The Genome of *Chenopodium ficifolium*: Developing Genetic Resources and a Diploid Model System for Allotetraploid Quinoa

Clayton D. Ludwig[1], Peter J. Maughan[2], Eric N. Jellen[2], Thomas M. Davis[1]

Affiliations:

[1]Department of Agriculture, Nutrition, and Food Systems, University of New Hampshire, Durham, NH 03824, USA

[2]Brigham Young University, Department of Plant and Wildlife Sciences, College of Life Sciences, Provo, UT 84602, USA.

Abbreviations:

AED, annotation edit distance; EST, expressed sequence tag; GO, gene ontology; GWAS, genome wide association study; HCGS, UNH Hubbard Center for Genome Studies; HMW, high molecular weight; MSA, multiple sequence alignment; ONT, Oxford Nanopore Technology; PC, pseudochromosome

Abstract

High-quality nuclear, chloroplast, and preliminary mitochondrial genomes have been assembled and annotated for the B-genome diploid (BB: 2n = 2x = 18) figleaf goosefoot (*Chenopodium ficifolium*). The primary objective was to advance a simplified model system for genetic characterization and improvement of allotetraploid (AABB: 2n = 4x = 36) quinoa (*Chenopodium*

quinoa), a nutritionally valuable, halophytic orphan crop. In addition to its diploidy and favorably small genome size, the *C. ficifolium* model provides a shorter generational period and smaller overall plant size as compared to *C. quinoa*, while displaying relevant agronomic trait variations amenable to gene-trait association studies. The *C. ficifolium* 'Portsmouth' nuclear genome was sequenced using PacBio HiFi Long Read technology and assembled using Hifiasm. After manual adjustments, the final ChenoFicP_1.0 assembly consisted of nine pseudochromosomes spanning 711.5 Mbp, while 22,617 genes were identified and annotated. BUSCO analyses indicated a nuclear genome completeness of 97.5%, and a proteome and transcriptome completeness of 98.4 percent. The chloroplast genome assembly detected two equally represented structural haplotypes differing in the orientation of the Short Single Copy region relative to the Long Single Copy region. Phylogenetic and parentage analyses pointed to an unspecified AA diploid species and away from *C. ficifolium* as the likely maternal chloroplast and mitochondrial genome donor(s) during the initial tetraploidization event in the *C. quinoa* lineage. Using the new ChenoFicP_1.0 reference genome, a GWAS was performed on a previously studied *C. ficifolium* F2 population to further define region(s) implicated in the control of three key agronomic traits: days to flowering, plant height, and branch number. This analysis localized control of all three traits to a 7 Mb interval on pseudochromosome Cf4. This region contains approximately 770 genes, including the *FTL1* locus, thus confirming and extending our prior, single-marker analysis showing association of these three traits with an *FTL1* amplicon length polymorphism. The use of these data to further develop *C. ficifolium* as a model species for genetics and breeding of quinoa serves to expand knowledge and germplasm resources for quinoa improvement.

Introduction

Chenopodium ficifolium (figleaf goosefoot) is a weedy, Old-World B-genome diploid (BB: 2n = 2x = 18) relative of the cultivated allotetraploid (AABB: 2n = 4x = 36) pseudocereal crop quinoa (*Chenopodium quinoa* Willd.) (Walsh *et al.*, 2015; Kolano *et al.*, 2016) and its wild sister AABB tetraploid *Chenopodium berlandieri* (Walsh *et al.*, 2015). The genus *Chenopodium* now resides in the family Amaranthaceae, its former Chenopodiaceae family home having recently been subsumed into the Amaranthaceae (Chase *et al.*, 2016), which also hosts the relevant comparator *Beta vulgaris* (sugar beet). Putatively originating from Eurasia (Walsh *et al.*, 2015), figleaf goosefoot is now widespread in North America, both as a weed of human disturbance and as a naturalized part of disturbed ecosystems. It is one of two recognized potential contributors, or a close relative thereof, of the B subgenome to the *C. quinoa* lineage during its initial tetraploidization event, the other B subgenome candidate being diploid *C. suecicum* (Jellen *et al.*, 2011; Mandák *et al.*, 2012; Walsh *et al.*, 2015; Jarvis *et al.*, 2017). Because figleaf goosefoot potentially contributed the B subgenome to quinoa, or is an extant sister species of the contributor, it is a highly relevant candidate to explore as a diploid model species in relation to quinoa's genetic characterization and breeding, and its study may shed light on specific genes governing domestication-related traits in quinoa (Jellen *et al.*, 2013; Walsh *et al.*, 2015; Kolano *et al.*, 2016; Subedi, Neff and Davis, 2021).

Absent a history of intensive breeding, *C. quinoa* has been categorized as an 'orphan crop' (Lemmon *et al.*, 2018; Kumar and Bhalothia, 2020). However, quinoa is receiving increasing interest from consumers due to its high nutritional value and from agronomists and breeders due to its drought and salt tolerance (Gorinstein *et al.*, 2002; Rojas *et al.*, 2009; Vega-Gálvez *et al.*, 2010; Jacobsen, 2011). These and other attributes make quinoa an attractive candidate for

67 production in regions that cannot sustain the major grain crops. Furthermore, quinoa has the
68 potential to serve as a valuable secondary crop, thereby contributing to food security. Thus, it is
69 an attractive candidate for increased production, justifying investment in development of novel
70 genetics and methods to streamline the breeding process (Jacobsen, 2003; Bazile, Jacobsen and
71 Verniau, 2016).

72 Using *C. ficifolium* as a diploid model species has the potential to accelerate trait dissection and
73 gene identification due to its comparative genetic simplicity; shorter generation time of ~40 days
74 compared to quinoa's 90-120 days (Jacobsen, 1997); profuse flower, pollen, and seed production;
75 and smaller plant size facilitating growth chamber cultivation (Subedi et al., 2021). Building
76 upon prior, foundational studies (Štorchová *et al.*, 2015, 2019), the targeted development of *C.*
77 *ficifolium* as a diploid model species was initiated (Subedi, 2020; Subedi, Neff and Davis, 2021)
78 with a focus on: i) producing intraspecific hybrids and segregating progenies; ii) investigating
79 segregation in domestication-related traits of relevance to *C. quinoa* breeding programs; and iii)
80 detecting gene-trait associations for key genes and traits.

81 In an F_2 segregating population from an initial cross involving *C. ficifolium* accessions from
82 Portsmouth ('P') NH USA, and Quebec City ('QC') Quebec Canada, allelic variation in the
83 *Flowering Locus T-Like 1* (*FTL1*) marker locus was found to be associated with variation in days
84 to flower, plant height, and branch number (Subedi, 2020; Subedi, Neff and Davis, 2021). Using
85 previously designed primers (Cháb *et al.*, 2008; Štorchová *et al.*, 2015) to target and
86 conveniently genotype an indel polymorphism within the *FTL1* gene, it was found that F2 plants
87 homozygous for the 'P' allele flowered on average 12 days earlier than plants homozygous for the
88 'QC' allele, and two days earlier than heterozygotes (Subedi, 2020; Subedi, Neff and Davis,
89 2021). As a likely consequence of their decreased maturation time, 'P' allele homozygotes were

also significantly shorter and had fewer branches than plants of the alternate genotypes. These results strongly implicated the *FTL1* chromosomal region in the control of three agronomically important traits, but did not rule out the possibility of influence from other genomic regions. While genotype-phenotype relationships involving *FTL1* must be further studied at both the diploid and the tetraploid level, this research demonstrates the potential value of *C. ficifolium* as a model system in which to accelerate discovery and characterization of gene-trait relationships. The goal of the work reported here is to expand knowledge and establish foundational genomic resources for the *C. ficifolium* model system.

Materials and Methods

Germplasm

All plant cultivation and phenotyping activities were conducted at the New Hampshire Agricultural Experiment Station (NHAES) at the University of New Hampshire (UNH). Two figleaf goosefoot accessions were employed in this study, both of which have been deposited as seed samples with the USDA North Central Plant Introduction Station (NCRPIS) at Ames Iowa. The 'Portsmouth' ('P') accession (PI 698433) was collected by Erin Neff and Thomas Davis in Portsmouth, New Hampshire, USA (Neff, 2017), while the 'Quebec City' ('QC') accession (PI 698434) was collected by Thomas Davis from Quebec City, Quebec, Canada. Crosses between the 'P' and 'QC' accessions to generate F_1 and F_2 populations, and confirmation of parentage using informative *FTL1* amplicons, were as previously described (Subedi, Neff and Davis, 2021). For purposes of comparison, we also employed in-house-numbered accessions 302-A of AA diploid *C. foggii*, and RB6 of AABB allotetraploid *C. berlandieri* var. *macrocalycium*, which

had been collected by Erin Neff and Thomas Davis in southern Maine (Neff, 2017; Neff, Sullivan and Davis, 2018) and in Rye Beach New Hampshire, respectively; and *C. quinoa* accession QQ065 (PI 614880), obtained from USDA NPGS.

DNA Isolation and Genome Sequencing

The figleaf goosefoot 'Portsmouth' ('P') accession was used to construct the primary *de novo* reference genome assembly. Using leaf tissue provided by the UNH investigators, DNA from the 'P' accession was isolated at Brigham Young University (BYU) using methods designed to yield high molecular weight (HMW) DNA from plant tissues (Vaillancourt and Buell, 2019). The 'P' accession HMW DNA was sequenced at the BYU DNA Sequencing Center (DNASC) using the PacBio HiFi platform (Eid *et al.*, 2009). In addition, this DNA was used at UNH to generate Nanopore sequence on an Oxford Nanopore Technologies (ONT) R9.4.1 flow cell using a library prepared with a rapid sequencing kit (SQK-RAD004) following methods described by Nanopore entitled “Rapid sequencing gDNA – whole genome amplification”.

Illumina sequence was generated at the UNH Hubbard Center for Genome Studies (HCGS) for the following: for the 'P' and 'QC' accessions of *C. ficifolium*; for 35 F2 generation individuals descended from a 'P' x 'QC' cross that was previously studied and described (Subedi et al., 2021); the RB6 accession of *C. berlandieri* var. *macrocalycium*; and for the 302-A accession of *C. foggii*. DNA for Illumina sequencing was isolated at UNH (Subedi et al., 2021) using a CTAB method described by Torres *et al.* (1993).

Nuclear Genome Assembly

Using a common set of PacBio reads, nuclear genome assemblies were constructed in parallel at both collaborating institutions: at UNH using Hifiasm Version v0.18.5-r499 (Cheng *et al.*, 2021, 2022) with default settings, and at BYU using Hifiasm Version 19 for use in the Chenopodium Pangenome Project (Jaggi *et al.*, 2024). Unless otherwise noted, all methods and results reported herein refer to the UNH assembly. Inspector (Chen *et al.*, 2021) was used to detect and correct errors in the Hifiasm assembly followed by manual examination and adjustments, yielding the *C. ficifolium* Version 1.0 nuclear reference genome designated as ChenoFicP_1.0. Finally, BUSCO (Manni *et al.*, 2021) and dependencies (Stanke *et al.*, 2008; Camacho *et al.*, 2009; Eddy, 2011; Mirarab, Nguyen and Warnow, 2012; Levy Karin, Mirdita and Söding, 2020; Li, 2023) using default settings and the embryophyta_odb10 database was used to calculate nuclear genome completeness. Telomeres were identified within pseudochromosomes (PCs) with the quarTeT toolkit program TeloExplorer (Lin *et al.*, 2023) and within unplaced contigs using tidk (Brown, la Rosa and Mark, 2023). Centromeric regions were identified using the tool CentroMiner which is supplied with the quarTeT toolkit (Lin *et al.*, 2023). Nuclear genome assembly data was visualized using Circos (Krzywinski *et al.*, 2009) in combination with the deepStats toolkit (Richard, 2019) for track generation.

Nuclear Genome Annotation

Once ChenoFicP_1.0 was finalized, repetitive elements were identified using RepeatModeler (Flynn *et al.*, 2020). Those repetitive elements were then masked using RepeatMasker (Tarailo-graovac and Chen, 2009; Flynn *et al.*, 2020; Storer *et al.*, 2021) as the first step in an annotation

156 pipeline patterned after Card *et al.* (2019). The RepeatMasker scripts
157 calcDivergenceFromAlign.pl and createRepeatLandscape.pl were used to analyze output files,
158 determine repetitive element content, and produce data visualizations. A BRAKER IsoSeq
159 compatible singularity container (Bruuna, Gabriel and Hoff, 2024) was passed the fasta file
160 containing soft-masked repetitive elements and Pacific Biosciences IsoSeq data aligned to the
161 genome using minimap2 (Li, 2018). BRAKER flags were set to use the Viridiplantae OrthoDB
162 database (Kuznetsov *et al.*, 2023) for protein sequence, and the Embryophyta odb10 database
163 (Manni *et al.*, 2021) as the Busco lineage.

164 The resulting GFF3 annotation file was passed to MAKER (Holt and Yandell, 2011) as a
165 prediction GFF file, as were the cds-transcripts and proteins files which were used for EST
166 evidence and protein homology evidence, respectively. MAKER was used to update annotations
167 and to provide Annotation Edit Distance (AED) scores. AED values for the final annotation file
168 were produced using AED_cdf_generator.pl (https://github.com/mscampbell). Once complete,
169 the output files for each pseudochromosome were merged to create one contiguous file. This
170 output file contained gene predictions, however without functional annotations. To integrate
171 functional details and GO terms, BLASTP (Johnson *et al.*, 2008) using the UniProtKB/Swiss-
172 Prot database (Bateman *et al.*, 2015), as well as InterProScan (Jones *et al.*, 2014) were passed the
173 protein file produced by MAKER, and the resulting information was integrated into a single file
174 using the maker_functional_gff and ipr_update_gff commands, and sorted with AGAT (Dainat,
175 Hereñú and Pucholt, 2020), which yielded the final GFF3 annotation file.

176 A genome-wide synteny analysis was performed comparing the ChenoFicP_1.0 assembly to the
177 B subgenome of the *C. quinoa* V2 reference genome (Rey *et al.*, 2023) and to the *Beta vulgaris*
178 EL10.1 reference genome (McGrath *et al.*, 2023). As a co-member of the Amaranthaceae family,

179 diploid (2n = 2x = 18) *B. vulgaris* (sugar beet) has been used in previous genomic comparisons
180 with quinoa (Jarvis *et al.*, 2017; Rey *et al.*, 2023). To produce the synteny analysis, *C. quinoa*
181 annotations and protein fasta files were downloaded from CoGe (id60716), and *B. vulgaris* files
182 from NCBI (GCF_002917755.1). Annotation files were converted to BED format via
183 agat_convert_sp_gff2bed.pl (Dainat, Hereñú and Pucholt, 2020), before being manually
184 rearranged to comply with requirements for MCScanX_h (Wang *et al.*, 2012). BlastP was used to
185 compare the proteome of *C. ficifolium* with those of *B. vulgaris* and *C. quinoa* (Camacho *et al.*,
186 2009), yielding two blast output files which were merged and passed to MCScanX_h, along with
187 BED files containing positional information for the genes in the blast files. For this analysis,
188 MCScanX used default filtering parameters. SynVisio (Bandi and Gutwin, 2020) was used to
189 visualize the resulting files, illuminating rearrangements and structural variations.

190

191 **Chloroplast Genome Assembly and Annotation**

192 The 'P' chloroplast genome was initially assembled from extracted PacBio reads using
193 GetOrganelle v1.7.6.1 (Jin *et al.*, 2020) with dependencies SPAdes (Bankevich *et al.*, 2012),
194 Bowtie2 (Langmead and Salzberg, 2012), BLAST+ (Camacho *et al.*, 2009), and Bandage (Wick
195 *et al.*, 2015), and annotated with the online tool GeSeq (Tillich *et al.*, 2017). GetOrganelle was
196 called using get_organelle_from_reads.py and was supplied with raw PacBio reads using the
197 unpaired '-u' flag. Embplant_pt was selected as the genome type, and the target genome size was
198 set to 150 kb which is the approximate size of the *C. quinoa* chloroplast genome (Hong *et al.*,
199 2017). For the purpose of determining the relative abundance of each detected haplotype of the
200 'P' chloroplast genome based on the numbers of supporting Nanopore reads, Cp-hap (Wang,
201 Lanfear and Gaut, 2019) was passed reads produced via an ONT R9.4.1 flow cell as previously

202 noted. The final output file produced from GetOrganelle was passed to GeSeq
203 (https://chlorobox.mpimp-golm.mpg.de/geseq.html) (Tillich *et al.*, 2017) for visualization and
204 annotation.

205

206 **Mitochondrial Genome Assembly and Annotation**

207 A contig from the Hifiasm assembly was identified as mitochondrial by aligning all unplaced
208 contigs to the quinoa mitochondrial reference sequence NC_041093.1 (Maughan *et al.*, 2019)
209 with BWA MEM (Li and Durbin, 2009), and manually selecting a long, well-aligned contig. This
210 contig was used as a seed sequence and passed to a *de novo* organelle assembly pipeline
211 described by Jarvis et al., (2022) to assemble the mitochondrial genome. The pipeline was used
212 in two iterations, initially using the HiFi dataset which produced a contig that was passed to the
213 mitochondrial assembly pipeline in conjunction with the long read Nanopore dataset. The
214 Nanopore dataset was then aligned to this contig using minimap2 (Li, 2018), reads matching GC
215 content and length thresholds were stripped from the alignment file via SeqKit (Shen *et al.*,
216 2016), and these reads were passed to Canu (Koren *et al.*, 2017) for *de novo* assembly. Nanopore
217 reads were later used to resolve assembly structure via comparing alignments to the draft
218 mitogenome assembly. The output contig from Canu was manually circularized by locating
219 regions at the start and end of the file that shared 100 percent sequence identity. The circularized
220 version of the assembly was polished using NextPolish (Hu *et al.*, 2020) and the PacBio read set
221 to correct single nucleotide errors. Annotation was performed by GeSeq. Importantly, GeSeq was
222 set to pass third party mitochondrial NCBI RefSeqs from both *C. quinoa* (NC_041093.1) and
223 *Beta vulgaris* subsp. *vulgaris* (NC_002511.2) to BLAT (Kent, 2002) which used annotations
224 from those genomes to locate and annotate homologs within the *C. ficifolium* mitogenome.

225 **Chloroplast and Mitochondrial Inheritance Patterns in *C. ficifolium***

226 Chloroplast and mitochondrial reference-guided assemblies for 'QC', one F1 individual, and four
227 F2 individuals, were constructed and aligned to the respective 'P' organelle genomes using BWA
228 mem (Li and Durbin, 2009) to identify heritable organelle polymorphisms differentiating the two
229 parents. The resulting SAM files were then converted to BAM, sorted, indexed, and a consensus
230 was called using the respective samtools modules (Danecek *et al.*, 2021). Once BAM files for
231 each individual were converted to fasta, freebayes-parallel (Garrison and Marth, 2012; Tange,
232 2018) was used to call variants. Freebayes-parallel was invoked with options to produce a gVCF
233 file, include genotype qualities, use best n alleles set to 4, and cap read depth to 5000. VCF files
234 were filtered to remove low quality calls using vcffilter from vcflib (Garrison *et al.*, 2022) and
235 bcftools (Danecek *et al.*, 2021). Sites or calls were pruned from the chloroplast multisample vcf
236 file if: a) the average depth across all samples was below 100; b) or the site quality was below
237 998; c) or the genotype quality phred score was below 30; d) or depth for a given sample was
238 below 50; e) or more than half of individuals were not represented; f) or the allele balance was
239 below 25 percent or greater than 75 percent; g) or there was a discrepancy between the number
240 of paired and unpaired reads aligned. The vcf file generated for the mitochondrial genome using
241 the same samples was filtered using the same criteria with the addition of a cap of 3000 on the
242 average depth per sample. The filtering criteria and script were based on those provided by
243 http://www.ddocent.com/filtering/. The resulting filtered vcf files were visualized with IGV
244 (Thorvaldsdóttir, Robinson and Mesirov, 2013).

245

246

247 **Chloroplast and Mitochondrial Genome Ancestries**

248 To illuminate the possible role of *C. ficifolium* in the cpDNA ancestries of allotetraploids *C.*
249 *quinoa* and *C. berlandieri*, a minimal phylogenetic tree was generated using six chloroplast
250 genome (cpDNA) assemblies representing four *Chenopodium* species and including beet (*Beta*
251 *vulgaris*) as an outgroup. The *C. quinoa* and *B. vulgaris* chloroplast genomes used were NCBI
252 accessions MK159176.1 and OU343016.1, respectively. All other accessions included in this
253 analysis were assembled by us *de novo* from raw reads using GetOrganelle (Jin *et al.*, 2020). We
254 assembled the *C. berlandieri* chloroplast genome from Illumina 2x150 PE reads previously
255 uploaded to NCBI by Mark Samuels from the University of Montreal (Accession:
256 PRJNA895488) (Samuels *et al.*, 2023). The *C. ficifolium* 'P' chloroplast genome was *de novo*
257 assembled using sequence data extracted from the PacBio dataset as described above, while the
258 *C. ficifolium* 'QC' and the *C. foggii* 302-A chloroplast genomes were assembled from Illumina
259 data produced at the UNH-HCGS. These assemblies were aligned using Mafft (Katoh and
260 Standley, 2013) with default (auto) settings, and a tree was produced using RAxML V. 8.2.12
261 with the GTRCAT model (Stamatakis, 2014). Trees were visualized with TreeViewer (Bianchini
262 and Sánchez-Baracaldo, 2023).

263 For a parallel purpose, a minimal mitochondrial phylogeny was produced by aligning raw reads,
264 either generated by the HCGS for *C. foggii* 302-A, and 'P' and 'QC' *C. ficifolium* accessions or
265 downloaded from NCBI, to the quinoa mitochondrial reference genome NC_041093.1 (Maughan
266 *et al.*, 2019). *B. vulgaris* (SRR6305245) (McGrath *et al.*, 2023) and *C. berlandieri*
267 (PRJNA895488) (Samuels *et al.*, 2023) read sets were downloaded from NCBI. All read data
268 used in this analysis were trimmed and cleaned with Trimmomatic (Bolger, Lohse and Usadel,
269 2014) using the primer sequence file “TruSeq3-PE-2.fa” invoked with the following settings:

ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to visualize and validate the Trimmomatic output. These reads were then aligned to the quinoa mitochondrial reference genome (Maughan *et al.*, 2019) using BWA mem (Li and Durbin, 2009). Using Samtools view, sort, index, and consensus (Danecek *et al.*, 2021), the BWA produced SAM file was converted to a fasta file containing the consensus sequence. These six fasta files, those produced from reads plus the quinoa mitogenome file used as a reference in the alignment process, were merged into a multi-fasta and Mafft (Katoh and Standley, 2013) was used with default (auto) settings to produce an MSA. A tree was produced using RAxML V. 8.2.12 (Stamatakis, 2014) with the GTRCAT model and *B. vulgaris* specified as the outgroup.

Genome Wide Association Studies

A subset of 21 'P' x 'QC' F2 individuals chosen from the 35 sequenced at the HCGS on the basis of high read depth were used in a GWAS analysis to locate regions throughout the genome implicated in the control of branch angle, internode length, flowering time, plant height, and number of lateral branches. BWA MEM was used to align reads from F2 individuals to the *C. ficifolium* 'P' V1.0 reference genome. Resulting sam files were converted to bam, sorted, and indexed via samtools (Danecek *et al.*, 2021). GATK AddOrReplaceReadGroups (Van der Auwera and O'Connor, 2020) was called to change read group IDs to reflect the sample name to make variant calling possible. Freebayes-parallel (Garrison and Marth, 2012; Tange, 2018) was used to generate a vcf file containing both InDel and SNP variants for all 35 initial individuals. The resulting vcf file was filtered to remove low quality calls following guidelines from

293 http://www.ddocent.com/filtering/. The filtered vcf data in conjunction with phenotypic data
294 collected by Subedi (2020), and name and positional data extracted from the GFF3 annotation
295 file were supplied to the package vcf2gwas (Vogt, Shirsekar and Weigel, 2022) which produced
296 Manhattan plots and positional data for putatively causative SNPs.

297 A region of interest on Cf4 identified via vcf2gwas was then closely examined with respect to
298 gene content and local synteny and collinearity with the *C. quinoa* V2 assembly (Rey *et al.*,
299 2023). For this purpose, this approximately 7 Mb region of interest was manually excised from
300 the full annotation file associated with our ChenoFicP_1.0 assembly, Revigo (Supek *et al.*, 2011)
301 was used to analyze pathways and relationships of genes with attached GO terms, and gene IDs
302 were used to subset the proteome fasta file. BlastP and MCScanX were used to compare this
303 protein subset to the complete proteome of *C. quinoa* and create a homology file. This file, plus
304 the required concatenated BED files were passed to MCScanX_h, which used default filtering
305 criteria with the exception of MATCH_SIZE, which was set to 9 to only include matches with 10
306 or more collinear genes. The MCScanX pipeline developed by BDX consulting (Wang *et al.*,
307 2024) was used to convert data to a format compatible with Circos (Krzywinski *et al.*, 2009).

308 Results

309 Nuclear Genome Assembly

310 The PacBio HiFi read set from *C. ficifolium* 'P' comprised a total of 26.5 Gb of sequence with an
311 N50 of ~14 Kb. These data were supplied to Hifiasm Version 18, yielding 781 contigs of
312 cumulative length 759,656,873 bp (~760 Mb) with an average read depth of 34.8x (Table 1).

Table 1. Hifiasm assembly statistics.

Statistics of contigs:		Read to Contig alignment:	
Number of contigs	780	Mapping rate /%	100.00
Number of contigs > 10000 bp	780	Split-read rate /%	0.47
Number of contigs >1000000 bp	9	Depth	34.87
Total length	759656810	Mapping rate in large contigs /%	92.05
Total length of contigs >1000000bp	729997343	Split-read rate in large contigs /%	0.46
Longest contig	92986964	Depth in large contigs	33.45
Second longest contig length	85613771	Error and Quality Metrics	
N50	79863811	Small-scale assembly error /per Mbp	0.30
L50	5	QV	62.60

Of these contigs, the largest nine ranged in size from a maximum of 92,986,964 bp down to 70,697,261 bp, while the tenth largest (contig ptg000009l) was 8,313,269 bp (Table 2). The remaining 771 contigs were all much smaller: less than 1 Mbp. On the basis of these size discontinuities, and the prior expectation based on the *C. ficifolium* chromosome number (n = x = 9) (Walsh *et al.*, 2015), the nine largest Hifiasm contigs were defined as comprising the initial pseudochromosome (PC) assembly. These nine PCs were then numbered and oriented to

Table 2. Original contig names, pseudochromosome designations, and contig lengths of the primary assembly plus the two largest unplaced contigs after ptg000009l were appended to Cf7.

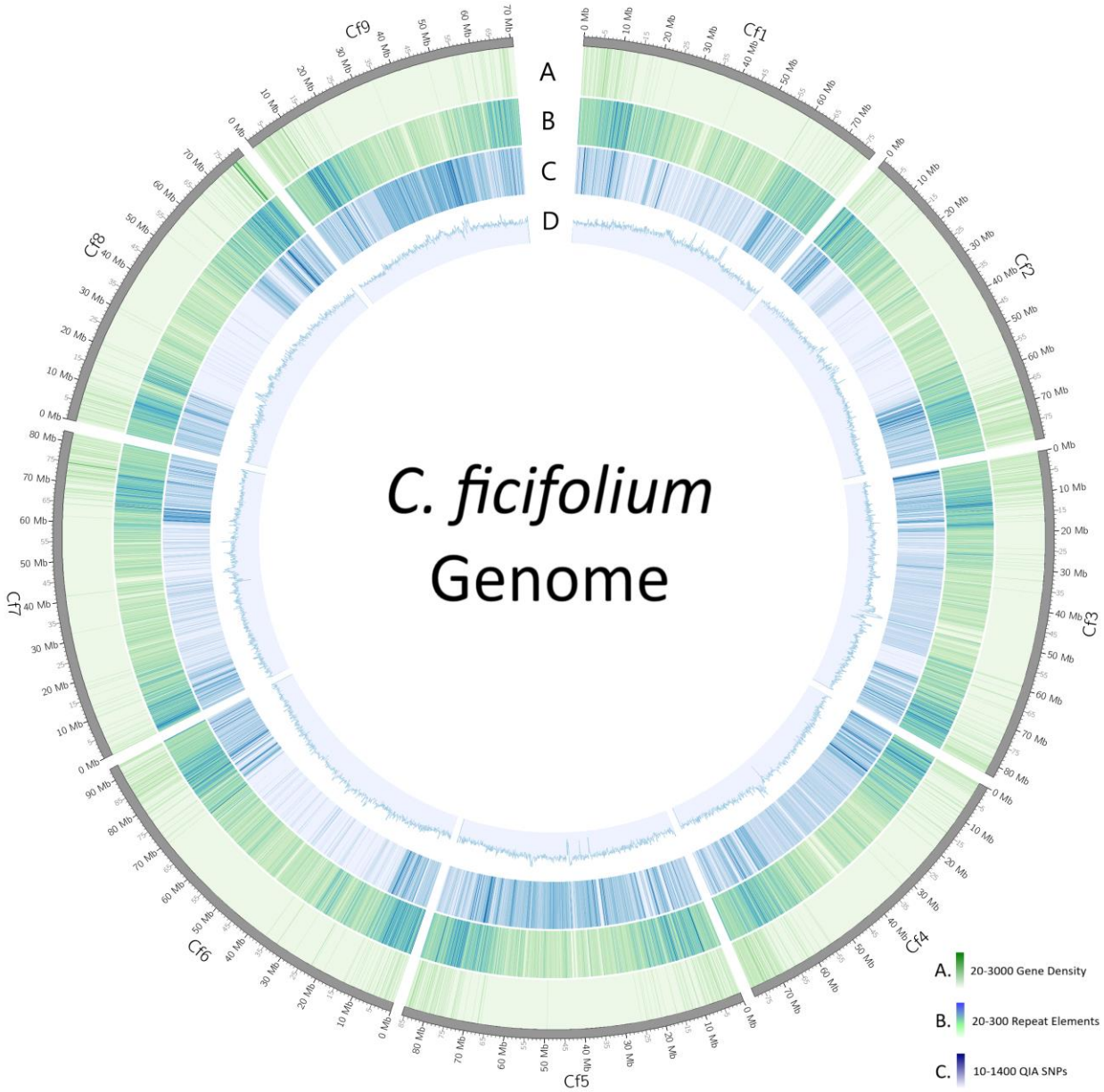
Name	Renamed	Initial Length	Final Length
ptg000002l	Cf6	92986964	92986964
ptg000006l	Cf5	85613789	85613789
ptg000005l	Cf3	82834534	82834534
ptg000008l	Cf7	74073315	82386584
ptg000003l	Cf2	79863818	79863818
ptg000001l	Cf8	79853931	79853931
ptg000011l	Cf1	77970451	77970451
ptg000004l	Cf4	77790002	77790002
ptg000007l	Cf9	70697261	70697261
ptg000009l	(Merged with Cf7)	8313269	----------
ptg000020l		871519	871519
ptg000025l		721674	721674

323 maximize correspondence with the *Beta vulgaris* EL10.1 reference genome (McGrath *et al.*,
324 2023), as was done previously in the assembly of the quinoa Version 1 assembly (Jarvis *et al.*,
325 2017). Inspector subsequently identified and corrected three structural errors, and 228 small-
326 scale assembly errors, reporting a QV score of 63.39 after correction for this assembly (Figure
327 1). The G/C content of this genome is 36.9 percent, and varies between approximately 20 and 80
328 percent per 100 Kb window (Figure 1 Ring D). G/C rich regions are often centrally located
329 within pseudochromosomes; however, centromeric regions on the whole have an average G/C
330 content, with pseudochromosome Cf1 having the highest content with 45.47 percent, and the
331 average across all centromeres being 37.75 percent. When comparing the *C. ficifolium* 'QC'
332 accession to the *C. ficifolium* 'P' assembly (Figure 1 Ring C), the number of SNPs differentiating
333 the two per 100 Kb window ranges from as low as 12 to a maximum of 2,321, or approximately
334 one SNP per 43 bases, as found on PC Cf1 from base 11,800,000 to base 11,900,000.

335 The initial Hifiasm assembly was telomere-to-telomere for PCs Cf3, Cf4, Cf5, Cf6, Cf8,
336 and Cf9, with Cf1 and Cf2 having only one telomere each, and Cf7 having a partially assembled
337 telomere which was oriented in a backwards direction. A homology search of contig ptg000009l,
338 the only unplaced contig larger than 1 Mbp, relative to the quinoa B subgenome (Rey *et al.*,
339 2023) revealed that this contig was syntenic to the distal end of quinoa chromosome Cq7B,
340 including a telomeric sequence, and thus directed its placement onto Cf7. Following this
341 adjustment, the final "ChenoFicP_1.0" assembly of nine pseudochromosomes had a total length
342 of 729,997,451 bp (~730 Mb). The CentroMiner package identified one region per chromosome
343 as being centromeric (Supplemental Table 1) and produced positional data. Centromeric and
344 telomeric positional data are provided in Supplemental Table 1. BUSCO completeness of the

345 reference genome without unplaced contigs is 96.7 percent, and with unplaced contigs is 97.2

346 percent (Figure 2B).



347

348 *Figure 1. This assembly of the C. ficifolium 'UNH_ChenoFicP_1.0' nuclear genome comprises*
349 *nine pseudochromosomes and spans 729.9 Mb.* ***A.*** *Gene density ranging from 20 to 3000 genes*
350 *represented per 100 kb window.* ***B.*** *Repetitive element density ranging from 20 to 300 elements*
351 *per 100 kb window* ***C.*** *SNPs present compared to C. ficifolium accession 'QC' per 100 kb*
352 *window.* ***D.*** *G/C content. Graph represents 20 - 55 percent G/C.*

353 The remaining 771 unplaced contigs comprised a total of about 29.6 Mb, of which 594 contigs
354 (19.6 Mb) and 6 contigs (626 Kb) were subsequently assigned on the basis of sequence
355 homology to the chloroplast and mitochondrial genomes, respectively, leaving a final total of 171
356 contigs (9.4 Mb) unplaced. Further investigation into unplaced contigs using tidk indicated that
357 contigs ptg000020 (871,519 bp) and ptg000025 (721,674 bp) both contain telomeric sequence.
358 These contigs were found by BWA to likely belong on the upstream ends of Cf2 and Cf7
359 respectively; however, they have not been included in the ChenoFicP_1.0 assembly pending
360 further confirmation based upon anticipated long read data.

361

362 **Repetitive Element Identification and Genome Annotation**

363 Data produced by RepeatModeler indicates that 69.9 percent of the ChenoFicP_1.0 genome is
364 comprised of repetitive elements, with the largest contributors being LTR/Gypsy (33.8 percent)
365 and LTR/Copia (11.2 percent) type repetitive sequences, while unknown repeats account for 17.5
366 percent of the total genome fraction. These repetitive elements tend to cluster towards telomeres
367 in (Figure 1 Ring B). Non-repetitive sequence accounts for 30.1 percent of the genome (Figure
368 2A).

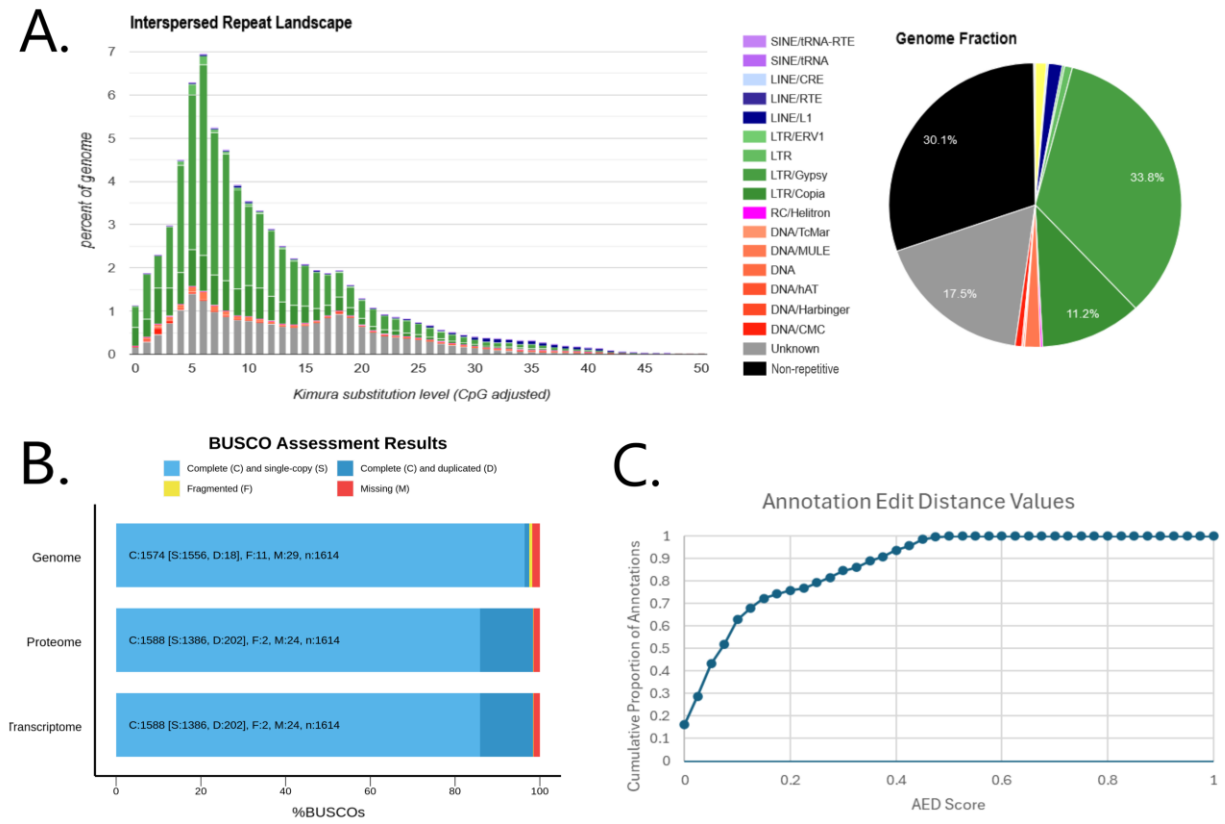


Figure 2 ***A****. Repetitive element analysis and summary graph produced by RepeatModeler. Green: LTR elements, Grey: Unknown repetitive elements, Black: Non-repetitive sequence.* ***B****. BUSCO analyses of completeness for the masked genome used in the annotation process (top), the proteome (middle), and transcriptome (bottom).* ***C****. AED (annotation edit distance) scores of gene reannotations produced by Maker. Lower AED values correspond with higher levels of agreement between the annotation and evidence, with 0 representing perfect agreement between the two.*

The final annotation file was used to calculate relative gene density (Figure 1 Ring A). The BUSCO analysis yielded a transcriptome and proteome completeness of 98.4 percent (Figure 2B). Resultant AED values for the final annotation are graphed in Figure 2C. In total, 22,617 genes were identified and annotated with the average gene length (including introns and UTRs, as defined by Maker utilizing unprocessed mRNA transcripts) being 3,958 bp.

Synteny with Quinoa B-subgenome Chromosomes

The MCScanX synteny analysis of the ChenoFicP_1.0 assembly relative to *B. vulgaris* EL10.1 (McGrath *et al.*, 2023) and the quinoa V2 assembly B-subgenome (Rey *et al.*, 2023) revealed numerous structural variations. In both comparisons, translocational differences tended to involve small distal regions near telomeres. Large inversional differences were revealed in ChenoFicP_1.0 chromosomes Cf1, Cf3, and Cf4 relative to the quinoa B-subgenome homologues, the largest of which was on Cf3 and spanned approximately 63.2 Mb (Figure 3). These inversions were positioned centrally, and in every case for chromosomes Cf1, Cf3, and Cf4 were pericentric.

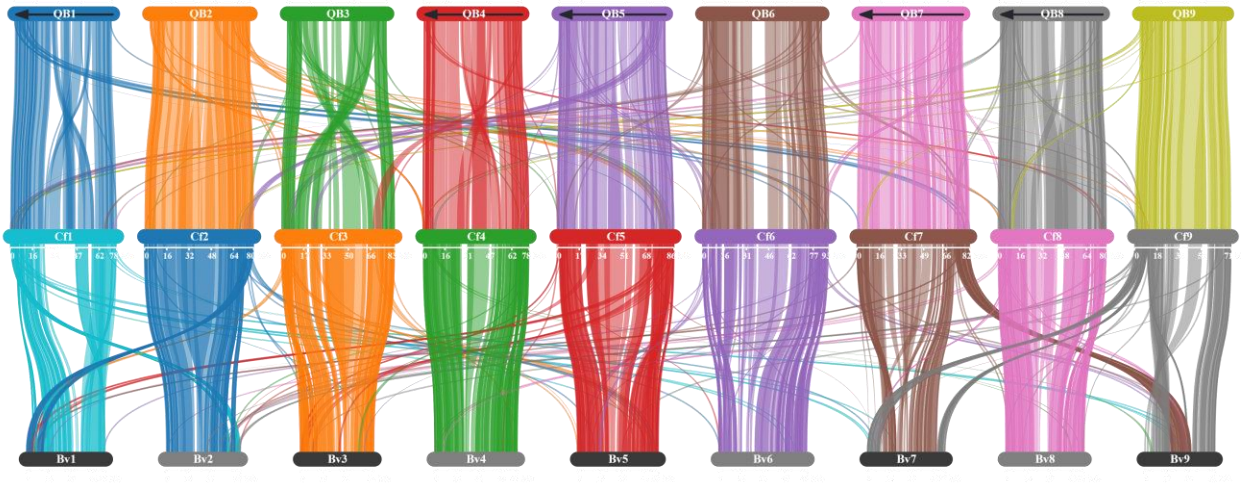


Figure 3. Synteny conservation diagram comparing the assembled pseudochromosomes of C. ficifolium (center) to the B subgenome of quinoa (top), and B. vulgaris EL10.1 (bottom). Chromosomes 1, 4, 5, 7, and 8 of the quinoa B subgenome were manually reoriented (flipped – indicated by arrows) to match those of C. ficifolium for illustrative purposes.

Chloroplast Genome Assembly

GetOrganelle produced two versions of the *C. ficifolium* chloroplast genome assembly, termed "Path 1" and "Path 2", both 151,826 bp in length, but differing in their orientations of the Short Single Copy (SSC) region relative to the Long Single Copy (LSC) region (Figure 4A). When supplied with approximately 263 Mb of Nanopore reads, Cp-hap (Wang, Lanfear and Gaut, 2019) returned approximately equal long read support for both Path 1 (248 reads) and Path 2 (242 reads) orientations, where the Path 2 orientation (Figure 4A, lower) coincided with that of the quinoa chloroplast assembly (MK159176.1) reported by Maughan et al. (2019).

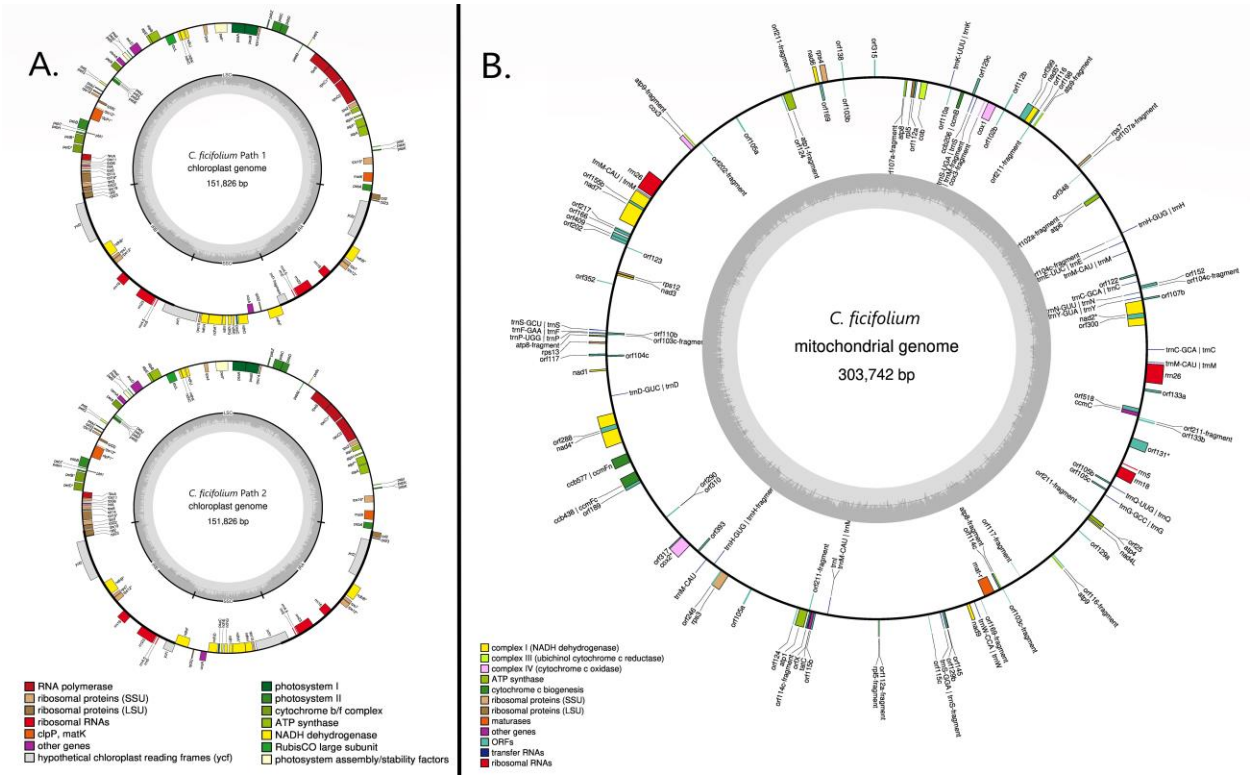


Figure 4 ***A****. Both assembly versions of the C. ficifolium chloroplast genome represented by 'Path 1' and 'Path 2'. These assemblies differ only in the direction of the Short Single Copy region. 'Path 2' (lower) was concordant with the C. quinoa chloroplast genome (MK159176). Genes inside the circle are transcribed in the clockwise direction, genes outside are transcribed counterclockwise.* ***B****. Gene map of the C. ficifolium mitogenome. The mitochondrial genome contains 118 protein coding genes, 23 tRNA genes, and is ~304 Kb in length.*

In both Path 1 and Path 2 orientations, the large single copy region is 83,900 bp, the short single copy is 18,168 bp, and each of the inverted repeat regions are 15,109 bases in length. The chloroplast genome regional boundaries are specified in Table 3. The assembly, including both inverted repeat regions, has a GC content of 37.3 percent and comprises 131 genes, of which there are 86 protein-encoding, 37 tRNA, and 8 rRNA genes. For comparison, the previously published (Kim et al., 2019) *C. ficifolium* chloroplast genome assembly (MK182725) had 129 genes, including 84 protein-encoding, 37 tRNA, and 8 rRNA.

Table 3. Start and end positions of the four regions of the chloroplast assembly. Note that these positions apply to both Path 1 and Path 2 haplotypes; however, the SSC region is inverted in Path 1 relative to Path 2.

Cp Region	Start	End	No. Genes	Protein Encoding
LSC	1	83670	85	63
IR	83671	108779	16	5
SSC	108780	126717	14	13
IR	126718	151826	16	5
		Total	131	86

Mitochondrial Genome Assembly

A 275 kb contig containing a partially assembled mitochondrial genome from the 'P' accession was initially identified from a preliminary Hifiasm assembly. In an iterative sequence seeding process, Nanopore reads were aligned to this contig, extracted from the resulting alignment, assembled *de novo*, and reads were once again aligned. After two rounds of this process, a contig was produced that was manually circularized, and polished with NextPolish using HiFi reads. The resulting 'P' mitochondrial genome assembly has a GC content of 43.8 percent and comprises 90 genes, of which 33 are protein-coding (Figure 4B). The *C. ficifolium* mitochondrial genome assembly is similar in length to that of *C. quinoa* (MK182703), at 303,742 bp versus

319,505 bp respectively, and in protein-coding genes, with 33 and 30 respectively (Maughan *et al.*, 2019), although these two assemblies display several structural differences.

Organelle Genomes' Modes of Transmission

When plastid genome mode of transmission was examined using multisample vcf files containing both parents, 'P' and 'QC', an F1, and four F2 individuals, two distinct sequence haplotypes were revealed. Of the respective polymorphisms, two were single nucleotide indels, while the remaining 17 were SNPs (Supplemental Table 2). The F1 and all F2 individuals share the same sequence haplotype as the maternal parent ('P') as distinct from the 'QC' paternal haplotype (Figure 5A), thereby demonstrating maternal transmission of the chloroplast genome in the 'P' x 'QC' cross used to generate the respective F1 and F2 progenies.

To determine the transmission pattern of the mitochondrial genome in the 'P' x 'QC' cross, 'QC' short read alignment to the 'P' mitochondrial reference genome detected nine polymorphisms, of which eight were informative, with the uninformative polymorphism being a SNP for which all individuals had a non-reference nucleotide. Seven of those informative polymorphisms were SNPs, while the eighth was a 22 bp indel (Figure 5B, Supplemental Table 2). These data demonstrate that transmission of the mitochondrial genome in the 'P' x 'QC' cross was maternal.

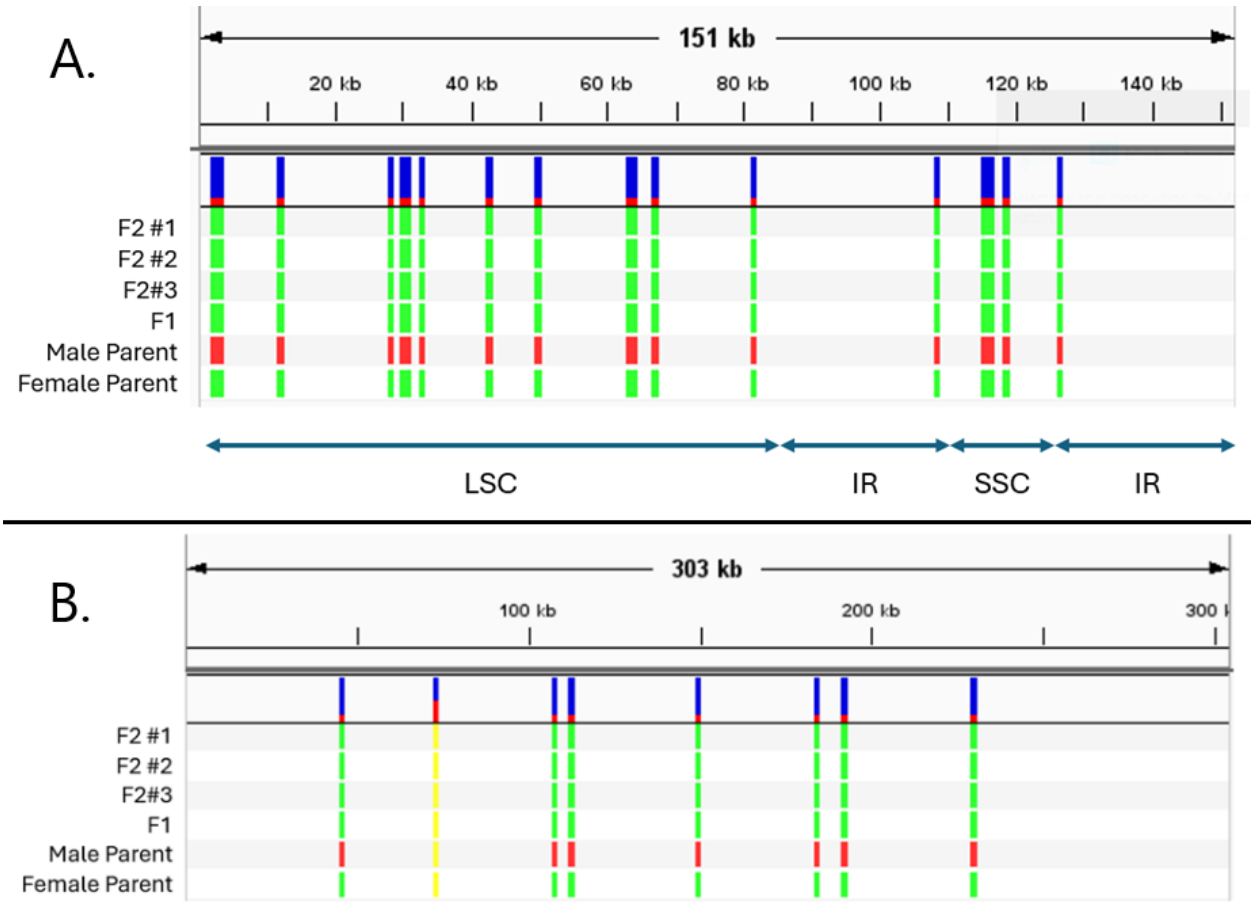


Figure 5 ***A****. Chloroplastic 'Path 2' polymorphisms found among three F2 individuals, one F1 individual, and the maternal ('QC') and paternal ('P') parents in the cross that yielded the F1 individual.* ***B****. Mitochondrial polymorphisms found among three F2 individuals, one F1 individual, and the male and female within the cross that yielded the F1 individual, 'QC' and 'P' respectively. Interestingly one polymorphism shared across all individuals is a heterozygous A/C SNP.*

Chloroplast and Mitochondrial Genome Ancestries

A phylogeny based on *de novo* assembled chloroplast genomes from *B. vulgaris, C. quinoa, C. berlandieri* var. *macrocalycium, C. foggii,* and both the 'P' and 'QC' *C. ficifolium* accessions (Figure 6A) revealed a close affinity of the two *Chenopodium* allotetraploids with the A-genome diploid *C. foggii*, and a more distant relationship with B-genome diploid *C. ficifolium*. Similarly, mitochondrial reference-based assemblies utilizing the same set of germplasm (Figure 6

B) in a narrowly focused phylogeny detected the same pattern of affinities. These outcomes support the prior hypothesis (Maughan *et al.*, 2019) that the organelle genomes of the AABB allotetraploids were both contributed by the AA ancestral diploid.

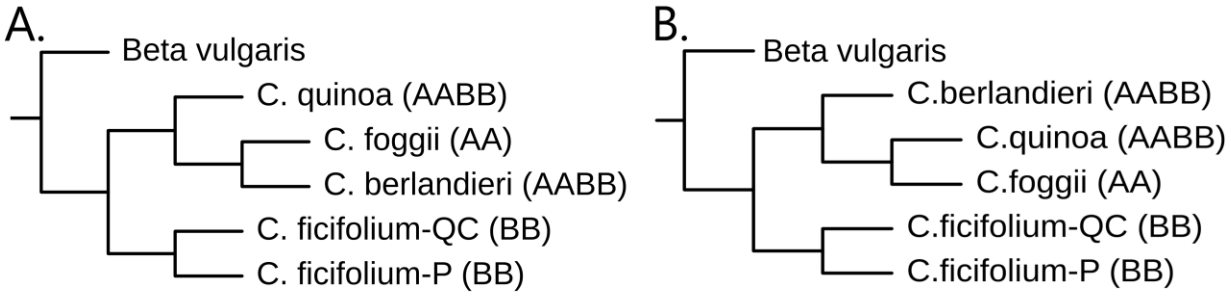


Figure 6 ***A****. Phylogenetic tree illustrating chloroplast genome ancestral relationships among Chenopodium AA and BB diploids and AABB tetraploids in our study. Phylogenetic tree illustrating chloroplast genome ancestral relationships among Chenopodium AA and BB diploids and AABB tetraploids in our study. This phylogenetic tree is based on chloroplast assemblies from C. foggii, both C. ficifolium accessions, C. quinoa, C. berlandieri, and Beta vulgaris which is set as the outgroup.* ***B****. Phylogenetic tree illustrating mitochondrial genome ancestral relationships among Chenopodium AA and BB diploids and AABB tetraploids. This phylogenetic tree was constructed using reference guided assemblies with the C. quinoa mitogenome (NC_041093.1) selected as the reference genome. Beta vulgaris is set as the outgroup.*

Genome Wide Association Study

Of the five trait-specific GWAS analyses performed, two (branch angle and internode length) failed to detect genomic associations. The remaining three traits – days to flowering, plant height, and branch number – were found by GWAS to correlate with a single common region of chromosome Cf4 (Figure 7 A-C). Of these three traits, the correlated SNPs for branch number covered the largest region and fully contained those implicated in the plant height and days to flower analyses. The implicated region spans 7,020,561 bp between positions 70,765,247 and 77,785,808 on Cf4. Features within this region were pulled from the annotation file to create a list of gene names and GO terms associated with those genes. There are 770 annotated genes

within the defined region. Of those, 132 were of unknown function. GO terms were assigned to 403 genes, including *FTL1*, which at position 74,807,401 – 74,811,571 is number 425 on the list of 770 genes, and is annotated as *Vernalization 3*. *FTL1* had one product classification identified by Interpro (Paysan-Lafosse *et al.*, 2023), being a phosphatidylethanolamine-binding protein (IPR008914).

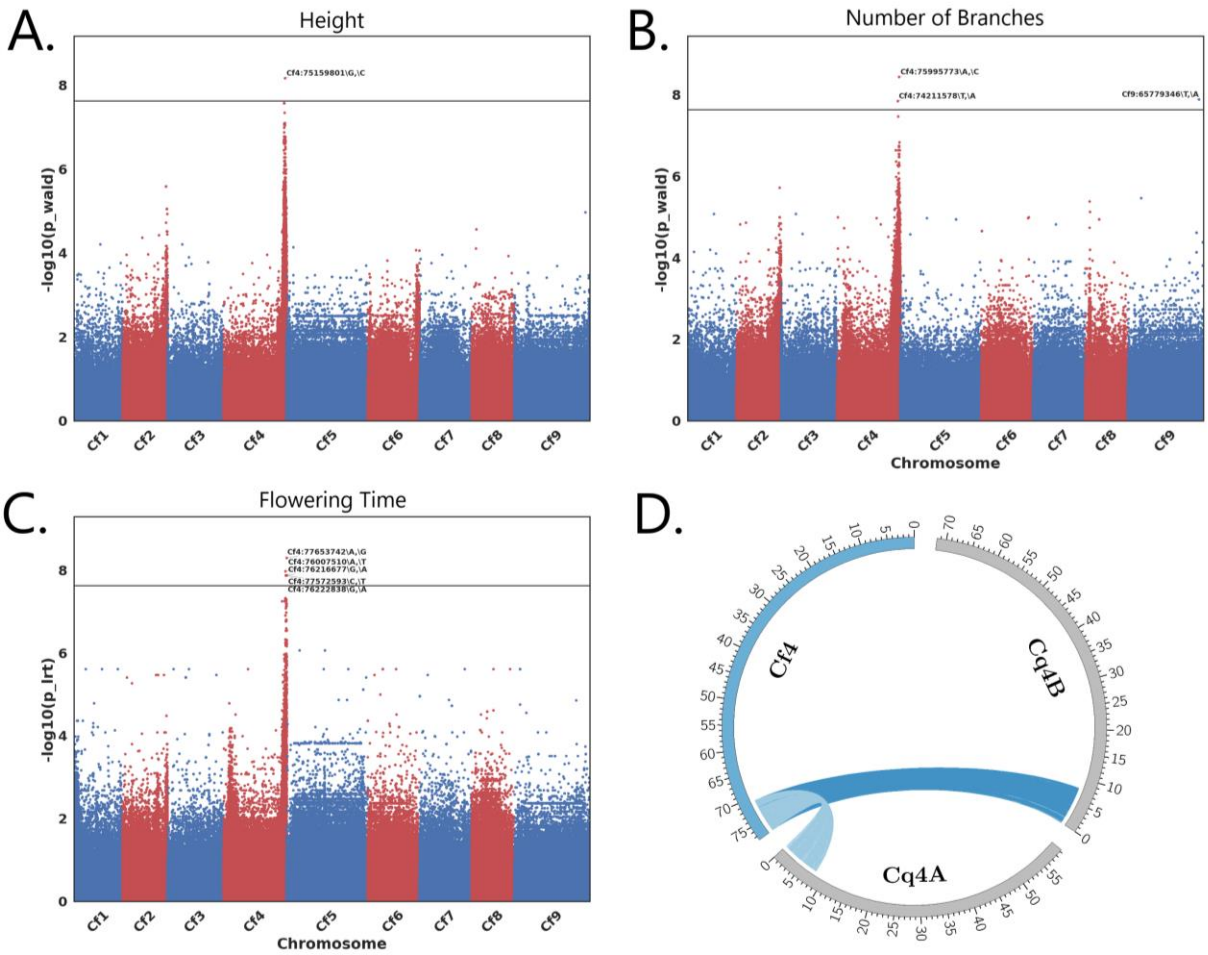


Figure 7 ***A-C****. GWAS Manhattan plots indicating regions of the genome implicated in the control of plant height, number of branches, and flowering time.* ***D****. Synteny analysis of the ~7 Mb region found in C. ficifolium Cf4 (Left) implicated by the genome wide association study as compared to C. quinoa chromosomes Cq4A and Cq4B as oriented in Rey et al., (2023). Both quinoa chromosomes Cq4A and Cq4B are in reverse orientation compared to C. ficifolium chromosome Cf4.*

503 A microsynteny analysis focusing on the ~7 Mb window implicated in this GWAS was
504 performed to identify homeologs within both *C. quinoa* subgenomes and their locations (Figure 7
505 D). MCScanX identified nine collinear blocks of 10 or more genes from this region in *C.*
506 *ficifolium* that also occur, but in reverse orientation, on quinoa chromosomes Cq4A and Cq4B of
507 the Rey et al., (2023) assembly, as will be discussed later. The analysis of these data indicates
508 that for the 770 genes found in *C. ficifolium* annotation file at this ~7 Mb region, there were 581
509 homoeologous genes found within the *C. quinoa* A subgenome and 610 found within the B
510 subgenome. In total, 107 of the *C. ficifolium* genes within this region are of unknown function
511 and of those, only six have associated GO terms.

512

513 **Discussion**

514 **Nuclear Genome**

515 As described here the ChenoFicP_1.0 assembly provides a valuable component to the *C.*
516 *ficifolium* diploid (BB) model system, as will be expanded upon in subsequent reports on
517 ongoing investigations at the diploid and allotetraploid (AABB) levels in *Chenopodium*.
518 Pseudochromosomes within this *C. ficifolium* assembly have been numbered to correspond with,
519 and oriented to be syntenic and collinear with, the previously published *Beta vulgaris* (2n = 2x =
520 18) EL10.1 genome (McGrath *et al.*, 2023), thereby following the example of Jarvis et al.,
521 (2017) in which the quinoa V1 assembly was similarly oriented. The fact that the quinoa V2
522 assembly (Rey *et al.*, 2023) did not follow this orientation pattern opens the possibility for some
523 confusion, especially with respect to chromosomes Cq1B, Cq4B, Cq5B, Cq7B and Cq8B, which
524 are presented and numbered in reverse orientation in quinoa V2 as compared with quinoa V1 and

525 beet EL10.1. We have followed the initial orientation approach in conjunction with the recently
526 published *Chenopodium* pangenome analysis (Jaggi *et al.*, 2024), which employs the original
527 orientation pattern.

528 Beet was initially selected as the genome comparator for *Chenopodium* because it was the first
529 species to have a published reference genome within the subfamily Chenopodioideae. While the
530 diploid *Spinacia oleracea* (2n = 2x = 12) now has a published genome (Cai *et al.*, 2021), its
531 nuclear genome has only six chromosomes, making comparisons with it suboptimal for
532 identifying syntenic relationships and orienting chromosomes of newly sequenced species.
533 Genome assemblies have been reported for three other *Chenopodium* diploids: the AA diploids
534 *C. pallidicaule* (Jarvis *et al.*, 2017), *C. watsonii* (Young *et al.*, 2023), and the BB diploid *C.*
535 *suecicum* versions 1 (Jarvis et al., 2017) and 2 (Rey et al., 2023). However, to our knowledge, no
536 gene-trait association studies have yet been conducted in these species.

537 The ChenoFicP_1.0 nine-pseudochromosome assembly size of ~730 Mbp incorporates 96% of
538 the 760 Mbp total assembly for the 781 contigs generated by Hifiasm. Cytometrically determined
539 C-values of *C. ficifolium* 'Portsmouth' predict a genome size in the range of 821 to 831 Mbp
540 (Neff, 2017). For comparison, the nine B-subgenome pseudochromosomes of the quinoa
541 assembly sum to 670 Mbp in length (Rey *et al.*, 2023), or approximately 8% smaller than the
542 ChenoFicP_1.0 assembly. Similarly, the nine pseudochromosomes of the quinoa A subgenome
543 sum to 531 Mbp (Rey *et al.*, 2023), which is 3% smaller than the diploid *C. watsonii* 548 Mbp A-
544 genome assembly (Young *et al.*, 2023). An earlier, broad survey of *Chenopodium* germplasm
545 reported C values of 893 Mbp for a Eurasian accession of *C. ficifolium*, 645 Mbp for *C. watsonii*,
546 and 1.545 Gbp for *C. quinoa* (Mandák *et al.*, 2016) Thus, all three of the aforementioned
547 *Chenopodium* genome assemblies are somewhat smaller than the size predicted by cytometric

548 assay. It has been speculated that this pattern of discrepancy may be due to incomplete assembly
549 coverage of some repetitive regions (Young et al., 2023), or it could be due to a consistent
550 technical artifact of flow cytometric analyses.

551 The ChenoFicP_1.0 genome contains 22,617 genes (total genes), of which 5,060 encode proteins
552 that have yet to be characterized within the UniProtKB/Swiss-Prot database. Gene density is
553 greatest in the distal regions of chromosomes (Figure 1 Ring A). For comparison, the total gene
554 counts of the *C. quinoa* A and B subgenomes are 25,913 and 26,942 respectively (Rey *et al.*,
555 2023), the AA diploid *C. watsonii* count was 30,325 (Young *et al.*, 2023), and the BB diploid *C.*
556 *suecicum* count was 29,702 (Rey *et al.*, 2023). Compared with the ChenoFicP_1.0 assembly, the
557 total number of annotated genes within the *C. quinoa* B subgenome is approximately 19.1
558 percent higher. Given that different gene count methodologies were used in these various studies,
559 the observed differences in gene counts among subgenomes and species should be interpreted
560 with caution.

561 RepeatModeler indicated that repetitive elements account for approximately 70 percent (510
562 Mbp) of the *C. ficifolium* genome. For comparison, the quinoa repetitive element content was
563 approximately 65% for the total genome (Zou *et al.*, 2017; Rey *et al.*, 2023), and approximately
564 63% (421 Mbp) for the B subgenome, with the largest contributors being long terminal repeat
565 retrotransposons representing 46 percent of the total genome size. Over half of the *C. ficifolium*
566 repetitive elements are retrotransposons that cluster towards the center of chromosomes (Figure 1
567 Ring B). These data indicate that there has been some change in the repetitive element landscape
568 in *C. ficifolium* and/or the *C. quinoa* B subgenome subsequent to the quinoa allotetraploid
569 speciation event.

570

571 **Organelle Genome Assemblies**

572 The generation of two different chloroplast genome assembly versions, designated Paths 1 and 2,
573 was an unexpected outcome. Prior chloroplast genome assemblies in *Chenopodium*, including *C.*
574 *quinoa* (Maughan *et al.*, 2019; Gao *et al.*, 2021) and *C. ficifolium* (Kim, Chung and Park, 2019),
575 and in the broader Amaranthaceae family (Chaney *et al.*, 2016; Dong *et al.*, 2016; Kim, Chung
576 and Park, 2019; Ding *et al.*, 2021) including sugar beet (McGrath *et al.*, 2023), have reported
577 only a single chloroplast genome configuration, corresponding to our Path 2 assembly, although
578 Kim et al. (2019b) mention the need to reposition the SSC and LSC regions of the quinoa
579 accession 'Real Blanca' for their alignment procedure, indicating a disagreement in SSC and/or
580 LSC orientation when compared to other assemblies. However, recent literature (Wang, Lanfear
581 and Gaut, 2019) (and citations therein) has established that chloroplast genomes do indeed exist
582 in two equally represented alternate configurations in many if not most plant species, as has now
583 become evident in *Chenopodium*, and that the detection of these alternate forms is enhanced by
584 employment of long read sequencing technologies such as PacBio HiFi and Nanopore. The very
585 long and accurate reads now available from Nanopore technology will be of particular value in
586 relation to mitochondrial genomes, the potential alternate forms of which are often challenging to
587 assemble (Kozik *et al.*, 2019). Given the known complexities of mitochondrial genome
588 organization, the assembly configuration that we present here should be viewed as preliminary
589 and subject to revision.

590

591

592

593 **Organelle Transmission Patterns**

594 In the absence of evidence to the contrary, organelle genomes are generally assumed to be
595 transmitted maternally in Angiosperms: however, exceptions have been reported (Davis *et al.*,
596 2010), including paternal chloroplast transmission in some *Actinidia* hybrids (Testolin and
597 Cipriani, 1997). Until very recently (Maughan *et al.*, 2024), empirical documentation of modes
598 of organelle inheritance in *Chenopodium* appears to have been limited to a single study involving
599 atrazine resistance in *Chenopodium album* (Warwick and Black, 1980; Corriveau and Coleman,
600 1988; Harris and Ingram, 1991; Kolano *et al.*, 2016). Our analysis agrees with this prior study in
601 its finding of maternal transmission of both chloroplast and mitochondrial markers in our diploid
602 level *C. ficifolium* cross. Similarly, Maughan et al. (2024) have documented the maternal
603 transmission of the organelle genomes in several tetraploid level *Chenopodium* crosses.

604

605 **Organelle Genome Ancestries**

606 Although *C. ficifolium* is a candidate B nuclear genome donor in the initial hybridization that
607 formed quinoa 3.3 to 6.3 million years ago (Štorchová *et al.*, 2015; Kolano *et al.*, 2016; Jarvis *et*
608 *al.*, 2017; Maughan *et al.*, 2019), prior analyses indicate that a BB diploid was not the ancestral
609 donor of the allotetraploids' organelle genomes (Maughan *et al.*, 2019). Instead, that role has
610 been assigned to the A nuclear genome donor (Maughan *et al.*, 2019). That assignment is
611 supported by our chloroplast and mitochondrial minimal phylogenies, which both place *C.*
612 *ficifolium* as sister to a clade consisting of AABB tetraploids and an AA diploid.

613 It is considered likely that, in the initial AABB hybridization event, the ancestral AA
614 diploid served as female and maternal organelle genome donor (Maughan *et al.*, 2019).

615 Documentation of such in our *C. ficifolium* cross and in tetraploid level crosses (Maughan *et al.*,
616 2024) strengthens the thesis that the original hybridization event was of the form AA x BB, i.e.,
617 with the AA diploid ancestor serving as female. Final validation of this model will await the
618 analysis of organelle genome transmission in reciprocal crosses between AA and BB diploids.

619

620 **Genome Wide Association Studies**

621 The data produced from these genome-wide analyses indicate that, in the studied segregating
622 population, variation in the agronomically important traits of days to flower, plant height, and
623 branch number were predominantly, if not entirely, under the control of one or more genes within
624 a single, 7 Mb region located towards the bottom of pseudochromosome Cf4 and containing the
625 *FTL1* locus previously employed as a marker by Subedi et al. (2021). Our genome-wide analysis
626 extends the findings of Subedi *et al*. (2021) by focusing attention on this Cf4 "hot spot" and the
627 genes residing therein. Of note, the association of branch angle with a single SNP on Cf9 was not
628 further investigated due to its presumed spurious nature. Manhattan plots will typically associate
629 a number of SNPs within and surrounding a causative region, and this questionable SNP did not
630 have any significantly associated markers adjacent to it, nor was it seen in the other two analyses
631 investigating plant height or days to flower.

632 Gene ontology data submitted to Revigo indicated the presence of a myriad of genes responsible
633 for controlling functions that are not intrinsically associated with these traits of interest; however,
634 this region contains *Flowering Locus T-Like 1* (*FTL1*), which was automatically annotated as
635 *Vernalization 3* (*VRN3*) within the GFF file. This gene is between positions 74,807,401 and
636 74,811,571, placing it centrally within the denoted significance region on Cf4. *FTL1* had one

product classification identified by Interpro (Paysan-Lafosse *et al.*, 2023), being a phosphatidylethanolamine-binding protein (IPR008914).

To further understand relationships between *C. ficifolium* and *C. quinoa,* and functionally utilize *C. ficifolium* as a model species, BlastP and MCScanX microsynteny analyses were performed to investigate relative gene locations across species and subgenomes (Figure 7D). When comparing the *C. ficifolium* region of interest to both of the quinoa subgenomes, all homologous genes in syntenic blocks of ten or more genes were found to reside in corresponding distal regions of quinoa chromosomes Cq4A and Cq4B, reflecting substantial gene order and location conservation within the respective genomic window. This analysis identified quinoa genes CQ035263 (A subgenome) and CQ019992 (B subgenome) within the *C. quinoa* V2 annotation (GFF3) file as being homologous to *C. ficifolium VRN3* (*FTL1*). The product of CQ035263 was queried with BlastP and has an 83 percent sequence identity match to XP_021772286.1, *C. quinoa FLOWERING LOCUS T-like*. Similarly, the product of gene CQ019992 is identical to XP_021734330.1. *C. quinoa FLOWERING LOCUS T-like* has a 97% protein identity with the annotated *C. ficifolium* gene Cfic_20093. The data produced using *C. ficifolium* as a model organism for *C. quinoa* via this pipeline has been demonstrated to be an efficacious way of establishing gene-trait relationships via GWAS and locating syntenic regions across species that may be worth investigating to understand agronomically relevant traits. Further investigation of genes within this region is required, perhaps in conjunction with more high-depth sequence data and additional plants to facilitate a fine-mapping study; however, these data are strongly indicative of one or more genes within this region as being implicated in the control of these traits. Further research, likely employing gene editing or transformation methods, is required to test this hypothesis and firmly establish this relationship.

660 The results of this study, in particular the synteny and collinearity analysis comparing *C.*
661 *ficifolium* and quinoa 'QQ74' chromosomes, presents intriguing hints regarding the potential
662 usefulness of figleaf goosefoot as a breeding resource for quinoa improvement. The primary
663 structural variants differentiating the B genome chromosomes in these two species are three large
664 pericentric inversions. However, the largest of these, on chromosome Cq3B, was a
665 rearrangement present only in a subset of quinoa lines derived from lowland Chilean germplasm,
666 with at least 80% of quinoa lines expected to have collinear Cq3B in comparison with figleaf
667 goosefoot (Rey *et al.* 2023). The remaining two inversions, on Cq1B and Cq4B, appear to have
668 breakpoints extending less than halfway out into their respective chromosome arms, such that
669 they do not encompass recombinogenic regions (Maughan *et al.* 2024). Consequently, while
670 triploid (AB_qB_f) figleaf goosefoot x quinoa hybrids are expected to be highly sterile, backcross
671 hybrids might theoretically be recoverable and, if so, we might expect a high degree of pairing
672 and recombination between their B genomes. This approach – crossing of the polyploid crop to
673 its wild, diploid progenitor - has been successfully employed to introduce numerous resistance
674 genes from the wild progenitor goatgrass (*Aegilops tauschii*, DD) into bread wheat (*Triticum*
675 *aestivum*, AABBDD) while restricting recombination, and consequently undesirable linkage
676 drag, to a single subgenome (Cox, Raupp and Gill, 1994).

677 **Conclusion**

678 The development of high-quality annotated reference nuclear and chloroplast genomes
679 and a preliminary mitochondrial genome for *C. ficifolium* serves as a foundational advance in the
680 development of this species as a BB diploid model system for related allotetraploids of interest,
681 with a primary focus on *C. quinoa* and *C. berlandieri.* Looking ahead, next steps in the
682 development and implementation of this model system include: broadening the germplasm base

683 via plant collection in diverse environments; establishing new hybrids and populations
684 segregating for additional traits of interest including male sterility (Ward, 1998), and disease
685 resistance, specifically downy mildew (Nolen *et al.*, 2022); and the establishment of effective
686 protocols for gene editing and *in vitro* culture and regeneration for both quinoa and its diploid
687 model(s), with immediate focus on manipulation of *FTL1*. The fact that the organelle genomes
688 present in allotetraploid quinoa were likely derived via maternal transmission from its AA
689 diploid ancestor, not from its BB ancestor, points to the need for parallel development of an AA
690 diploid model system in *Chenopodium*. Foundational steps in this direction have already been
691 taken via the generation of genome assemblies for AA diploids *C. pallidicaule* (Mangelson *et al.*,
692 2019) and *C. watsonii* (Young *et al.*, 2023), and useful germplasm collection in *C. foggii* (Neff,
693 Sullivan and Davis, 2018). As yet, to our knowledge, no genetic crosses or segregating progeny
694 populations have yet been generated in an AA diploid, so this should be a top priority going
695 forward. Finally, opportunity exists for breeding of quinoa with figleaf goosefoot as an avenue
696 for gene introgression from the diploid to the tetraploid level.

697

698

699

700

701

702

703

Acknowledgements

The research reported in this publication was partially funded by the New Hampshire Agricultural Experiment Station. This is Scientific Contribution Number 3032. This work is/was supported by awards to Thomas Davis, UNH, from the USDA National Institute of Food and Agriculture (Hatch) Project NH00678 (accession number 1019990), and by USDA NIFA project 110390. Ed Wilcox and the BYU DNASC (RRID: SCR_017781) provided PacBio sequencing services, as compensated by USDA NIFA project 110390.

Conflicts of Interest

These authors have not declared any conflicts of interest.

ORCID iDs

Clayton D. Ludwig https://orcid.org/0009-0004-8921-2479

Peter J. Maughan https://orcid.org/0000-0003-3714-3411

Eric N. Jellen https://orcid.org/0000-0002-7906-4845

Thomas M. Davis https://orcid.org/0000-0001-5455-0524

Data Availability

Genome and sequence datasets are available from NCBI using IDs SUB12919328 and SRR28959327, respectively. This genome can be found using the search term "UNH_ChenoFicP_1.0". This assembly, annotation data in GFF3 format, and supplemental annotation files can be accessed via figshare.com: https://doi.org/10.6084/m9.figshare.c.7590620

References

Van der Auwera, G.A. and O'Connor, B.D. (2020) *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Bandi, V. and Gutwin, C. (2020) 'Interactive exploration of genomic conservation', *Proceedings - Graphics Interface*, 2020-May.

Bankevich, A. *et al.* (2012) 'SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing', *Journal of Computational Biology*, 19(5), pp. 455–477. Available at: https://doi.org/10.1089/cmb.2012.0021.

Bateman, A. *et al.* (2015) 'UniProt: A hub for protein information', *Nucleic Acids Research*, 43(D1), pp. D204–D212. Available at: https://doi.org/10.1093/nar/gku989.

Bazile, D., Jacobsen, S.-E. and Verniau, A. (2016) 'The Global Expansion of Quinoa: Trends and Limits', *Frontiers in Plant Science*, 7(May), pp. 1–6. Available at: https://doi.org/10.3389/fpls.2016.00622.

Bianchini, G. and Sánchez-Baracaldo, P. (2023) 'TreeViewer Version 2.1.0'. Zenodo. Available at: https://doi.org/10.5281/zenodo.7768344.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. Available at: https://doi.org/10.1093/bioinformatics/btu170.

Brown, M., la Rosa, P.M. and Mark, B. (2023) 'A Telomere Identification Toolkit'. Zenodo. Available at: https://doi.org/10.5281/zenodo.10091385.

Bruuna, T., Gabriel, L. and Hoff, K.J. (2024) 'Navigating Eukaryotic Genome Annotation Pipelines: A Route Map to BRAKER, Galba, and TSEBRA', *arXiv preprint arXiv:2403.19416* [Preprint].

Cai, X. *et al.* (2021) 'Genomic analyses provide insights into spinach domestication and the genetic basis of agronomic traits', *Nature Communications*, 12(1), p. 7246. Available at: https://doi.org/10.1038/s41467-021-27432-z.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10, pp. 1–9. Available at: https://doi.org/10.1186/1471-2105-10-421.

Card, D.C. *et al.* (2019) 'GBE Genomic Basis of Convergent Island Phenotypes in Boa Constrictors', 11(11), pp. 3123–3143. Available at: https://doi.org/10.1093/gbe/evz226.

Cháb, D. *et al.* (2008) 'Two FLOWERING LOCUS T (FT) homologs in Chenopodium rubrum differ in expression patterns', *Planta*, 228(6), pp. 929–940. Available at: https://doi.org/10.1007/s00425-008-0792-3.

Chaney, L. *et al.* (2016) ' The complete chloroplast genome sequences for four Amaranthus species (Amaranthaceae) ', *Applications in Plant Sciences*, 4(9), pp. 4–9. Available at: https://doi.org/10.3732/apps.1600063.

Chase, M.W. *et al.* (2016) 'An update of the Angiosperm Phylogeny Group classification for the

764 orders and families of flowering plants: APG IV', *Botanical Journal of the Linnean Society*,
765 181(1), pp. 1–20. Available at: https://doi.org/10.1111/boj.12385.

766 Chen, Y. *et al.* (2021) 'Accurate long-read de novo assembly evaluation with Inspector',
767 *Genome Biology*, 22(1), pp. 1–21. Available at: https://doi.org/10.1186/s13059-021-02527-4.

768 Cheng, H. *et al.* (2021) 'Haplotype-resolved de novo assembly using phased assembly graphs
769 with hifiasm', *Nature Methods*, 18(2), pp. 170–175. Available at:
770 https://doi.org/10.1038/s41592-020-01056-5.

771 Cheng, H. *et al.* (2022) 'Haplotype-resolved assembly of diploid genomes without parental data',
772 *Nature Biotechnology*, 40(9), pp. 1332–1335. Available at: https://doi.org/10.1038/s41587-022-
773 01261-x.

774 Corriveau, J.L. and Coleman, A.W. (1988) 'Rapid screening method to detect potential
775 biparental inheritance of plastid DNA and results for over 200 angiosperm species', *American*
776 *Journal of Botany*, 75(10), pp. 1443–1458. Available at:
777 https://doi.org/https://doi.org/10.1002/j.1537-2197.1988.tb11219.x.

778 Cox, T.S., Raupp, W.J. and Gill, B.S. (1994) 'Leaf rust-resistance genes Lr41, Lr42, and Lr43
779 transferred from Triticum tauschii to common wheat', *Crop science*, 34(2), pp. 339–343.

780 Dainat, J., Hereñú, D. and Pucholt, P. (2020) 'AGAT: Another Gff Analysis Toolkit to handle
781 annotations in any GTF', *GFF format. Zenodo*, 431.

782 Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2), pp. 1–
783 4. Available at: https://doi.org/10.1093/gigascience/giab008.

784 Davis, T.M. *et al.* (2010) 'Chloroplast DNA inheritance, ancestry, and sequencing in Fragaria',
785 *Acta Hort*, 859, pp. 221–228. Available at: http://www.actahort.org/books/859/859_25.htm.

786 Ding, D. Bin *et al.* (2021) 'Characterization and phylogenetic analysis of the complete
787 chloroplast genome of Amaranthus viridis (Amaranthaceae)', *Mitochondrial DNA Part B:*
788 *Resources*, 6(9), pp. 2610–2612. Available at: https://doi.org/10.1080/23802359.2021.1961631.

789 Dong, W. *et al.* (2016) 'Comparative analysis of the complete chloroplast genome sequences in
790 psammophytic Haloxylon species (Amaranthaceae)', *PeerJ*, 2016(11), pp. 1–21. Available at:
791 https://doi.org/10.7717/peerj.2699.

792 Eddy, S.R. (2011) 'Accelerated profile HMM searches', *PLoS Computational Biology*, 7(10).
793 Available at: https://doi.org/10.1371/journal.pcbi.1002195.

794 Eid, J. *et al.* (2009) 'Real-time DNA sequencing from single polymerase molecules', *Science*,
795 323(5910), pp. 133–138. Available at: https://doi.org/10.1126/science.1162986.

796 Flynn, J.M. *et al.* (2020) 'RepeatModeler2 for automated genomic discovery of transposable
797 element families', *Proceedings of the National Academy of Sciences of the United States of*
798 *America*, 117(17), pp. 9451–9457. Available at: https://doi.org/10.1073/pnas.1921046117.

799 Gao, M.Z. *et al.* (2021) 'Complete chloroplast genome of the grain Chenopodium quinoa Willd.,
800 an important economical and dietary plant', *Mitochondrial DNA Part B: Resources*, 6(1), pp.
801 40–42. Available at: https://doi.org/10.1080/23802359.2020.1845107.

802 Garrison, E. *et al.* (2022) 'A spectrum of free software tools for processing the VCF variant call
803 format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar', *PLoS Computational Biology*, 18(5), pp. 1–
804 14. Available at: https://doi.org/10.1371/journal.pcbi.1009123.

805 Garrison, E. and Marth, G. (2012) 'Haplotype-based variant detection from short-read
806 sequencing', pp. 1–9. Available at: http://arxiv.org/abs/1207.3907.

807 Gorinstein, S. *et al.* (2002) 'Characterisation of pseudocereal and cereal proteins by protein and
808 amino acid analyses', *Journal of the Science of Food and Agriculture*, 82(8), pp. 886–891.
809 Available at: https://doi.org/10.1002/jsfa.1120.

810 Harris, S.A. and Ingram, R. (1991) 'Chloroplast DNA and biosystematics: The effects of
811 intraspecific diversity and plastid transmission', *TAXON*, 40(3), pp. 393–412. Available at:
812 https://doi.org/https://doi.org/10.2307/1223218.

813 Holt, C. and Yandell, M. (2011) 'MAKER2: An annotation pipeline and genome-database
814 management tool for second-generation genome projects', *BMC Bioinformatics*, 12(1), p. 491.
815 Available at: https://doi.org/10.1186/1471-2105-12-491.

816 Hong, S.Y. *et al.* (2017) 'Complete chloroplast genome sequences and comparative analysis of
817 Chenopodium quinoa and C. Album', *Frontiers in Plant Science*, 8(October), pp. 1–12.
818 Available at: https://doi.org/10.3389/fpls.2017.01696.

819 Hu, J. *et al.* (2020) 'NextPolish: A fast and efficient genome polishing tool for long-read
820 assembly', *Bioinformatics*, 36(7), pp. 2253–2255. Available at:
821 https://doi.org/10.1093/bioinformatics/btz891.

822 Jacobsen, S.E. (1997) 'Adaptation of quinoa (Chenopodium quinoa) to Northern European
823 agriculture: Studies on developmental pattern', *Euphytica*, 96(1), pp. 41–48. Available at:
824 https://doi.org/10.1023/A:1002992718009.

825 Jacobsen, S.E. (2003) 'The worldwide potential for quinoa (Chenopodium quinoa Willd.)', *Food*
826 *Reviews International*, 19(1–2), pp. 167–177. Available at: https://doi.org/10.1081/FRI-
827 120018883.

828 Jacobsen, S.E. (2011) 'The Situation for Quinoa and Its Production in Southern Bolivia: From
829 Economic Success to Environmental Disaster', *Journal of Agronomy and Crop Science*, 197(5),
830 pp. 390–399. Available at: https://doi.org/10.1111/j.1439-037X.2011.00475.x.

831 Jaggi, K.E. *et al.* (2024) 'A Pangenome Reveals LTR Repeat Dynamics as a Major Driver of
832 Genome Evolution in Chenopodium', *The Plant Genome* [Preprint].

833 Jarvis, D.E. *et al.* (2017) 'The genome of Chenopodium quinoa', *Nature*, 542(7641), pp. 307–
834 312. Available at: https://doi.org/10.1038/nature21370.

835 Jarvis, D.E. *et al.* (2022) 'Chromosome-Scale Genome Assembly of the Hexaploid Taiwanese
836 Goosefoot "Djulis" (Chenopodium formosanum)', *Genome Biology and Evolution*, 14(8), pp. 1–
837 7. Available at: https://doi.org/10.1093/gbe/evac120.

838 Jellen, E.N. *et al.* (2011) 'Chenopodium', in *Wild Crop Relatives: Genomic and Breeding*
839 *Resources*. Berlin, Heidelberg, Heidelberg: Springer Berlin Heidelberg, pp. 35–61. Available at:
840 https://doi.org/10.1007/978-3-642-14387-8_3.

Jellen, E.N. *et al.* (2013) 'Botany , Phylogeny and Evolution', in *State of the Art Report on Quinoa around the World*. Rome: FAO & CIRAD, pp. 12–23.

Jin, J.J. *et al.* (2020) 'GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes', *Genome Biology*, 21(1), pp. 1–31. Available at: https://doi.org/10.1186/s13059-020-02154-5.

Johnson, M. *et al.* (2008) 'NCBI BLAST: a better web interface.', *Nucleic acids research*, 36(Web Server issue), pp. 5–9. Available at: https://doi.org/10.1093/nar/gkn201.

Jones, P. *et al.* (2014) 'InterProScan 5: Genome-scale protein function classification', *Bioinformatics*, 30(9), pp. 1236–1240. Available at: https://doi.org/10.1093/bioinformatics/btu031.

Katoh, K. and Standley, D.M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. Available at: https://doi.org/10.1093/molbev/mst010.

Kent, W.J. (2002) ' BLAT —The BLAST -Like Alignment Tool ', *Genome Research*, 12(4), pp. 656–664. Available at: https://doi.org/10.1101/gr.229202.

Kim, Y., Chung, Y. and Park, J. (2019) 'The complete chloroplast genome of Chenopodium ficifolium Sm. (Amaranthaceae)', *Mitochondrial DNA Part B: Resources*, 4(1), pp. 872–873. Available at: https://doi.org/10.1080/23802359.2019.1573122.

Kim, Y., Park, J. and Chung, Y. (2019) 'The complete chloroplast genome of Suaeda japonica Makino (Amaranthaceae)', *Mitochondrial DNA Part B: Resources*, 4(1), pp. 1505–1507. Available at: https://doi.org/10.1080/23802359.2019.1601039.

Kolano, B. *et al.* (2016) 'Molecular and cytogenetic evidence for an allotetraploid origin of Chenopodium quinoa and C. berlandieri (Amaranthaceae)', *Molecular Phylogenetics and Evolution*, 100, pp. 109–123. Available at: https://doi.org/10.1016/j.ympev.2016.04.009.

Koren, S. *et al.* (2017) 'Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation', *Genome Research*, 27(5), pp. 722–736. Available at: https://doi.org/10.1101/gr.215087.116.

Kozik, A. *et al.* (2019) 'The alternative reality of plant mitochondrial DNA: One ring does not rule them all', *PLoS Genetics*, 15(8), pp. 1–30. Available at: https://doi.org/10.1371/journal.pgen.1008373.

Krzywinski, M.I. *et al.* (2009) 'Circos: An information aesthetic for comparative genomics', *Genome Research* [Preprint]. Available at: https://doi.org/10.1101/gr.092759.109.

Kumar, B. and Bhalothia, P. (2020) 'Orphan crops for future food security', *Journal of Biosciences*, 45(1). Available at: https://doi.org/10.1007/s12038-020-00107-5.

Kuznetsov, D. *et al.* (2023) 'OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity', *Nucleic Acids Research*, 51(1 D), pp. D445–D451. Available at: https://doi.org/10.1093/nar/gkac998.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature*

Methods, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Lemmon, Z.H. *et al.* (2018) 'Rapid improvement of domestication traits in an orphan crop by genome editing', *Nature Plants*, 4(10), pp. 766–770. Available at: https://doi.org/10.1038/s41477-018-0259-x.

Levy Karin, E., Mirdita, M. and Söding, J. (2020) 'MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics', *Microbiome*, 8(1), pp. 1–15. Available at: https://doi.org/10.1186/s40168-020-00808-x.

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

Li, H. (2023) 'Genome analysis Protein-to-genome alignment with miniprot', 39(January), pp. 1–6.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. Available at: https://doi.org/10.1093/bioinformatics/btp324.

Lin, Y. *et al.* (2023) 'QuarTeT: A telomere-To-Telomere toolkit for gap-free genome assembly and centromeric repeat identification', *Horticulture Research*, 10(8), pp. 1–7. Available at: https://doi.org/10.1093/hr/uhad127.

Mandák, B. *et al.* (2012) 'Is hybridization involved in the evolution of the Chenopodium album aggregate? An analysis based on chromosome counts and genome size estimation', *Flora: Morphology, Distribution, Functional Ecology of Plants*, 207(7), pp. 530–540. Available at: https://doi.org/10.1016/j.flora.2012.03.010.

Mandák, B. *et al.* (2016) 'How genome size variation is linked with evolution within Chenopodium sensu lato', *Perspectives in Plant Ecology, Evolution and Systematics*, 23, pp. 18–32. Available at: https://doi.org/10.1016/j.ppees.2016.09.004.

Mangelson, H. *et al.* (2019) 'The genome of Chenopodium pallidicaule: An emerging Andean super grain', *Applications in Plant Sciences*, 7(11), pp. 1–12. Available at: https://doi.org/10.1002/aps3.11300.

Manni, M. *et al.* (2021) 'BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes', *Molecular Biology and Evolution*, 38(10), pp. 4647–4654. Available at: https://doi.org/10.1093/molbev/msab199.

Maughan, P.J. *et al.* (2019) 'Mitochondrial and chloroplast genomes provide insights into the evolutionary origins of quinoa (Chenopodium quinoa Willd.)', *Scientific Reports*, 9(1), pp. 1–11. Available at: https://doi.org/10.1038/s41598-018-36693-6.

Maughan, P.J. *et al.* (2024) 'North American pitseed goosefoot (Chenopodium berlandieri) is a genetic resource to improve Andean quinoa (C. quinoa)', *Scientific Reports*, 14(1), pp. 1–13. Available at: https://doi.org/10.1038/s41598-024-63106-8.

McGrath, J.M. *et al.* (2023) 'A contiguous de novo genome assembly of sugar beet EL10 (Beta vulgaris L.)', *DNA research : an international journal for rapid publication of reports on genes*

and genomes, 30(1), pp. 1–14. Available at: https://doi.org/10.1093/dnares/dsac033.

Mirarab, S., Nguyen, N. and Warnow, T. (2012) 'SEPP: SATé-enabled phylogenetic placement', in *Biocomputing 2012*. World Scientific, pp. 247–258.

Neff, E. (2017) *Developing a Molecular Pipeline To Identify Chenopodium Species in New England. Master's thesis, University of New Hampshire*. University of New Hampshire.

Neff, E., Sullivan, J.R. and Davis, T.M. (2018) 'Enhanced Documentation of Chenopodium foggii (Amaranthaceae) in Northern New England', *Rhodora*, 120(983), pp. 257–259. Available at: https://doi.org/10.3119/18-02.

Nolen, H. *et al.* (2022) 'Evaluation of Disease Severity and Molecular Relationships Between Peronospora variabilis Isolates on Chenopodium Species in New Hampshire', *Plant Disease*, 106(2), pp. 564–571. Available at: https://doi.org/10.1094/PDIS-06-21-1150-RE.

Paysan-Lafosse, T. *et al.* (2023) 'InterPro in 2022', *Nucleic Acids Research*, 51(D1), pp. D418–D427. Available at: https://doi.org/10.1093/nar/gkac993.

Rey, E. *et al.* (2023) 'A chromosome-scale assembly of the quinoa genome provides insights into the structure and dynamics of its subgenomes', *Communications Biology*, 6(1), pp. 1–14. Available at: https://doi.org/10.1038/s42003-023-05613-4.

Richard, G. (2019) 'gtrichard/deepStats: deepStats 0.3. 1 (Version 0.3. 1)', *Zenodo. doi*, 10.

Rojas, W. *et al.* (2009) 'From neglect to limelight: issues, methods and approaches in enhancing sustainable conservation and use of Andean grains in Peru and Bolivia', *Journal of Agriculture and Rural Development in the Tropics and Subtropics*, 92, p. 87.

Samuels, M.E. *et al.* (2023) 'Genomic Sequence of Canadian Chenopodium berlandieri: A North American Wild Relative of Quinoa', *Plants*, 12(3), pp. 1–23. Available at: https://doi.org/10.3390/plants12030467.

Shen, W. *et al.* (2016) 'SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation', *PLoS ONE*, 11(10), pp. 1–10. Available at: https://doi.org/10.1371/journal.pone.0163962.

Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. Available at: https://doi.org/10.1093/bioinformatics/btu033.

Stanke, M. *et al.* (2008) 'Using native and syntenically mapped cDNA alignments to improve de novo gene finding', *Bioinformatics*, 24(5), pp. 637–644. Available at: https://doi.org/10.1093/bioinformatics/btn013.

Štorchová, H. *et al.* (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(6), pp. 913–925. Available at: https://doi.org/10.1007/s10722-014-0200-8.

Štorchová, H. *et al.* (2019) 'Chenopodium ficifolium flowers under long days without upregulation of FLOWERING LOCUS T (FT) homologs', *Planta* [Preprint]. Available at:

https://doi.org/10.1007/s00425-019-03285-1.

Storer, J. *et al.* (2021) 'The Dfam community resource of transposable element families, sequence models, and genome annotations', *Mobile DNA*, 12(1), p. 2. Available at: https://doi.org/10.1186/s13100-020-00230-y.

Subedi, M. (2020) *Developing Chenopodium ficifolium as a diploid model system relevant to genetic characterization and improvement of allotetraploid C. quinoa Master's Theses and Capstones. 1407.* Available at: https://scholars.unh.edu/thesis/1407.

Subedi, M., Neff, E. and 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(1), pp. 1–18. Available at: https://doi.org/10.1186/s12870-021-03270-5.

Supek, F. *et al.* (2011) 'Revigo summarizes and visualizes long lists of gene ontology terms', *PLoS ONE*, 6(7). Available at: https://doi.org/10.1371/journal.pone.0021800.

Tange, O. (2018) *GNU Parallel 2018*, *Isbn 9781387509881*. Available at: http://ole.tange.dk.

Tarailo-graovac, M. and Chen, N. (2009) 'Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences', (March), pp. 1–14. Available at: https://doi.org/10.1002/0471250953.bi0410s25.

Testolin, R. and Cipriani, G. (1997) 'Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in the genus Actinidia', *Theoretical and Applied Genetics*, 94(6–7), pp. 897–903. Available at: https://doi.org/10.1007/s001220050493.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013) 'Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration', *Briefings in Bioinformatics*, 14(2), pp. 178–192. Available at: https://doi.org/10.1093/bib/bbs017.

Tillich, M. *et al.* (2017) 'GeSeq - Versatile and accurate annotation of organelle genomes', *Nucleic Acids Research*, 45(W1), pp. W6–W11. Available at: https://doi.org/10.1093/nar/gkx391.

Torres, A.M., Weeden, N.F. and Martín, A. (1993) 'Linkage among isozyme, RFLP and RAPD markers in Vicia faba', *Theoretical and Applied Genetics*, 85(8), pp. 937–945. Available at: https://doi.org/10.1007/BF00215032.

Vaillancourt, B. and Buell, C.R. (2019) 'High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing', *bioRxiv*, p. 783159. Available at: https://doi.org/10.1101/783159.

Vega-Gálvez, A. *et al.* (2010) 'Nutrition facts and functional potential of quinoa (Chenopodium quinoa willd.), an ancient Andean grain: A review', *Journal of the Science of Food and Agriculture*, 90(15), pp. 2541–2547. Available at: https://doi.org/10.1002/jsfa.4158.

Vogt, F., Shirsekar, G. and Weigel, D. (2022) 'vcf2gwas: Python API for comprehensive GWAS analysis using GEMMA', *Bioinformatics*, 38(3), pp. 839–840. Available at: https://doi.org/10.1093/bioinformatics/btab710.

994 Walsh, B.M. *et al.* (2015) ‘ Chenopodium polyploidy inferences from Salt Overly Sensitive 1 (
995 SOS1) data ’, *American Journal of Botany*, 102(4), pp. 533–543. Available at:
996 https://doi.org/10.3732/ajb.1400344.

997 Wang, W., Lanfear, R. and Gaut, B. (2019) ‘Long-Reads Reveal That the Chloroplast Genome
998 Exists in Two Distinct Versions in Most Plants’, *Genome Biology and Evolution*, 11(12), pp.
999 3372–3381. Available at: https://doi.org/10.1093/gbe/evz256.

1000 Wang, Y. *et al.* (2012) ‘MCScanX: A toolkit for detection and evolutionary analysis of gene
1001 synteny and collinearity’, *Nucleic Acids Research*, 40(7), pp. 1–14. Available at:
1002 https://doi.org/10.1093/nar/gkr1293.

1003 Wang, Y. *et al.* (2024) ‘Detection of colinear blocks and synteny and evolutionary analyses
1004 based on utilization of MCScanX’, *Nature Protocols*, 19(7), pp. 2206–2229. Available at:
1005 https://doi.org/10.1038/s41596-024-00968-2.

1006 Ward, S.M. (1998) ‘A new source of restorable cytoplasmic male sterility in quinoa’, *Euphytica*,
1007 101(2), pp. 157–163. Available at: https://doi.org/10.1023/A:1018396808387.

1008 Warwick, S.I. and Black, L. (1980) ‘Uniparental Inheritance of Atrazine Resistance in
1009 Chenopodium album’, *Canadian Journal of Plant Science*, 60(2), pp. 751–753. Available at:
1010 https://doi.org/10.4141/cjps80-108.

1011 Wick, R.R. *et al.* (2015) ‘Bandage: Interactive visualization of de novo genome assemblies’,
1012 *Bioinformatics*, 31(20), pp. 3350–3352. Available at:
1013 https://doi.org/10.1093/bioinformatics/btv383.

1014 Young, L.A. *et al.* (2023) ‘A chromosome-scale reference of Chenopodium watsonii helps
1015 elucidate relationships within the North American A-genome Chenopodium species and with
1016 quinoa’, *Plant Genome*, 16(3): e20(April), pp. 1–17. Available at:
1017 https://doi.org/10.1002/tpg2.20349.

1018 Zou, C. *et al.* (2017) ‘A high-quality genome assembly of quinoa provides insights into the
1019 molecular basis of salt bladder-based salinity tolerance and the exceptional nutritional value’,
1020 *Cell Research*, 27(11), pp. 1327–1340. Available at: https://doi.org/10.1038/cr.2017.124.

1021

1022

1023

1024

1025

1026

1027

1028

1029

Supplemental Tables

Supplemental Table 1. Centromere positions and lengths as identified by quarTeT CentroMiner and telomere lengths and directions as identified by quarTeT TeloExplorer.

ID	Centromere			Telomere				
	Start	End	Length	Present	Upstream Length	Dir.	Downstream Length	Dir.
Cf1	58004969	58247898	242930	left	3668	+	0	
Cf2	79538569	79657117	118549	right	0		3941	-
Cf3	49844598	51718346	1873749	both	9779	+	3549	-
Cf4	39049495	40750452	1700958	both	20412	+	5019	-
Cf5	41797451	42801417	1003967	both	17843	+	3731	-
Cf6	6996291	7223985	227695	both	5397	+	4382	-
Cf7	38832653	41378639	2545987	both	882	-	5054	-
Cf8	34002546	34566367	563822	both	4480	+	5243	-
Cf9	46653552	46799075	145524	both	3710	+	4753	-

Supplemental Table 2. Positions and alleles of SNPs found in mitochondrial and chloroplast genomes used to establish maternal cytoplasmic inheritance.

Genome	Position	Reference	Alternate
Mitochondria	45844	AAAAGAATTCATTAAGAATTCACA	AA
	73300	A	C
	107870	A	C
	112616	A	C
	112987	A	C
	149592	C	T
	184591	T	A
	192086	G	T
	229731	C	A
Chloroplast	2238	T	G
Path 2	3171	C	A
	12101	C	A
	28316	CTTTTTTTTCTTTGA	CTTTTTTTCTTTGA
	30112	G	C
	30993	C	G
	33109	T	G
	42745	G	T
	49677	T	G
	50161	C	A
	63570	CAAAAAAGAAA	CAAAAAAAAAA
	64005	C	T
	67316	T	C
	81756	ATTTTTTTTTTGTT	ATTTTTTTTTGTT
	108762	G	T
	115629	A	C
	116544	A	T
	118756	C	A
	126735	C	A