

The first reference genome assembly of the Chilean sea fig (*Carpobrotus chilensis*)

Haein Lee^{1,2,3}, Carla M. D'Antonio^{1,3*}, Soojin V. Yi^{1,2,4*}

¹ Department of Ecology Evolution, and Marine Biology, University of California, Santa Barbara, CA 93106, USA

² Neuroscience Research Institute, University of California, Santa Barbara, CA 93106, USA

³ Environmental Studies Program, University of California, Santa Barbara, CA 93106, USA

⁴ Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, CA 93106, USA

*Corresponding authors:

Carla M. D'Antonio (carla_dantonio@ucsb.edu)

Soojin V. Yi (soojinyi@ucsb.edu)

Abstract

Carpobrotus chilensis (Chilean sea fig) is a coastal succulent of uncertain origin that has naturalized along the California coast, where it co-occurs and hybridizes with the invasive species, *Carpobrotus edulis*. Despite their ecological importance and widely supported hybridization, genomic resources for this genus remain scarce. Here, we present a draft genome assembly of *C. chilensis* generated from PacBio HiFi long reads. The assembled nuclear genome spans 981.7 Mb across 178 contigs. The contig N50 was 73.0 Mb, and BUSCO completeness was 96.3%. K-mer and SNP-based analyses indicate extremely low heterozygosity (3.4×10^{-4}), reduced genetic diversity in this population. The genome is highly repetitive, with 81.67% of the sequences composed of transposable elements, predominantly long terminal repeat (LTR) retrotransposons. Gene prediction identified 21,744 protein-coding genes, with BUSCO completeness of 95.8%. Comparative analysis with *C. edulis* identified 8,783 single-copy orthologous gene pairs, with a median synonymous substitution rate (dS) of 0.019, indicating low sequence divergence between the two species. This genome assembly provides a foundational resource for investigating the genomic basis of hybridization and invasion in *Carpobrotus*.

Keywords: *Carpobrotus chilensis*; Aizoaceae; genome assembly; PacBio HiFi

Introduction

Carpobrotus species are mat-forming succulents in the family Aizoaceae, most of which are native to South Africa (Hartmann, 2002). While they are traded and planted because they provide

societal benefits such as slope stabilization and ornamental use (Campoy et al., 2018), they have also become invasive in many regions of introduction because they spread rapidly beyond planted sites, overgrow native species, alter soil chemistry and change food web relationships (e.g. Molinari et al., 2007; Campoy et al., 2018). Hybridization has been reported among *Carpobrotus* species, both between introduced (Suehs et al., 2004; Campoy et al., 2019; Novoa et al., 2023) and potentially native congeners (Albert et al., 1997; Gallagher et al., 1997), often resulting in extensive hybrid swarms.

Along the California coast, hybridization between a species reported as *C. edulis* (Jepson eFlora, 2026) and *C. chilensis* (Figure 1A–B), a species whose native status remains debated (Bicknell & Mackey, 1988; Novoa et al., 2023), appears to be widespread, resulting in hybrid swarms throughout coastal habitats (Albert et al., 1997; Gallagher et al., 1997). Ongoing hybridization is considered likely to contribute to genetic introgression and the potential loss of genetically distinct *C. chilensis* populations (Albert et al., 1997; Gallagher et al., 1997; Schierenbeck et al., 2005).

Hybridization between *C. edulis* and *C. chilensis* has been investigated using morphological analyses (Albert et al., 1997), allozyme markers (Gallagher et al., 1997), chloroplast DNA (Schierenbeck et al., 2005), and microsatellite-based approaches (Novoa et al., 2023), all of which have provided evidence for introgression in natural populations, but are limited in resolution for characterizing genome-wide patterns of divergence and admixture. As a result, key questions regarding the genomic structure of hybrid populations, the extent and direction of introgression, and the distribution of ancestry across the genome remain unresolved. Addressing these questions requires genome-wide data that enable direct comparison between parental species and their hybrids. Genome resources for the genus *Carpobrotus* remain limited. Recently, a draft reference genome for *C. edulis* has become available (NCBI Assembly GCA_965788445.1). However, the lack of a corresponding genome for *C. chilensis* has restricted genome-wide investigations of divergence and introgression, as well as accurate inference of ancestry in hybrid populations. This asymmetry has constrained comparative analyses between the two species at the whole-genome level.

Here, we present a draft genome assembly of *Carpobrotus chilensis* generated using PacBio HiFi sequencing from an isolated population on San Miguel Island, Channel Islands National Park in Santa Barbara County, California, one of the few locations where *C. chilensis* is abundant and where *C. edulis* has not been reported (Hasenstab-Lehman & Guilleims, 2021), allowing sampling of individuals with minimal expected hybrid influence. This assembly provides a foundational genomic resource for the genus and enables direct comparison between the two parental species, facilitating future investigations of genomic divergence, hybrid ancestry, and introgression dynamics in California coastal populations.

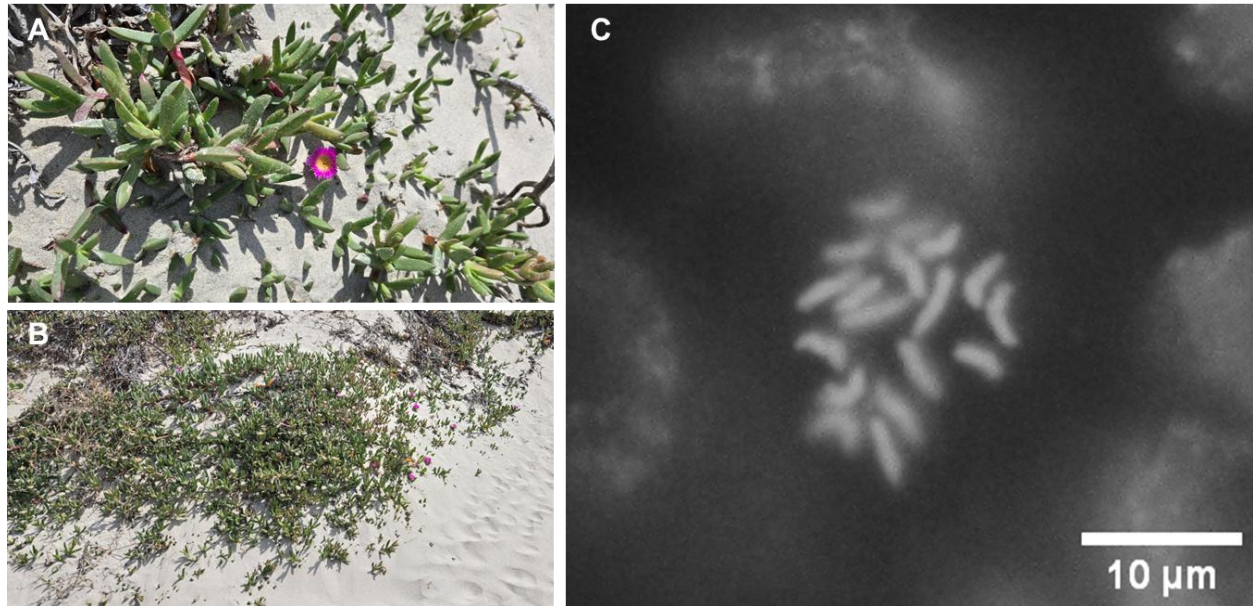


Figure 1. *Carpobrotus chilensis* and its growth habit at San Miguel Island. (A) Representative morphology of *C. chilensis*. (B) Growth habit of *C. chilensis* in its natural coastal habitat on San Miguel Island, Channel Islands National Park, Santa Barbara County, California, USA. (C) Chromosome count of *Carpobrotus chilensis*. Representative chromosome squash showing diploid chromosome number ($2n = 18$).

Materials and methods

Biological material

Biological material used for genome assembly was collected from a single individual of *C. chilensis* on San Miguel Island (34.032950°N, 120.363932°W) in Channel Islands National Park, Santa Barbara County, California, USA. San Miguel Island was selected because *C. edulis* and hybrid populations are absent, reducing the likelihood of sampling admixed individuals. The collected individual was propagated in a greenhouse, and fresh young leaves were harvested and snap-frozen in liquid nitrogen prior to DNA extraction.

Chromosome preparation and counting

Actively growing root tips were collected in the morning and pretreated in ice cold water for 24 hours. Samples were fixed in 3:1 (v/v) mixture of ethanol and acetic acid for 24 hours and stored in 70% ethanol at 4°C until use. Root tips were macerated in 1N HCl at 60°C for 10 minutes, stained with 2% acetocarmine for 1 hour, and the apical meristem was excised. Cells were squashed in 1% acetocarmine under a coverslip and used for chromosome observation and

counting. Chromosomes were visualized using fluorescence microscopy with a TRITC filter set, taking advantage of the intrinsic fluorescence of carmine-stained chromatin (Stockert et al., 1990).

High-molecular-weight DNA extraction for PacBio HiFi sequencing

High-molecular-weight (HMW) genomic DNA was extracted from approximately 1 g of young leaves (approximately 1 cm in length) that were stored at -80°C and used without further sectioning. DNA extraction followed the sorbitol pre-wash and high-salt CTAB protocol described by Inglis et al. (2018), including their recommendations for HMW DNA handling. Several modifications described below were introduced to optimize HMW DNA extraction from succulent *Carpobrotus* tissue. Extraction of HMW DNA from succulent species such as *Carpobrotus* (Aizoaceae) can be particularly challenging due to high levels of polysaccharides and polyphenols that interfere with DNA extraction and downstream applications (Diadema et al., 2003). To reduce the effects of these compounds, sorbitol wash steps were repeated one to three times until the supernatant became sufficiently clear. Frozen tissue was ground in liquid nitrogen using a pre-chilled mortar and pestle until a fine powder was obtained, with intermittent re-freezing during grinding to minimize thawing. All extraction steps, except heat incubation, were performed on ice or at 4°C . To minimize mechanical shearing, samples were handled gently throughout the protocol, avoiding vortexing and using wide-bore pipette tips for resuspension and transfer. Proteinase K was added during lysis at a final concentration of 0.2 mg/mL, followed by RNase A treatment at a final concentration of 2 mg/mL prior to chloroform:isoamyl alcohol (24:1, v/v) extraction. Chloroform:isoamyl alcohol extraction was performed twice to further improve DNA purity. When visible DNA precipitates were observed, they were carefully transferred using wide-bore pipette tips while retaining a small volume of solution to minimize fragmentation. Following these steps, genomic DNA was resuspended in EB buffer (10 mM Tris-Cl, pH 8.5) and incubated at 4°C overnight to facilitate dissolution of HMW DNA. When DNA purity did not meet the quality requirements for PacBio library preparation, an additional salt-chloroform purification protocol was used to remove polysaccharides and other contaminants (Flores, 2020). DNA quality and purity were assessed using a spectrophotometer (NanoDrop, Thermo Fisher Scientific, Waltham, MA, USA) by measuring absorbance at 260, 280, and 230 nm, and DNA concentration was quantified using a fluorometer (Qubit, Thermo Fisher Scientific, Waltham, MA, USA). DNA integrity and fragment size distribution were evaluated using 0.8% agarose gel electrophoresis and a Femto Pulse system (Agilent Technologies, Santa Clara, CA, USA), respectively. To minimize DNA fragmentation, HMW DNA was maintained under refrigerated conditions prior to library preparation.

PacBio HiFi library preparation and sequencing

PacBio HiFi library preparation and sequencing were performed at the UC Davis Genome Center. SMRTbell libraries were constructed from purified genomic DNA using the SMRTbell Prep Kit 3.0 following the manufacturer's protocol. Gel-based size selection was performed using a LightBench instrument to enrich for long DNA fragments prior to sequencing. Sequencing was

carried out on a PacBio Revio platform. Circular consensus sequencing (CCS) reads were generated using the PacBio SMRT Link pipeline to produce highly accurate HiFi reads.

De novo genome assembly

HiFi reads were assembled de novo using Hifiasm (v0.23.0-r691) with default parameters to generate a primary assembly and alternative haplotigs (Cheng et al., 2021). Redundant haplotypic sequences were further removed using purge_dups (v1.2.6) (Guan et al., 2020). Read-depth statistics were calculated using the purge_dups utilities (pbcstat and calcuts) based on PacBio HiFi read alignments to the assembly, and coverage thresholds for haplotig purging were determined from the read-depth distribution. The resulting purged assembly was used for downstream analyses.

Contamination screening and organelle separation

To obtain a nuclear genome assembly, putative plastid- and mitochondrial-derived contigs were identified and removed using sequence similarity and read-depth criteria. PacBio HiFi reads were mapped to the purge_dups-filtered primary assembly using minimap2 (Li, 2018), and per-contig sequencing depth was calculated using samtools (Danecek et al., 2021). Reads were also mapped separately to chloroplast and mitochondrial references from the *C. edulis* reference assembly (NCBI Assembly GCA_965788445.1) to estimate organelle-derived read depth. Sequence similarity between assembly contigs and organelle reference genomes was assessed using BLASTN from BLAST+ (Camacho et al., 2009). For each contig, alignment coverage was calculated as the proportion of aligned length relative to the contig length.

Contigs were classified as putative organelle-derived if they satisfied all of the following criteria: (i) at least 50% alignment coverage to plastid or mitochondrial reference genomes, (ii) sequencing depth at least threefold higher than the nuclear median, and (iii) plastid- or mitochondrial-mapped read depth ratios of ≥ 0.10 or ≥ 0.05 , respectively. To further reduce residual plastid-like contigs, the filtered assembly was additionally screened against the assembled chloroplast genome using BLASTN, and contigs showing $\geq 99\%$ sequence identity and $\geq 99\%$ query coverage were removed from the final nuclear assembly.

Following organelle filtering, the assembly was further screened using Tiara (v1.0.3) (Karlicki et al., 2022) to identify residual organelle-derived and other non-nuclear contigs, which were removed from the assembly. The resulting assembly was also screened for foreign contamination using the NCBI Foreign Contamination Screen (FCS-GX, v0.5.5) (Astashyn et al., 2024). Contigs were compared against the NCBI reference database to identify potential non-target sequences based on sequence similarity and taxonomic assignment. No contigs were flagged for removal by FCS-GX.

Genome characterization

Genome characteristics were estimated from PacBio HiFi reads using k-mer frequency spectra generated with Jellyfish (v2.2.10) at $k = 21$ (Marçais & Kingsford, 2011), and the resulting k-mer histogram was analyzed using GenomeScope 2.0 with default parameters (Ranallo-Benavidez et al., 2020). Genome-wide heterozygosity was estimated from PacBio HiFi reads mapped to the final nuclear assembly using minimap2. Single nucleotide polymorphisms (SNPs) were identified using bcftools mpileup and call (v1.22) (Danecek et al., 2021), retaining only biallelic sites. Variants were filtered by mapping quality (≥ 20), base quality (≥ 20), variant quality (QUAL ≥ 20), and sequencing depth (5 to $2.5\times$ the mean coverage). Heterozygosity was calculated as the number of heterozygous SNPs divided by the total number of callable bases defined by the same depth thresholds.

Assembly evaluation

Assembly contiguity statistics, including total assembly size, contig N50, and GC content, were calculated from the contig-level assembly using QUAST (v5.3.0) (Gurevich et al., 2013). Gene space completeness was evaluated using BUSCO (v5.8.3) with the eudicotyledons_odb12 lineage dataset in genome mode (Manni et al., 2021; Tegenfeldt et al., 2025). For k-mer-based assembly evaluation, k-mer databases were generated from PacBio HiFi reads using meryl ($k = 21$) and used as input for Merqury, from which consensus accuracy (QV) and k-mer completeness were estimated (Rhie et al., 2020). Assembly contiguity and composition were visualized using the snailplot implemented in BlobToolKit (v4.5.1) (Challis et al., 2020).

Chloroplast genome assembly and annotation

The chloroplast genome of *C. chilensis* was assembled using Oatk (v1.0) (Zhou et al., 2025). The chloroplast genome was annotated using GeSeq with the chloroplast reference set for land plants (Tillich et al., 2017).

Repeat annotation

Repetitive elements in the *C. chilensis* genome were identified using a de novo and homology-based approach implemented in the Extensive de novo TE Annotator (EDTA, v2.2.2) (Ou et al., 2019). EDTA was run in full annotation mode (--step all) with species parameter set to "others", allowing detection of a broad range of transposable element (TE) families without lineage-specific priors. To improve TE classification and reduce false positives in coding regions, a non-redundant coding sequence (CDS) set derived from BRAKER3 gene predictions was supplied using the --cds option. TE annotation was enabled using --anno 1, which integrates structural and homology-based evidence to produce a curated TE library and genome-wide TE annotation. To produce a soft-masked genome for downstream analyses, RepeatMasker v4.1.9 was run using the EDTA-derived TE library with the -xsmall option. RepeatMasker was executed with RMBlast (rmbblastn

v2.14.1+) and default parameters, and repeat annotation statistics were summarized from the resulting .out and .tbl files.

RNA extraction, library preparation, and sequencing

RNA was isolated from approximately 300 mg of mature leaves that had been snap-frozen in liquid nitrogen and stored at -80°C using the NucleoSpin RNA Plant and Fungi kit (Macherey-Nagel, Düren, Germany), following the manufacturer's instructions. RNA concentration and purity were assessed using a spectrophotometer (NanoDrop, Thermo Fisher Scientific, Waltham, MA, USA) by measuring absorbance at 230, 260, and 280 nm. RNA integrity was initially evaluated by electrophoresis on a 1.5% agarose gel and further assessed using a TapeStation system (Agilent Technologies, Santa Clara, CA, USA). RNA-seq libraries were prepared using a strand-specific protocol with poly(A) selection (NEBNext Ultra II Directional RNA Library Prep Kit, New England Biolabs, Ipswich, MA, USA) and sequenced on an Illumina NovaSeq X Plus platform (2 × 150 bp paired-end reads) by Admera Health, yielding approximately 34 million read pairs per sample.

RNA read processing and alignment

Raw RNA reads were processed using fastp (v1.0.1) to remove adapter sequences and low-quality bases (Chen et al., 2018). Adapter sequences were automatically detected for paired-end reads, and polyG tails were trimmed. Quality trimming was performed using a sliding window approach (window size = 4, mean quality threshold = 20), and reads shorter than 35 bp were discarded. Trimmed reads were aligned to the soft-masked genome assembly using HISAT2 (v2.2.2) with the --dta option (Kim et al., 2019). The resulting alignments were converted to sorted BAM files and indexed using SAMtools (v1.22.1) (Barnett et al., 2011; Li et al., 2009).

Gene prediction and annotation

For both *C. chilensis* and *C. edulis*, gene prediction was performed using the same annotation pipeline to ensure consistency in downstream comparative analyses. The genome assembly of *C. edulis* was obtained from publicly available resources, and repeat annotation and gene prediction were conducted following the same procedures applied to *C. chilensis*.

Protein-coding genes were predicted using BRAKER (v3.0.8), which integrates GeneMark-ETP and AUGUSTUS in an evidence-guided annotation framework (Brûna et al., 2024; Gabriel et al., 2024; Stanke et al., 2006, 2008). The soft-masked genome generated from EDTA and RepeatMasker was used as input. The coordinate-sorted BAM file derived from RNA-seq alignments was provided as transcriptomic evidence to BRAKER. Protein evidence was generated by combining Viridiplantae OrthoDB v12 protein sequences (Tegenfeldt et al., 2025) with additional protein sequences from *Carpobrotus edulis* (NCBI Assembly GCA_965788445.1). The combined dataset was clustered with CD-HIT at 95% sequence identity to generate a non-

redundant protein database (L. Fu et al., 2012), which was used as homology evidence for GeneMark-ETP-guided annotation (Buchfink et al., 2015). BRAKER used GeneMark-ETP to integrate transcriptomic and protein evidence for training, followed by gene prediction with AUGUSTUS. TSEBRA was applied internally to select high-confidence gene models by reconciling predictions from GeneMark-ETP and AUGUSTUS based on extrinsic evidence support (Gabriel et al., 2021).

From the BRAKER annotation, the longest isoform per gene was retained using AGAT (v1.5.1) to generate a non-redundant representative gene set (Dainat et al., 2025). Annotation completeness was evaluated using BUSCO (v5.8.3) in protein mode with the eudicotyledons_odb12 lineage dataset based on this longest isoform set (Manni et al., 2021; Tegenfeldt et al., 2025).

Functional annotation of predicted protein-coding genes was performed using eggNOG-mapper (v2.1.13) (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019). Protein sequences from the longest isoform set were used as input, and annotations were assigned based on orthology using the eggNOG database. Sequence searches were conducted in DIAMOND mode, and Gene Ontology (GO) terms and KEGG ortholog annotations were retrieved for each gene where available (Buchfink et al., 2021).

Estimation of substitution rates and identification of putatively selected genes

Orthologs between *C. chilensis* and *C. edulis* were inferred using OrthoFinder (Emms & Kelly, 2019), using amino acid sequence FASTA from the longest predicted isoform of each protein-coding gene. Single-copy orthogroups shared between the two species were extracted, and one-to-one ortholog pairs were retained for downstream analyses. Amino acid sequences for each ortholog pair were aligned using MAFFT (Kato & Standley, 2013), and the corresponding coding sequences were projected onto the amino acid alignments using PAL2NAL to generate codon alignments (Suyama et al., 2006). Codon alignments were converted to the AXT format required by KaKs_Calculator (Wang et al., 2010), and synonymous (dS) and nonsynonymous (dN) substitution rates were estimated under the YN model (Yang & Nielsen, 2000). Ortholog pairs were filtered to remove unreliable estimates, including pairs with non-finite values, dS values outside the range of 0.001 to 2, dN/dS > 10, amino acid alignments shorter than 50 residues, codon alignments shorter than 150 nucleotides, or internal stop codons. Ortholog pairs with dN/dS > 1 and dS ≥ 0.01 were designated as candidates for putative positive selection. The minimum dS threshold was applied to reduce the influence of inflated dN/dS estimates associated with very low synonymous divergence. Gene Ontology (GO) annotations were obtained from eggNOG-mapper annotations, and GO enrichment analyses were performed using the R package topGO (Alexa & Rahnenführer, 2026). Enrichment was tested separately for biological process, molecular function, and cellular component ontologies using Fisher's exact test with the weight01 algorithm (Alexa et al., 2006), with all filtered ortholog pairs used as the background set. Terms with $p < 0.05$ were considered significant.

Results

Genome size and characteristics

PacBio HiFi sequencing generated 96.3 Gb of reads from 6.38 million reads, with a mean read length of 15.1 kb and a read N50 of 13.8 kb (Supplementary Figure S1), corresponding to approximately 102× genome coverage for *Carpobrotus chilensis*. A total of 93.7% of bases had quality scores $\geq$ Q30. K-mer spectra showed a single dominant peak at approximately 101× coverage, indicating extremely low heterozygosity (Supplementary Figure S2). SNP-based analysis further estimated a genome-wide heterozygosity of 3.4×10^{-4} . K-mer-based analysis suggested a genome size of approximately 940 Mb (Supplementary Figure S2). Chromosome counting confirmed a diploid complement ($2n = 18$), supporting that the genome of *C. chilensis* is diploid (Figure 1C).

Genome assembly and quality

The nuclear genome assembly of *C. chilensis* was 981.7 Mb in length and consisted of 178 contigs, with a contig N50 of 73.0 Mb and an L50 of six contigs (Figure 2 and Table 1). The assembly span was broadly similar to the k-mer-based genome size estimate of approximately 940 Mb (Supplementary Figure S2). The largest contig reached 94.2 Mb. The GC content of the assembly was 41.8%. BUSCO analysis using the eudicotyledons_odb12 dataset recovered 96.3% complete genes (93.8% single-copy and 2.5% duplicated), with 1.8% fragmented and 1.9% missing BUSCOs. Consensus quality value (QV) estimated using Merqury was 65.9, and k-mer completeness was 98.4% (Supplementary Figure S3).

The chloroplast genome of *C. chilensis* was assembled as a single circular sequence of 148,091 bp, which is similar to the sizes of the chloroplast genomes of *C. edulis* (148,807 bp) and other Aizoaceae plants (NCBI Assembly GCA_965788445.1; W. Fu et al., 2024; Xu, 2019). A total of 105 unique genes were annotated, including 78 protein-coding genes, 23 tRNA genes, and 4 rRNA genes (Supplementary Figure S4).

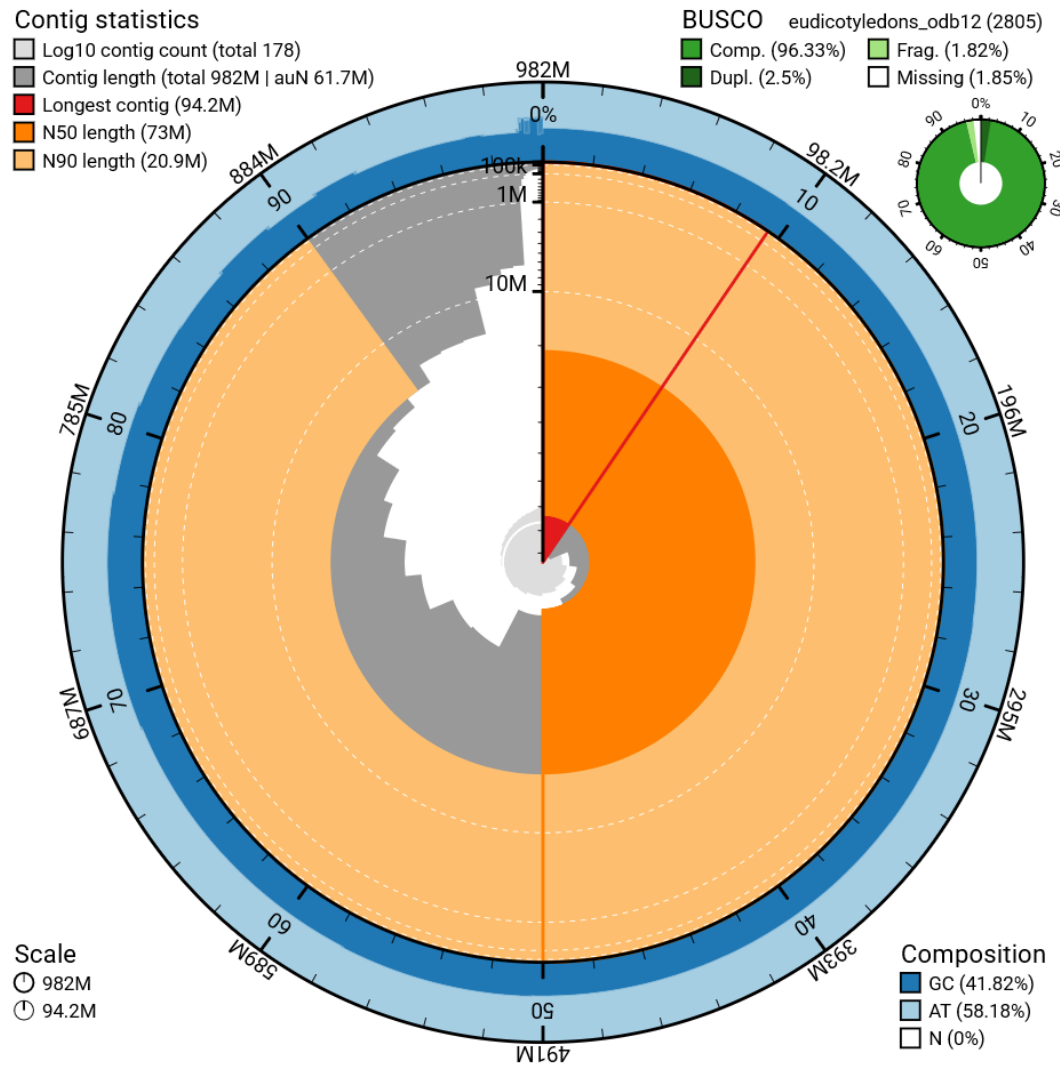


Figure 2. Assembly statistics and quality assessment of the *Carpobrotus chilensis* genome. The snail plot shows contig length distribution, assembly continuity metrics, nucleotide composition, and BUSCO completeness. The assembly spans 981.7 Mb across 178 contigs, with a largest contig of 94.2 Mb and a contig N50 of 73.0 Mb.

Genome Assembly Quality Metrics	
HiFi Read coverage	101×
Number of contigs	178
Contig N50 (bp)	73,006,540
Contig L50	6
Largest contig (bp)	94,225,371
Total length (bp)	981,733,167
Mercury QV (Phred)	65.9
Mercury completeness (%)	98.4

Table 1. Summary statistics of the *Carpobrotus chilensis* genome assembly.

Repeat annotation and gene prediction

Repeat annotation suggested a highly repetitive *C. chilensis* genome, with 81.67% of the genome assembly masked as repetitive DNA (Table 2). Interspersed repeats accounted for 80.76% of the genome, with retroelements comprising 70.48%, predominantly long terminal repeat (LTR) elements (69.28%). Among these, *Gypsy* and *Copia* families represented 28.52% and 4.18% of the genome, respectively. DNA transposons accounted for 9.82%, and long interspersed nuclear elements (LINEs) contributed 1.20%. Other repeat categories collectively represented 1.37% of the genome, including unclassified repeats (0.45%), simple repeats (0.81%), and low-complexity regions (0.11%).

Gene prediction identified 21,744 protein-coding genes with 24,921 transcript isoforms (Table 3). BUSCO analysis of the longest isoform protein set recovered 95.8% complete genes (94.0% single-copy and 1.8% duplicated), with 0.8% fragmented and 3.5% missing BUSCOs.

Group	Number of elements	Length occupied (bp)	Percentage of sequence (%)
Retroelements	942,398	691,939,278	70.48
LINEs:	12,910	11,797,525	1.20
RTE/Bov-B	256	106,754	0.01
L1/CIN4	12,654	11,690,771	1.19
LTR elements:	929,488	680,141,753	69.28
Ty1/Copia	72,483	41,013,136	4.18

Gypsy/DIRS1	381,729	280,020,217	28.52
DNA transposons	367,195	96,448,601	9.82
Unclassified:	29,786	4,442,236	0.45
Total interspersed repeats:	-	792,830,115	80.76
Simple repeats:	136,479	7,955,335	0.81
Low complexity:	18,882	1,035,086	0.11

Table 2. Summary of repetitive elements in the *Carpobrotus chilensis* genome. Classification of transposable elements and other repeat types identified using EDTA and RepeatMasker, including the number of elements, total length occupied, and percentage of the genome.

Metric	Value
Number of genes	21,744
Number of transcripts	24,921
Number of exons	128,501
Number of introns	103,580
Mean mRNAs per gene	1.1
Mean exons per mRNA	5.2
Mean gene length (bp)	4,169
Mean CDS length (bp)	1,234
Mean exon length (bp)	239
Mean intron length (bp)	753

Table 3. Summary of gene prediction and structural annotation for the *Carpobrotus chilensis* genome. Statistics of predicted protein-coding genes, including gene, transcript, exon, and intron counts, as well as average gene and feature lengths derived from BRAKER3 annotation.

Divergence of protein-coding genes

A total of 13,254 single-copy one-to-one orthologous gene pairs were identified between *C. chilensis* and *C. edulis*, of which 8,783 were retained after quality filtering. The distribution of synonymous substitution rates (dS) showed a low median value (0.019; mean = 0.038), with the majority of values concentrated below 0.05 (Figure 3). These results indicate low sequence divergence between the two species. The distribution of dN/dS ratios showed a median of 0.290 (mean = 0.392), with most values below 1 and a small number of higher values (Supplementary Figure S5), consistent with purifying selection acting on the majority of orthologous genes.

A total of 163 ortholog pairs met the criteria for putative positive selection. Gene Ontology enrichment analysis identified terms associated with defense and signaling processes, including

defense response, signal transduction, and cell surface receptor signaling, as well as calcium ion transmembrane transport and leaf morphogenesis (Supplementary Figure S6).

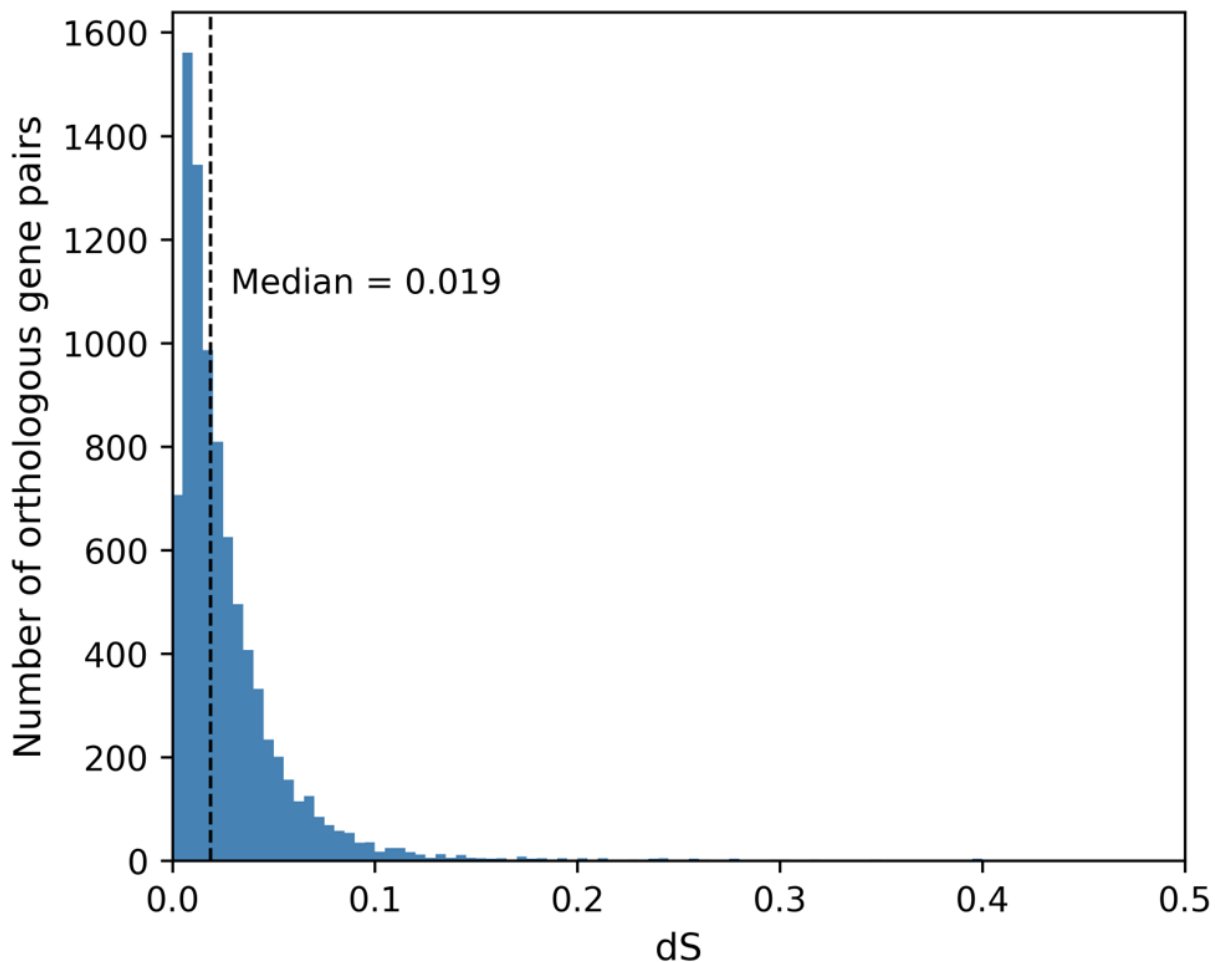


Figure 3. Distribution of synonymous substitution rates (dS) between *Carpobrotus chilensis* and *Carpobrotus edulis*. Frequency distribution of dS values estimated from 8,783 filtered single-copy orthologous gene pairs. The dashed vertical line indicates the median dS value (0.019).

Discussion

The *C. chilensis* genome assembly presented here represents the first nuclear genome assembly for this species and expands the currently limited genomic resources available for the genus *Carpobrotus*, a globally distributed and ecologically important genus (Campoy et al., 2018; Novoa et al., 2023). Together with the recently published genome of *C. edulis*, this assembly provides genomic references for both putative parental species involved in hybridization along the California coast and establishes a foundation for comparative genomic analyses within the genus.

The assembly also revealed a highly repetitive genome, with 81.67% of the genome composed of transposable elements, predominantly LTR retrotransposons.

A notable characteristic of the *C. chilensis* genome is its extremely low heterozygosity (3.4×10^{-4}). This low level of variation may reflect demographic history, self-fertility (Vilà et al., 1998), clonal propagation, or a combination of these factors. Because the inference of low heterozygosity is based on a single individual, these interpretations remain tentative. Broader sampling across the species' geographic range will be required to distinguish among demographic and reproductive explanations.

A comparison of protein-coding sequences between *C. edulis* and *C. chilensis* revealed a median dS of 0.019 across 8,783 single-copy ortholog pairs, indicating low synonymous divergence between the two species. This low divergence is consistent with a close evolutionary relationship and may contribute to the widespread hybridization observed in natural populations in California (Albert et al., 1997; Gallagher et al., 1997; Schierenbeck et al., 2005). The predominance of purifying selection across orthologous gene pairs (median dN/dS = 0.290) suggests that protein-coding sequences have remained largely conserved. Nevertheless, the identification of candidate genes under putative positive selection and their associated biological processes illustrates the utility of the *C. chilensis* genome resource for future comparative and evolutionary genomic studies within *Carpobrotus*.

Previous studies of hybridization in *Carpobrotus* have relied on morphological traits, allozyme markers, chloroplast DNA, and microsatellite-based approaches (Albert et al., 1997; Gallagher et al., 1997; Novoa et al., 2023; Schierenbeck et al., 2005), which provide limited resolution for genome-wide analyses of ancestry and introgression. Together with the existing *C. edulis* genome (NCBI Assembly GCA_965788445.1), the *C. chilensis* genome presented here provides a framework for genome-wide analyses of hybridization, such as SNP-based inference of admixture, characterization of introgressed regions, and more precise estimation of divergence between parental species. The chloroplast genome generated in this study provides an additional resource for future phylogeographic and hybridization studies in *Carpobrotus*.

This assembly has several limitations. First, the genome remains at the contig level, which restricts chromosome-scale comparative analyses with the chromosome-level assembly of *C. edulis*. Second, the genome was assembled from a single individual and thus does not capture genetic variation across the species. Third, gene annotation was performed using BRAKER with leaf RNA-seq and protein homology evidence; therefore, genes primarily expressed in other tissues or at other time points may not be fully represented. Despite these limitations, this *C. chilensis* genome provides a valuable resource for future studies of hybridization, divergence, and ecological adaptation in *Carpobrotus*.

Data Availability

The genome assembly, PacBio HiFi reads, and RNA-seq data generated in this study have been deposited in NCBI under BioProject PRJNA1468704. The genome assembly is available at DDBJ/ENA/GenBank under accession JBZCIG0000000000. Genome annotation resources, including structural and functional gene annotations, repeat annotations, and chloroplast genome annotation files, are available through Zenodo (DOI: 10.5281/zenodo.20687645).

The *Carpobrotus edulis* genome assembly used for comparative analyses is publicly available through the International Nucleotide Sequence Database Collaboration (INSDC) under accession GCA_965788445.1 (BioSample: SAMEA110450299), generated by the Darwin Tree of Life project. Public RNA-seq datasets used for gene prediction are available from the European Nucleotide Archive (ENA) under accessions ERR15315144 and ERR15315142.

Acknowledgements

We thank the Channel Islands National Park for assistance with collection as well as P. Schuyler for transportation to San Miguel Island. We also thank C. Hannah-Bick for assistance growing specimens in the UCSB Greenhouse. We thank the NRI-MCDB Microscopy Facility at UCSB for access to microscopy instrumentation and Ben Lopez for assistance with microscopy. We acknowledge the Darwin Tree of Life Project for generating and making publicly available the genome assembly of *Carpobrotus edulis* used in this study.

Funding

This research was funded by the Schuyler Endowment at UCSB (Dept. of Environmental Studies), start-up funds to S.V.Y., and the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2024-00408988).

Conflicts of interest

The authors declare no conflicts of interest.

Reference

- Albert, M. E., D'Antonio, C. M., & Schierenbeck, K. A. (1997). Hybridization and introgression in *Carpobrotus* spp. (Aizoaceae) in California. I. Morphological evidence. *American Journal of Botany*, 84(7), 896–904. <https://doi.org/10.2307/2446279>
- Alexa, A., & Rahnenführer, J. (2026). *topGO: Enrichment Analysis for Gene Ontology* [Computer software]. <https://bioconductor.org/packages/topGO>
- Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13), 1600–1607. <https://doi.org/10.1093/bioinformatics/btl140>
- Astashyn, A., Tvedte, E. S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strobe, P. K., Sylla, P. M., Wagner, L., Bidwell, S. L., Brown, L. C., Clark, K., Davis, E. W., Smith-White, B., Hlavina, W., Pruitt, K. D., Schneider, V. A., & Murphy, T. D. (2024). Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biology*, 25(1), 60. <https://doi.org/10.1186/s13059-024-03198-7>
- Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12), 1691–1692. <https://doi.org/10.1093/bioinformatics/btr174>
- Bicknell, S. H., & Mackey, E. M. (1988). *Evidence of early occurrence of *Carpobrotus aequilaterus* N. E. Br. At Marina Beach State Park* (Nos. 05-10–065). California Department of Parks and Recreation.
- Brūna, T., Lomsadze, A., & Borodovsky, M. (2024). GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Research*, 34(5), 757–768. <https://doi.org/10.1101/gr.278373.123>

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368. <https://doi.org/10.1038/s41592-021-01101-x>

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 421. <https://doi.org/10.1186/1471-2105-10-421>

Campoy, J. G., Acosta, A. T. R., Affre, L., Barreiro, R., Brundu, G., Buisson, E., González, L., Lema, M., Novoa, A., Retuerto, R., Roiloa, S. R., & Fagúndez, J. (2018). Monographs of invasive plants in Europe: Carpobrotus. *Botany Letters*, 165(3–4), 440–475. <https://doi.org/10.1080/23818107.2018.1487884>

Campoy, J. G., Roiloa, S. R., Santiso, X., & Retuerto, R. (2019). Ecophysiological differentiation between two invasive species of Carpobrotus competing under different nutrient conditions. *American Journal of Botany*, 106(11), 1454–1465. <https://doi.org/10.1002/ajb2.1382>

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12), 5825–5829. <https://doi.org/10.1093/molbev/msab293>

Carpobrotus edulis genome assembly GCA_965788445.1 (Genome Assembly GCA_965788445.1; Version 1). (n.d.). NCBI Genome. Retrieved May 11, 2026, from https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_965788445.1/

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics*, 10(4), 1361–1374. <https://doi.org/10.1534/g3.119.400908>

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2), 170–175. <https://doi.org/10.1038/s41592-020-01056-5>

Dainat, J., Cannoodt, R., Soares, A., Ruano, D. G., Hereñú, D., Murray, D. K. D., Davis, E., Ugrin, I., Crouch, K., Soler, L., Agostinho, N., pascal-git, Zollman, Z., & tayyrov. (2025). *NBISweden/AGAT: AGAT v1.5.1* [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.16317950>

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. <https://doi.org/10.1093/gigascience/giab008>

Diadema, K., Baumel, A., Lebris, M., & Affre, L. (2003). Genomic DNA isolation and amplification from callus culture in succulent plants, *Carpobrotus* species (Aizoaceae). *Plant Molecular Biology Reporter*, 21(2), 173–173. <https://doi.org/10.1007/BF02774244>

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>

Flores, M. (2020). *Guidelines for Using a Salt:Chloroform Wash to Clean Up HMW DNA for PacBio*. <https://www.protocols.io/view/guidelines-for-using-a-salt-chloroform-wash-to-cle-ewov182e2gr2/v1>

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>

Fu, W., Li, L., Li, S., Li, Y., Sun, J., Zhang, L., & Zou, Y. (2024). Characteristics and phylogenetic analysis of the complete chloroplast genome of *Mesembryanthemum cordifolium* L. F. (Aizoaceae). *Mitochondrial DNA Part B*, 9(9), 1258–1262. <https://doi.org/10.1080/23802359.2024.2398180>

Gabriel, L., Brûna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*, 34(5), 769–777. <https://doi.org/10.1101/gr.278090.123>

Gabriel, L., Hoff, K. J., Brûna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: Transcript selector for BRAKER. *BMC Bioinformatics*, 22(1), 566. <https://doi.org/10.1186/s12859-021-04482-0>

Gallagher, K. G., Schierenbeck, K. A., & D'Antonio, C. M. (1997). Hybridization and Introgression in *Carpobrotus* spp. (Aizoaceae) in California II. Allozyme Evidence. *American Journal of Botany*, 84(7), 905–911. <https://doi.org/10.2307/2446280>

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics (Oxford, England)*, 36(9), 2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>

Hartmann, H. E. K. (2002). Aizoaceae. In H. E. K. Hartmann (Ed.), *Illustrated Handbook of Succulent Plants: Aizoaceae A-E* (pp. 9–268). Springer. https://doi.org/10.1007/978-3-642-56306-5_2

Hasenstab-Lehman, K. E., & Guilliams, C. M. (2021). *CCH2 Portal Checklist: Checklist of the Vascular Flora of San Miguel Island, California*. <https://www.cch2.org/portal/checklists/checklist.php?clid=22>

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1), D309–D314. <https://doi.org/10.1093/nar/gky1085>

Inglis, P. W., Pappas, M. de C. R., Resende, L. V., & Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLOS ONE*, 13(10), e0206085. <https://doi.org/10.1371/journal.pone.0206085>

Jepson eFlora. (2026). *Jepson Flora Project* (eds.). https://ucjeps.berkeley.edu/eflora/eflora_display.php?tid=17975

Karlicki, M., Antonowicz, S., & Karnkowska, A. (2022). Tiara: Deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, 38(2), 344–350. <https://doi.org/10.1093/bioinformatics/btab672>

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), 907–915. <https://doi.org/10.1038/s41587-019-0201-4>

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (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), 4647–4654. <https://doi.org/10.1093/molbev/msab199>

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6), 764–770. <https://doi.org/10.1093/bioinformatics/btr011>

Molinari, N., D'Antonio, C., & Thomson, G. (2007). 7 *Carpobrotus* as a case study of the complexities of species impacts. *Theoretical Ecology Series*, 4, 139–II.

Novoa, A., Hirsch, H., Castillo, M. L., Canavan, S., González, L., Richardson, D. M., Pyšek, P., Rodríguez, J., Borges Silva, L., Brundu, G., D'Antonio, C. M., Gutiérrez, J. L., Mathese, M., Levin, S., Silva, L., & Le Roux, J. J. (2023). Genetic and morphological insights into the *Carpobrotus* hybrid complex around the world. *NeoBiota*, 89, 135–160. <https://doi.org/10.3897/neobiota.89.109164>

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, 20(1), 275. <https://doi.org/10.1186/s13059-019-1905-y>

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1), 1432. <https://doi.org/10.1038/s41467-020-14998-3>

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21(1), 245. <https://doi.org/10.1186/s13059-020-02134-9>

Schierenbeck, K. A., Symonds, V. V., Gallagher, K. G., & Bell, J. (2005). Genetic variation and phylogeographic analyses of two species of *Carpobrotus* and their hybrids in California. *Molecular Ecology*, 14(2), 539–547. <https://doi.org/10.1111/j.1365-294X.2005.02417.x>

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5), 637–644. <https://doi.org/10.1093/bioinformatics/btn013>

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7(1), 62. <https://doi.org/10.1186/1471-2105-7-62>

Stockert, J. C., Llorente, A. R., Castillo, P. D., & Gómez, A. (1990). Chromatin Fluorescence after Carmine Staining. *Stain Technology*, 65(6), 299–302. <https://doi.org/10.3109/10520299009105621>

Suehs, C. M., Affre, L., & Médail, F. (2004). Invasion dynamics of two alien *Carpobrotus* (Aizoaceae) taxa on a Mediterranean island: II. Reproductive strategies. *Heredity*, 92(6), 550–556. <https://doi.org/10.1038/sj.hdy.6800454>

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl_2), W609–W612. <https://doi.org/10.1093/nar/gkl315>

Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E. M., & Kriventseva, E. V. (2025). OrthoDB and BUSCO update: Annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research*, 53(D1), D516–D522. <https://doi.org/10.1093/nar/gkae987>

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, 45(W1), W6–W11. <https://doi.org/10.1093/nar/gkx391>

Vilà, M., Weber, E., & D'Antonio, C. M. (1998). Flowering and mating system in hybridizing *Carpobrotus* (Aizoaceae) in coastal California. *Canadian Journal of Botany*, 76(7), 1165–1169. <https://doi.org/10.1139/b98-130>

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., & Yu, J. (2010). KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genomics, Proteomics & Bioinformatics*, 8(1), 77–80. [https://doi.org/10.1016/S1672-0229\(10\)60008-3](https://doi.org/10.1016/S1672-0229(10)60008-3)

Xu, H. (2019). The complete chloroplast genome of newly alien medicinal and toxic species, *Zaleya pentandra* (L.) C.Jeffrey (Aizoaceae), in China. *Mitochondrial DNA Part B*, 4(2), 3412–3413. <https://doi.org/10.1080/23802359.2019.1674709>

Yang, Z., & Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Molecular Biology and Evolution*, 17(1), 32–43. <https://doi.org/10.1093/oxfordjournals.molbev.a026236>

Zhou, C., Brown, M., Blaxter, M., McCarthy, S. A., Durbin, R., & Darwin Tree of Life Project Consortium. (2025). Oatk: A de novo assembly tool for complex plant organelle genomes. *Genome Biology*, 26(1), 235. <https://doi.org/10.1186/s13059-025-03676-6>