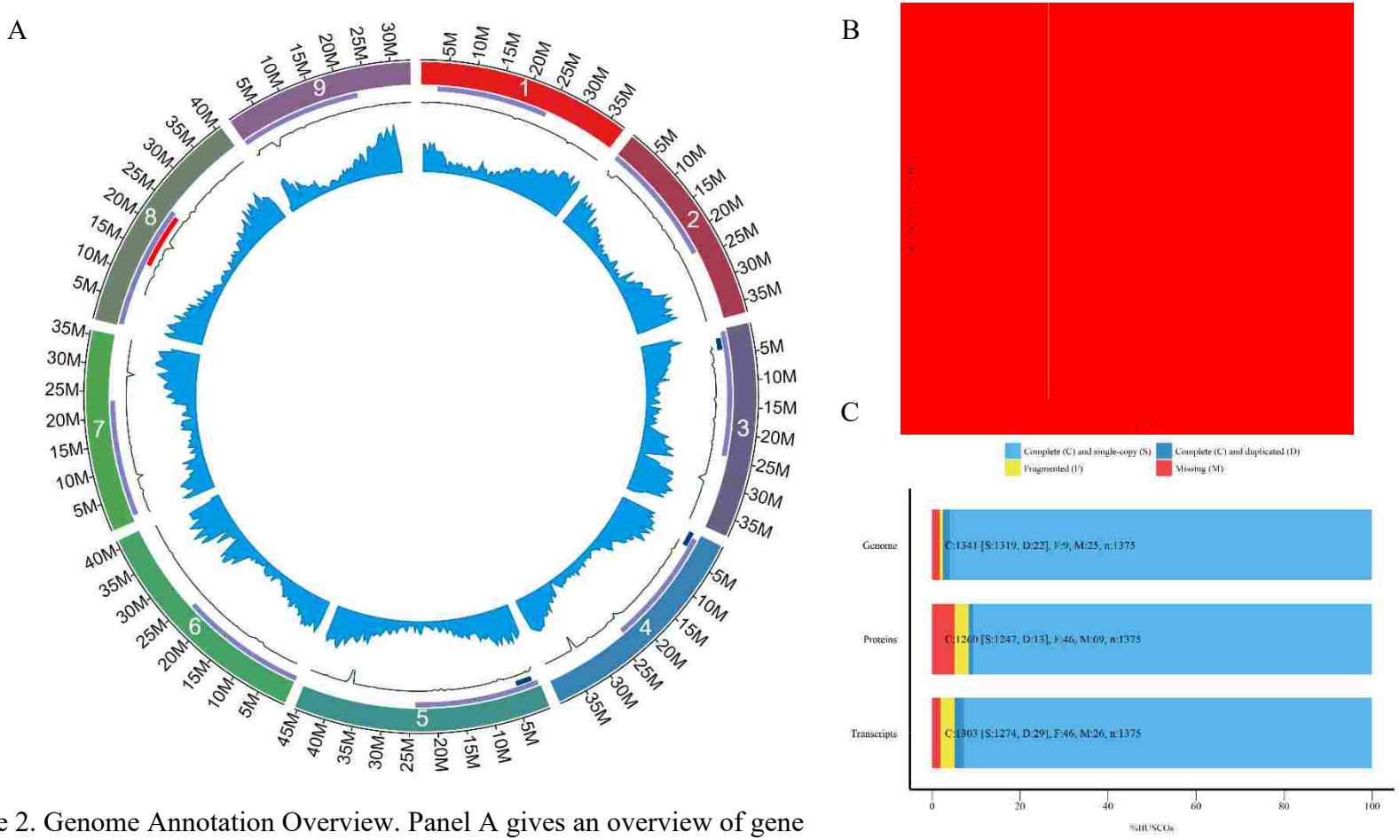

Figure 2. Genome Annotation Overview. Panel A gives an overview of gene and repetitive element annotations. Track 1: Pseudochromosome names and sizes; Track 2: Frequency of pericentromeric 12-13P repetitive elements (purple); Track 3: Frequency of 18-24J repetitive element (blue) and the 5S rRNA locus (red); Track 4: Frequency of canonical telomeric repeat; Track 5: Gene density. Panel B shows the distribution of annotation edit distance (AED) metrics for features annotated by MAKER. Annotations with an AED value < 0.25 are considered high-quality. Panel C compares BUSCO assessments of PGA2, protein annotations, and transcript annotations.

A

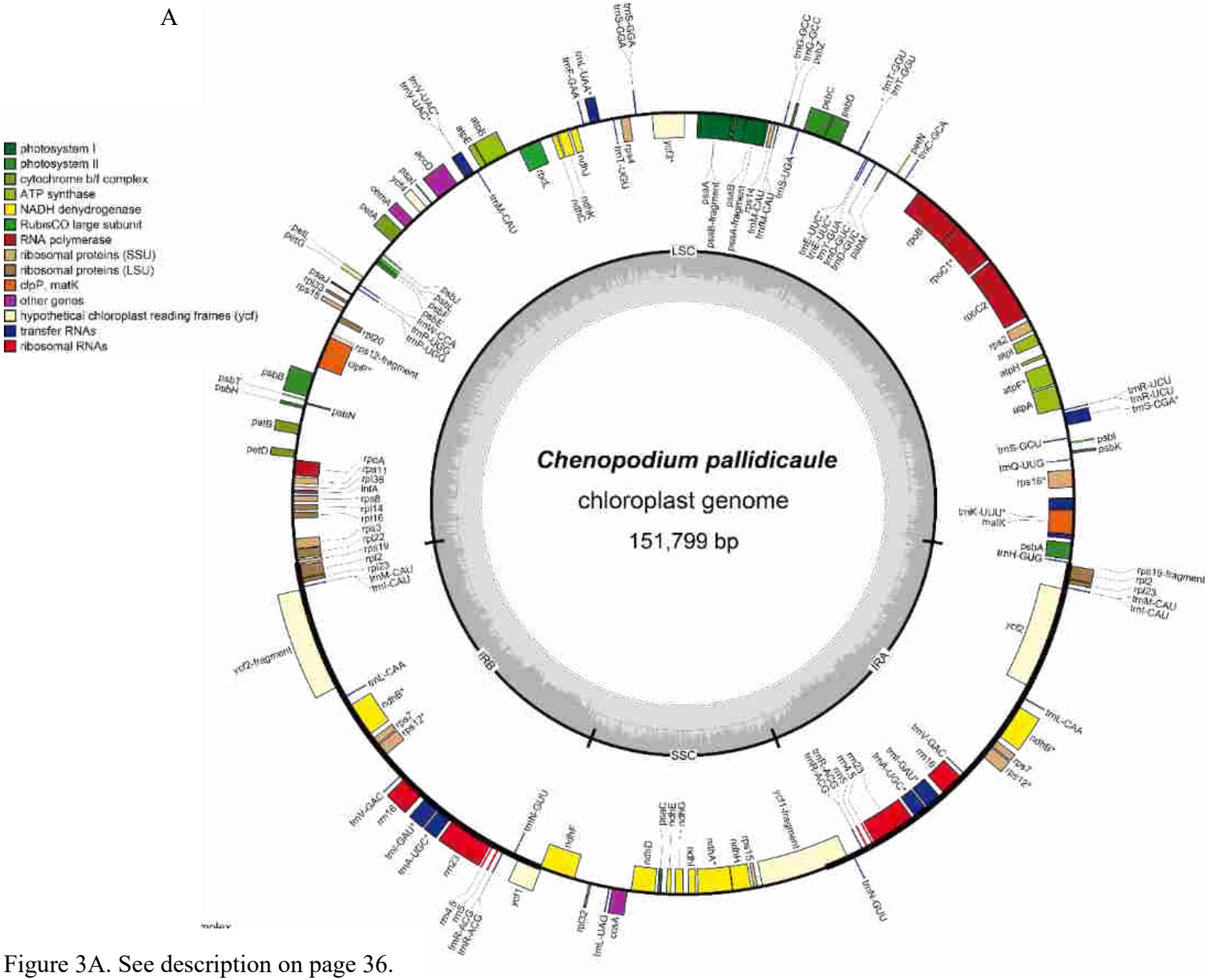

Figure 3A. See description on page 36.

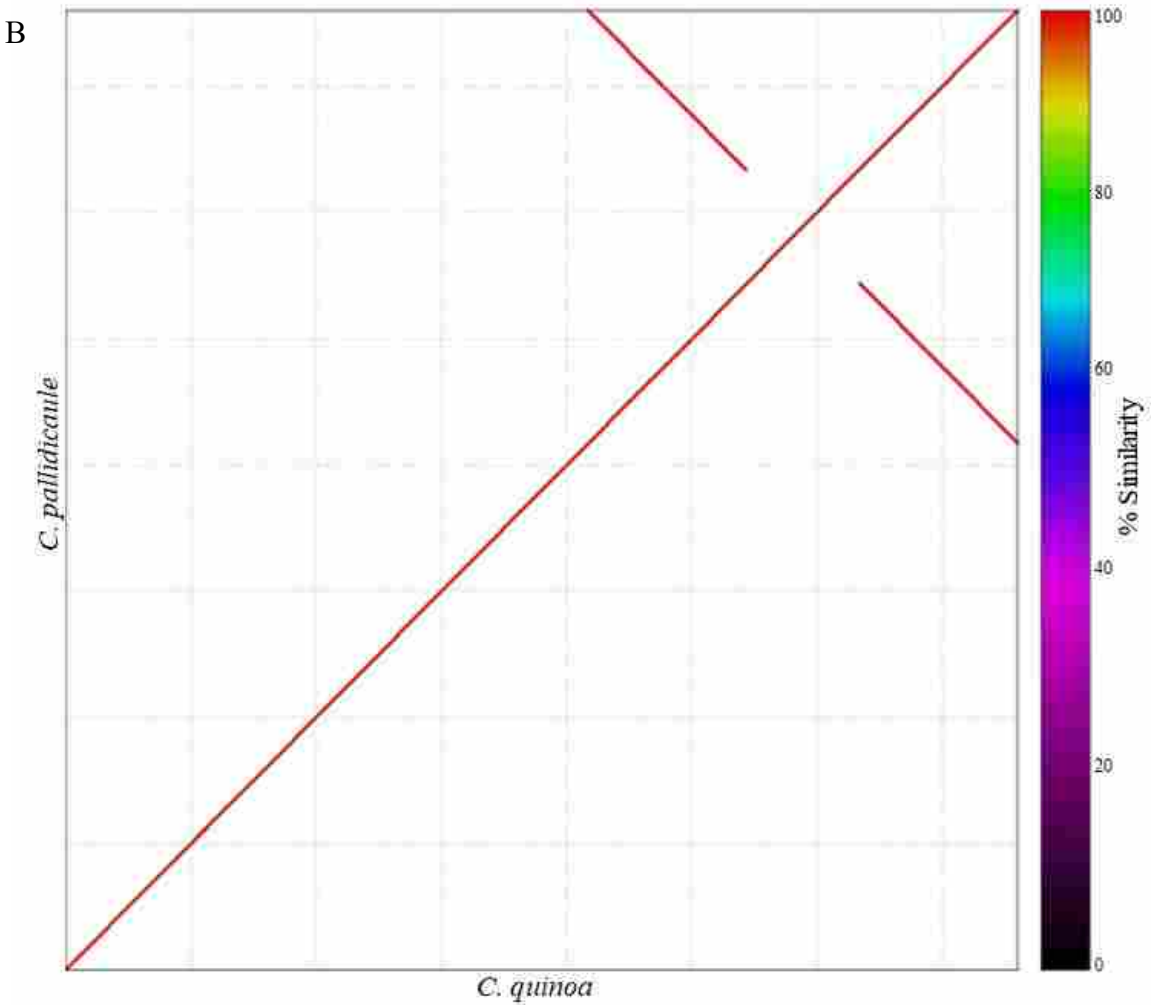

Figure 3B. Assembly and Annotation of Cañahua Chloroplast. In panel A, the outside track shows genes transcribed in a clockwise direction. The second track shows genes transcribed in a counterclockwise direction and the inside track shows G/C content levels. Annotation reveals a quadripartite structure, including two copies of the IR (bolded line) dividing large and small single-copy regions. Panel B is a comparison of the cañahua and quinoa chloroplast genomes generated by MUMmer. Dark red indicates regions of homology.

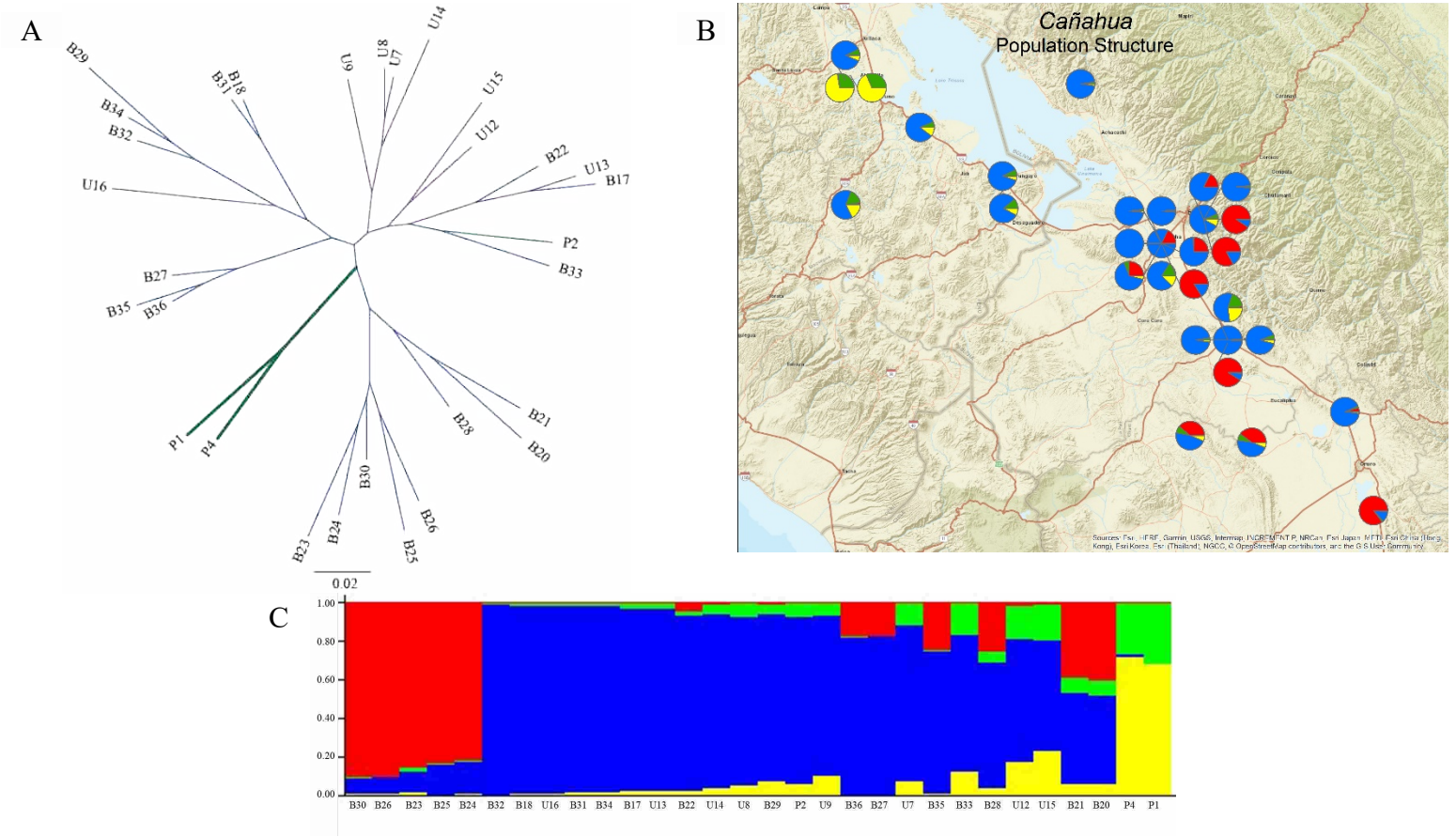

Figure 4. Diversity Panel. The unrooted tree in panel A was designed using 16,194 SNPs filtered to remove SNPs with > 10% missing data, minor allele frequency < 5%, and LD < 40%. Colors represent the collection source (purple = USDA, green = UNALM, blue = UMSA), and bolded lines indicate wild accessions. Panel B shows geographic location (see Table 1 for passport information) combined with population structure information developed by Structure with K = 4. There is no significant correlation between collection site and genetic distance (p = 0.837). Panel C further illustrates population structure in the diversity panel.

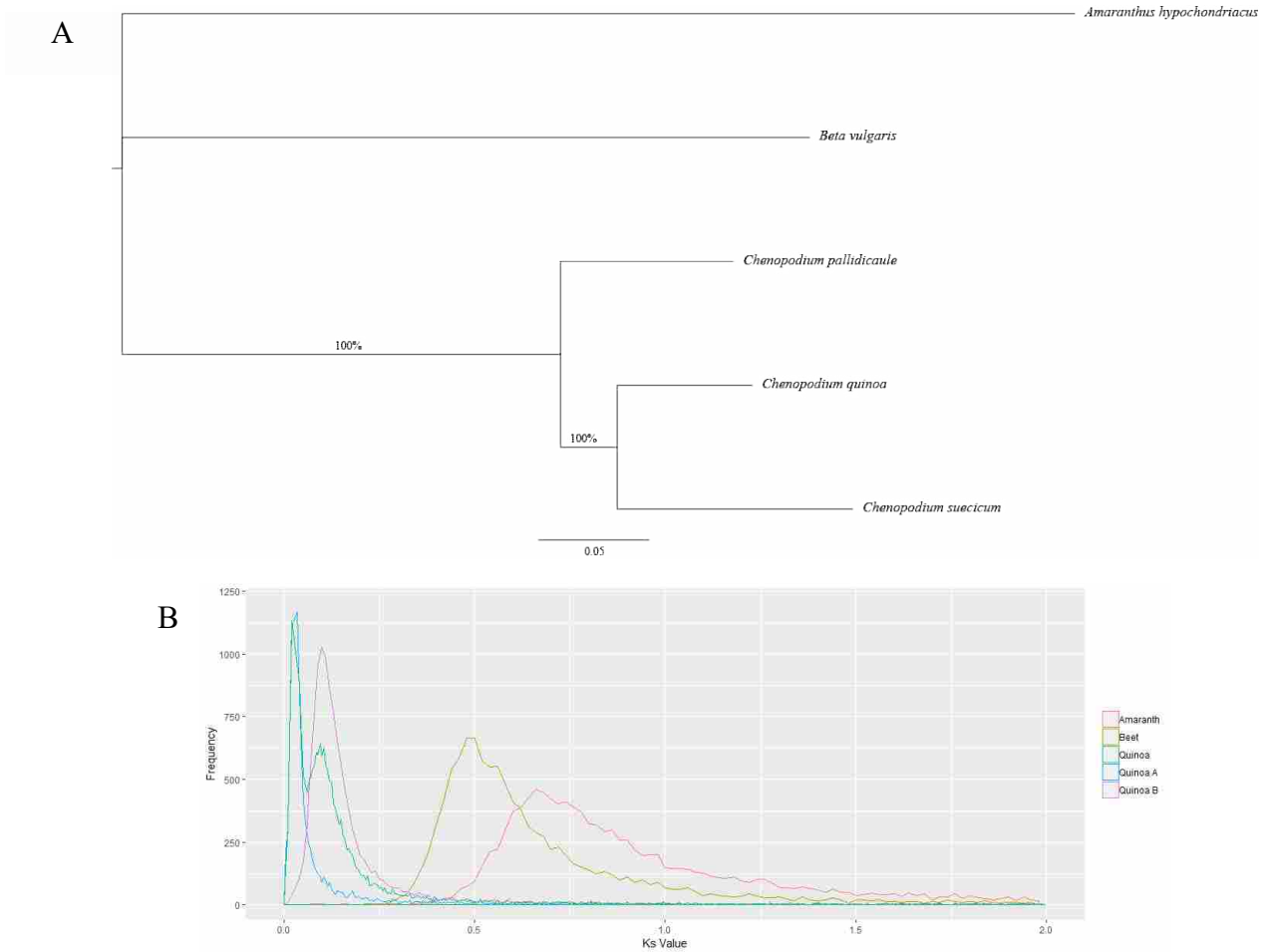

Figure 5. *Amaranthaceae* Relationships. Panel A. A phylogeny developed by IQ-TREE using the VT+F+G4 model and including several important members of the *Amaranthaceae* family was developed using 254 conserved genes. Percentages at two nodes reflect the percent agreement after 1,000 rounds of boostrapping. Branch lengths are calculated by number of nucleotide substitutions per codon site. Panel B provides Ks value distributions in comparison to amaranth (red), beet (yellow-brown), tetraploid quinoa (green), the A-subgenome of quinoa (blue), and the B-subgenome of quinoa (purple).

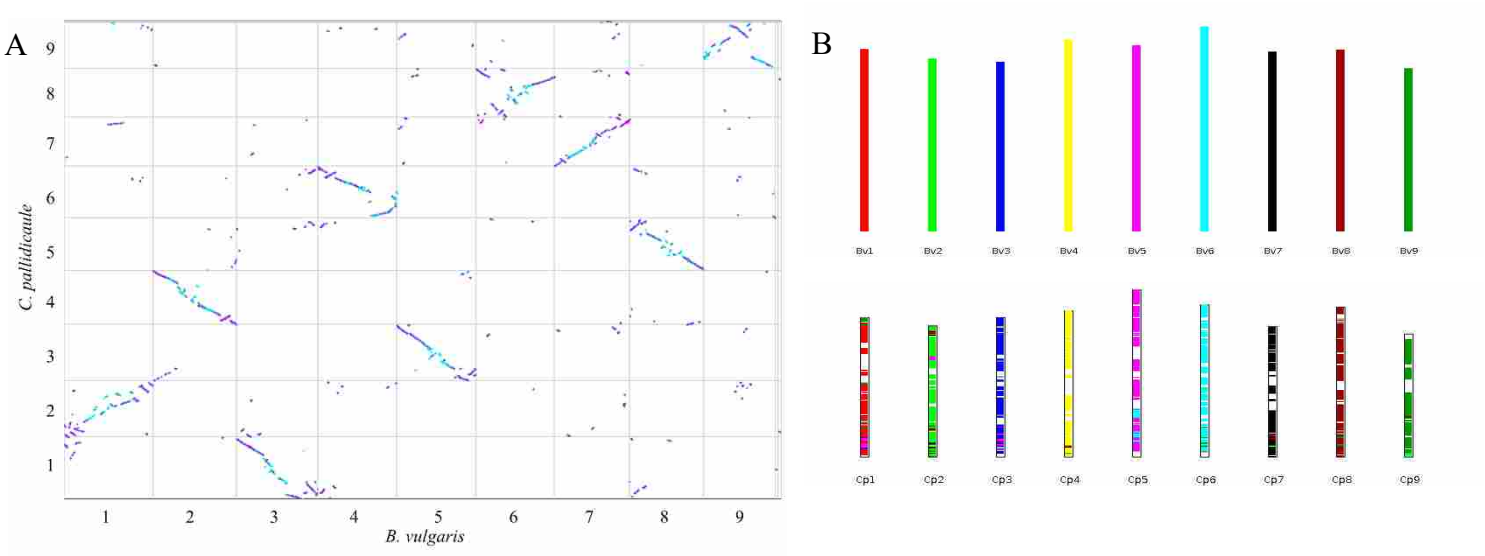

Figure 6. Genomic Comparison with Beet. Panel A shows collinearity between cañahua and beet output by CoGe. Darker color indicates greater homology. Panel B is an MCScanX bar chart comparison of the beet (top) and cañahua (bottom) chromosomes. This comparison was used to name the cañahua chromosomes.

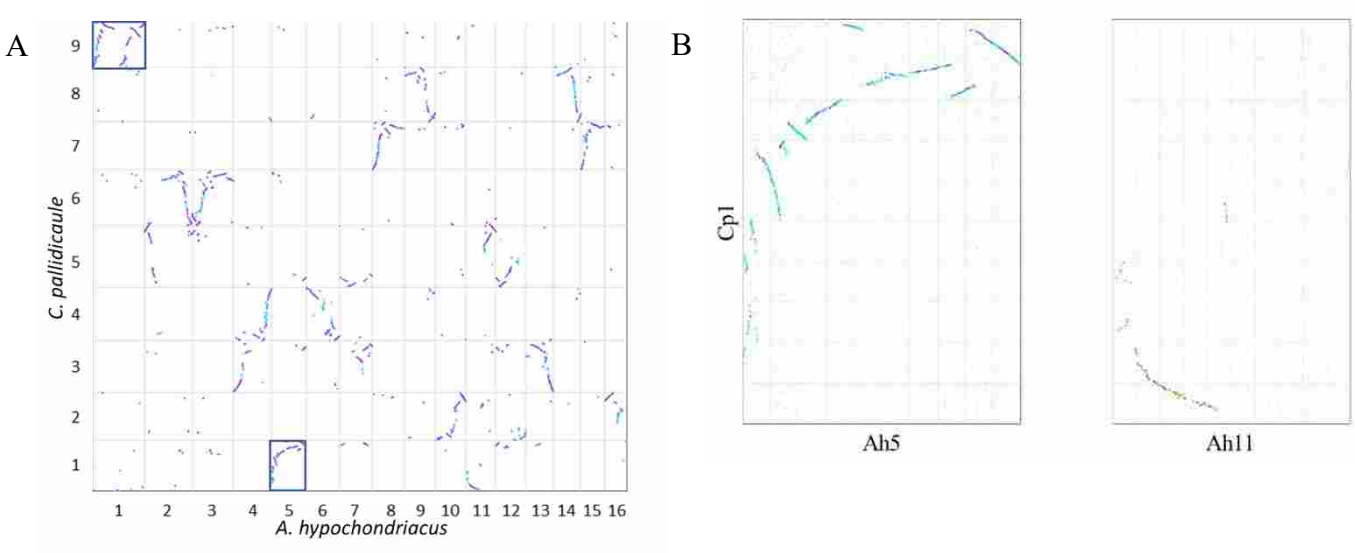

Figure 7. Genomic Comparison with Amaranth. A CoGe dot plot showing syntenous regions between cañahua and amaranth coding sequence is shown in panel A. A close-up image of amaranth chromosomes 5 and 11 is shown in comparison to cañahua chromosome 1 in panel B. Increasing color intensity is associated with increasing homology.

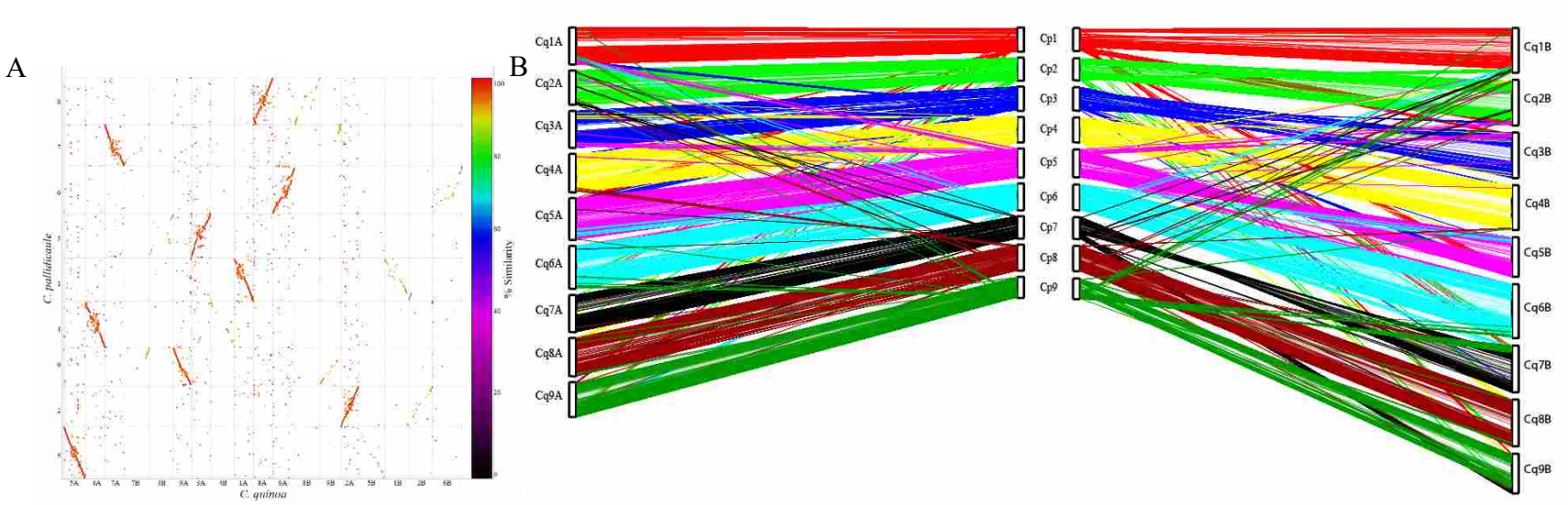

Figure 8. Genomic Comparison with Quinoa. Panel A gives a MUMmer dotplot comparison of cañahua and quinoa whole genomes. Areas of high homology are dark red. The ribbon chart in panel B divides the quinoa genome into A (left) and B (right) subgenomes with cañahua in the center.

TABLES

Table 1. Passport and ecotype information for plant materials used.

Name	Collection	Accession ID	Collection Location	Altitude (masl[a])	Ecotype
P1	UNALM	BYU 1780	-15.6967, -70.20510	3,830	wild
P2	UNALM	BYU 1781	-15.7268, -70.23560	3,838	NA
P4	UNALM	BYU 1785	-15.7693, -70.27050	3,860	wild
U7	USDA	PI 510525	-16.36284270, -69.27651950	NA	NA
U8	USDA	PI 510526	-16.28333333, -69.28333333	NA	NA
U9	USDA	PI 510527	-16.00000000, -69.78333333	3,810	NA
U12	USDA	PI 510530	-16.45000000, -70.23333333	NA	NA
U13	USDA	PI 665279	-17.233333, -67.91666667	3,700	NA
U14	USDA	PI 665280	-17.233333, -67.91666667	3,700	NA
U15	USDA	PI 665281	-17.23333333, -67.91666667	3,700	NA
U16	USDA	PI 665282	-17.23333333, -67.91666667	3,700	NA
B17	UMSA	Bol-1.1	-15.74722222, -68.80916667	3,845	saguia
B18	UMSA	Bol-3.1	-16.53444444, -68.06222222	3,445	saguia
B20	UMSA	Bol-19.1	-17.82416667, -67.77027778	3,721	saguia
B21	UMSA	Bol-20.123	-17.785, -68.14472222	4,025	saguia
B22	UMSA	Bol-21.123	-17.64833333, -67.20722222	3,777	saguia
B23	UMSA	Bol-22.123	-18.216666667, -67.0333333	3,707	saguia
B24	UMSA	Bol-23.123	-16.53444444, -68.06222222	3,445	lasta
B25	UMSA	Bol-24.123	-16.67402778, -68.31833343	3,900	saguia
B26	UMSA	Bol-25.123	-16.53444444, -68.06222222	3,445	saguia
B27	UMSA	Bol-26.123	-16.53444444, -68.06222222	3,445	saguia
B28	UMSA	Bol-28.123	-16.67402778, -68.31833343	3,900	saguia
B29	UMSA	Bol-29.123	-16.53444444, -68.06222222	3,445	saguia
B30	UMSA	Bol-30.123	-17.25, -67.91666667	3,800	saguia
B31	UMSA	Bol-4.3	-16.67402778, -68.31833333	3,900	saguia
B32	UMSA	Bol-6.2	-16.67402778, -68.31833334	3,900	saguia
B33	UMSA	Bol-7.1	-16.67402778, -68.31833336	3,900	saguia
B34	UMSA	Bol-8.1	-16.67402778, -68.31833337	3,900	saguia
B35	UMSA	Bol-13.3	-16.67402778, -68.31833342	3,900	saguia
B36	UMSA	Bol-27.123	-16.67402778, -68.31833343	3,900	saguia
Reference	USDA	PI 478407	-17.23333333, -67.91666667	3,800	NA
C. sonorensis	BYU	BYU 17220	31.6104, -111.0512	NA	NA
C. watsonii	BYU	BYU 873	34.51477, -112.00698	NA	NA

[a]Altitude reported in meters above sea level

NA indicates missing data.

Table 2. Sequencing statistics for the thirty-accession diversity panel.

Accession	Paired Reads	Length of Reads (Gb[a])	Coverage
P1	28,567,002	4.29	12.04
P2	27,965,508	4.19	11.78
P4	33,456,662	5.02	14.10
U7	43,367,360	6.51	18.27
U8	37,120,164	5.57	15.64
U9	31,141,872	4.67	13.12
U12	32,760,174	4.91	13.80
U13	33,520,294	5.03	14.12
U14	29,881,178	4.48	12.59
U15	32,163,912	4.82	13.55
U16	25,623,210	3.84	10.80
B17	28,533,140	4.28	12.02
B18	33,511,726	5.03	14.12
B20	27,702,300	4.16	11.67
B21	27,404,206	4.11	11.55
B22	28,864,446	4.33	12.16
B23	21,456,466	3.22	9.04
B24	26,451,270	3.97	11.15
B25	23,829,974	3.57	10.04
B26	28,803,002	4.32	12.14
B27	32,251,040	4.84	13.59
B28	35,815,108	5.37	15.09
B29	27,852,768	4.18	11.74
B30	37,193,432	5.58	15.67
B31	34,301,400	5.15	14.45
B32	31,909,762	4.79	13.45
B33	34,695,570	5.20	14.62
B34	34,158,064	5.12	14.39
B35	29,245,244	4.39	12.32
B36	34,029,356	5.10	14.34

[a]Read length was measured in gigabases

Table 3. Assembly statistics for the ASRA, PGA1, PBJelly2, and PGA2 assemblies.

Assembly Name	**ASRA[a]**	**PGA1[b]**	**PGA1.5[c]**	**PGA2[d]**
Assembly size (Mb)	337	337	363	363
Number of scaffolds	3,015	623	591	4,633
Scaffold N50 size (Mb)	0.357	35.6	37.8	38.1
Scaffold L50 count	243	5	5	5
Longest scaffold (Mb)	2.95	40.4	43.2	45.5
Number of contigs	8,984	8,984	2,580	8,210
Contig N50 size (Mb)	0.0831	0.0831	0.516	0.236
Contig L50 count	1,096	1,096	168	401
% missing bases	2.53	2.6	0.23	0.1
Assembly size (Mb) in top 9 scaffolds	19.6	321	344	350
Assembly % in top 9 scaffolds	5.82	95.4	94.8	96.5

[a]ALLPATHS-LG Short-Read Assembly
[b]Proximity-Guided Assembly 1
[c]Proximity-Guided Assembly 1.5
[d]Proximity-Guided Assembly 2

Table 4. PacBio SMRT cell statistics.

	Total Reads	**Mean Read Length (Kb[a])**	**Total Size (Gb[b])**
Cell 1	218,650	8.55	1.87
Cell 2	429,650	9.37	4.02
Cell 3	452,902	9.53	4.32
Merged	1,101,202	9.27	10.21

[a]Mean read length measured in kilobases
[b]Total size of cell output measured in gigabases

Table 5. Length and contig number for each chromosome-scale scaffold in PGA2.

Scaffold Name	Contigs	Length (Mb[a])
Cp1	366	37.93
Cp2	376	35.65
Cp3	347	38.12
Cp4	413	39.85
Cp5	474	45.40
Cp6	423	41.46
Cp7	376	35.49
Cp8	480	40.69
Cp9	331	33.52
Remaining Contigs	4,632	14.40
Total	8,218	362.51

[a]Scaffold length reported in megabases

Table 6. RNA-seq summary statistics for six unique tissue and treatment combinations used for transcriptome assembly.

Tissue	Treatment	Reads	Total Size (Gb[a])
Root	Control	114,255,878	11.4
Root	Salt	117,615,336	11.8
Leaf	Control	102,807,950	10.3
Leaf	Salt	114,209,984	11.42
Apical Meristem	Control	113,348,714	11.3
Flower	Control	101,256,094	10.1
Mean	--	110,582,326	11.1
Total	--	663,493,956	66.3

[a]Size of combined RNA-seq reads reported in gigabases.

Table 7. Repetitive element classification statistics for PGA2 as ouput by RepeatMasker.

Repeat Class	Repeat Name	Count	bp Masked	% Masked
DNA		237	61604	0.0002
	CMC-EnSpm	18730	7688255	2.12%
	Ginger	145	12961	0.00%
	MULE-MuDR	9015	4316160	1.19%
	MuLE-MuDR	6356	4827397	1.33%
	Novosib	199	22859	0.01%
	PIF-Harbinger	2200	804514	0.22%
	TcMar-Stowaway	12969	2351479	0.65%
	hAT	105	16558	0.00%
	hAT-Ac	8804	4597051	1.27%
	hAT-Charlie	130	40735	0.01%
	hAT-Tag1	3105	968983	0.27%
	hAT-Tip100	452	158540	0.04%
LINE		--	--	--
	Ambal	32	9891	0.00%
	CRE-II	1180	1110118	0.31%
	Jockey	302	146171	0.04%
	L1	4538	2181207	0.60%
	L1-Tx1	300	286875	0.08%
	L2	159	174531	0.05%
	R1	116	81049	0.02%
	RTE-BovB	3072	765582	0.21%
LTR		1739	373682	0.10%
	Caulimovirus	115	147914	0.04%
	Copia	24753	30898328	8.53%
	ERV1	990	99174	0.03%
	Gypsy	46095	67242074	18.57%
RC		--	--	--
	Helitron	512	108896	0.03%
Retroposon		146	42357	0.01%
Unknown		201582	60630384	16.75%
Total		348078	1.9E+08	52.52%
Low complexity		17556	931439	0.26%
Satellite		264	147437	0.04%
Simple repeat		100039	5940276	1.64%

Table 8. Comparison of gene synteny, mutation rates, and divergance of last common ancestor in *Amaranthaceae* species.

	Amaranth	**Beet**	**Quinoa A-subgenome**	**Quinoa B-subgenome**
Synteny Blocks	802	522	990	993
Total Syntenous Features	15,153	13,436	15,282	14,646
Percent of Total Features	64.00%	55.40%	71.60%	65.58%
Ks[a] Peak Value	0.64	0.48	0.025	0.05
Divergance of Last Common Ancestor (MYA[b])	21.33 - 39.51	16 - 29.63	0.830 - 1.54	1.67 - 3.09

[a]The Ks value represents synonymous substitutions per synonymous site.
[b]Divergance of last common ancestor reported as million years ago.

Table 9. Comparison of gene synteny between cañahua and the two subgenomes of quinoa.

Quinoa Chromosome	**Syntenous Blocks**	**Total Syntenous Genes**
Cq1A	39	1,456
Cq1B	29	829
Cq2A	32	1,409
Cq2B	32	973
Cq3A	27	1,556
Cq3B	32	1,505
Cq4A	41	1,444
Cq4B	41	1,491
Cq5A	39	1,802
Cq5B	34	1,695
Cq6A	31	1,712
Cq6B	38	1,597
Cq7A	33	1,444
Cq7B	32	713
Cq8A	32	1,428
Cq8B	36	1,364
Cq9A	26	1,323
Cq9B	19	536
A-Subgenome Total	300	13,574
B-Subgenome Total	293	10,703

Table 10. Summary of BWA alignments of three *Chenopodium* A-genome diploids to quinoa.

	C. pallidicaule	***C. watsonii***	***C. sonorensis***
Aligned Reads (%)	95.35	98.36	98.34
Mismatch (%)	3.46	3.19	3.41
Error Rate (%)	3.55	3.03	3.26
Reads Aligned in Pairs (%)	98.53	99.31	99.23

SUPPLEMENTAL MATERIAL

Supplemental Material 1: QIAGEN Genomic Tip Altered Protocol (midi tip 100/G)

1. Prepare buffers G2, QBT, GC, and QF according to instructions

2. For each prep, add 19ul of RNase A stock solution (100mg/ml) to a 9.5ml aliquot of Buffer G2.

3. Grind tissue (about 100mg) to a fine powder using liquid nitrogen in a precooled mortar and pestle. Grind as thoroughly as possible.

4. Transfer the ground tissue from step 3 to a 50ml screw-cap tube. Add 9.5ml of Buffer G2 (with RNase A) and .5 ml of Proteinase K stock solution. Mix well by vortexing.

5. Incubate at 20ºC overnight. Lysate should be clear after incubation to avoid clogging the tip. Centrifuge at 5000 x g for 10 minutes at 4oC to remove any particulate matter before loading. Take a 300ul aliquot and save for an analytical gel (aliquot 1).

6. Equilibrate QIAGEN genomic-tip 100/G with 4ml of Buffer QBT and allow QIAGEN genomic-tip to empty by gravity flow. Do not force out remaining buffer – a small amount will remain to keep the tip hydrated.

7. Vortex the sample for 10 seconds at maximum speed and apply it to the equilibrated QIAGEN genomic-tip. Allow it to enter the resin by gravity flow (less vortexing may result in longer genomic DNA segments). Take a 300ul aliquot and save for analytical gel (aliquot 2).

8. Wash QIAGEN genomic-tip with 3 x 7.5ml of Buffer QC. Take a 600ul aliquot of the flow-through and save for an analytical gel (aliquot 3).

9. Elute genomic DNA with 5ml of Buffer QF (pre-warmed to 50ºC) into a clean 10ml collection tube (preferably not polycarbonate).

10. Precipitate DNA by inverting the tube 10-20 times. Centrifuge immediately at >5000 x g for at least 15 minutes at 4ºC. Carefully remove supernatant.

11. Wash the pellet with 2ml of cold 70% ethanol. Centrifuge at >5000 x g for 10 minutes at 4ºC. Carefully remove supernatant without disturbing pellet. Air-dry for 5-10 minutes and resuspend in .1 ml of TE (pH 7.5) buffer. Dissolve DNA overnight on a shaker.

12. An analytical gel can be run using the aliquots to determine the source of errors if results are not good.