Comparative analysis of complete chloroplast genome of the Peruvian landrace of Capsicum chinense, arnaucho chili pepper, and related species of the Capsiceae tribe

Gianmarco Castillo
Universidad Nacional José Faustino Sánchez Carrión (UNJFSC)
Kevin R. Quiroz-Hidalgo
Universidad Nacional José Faustino Sánchez Carrión (UNJFSC)
Diego H. Takei-Idiaquez
Universidad Nacional José Faustino Sánchez Carrión (UNJFSC)
Julio E. Ballen-Gavidia
Universidad Nacional José Faustino Sánchez Carrión (UNJFSC)
Yhovana A. Changanaqui-Rengifo
Universidad Nacional José Faustino Sánchez Carrión (UNJFSC)
Sergio Contreras-Liza
Universidad Nacional José Faustino Sánchez Carrión (UNJFSC)
Carlos I. Arbizu
carlos.arbizu@untrm.edu.pe
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)
Pedro M. Rodriguez-Grados
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Research Article

Keywords: In-silico, Germplasm, Chili pepper, Bioinformatics

Posted Date: December 21st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5657151/v1

Additional Declarations: No competing interests reported.

Abstract

Although many complete chloroplast (cp) genomes of different types of peppers have already been published, there has been no comprehensive study that summarizes all the characteristics of the Peruvian landrace “arnaucho” chili pepper (ACP) comparing it with other types of genomes in its Capsiceae tribe. In this study, a comprehensive analysis was conducted using data from cp genomes obtained from NCBI GenBank. These 14 genomes were annotated using Geseq, followed by genomic comparisons, chloroplast structure analysis, phylogeny, and repetitive sequence analysis, employing a variety of bioinformatics tools. The findings revealed length variations among the cp genomes, ranging from 156,583 bp in *C. lycianthoides* to 157,390 bp in *C. pubescens*, with a GC content of 37% across all genomes. The comparative genome analysis revealed that the greatest variation among the 14 genomes occurred in the non-coding regions. Arnaucho chili pepper exhibited greater divergence in coding regions with *C. lycianthoides*, specifically in the genes *accD*, *rpl20*, *rps12*, *clpP*, *ycf2*, *ndhF*, *ndhA*, *ycf1*, and *rpl2*. The results of the phylogeny and pairwise distance analysis in this study support that the arnaucho chili pepper clusters with *C. galapagoense*, with an average distance value of 0.00002733. Additionally, the repetitive sequence analysis determined that ACP maintains a number of repetitive sequences similar to other *Capsicum* species but possesses a lower number of SSRs (33). Finally, it was determined that the junction regions of ACP have a total length of 156,931 bp, similar to *C. galapagoense* with 156,959 bp. The four boundary regions exhibited consistent gene patterns, except for the JSB region, where the *ycf1* gene in ACP was located only in the IRb region, whereas it was absent in other *Capsicum* species. This research provides additional effective evidence for characterizing the entire cp genome and classifying species and genera within the Capsiceae tribe.

Introduction

The genus *Capsicum* is the most important in the Solanaceae family. This genus comprises a total of 30 species, including sweet varieties like bell peppers and spicy ones like chili peppers. Both varieties come in various colors, such as red, yellow, green, orange, and, in some cases, purple. They also display heart-shaped or bell-like forms (Antonio et al. 2018 & Dong et al. 2013). It is estimated that *Capsicum* appeared in human history around 7500 B.C. and began to be cultivated between 5200 and 3400 B.C. The use of these fruits was not only for their flavor but also for ethnobotanical purposes, such as medicinal preparations, treatment of wounds, stomach aches, fever, as an anti-infective agent, and for treating hypertension when chewed (Florencia el al. 2022). This genus includes domesticated and wild species. Fossil studies of *Capsicum* spp. starch found seven sites dating back 6,000 years, ranging from the Bahamas to southern Peru (Perry et al. 2007). Currently, the production, use, and domestication of *Capsicum* have spread worldwide. Since the 1650s, when it was introduced by the Spanish, it has continued to the present day, leading to selective breeding associated with different environmental conditions (Hernandez et al. 2022 & Aguilar-Melendez et al 2009).

These fruits are an important ingredient in the daily diet of millions of people, used in various dishes as they are a great source of vitamin C and antioxidants. Additionally, Peru is a country that harbors the greatest biological diversity of these fruits in South America, with a potential source of genes with traits of interest for germplasm conservation (Minguez-Mosquera, 1992; Meckelmann, 2013 & Ibiza et al, 2012). Peru has various varieties, such as “mochero” chili pepper, “panca” chili pepper, “dulce” chili pepper, and “arnaucho” or “supano” chili pepper (Lopez et al. 2020). The arnaucho chili pepper (ACP) is a cultivar native to the Supe Valley in the province of Barranca, about 200 km north of Lima. This fruit is sold in markets in the provinces of Barranca, Huaura, and the district of Huacho. It is identified by its purple color and triangular or globular shape (Arbizu et al. 2022).

Next-generation sequencing (NGS) studies have accelerated the generation of high-throughput data and genomic information, including the identification of DNA markers such as single nucleotide polymorphisms, among others.

Moreover, the reduction in genome sequencing costs allows for the discovery of thousands of markers in a single step. Thanks to this technology, it is gaining widespread acceptance in the field of crop breeding (Ray, 2014). With this technology, several studies have been conducted on the sequencing of *Capsicum*, such as the study by Sebastian et al. (2024), which demonstrates that the most domesticated species of *C. annuum* exhibited a higher degree of divergence, very similar to their divergent genotypes, compared to their counterparts. Another study of the chloroplast genome of *C. chinense* revealed that the species was more closely related to *C. annuum* var. *glabriusculum* (Paar, et al. 2016). Additionally, the study of *C. eximium*, native to southern Bolivia, identified 113 unique genes, including 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. Of all these genes, 21 are duplicated in the inverted region, and in the phylogenetic analysis, it was grouped within the *Capsicum* clade. (Sebastian, et al. 2019). On the other hand, studies of the draft genome of *C. chinense* using Illumina HiSeq 2500 sequencing technology identified 71.96% of the repetitive DNA in the assembled genome, with a lower number of SSRs in the genome compared to other *Capsicum* species, reporting the first draft genome of the ACP (Estrada, et al. 2024).

Bioinformatics tools today are useful for understanding the functional and structural analysis of genomes, transcriptomes, proteomes, and more. Due to the use of these tools, we have been able to gain a deeper understanding of biomedical and plant sciences. They have become essential for predictive analysis in the era of big data, having a significant impact on research (Mansoor, et al. 2024). The advent of high-throughput sequencing tools, cutting-edge biotechnological techniques, and bioinformatics analyses has led to greater ease in characterizing plant traits (Roychowdhury, et al. 2023). There is a large volume of accumulated data, both in NCBI and EMBL, and therefore, the management and extraction of this data have become of great interest. This information is manageable for conducting further studies, utilizing algorithms based on mathematical and statistical models to handle these biological data (Li, et al. 2013). Bioinformatics and in silico analysis allowed us to understand the processes occurring at the cellular level, as it facilitates the identification of genes. This enables the enhancement of precision agriculture, optimizing natural resources, improving the efficiency of the production process, and reducing costs (Angel Gaviria, 2023).

Studies conducted by Arbizu et al. (2022) presented the first complete chloroplast genome (cp) of this local cultivar of Peruvian chili pepper (arnaucho) using high-throughput sequencing technology, providing us with a phylogenetic result of this species with its closest relative, *C. galapagoense*. In this work, we continue the studies on the ACP, revealing its structure, divergence, and evolution, using bioinformatics tools to gain a deeper understanding of the chloroplast genome of this species, which is endemic to Peru and belongs to the province of Barranca.

Materials and methods

Sequence annotation and comparison of chloroplast genomes

A total of twelve complete chloroplast genome sequences were retrieved from the Capsiceae plant tribe via the Organelle Genome Resources of NCBI (accessed on September 29, 2024), with the following GenBank accession numbers: MZ379791.1 (*C. chinense*), NC_018552.1 (*C. annuum*), KX913219.1 (*C. tovarii*), KR078312.1 (*C. frutescens*), NC_033524.1 (*C. galapagoense*), KX013217.1 (*C. chinense*), KU041709.1 (*C. chinense*), MH559321.1 (*C. chinense*), NC_030543.1 (*C. chinense*), KX913220.1 (*C. eximium*), MH559320.1 (*C. baccatum*), NC_039694.1 (*C. pubescens*), NC_026551.1 (*C. lycianthoides*), and NC_033525.1 (*C. chapecoense*). The Geseq program (https://chlorobox.mpimp-golm.mpg.de/geseq.html Frazer, 2004) was used to annotate and locate the genes of all cp genomes from the Capsiceae tribe, with default options selected (CDS, tRNA, rRNA, IR, and the interspaced *rps12* gene) (Tillich et al., 2017). Additionally, tRNA-scan v2.0 (Chan and Lowe, 2019), available on the same server, was used as a secondary tRNA annotator. The GC content calculator provided by JaimeMcGowan (2024) was also employed via its web service (available online: https://jamiemcgowan.ie/bioinf/gc_content.html) to analyze the GC content in the 12 cp genomes (Table 1). Furthermore, we used the annotated cp genome of *C. chinense* (MZ379791.1) as the reference genome to

compare with the other cp genomes of the Capsiceae plant tribe. The VISTA program (https://genome.lbl.gov/vista/index.shtml, accessed on September 29, 2024) was employed in Shuffle-LAGAN mode (Brudno et al., 2003) to analyze the sequence similarity within the Capsiceae tribe. Finally, we visualized the cp genome architecture of the ACP using program OGDRAW v1.3.1 (Greiner, 2019). A prior analysis of the genetic content was conducted to identify the functions of its chloroplast genome (Table 2)

Codon usage of protein-coding sequences of *C. chinense*

For this analysis, the 86 protein-coding genes of *C. chinense* (MZ379791) were used. The number of codons (Nc) and the relative synonymous codon usage (RSCU) were calculated using the MEGA11 program (Tamura, 2021) and the RSCU calculator available via JaimeMcGowan's web service (available online: https://jamiemcgowan.ie/bioinf/rscu.html, McGowan, 2024). Subsequently, codon usage and the relative synonymous codon usage (RSCU) values were manually analyzed by comparing the results from both programs. The Nc provides an intuitively meaningful measure of codon preference within a gene (Wright, 1990). Additionally, an RSCU value greater than 1.00 indicates a lack of bias, while a value less than 1.00 is considered indicative of bias, as described by the research of Sharp and Li (1987). The result was visualized using R v4.4.1 (R Core Team, 2024) with the ggplot2 and tidyr packages.

Phylogenomic analysis and pairwise distances

To reveal the evolutionary relationship of the Capsiceae tribe with a total of fifteen annotated complete cp genomes, a sequence alignment of the 14 species was performed. The genomes of the Capsiceae tribe include two additional species from the Solanaceae family. In this case, we used *Solanum lycopersicum* NC_007898 and *Physalis peruviana* MH019242 as outgroup. The sequences were retrieved from the GenBank NCBI database (https://www.ncbi.nlm.nih.gov/genbank/) and aligned using the MAFFT v7 web server (Katoh and Stanley, 2013; https://mafft.cbrc.jp/alignment/server/index.html) with default options. The alignment was then manually refined using Aliview v1.28 (Larsson, 2014). After the genomes were previously aligned and refined, a nucleotide substitution analysis was performed. In this case, the best-fitting model was TVM+I+G, determined using the Akaike information criterion (AIC) by the JModelTest v2.1.10 program (https://github.com/ddarriba/jmodeltest2). The phylogenetic reconstruction was performed using the Maximum Likelihood (ML) method using MEGA11 (Tamura, 2021) with 1000 Bootstrap repetitions, employing the Kimura 2 model and complete deletion. Additionally, the tree was visualized using the iTol v6 tool (https://itol.embl.de/, Letunic, 2024). Additionally, an alignment of only the Capsiceae tribe was performed to study the pairwise distances of the 14 cp genomes. We used MEGA11 with 1000 Bootstrap repetitions, a nucleotide substitution model, the p-distance method, and complete deletion of missing data. These results were visualized in a raincloud plot using R, employing the tidyr, ggplot2, and ggrain packages.

Identification of repetitive sequences and SSR analysis

In this analysis of repetitive sequences, we used various programs, including REPuter v2.7.4 (available online: https://bibiserv.cebitec.uni-bielefeld.de/reputer/, Kurtz et al., 2001), to identify repetitive sequences (including forward, reverse, and complementary repeat sequences). For the identification of these repetitive sequences, the following parameters were used: a minimum repeat size of 30 bp and a Hamming distance of 3 (with 90% or more sequence identity). Additionally, for palindromic sequences, the RepEX web server (available online: http://bioserver2.physics.iisc.ac.in/RepEx/query.html, Michael et al., 2019) was used with the following input parameters: repeat type set to "Palindrome" and a repeat length of more than 10. Additionally, for the detection of tandem repeats, the web tool (https://tandem.bu.edu/trf/home, Benson, 1999) was used with the default parameters in its advanced option. Finally, simple sequence repeats (SSR) or microsatellites in the genomes were examined using

the MISA web tool (https://webblast.ipk-gatersleben.de/misa/, Bier et al., 2017) with SSR search parameters set for repeat lengths with a minimum threshold of 10, 5, 4, 3, 3, and 3 repeat units for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively. The results were visualized in bar charts, pie charts, balloon plots, and heatmaps using R v4.4.1 with the tidyr, ggplot2, ggpubr, pheatmap, and RColorBrewer packages.

Analysis of shifts at the boundaries between the single-copy regions (LSC and SSC) and the inverted repeats (IRA and IRB)

A comparison of the positions of the boundaries between the single-copy regions (LSC and SSC) and the inverted regions (IRA and IRB) was made across the 14 species of the Capsiceae tribe. These boundaries were named JLB (boundary between LSC-IRB), JSB (boundary between IRB-SSC), JSA (boundary between SSC-IRA), and JLA (boundary between IRA-LSC). For the positions of the boundaries, the distance between these and the nearest genes to these regions or located on their boundary was calculated using the online program (https://irscope.shinyapps.io/irapp/, Amiryousefi, 2018).

Results

Sequence annotation and comparison of chloroplast genomes

Through a search in the NCBI database, complete cp genome sequences of 14 species from the *Capsicum* genus were obtained for subsequent comparative analysis and prior genome annotation. The Geseq program was used to obtain the structural features and genetic content of the cp genome of the ACP, as shown in Fig. 1. The results obtained with the other cp genomes of the *Capsicum* genus showed a four-part structure, consisting of an LSC region, an SSC region, and two IRs. In the IR regions, each of which ranges from 25,624 bp (*C. lycianthoides*) to 25,910 bp (*C. baccatum*). The IRs were separated by an LSC region, which ranged from 87,688 bp (*C. pubescens*) to 86,813 bp (*C. lycianthoides*), and an SSC region, which ranged from 17,849 bp (*C. annuum*) to 18,522 bp (C. lycianthoides). The GC content in the 12 cp genomes of the Capsicum tribe was similar, with an average of 37%. Additionally, the size of the cp genomes in the 14 genomes of the Capsiceae tribe ranged from 156,583 bp in *C. lycianthoides* to 157,390 bp in *C. pubescens* (Table 1). The ACP had a genome size of 156,931 bp, and the average genome size for the Capsiceae tribe was a total of 156,741 bp. The genome size of the Capsicum tribe, specifically of Arnaucho chili pepper, is similar to the first genome studied in this tribe, as well as the cp genomes used in this study, approximately 156 kb, except for *C. baccatum* and *C. pubescens*, which both had 157 kb (Raveendar et al., 2024).

Table 1
Comparación de datos de secuencias de los genomas completos de cloroplastos de la tribu Capsiceae

Genbank Number	Scientific name	Genome size bp	LSC región (bp)	IRs región (bp)	SSC región (bp)	Total genes	CDS	GC%	rRNA	tRNA
NC_018552	C. annuun	156,781	87,366	25,783	17,849	132	87	37,37	8	37
KX913219	C. *tovarii*	156,816	87,379	25,792	17,853	132	87	37,72	8	37
KR078312	*C. frutescens*	156,817	87,380	25,792	17,853	131	87	37,72	8	37
MZ379791	*C. chinense*	156,931	87,325	25,847	17,912	133	86	37,71	8	37
NC_033524	*C. galapagoense*	156,959	87,347	25,847	17,918	132	87	37,71	8	37
KX913217	*C. chinense*	156,936	87,330	25,847	17,912	132	87	37,72	8	37
KU041709	*C. chinense*	156,807	87,290	25,803	17,911	113	79	37,70	4	30
MH559321	*C. chinense*	156,858	87,288	*25,855*	17,860	132	86	37,43	4	30
NC_030543	*C. chinense*	156,807	87,290	*25,803*	17,911	113	79	37,73	4	30
KX913220	*C. eximium*	156,947	87,341	25,847	17,912	132	87	37,72	8	37
MH559320	*C. baccatum*	157,144	87,351	25,910	17,973	113	79	37,66	4	39
NC_039694	*C. pubescens*	157,390	87,688	25,887	17,928	131	86	37,69	8	37
NC_026551	*C. lycianthoides*	156,583	86,813	25,624	18,522	127	83	37,76	8	36
NC_033525	*C. chacoense*	156,959	87,379	25,859	17,898	132	87	37,68	8	37

The cp genome of the ACP pepper was annotated to predict a total of 133 genes, including 86 protein-coding genes, 37 tRNA genes, 8 rRNA genes, and 2 pseudogenes. Among the annotated genes, a total of 11 large ribosomal proteins (*rpl*), 14 small ribosomal proteins (*rps*), 8 different rRNA genes, 4 DNA-dependent RNA polymerases (*rpo*), 38 different tRNA genes (*trn*), 5 genes encoding the photosystem I protein complex (*psa*), 14 genes encoding the photosystem II protein complex (*psb*), 6 genes encoding the cytochrome b6/f complex (*pet*), 12 genes encoding NADH dehydrogenase proteins (*ndh*), 6 genes encoding ATP synthase proteins (*atp*), the ATP-dependent Clp protease proteolytic subunit (*clpP*), the large subunit of ribulose-1,5-bisphosphate carboxylase (*rbcL*), maturase K (*matK*), genes encoding cytochrome C biogenesis proteins (*ccsA*), the beta subunit of acetyl-CoA carboxylase (*accD*), the membrane envelope protein (*cemA*), 4 ycf1 genes (*ycf*), the gene encoding translation initiation factor 1 (*infA*), and a gene of unknown function (*pafI*) were found. In addition, the annotation results revealed that 24 of these genes contain introns (*ndhB*, *petB*, *petD*, *rpl2*, *rps12*, *trnA-UGC*, *trnG-UCC*, *trnI-GAU*, *trnL-UAA*, *atpF*, *clpP1*, *ndhA*, *ndhB*, *rpl16*, *rpl2*, *rpoC1*, *rps12*-fragment, *rps12*, *rps16*, *trnA*-UGC, *trnI*-GAU, *trnK*-UUU, *trnV*-UAC, *pafI*), and four of them contain two introns (*pafI*, *clpP1*, and the two *rps12* structures). The majority of the genes are expressed as a single copy, with the exception of 18 genes that were duplicated in the IR regions.

Table 2
Genetic content and its functions of the chloroplast genome of *C. chinense* Arnaucho chili pepper.

Category	Group of genes	Name of genes
	Large subunit of ribosomal proteins	*rpl2*[a], *rpl4, rpl16*[a], *rpl20, rpl22, rpl23, rpl32, rpl33 rpl36*
	Pequeño subconjunto de proteínas ribosómicas.	*rps2, rps3, rps4, rps7, rps8, rps11, rps12*[a], *rps14, rps15, rps16, rps18, rps19*
	ARN polimerasa dependiente de ADN	*rpoA, rpoB, rpoC*[a]
Sef-replication	Ribosomal RNA genes	*rrn4.5, rrn5, rrn16, rrn23*
		trnC-GCA, trnG-GCC, trnG-UCC[a], *trnG-UCC, trnL-CAA, trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU*
	Transfer RNA genes	*trnA-UGC, trnD-GUC, trnE-UUC, trnH-GUG, trnI-CAU, trnI-GAU, trnK-UUU, trnR-ACG, trnR-UCU*
		trnS-GGA,trnT-GGU, trnV-GAC,trnL-CAA,, trnP-UGG, *trnQ-UUG, trnR-ACG, trnS-GCU, trnS-UGA, trnT-UGU, trnV-GAC, trnV-UAC*[a], *trnW-CCA, trnY-GUA, trnfM-CAU*
	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbT, psbZ*
Genes for photosynthesis	RUBISCO	*rbcL*
	Subunits of ATPsynthase	*atpA, atpB, atpE, atpF*[a], *atpH, atpI*
	Subunit of NADH-dehydrogenase	*ndhA*[a], *ndhB*[a],*ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Cytochrome b/f complex	*petA, petB*[a], *petD*[a], *petG, petL, petN*
	Protease	*clpP*[b]
	Mutrase	*matK*
Other genes	Envelope membrane protein	*cemA*
	Translation initiation factor	*infA*
	C-type cytochrome synthesis gene	*ccsA*
	Subunit of Acetyl-CoA-carboxylase	*accD*
Genes of unknown function	Conserved hypothetical chlroplast	*ycf1, ycf2*,
		pafI

Comparative analysis of the cp genome sequences of the *Capsiceae* tribe

Finally, the results from the mVista web server provided the information needed to assess the divergence among the 14 cp genome sequences in the Capsiceae tribe, using ACP as a reference. It was observed that the highest divergence was found in *C. lycianthoides* compared to our reference genome. These results provided information that the sequence variation was higher in the non-coding regions across all genomes, with lower variation observed in the coding regions of the genes. However, *C. lycianthoides* exhibited variation in more than one coding gene, including the genes *accD*, *rpl20*, *rps12*, *clpP*, *ycf2*, *ndhF*, *ndhA*, *ycf1*, and *rpl2*. In addition, the four *C. chinense* species' cp genomes showed significant variation in the *accD* gene, along with *C. eximium*, *C. pubescens* and *C. baccatum* which also exhibited variation in the same gene. On the other hand, the species *C. galapagoense* showed significant variation in the *rpl20* gene, along with *C. annuum*, *C. baccatum*, *C. pubescens*, *C. lycianthoides* and *C. chacoense*. These results justify further research to better determine and understand the clade of *C. chinense*, as it is unclear whether this fragment was altered in its cp genome due to a mutation caused by environmental factors (since the species are found in different habitats), or if it was simply a result of incomplete genome assembly.

Codon usage analysis in the Arnaucho chili pepper

For this codon analysis, the software Mega11 was used, and the number of times each codon is repeated and the relative synonymous codon usage (RSCU) were calculated from the protein-coding sequences (CDS). The most used codon in the *C. chinense* genome was AAA, which codes for lysine (2140), followed by UUU, which codes for phenylalanine (2118), and lastly AUU, which codes for isoleucine (1864). In the stop codons of *C. chinense*, the most used codons were UAA (1136) and UGA (982). Finally, the least present amino acid codons in the *C. chinense* genome were GGG, which codes for glycine (578), and ACC, which codes for threonine (576).

Phylogenetic analysis and pairwise distance

The results obtained from the ML phylogenetic analysis showed that different cultivars of *C. chinense* grouped with others of the same species, forming monophyletic groups. It can also be observed that there is a genetic relationship between our *C. chinense* and *C. galapagoense*, demonstrating that the ACP is a sister species to *C. galapagoense*, a result similar to the findings of Arbizu et al. (2022). Additionally, the results obtained at each node were very similar to the recent studies conducted by Sebastián et al. (2024), showing that the *C. chinense* clade is related to *C. galapagoense*.

In the pairwise distance results of the 14 cp genomes using the p-distance algorithm, the ACP was compared as the reference with the other genomes of the Capsiceae tribe. As expected, the highest average distance was found between Arnaucho chili pepper and *C. lycianthoides* (0.00021523), while the lowest was between *C. chinense* and *C. galapagoense* (0.00002733). Additionally, the following cp genomes were observed: *C. eximium* with a distance of 0.00003181, and the other four *C. chinense* genomes with average distances of 0.00003422, 0.00003551, and two with 0.00003822. Thanks to these results, the distributions and densities between the distances within the Capsiceae tribe can be observed. Similarly, as seen in the phylogenetic results, the five *C. chinense* genomes, *C. galapagoense*, and *C. eximium* exhibit a similar distribution and probability density trend, in contrast to the other plastid genomes of their tribe.

Analysis of repetitive sequences

The analysis detected different types of repetitive sequences within the genome of the Arnaucho chili pepper, as well as across the entire tribe, consisting of 14 cp genome sequences. These simple sequence repeats (SSRs) are tandemly repeated DNA of 1 to 6 base pairs in length, commonly used as molecular markers to identify species, revealing gene rearrangements and losses during their evolution (Nashima et al. 2020; Mcdonald et al. 2011). The results obtained in the ACP genome predicted a total of 36 SSRs, consisting of 33 mononucleotides, 1 trinucleotide,

and 2 pentanucleotides, revealing a significantly different pattern compared to all fourteen cp genomes of the Capsicum tribe, including the cp genomes of the four *C. chinense* and *C. galapagoense*, which belong to the same clade (Raveendar et al. 2015). In addition, Reputter, RepEX, and Tandem Repeats Finder characterized a total of 172 palindromic sequences, 41 forward sequences, 3 reverse sequences, 0 complementary sequences, and 55 tandem repeats, maintaining similarity with the other genomes in these repetitive sequences (Fig. 6). Additionally, the other repetitive sequences from the 14 chloroplast genomes were also characterized and subsequently compared in a balloon plot. These results can be seen in (Fig. 07) and (Table 03).

Table 3
The different types of repetitive sequences and SSRs in the 14 chloroplast genomes of the Capsiceae tribe

Scientific name	Genbank Number	Palindromos	Forward	Reverse	Tandem	SSR
C. annun	NC_018552	175	44	5	62	44
C. tovarii	KX913219	175	39	1	62	50
C. frutescens	KR078312	175	29	1	62	50
C. chinense	MZ379791	172	41	3	55	36
C. galapogense	NC_033524	172	39	2	59	50
C. chinense	KX913217	172	39	3	57	50
C. chinense	KU041709	163	32	3	57	49
C. eximium	KX913220	172	45	3	59	50
C. baccanturn	MH559320	174	49	0	64	50
C. pubescens	NC_039694	173	49	0	66	52
C. lycianthoides	NC_026551	181	31	1	50	54
C. chacoense	NC_033525	178	41	1	60	55
C. chinense	MH559321	172	33	1	57	50
C. chinense	NC_030543	172	16	1	57	51

In the detection of simple sequence repeats (SSRs), a heatmap was created for the comparative analysis of SSR motif frequencies. There was a higher frequency of the thymine (T) motif in the Capsiceae tribe, ranging from 14 in *C. annuum* to 26 in *C. lycianthoides*, followed by the adenine (A) motif. It was noted that its frequency was identical across all genomes, except for *C. annuum*, which had a total of 11, and *C. lycianthoides*, which had 14. This result can also be observed with the TTA (trimer) motifs, where the cp genomes of *C. baccatum* and *C. lycianthoides* do not show these motifs, while the others all exhibited the same frequency of 1. Studies on SSR in plants indicate that trimers are the most frequent in most groups of higher plants, such as monocots and dicots. This is consistent with the results described for this tribe, confirming studies outlined by Victoria et al. (2011). In addition, the frequency of motifs such as AT, CT, TA, TTC, AAAC, AAAT, AATT, ATAA, CTAT, TTTG, TTATT, TTTTA, TATAG, TTCATT, TAAT, G, and TTTTAA in the identified SSRs was entirely different, where all exhibited varying frequencies, except for the ACP, as previously described in the analysis. In the predicted number of SSRs, Arnaucho chili pepper showed the lowest amount of SSRs according to MISAweb, compared to *C. galapagoense* and the other two *C. chinense* individuals, which exhibited similar quantities ranging from 51 to 49 SSRs (Fig. 7).

Comparative analysis in the binding regions

The chloroplast genomes of the Capsiceae tribe showed a magnitude of displacement between the unique copy regions and the inverted repeats, considering their position relative to the nearest genes (Fig. 7). These boundaries were named JLB (LSC-IRB), JSB (IRB-SSC), JSA (SSC-IRA), and JLA (IRA-LSC) (Caycho et al. 2023). The length of the IR in the Capsiceae tribe varies from 156,583 bp to 157,390 bp across the 14 plastid genomes. Our Arnaucho chili pepper possesses a length of 156,931 bp, similar to *C. galapagoense* with 156,959 bp and *C. chinense* (KU041709) with 156,936 bp, respectively. This indicates that the IR expansion was almost identical, with no significant expansion observed among these plastid genomes. It was very different with the other three *C. chinense* genomes, which have 156,858 bp (MH559321) and two with 156,807 bp (KU041709 and NC_030543), where there were apparently only two small contractions of 124 bp and 73 bp. These expansions and contractions of the IR boundary contribute to the size and variations of the plastid genomes (Zhang et al., 2023). Within the 14 plastid genomes of the Capsiceae tribe examined, the four distinct boundaries and their genes located near these boundaries were identified, showing consistent patterns, as well as some absences in the analyzed genomes (Fig. 08). Specifically, regarding the junction between LSC and IRb (JLB), where all plastid genomes showed the structure of the rps19 gene spanning from the LSC region to IRb, there was a difference in length ranging from 206 bp to 218 bp. Additionally, Arnaucho chili pepper and the other four *C. chinense* species, including *C. galapagoense*, exhibited the same position and size for the *rps19* gene. The *rpl22* and *rpl2* genes are located in the LSC and IRB regions.

Similarly, for the JSB boundary, where the union between the SSC and IRB regions occurs, two genes, *ycf1* and *ndhF*, were positioned. The *ndhF* gene was located in the SSC regions, while it was observed that *ycf1* crossed from the SSC region to the IRB region, except in Arnaucho chili pepper and *C. lycianthoides*. Additionally, it was absent in *C. pubescens*, *C. baccatum*, and *C. chinense* (MH559321). The *ycf1* genes that spanned both the SSC and IRB regions had the same size of 1160 bp, while those located only in the IRB region had a size of 1127 bp in *C. chinense* and 1043 bp in *C. lycianthoides*. The *ndhF* gene (2222 bp) in the plastid of Arnaucho chili pepper showed a slight positional difference compared to *C. galapagoense* and the other four *C. chinense* genomes, and it was located adjacent to the JSB boundary. At the JSA boundary between SSC and IRA, the closest genes in the Capsiceae tribe showed that the *ycf1* gene was intact at the JSA boundary and crossed from SSC to the IRA region. However, the length of this gene varied from 5693 bp in *C. lycianthoides* to 5735 bp in *C. baccatum*. Additionally, four *C. chinense* showed the same length for the gene *ycf1* with a size of 5720 bp, except for *C. chinense* which had 5675 bp and *C. galapagoense* with 5726 bp. Finally, the analysis of the junction between the IRA and LSC regions (JLA) showed that the gene *rpl2* was present in IRA, except in *C. lycianthoides*, which did not contain this gene. The genes *trnH* and *psbA* were located at the LSC boundary and were present in all the chloroplast genomes.

Discussion

The native Arnaucho pepper, also known as “supano”, “campiñero” or “arnaucho campiñero”, is an ecotype from the Supe Valley in the Barranca province, and is very important in the Peruvian gastronomy, especially in the Lima. This *Capsicum chinense* cultivar has been cultivated since in family gardens for self-consumption sin the 1960s. Starting in the 1980s, it began to be cultivated in small plots with marketing intentions, where its distribution was through seeds and seedlings, from farmer to farmer, via family members or neighbors, and it continues to be distributed this way to this day (Aliaga et al. 2020). Currently, the first chloroplast genome sequencing has been carried out, with the objectives of characterizing the species and understanding its phylogenetic evolution (Arbizu et al. 2022). In this study, we collected the chloroplast genomes of different *Capsicum* species from the Organelle Genome Resources database of NCBI. The collected genomes were then annotated using the Geseq tool, and the chloroplast genomes were compared using the Arnaucho chili pepper as the reference genome to further study its evolution, structure, and

comparison of its cp genome. The cp genome of *C. chinense* Arnaucho chili pepper was 156,931 bp (156.9 Kb) in length, also exhibiting a classic quadripartite circular structure: a large single-copy region (LSC), a small single-copy region (SSC), and two inverted repeats (IRA and IRB). These regions were also present in the other *Capsicum* species. Additionally, we found that the cp genomes of the Capsiceae tribe varied slightly in size but maintained a GC content of 37%. This finding aligns with previous reports indicating that the chloroplast genome structure in most angiosperms is typically maternally inherited and remains largely unchanged (Birky, 1995). The lengths of the LSC region ranged from 86,813 bp (*C. lycianthoides*) to 87,688 bp (*C. pubescens*), while the SSC region ranged from 17,849 bp (*C. annum*) to 18,522 bp (*C. lycianthoides*), and the IR region ranged from 25,624 bp (*C. lycianthoides*) to 25,910 bp (*C. baccatum*). These results were similar to studies reported by Raveendar et al. (2024). The same has been found in other important Solanaceae crop species, such as *S. lycopersicum*, *S. tuberosum*, and *P. peruviana* (Zhao et al. 2019; Chen et al. 2021; Feng et al. 2020). The chloroplast genomic regions of Solanaceae maintain a similar size among closely related species, as in the case of the Arnaucho chili pepper, an important crop for human consumption.

Additionally, in another exotic Solanaceae plant, *N. physalodes*, studies described by Chen & Zhang (2019) revealed that the total length of its chloroplast genome was 156,729 bp (157.7 kb), showing a similar length to *Capsicum* species. However, some Solanaceae species exhibit regions of varying lengths. For example, studies conducted on other Solanaceae species reported *Datura stramonium* with a size of 155,871 bp and the African boxwood shrub *Lycium ferocissimum* with 155,894 bp, showing a contraction of around 1,000 bp (Yang et al. 2014 & Li et al. 2019). However, these species exhibited similar boundary regions to the Capsiceae tribe, specifically in the IR region, which measured 25,601 bp in *D. stramonium* and 25,476 bp in *L. ferocissimum*, and in the LSC region, which measured 86,302 bp in *D. stramonium* and 86,536 bp in *L. ferocissimum*. This suggests that the reduction in chloroplast genomes of these species occurs in the SSC regions, possibly due to genetic variations present in this family (Su et al. 2020). Additionally, in the annotated chloroplast genome of *C. chinense*, 133 genes were identified, including a total of 86 CDS, 8 rRNA, 37 tRNA, and 2 pseudogenes. Compared to other *Capsicum* species, this variety has more genes in its chloroplast but fewer protein-coding sequences. Most of the genes are present as a single copy, except for 18 genes that are duplicated in the IR regions. These 18 duplicated genes in the IR region have also been found in other Solanaceae species that are not *Capsicum* (Asaf et al. 2016). In the genomic sequence comparison analysis conducted by mVISTA, it was revealed that *C. lycianthoides* exhibited the greatest divergence in relation to *C. chinense* Arnaucho chili pepper. The analysis showed that the greatest variation occurred in the non-coding regions compared to the coding regions present in all genomes of the Capsiceae tribe. Additionally, the variations observed in the coding regions were present in all the genomes, specifically in the genes *accD*, *rpl20*, *ycf1*, *ycf2*, and *ndhA*. On the other hand, the alignment between the Arnaucho chili pepper and *C. galapagoense* showed only one significant variation, which occurred in the *rpl20* gene. Meanwhile, the alignment between the ACP and the other four *C. chinense* individuals showed variations in the *accD* gene (Fig. 02).

The analysis of codon usage is essential for understanding the complexities of genomic structure, dynamic evolution, and the selective pressure imposed on genes (Morton, 1998). We found that the most commonly used codon in the *C. chinense* genome was AAA, which codes for the amino acid lysine, followed by UUU, which codes for phenylalanine, and finally AUU, which codes for isoleucine. These patterns indicate a preference for codons that end in A or T/U. Previous studies have shown that higher RSCU values indicate similar results in extensive research on Solanaceae species, with a tendency to be rich in A/U or A/T in chloroplast genomes (Mehmetoğlu et al. 2023 and Wang et al. 2023). This means that there is already a selective pressure favoring the use of codons in chloroplast genomes, commonly detected at the third position of codons.

Phylogenetic trees are used to understand the evolutionary relationships between species, and they can be based on genomic regions or the complete genome sequence. The angiosperm tree of life has been primarily determined

through plastid genome analysis. These chloroplast genomes are highly conserved in both sequence and structure, making them highly valuable for taxonomic studies and plant classification (Amenu et al. 2022; Zuntini et al. 2024). Here, a phylogenetic tree reconstruction was performed using the complete genome sequence of the pepper along with 14 additional Capsicum genomes. The goal was to classify it within its corresponding tribe and determine its phylogenetic position. The tribe was divided into six groups (green, mustard, light blue, purple, red, and yellow), along with one additional group (pink) representing two Solanaceae genome sequences used to root the tree. The analysis revealed that the Arnaucho chili pepper and *C. galapagoense* clustered in the corresponding group alongside other *C. chinense* and *C. eximium*, showing a genetic relationship among these species. This previous result had already been reported in a study described by Arbizu et al. (2022). To determine the pairwise distance between chloroplast genomes, an analysis was conducted using Mega11. This analysis was performed using the p-distance method with complete removal of missing data. The results showed that the p-distance values ranged from 0.00002733 for the closest species, *C. galapagoense*, to 0.00021523 for the most distant species, *C. lycianthoides*, respectively. These p-distance studies have also been conducted in other families, such as the Styracaceae. Here, we present one of the first results of p-distance analysis in the Capsicum tribe (Song et al. 2022).

Within each genome, a large number of repetitive sequences, also known as SSRs (Simple Sequence Repeats), are present. These repetitive sequences exhibit high sequence repeatability, high variability, and co-dominant heterozygous inheritance. These sequences are molecular genetic markers suitable for species identification, as well as for ecological and evolutionary studies (Yang et al. 2017; Li et al. 2014). Analyses of these SSR sequences offer a perspective as potential markers specific to each genus (Shirasawa et al. 2010). In general, the most abundant nucleotides in chloroplast genomes are A and T, or A/T repeats, which are prevalent in angiosperms, rather than G and C or G/C repeats (Kurt et al. 2023). In this SSR analysis of Arnaucho chili pepper, a total of 36 SSR markers were identified, comprising 33 mononucleotides, 1 trinucleotide, and 2 pentanucleotides. This revealed a lower pattern compared to the different types of *Capsicum* species present in this study. The repetitive sequence repetitions, including palindromic sequences, reverse sequences, forward sequences, and tandem repeats, showed values of 172, 41, 3, and 55, respectively. In the comparison of the number of repetitive sequences across the 14 cp genomes, it was observed that none of the genomes contained complementary sequences. There was a variation in the number of reverse sequences, ranging from 1 to 5 in relative magnitude, with the exception of *C. baccatum* and *C. pubescens*, which did not present these repetitive sequences (shown in red). For forward sequences, the relative magnitude of repetitive sequences ranged from 16 to 49. In the complementary sequences, a relative magnitude of repetitive sequences was observed between 50 and 66 (shown in light blue), and lastly, for tandem sequences, the relative magnitude was greater than 150, ranging from 163 to 181. In comparison to the SSR sequences, these repetitive sequences appeared similarly to the repetitive sequences found in the other genomes, corroborating that this particular chili variety has fewer SSR sequences in its genome compared to the other Capsicum species detailed in this research.

The IR regions can indicate the size of cp genomes because they expand or contract, reflecting the distance between species to some extent (Zhao et al., 2023). These highly variable IR regions can provide molecular marker studies in cp genomes, which are important for research related to species identification, phylogeny, and population genetics (Xiao et al., 2024). Our analysis of the ACP genome revealed that it retains an IR region of 25,847 bp (25 kb), leading to a total genome length of 156,931 bp. It is important to highlight that the chloroplast genome length of this Peruvian landrace closely resembles that of *C. galapagoense* and *C. chinense* (GenBank KU041709), indicating similar IR expansion. However, its structure is distinctly different. In the Arnaucho chili pepper, the JSB boundary composed of IRb and SSC contains the *ycf1* gene, which is located exclusively within the IRb structure. In contrast, *C. galapagoense* and *C. chinense* possess the *ycf1* gene spanning both the IRb and SSC structures at the JSB boundary. The current taxonomy of this Peruvian landrace is questioned based on the present work, and also by Arbizu et al. (2022). The *ycf1*

gene in the ACP is 1127 bp long, similar in size to the 1043 bp *ycf1* gene in *C. lycianthoides*, positioned solely at the junction. This gene's variation plays a key role in the structural variation of chloroplast genomes within the Capsiceae tribe. This result regarding the IR structure indicates that the *ycf1* gene in most angiosperms is generally larger and more diverse. In some cases, the *ycf1* gene may be absent, which could lead to a reduction in protein-coding capacity and variation in the IRb/SSC boundary structure. As noted by Ge et al. (2019), this variation can influence the overall stability and function of the chloroplast genome, potentially affecting the plant's metabolic processes and adaptation to different environments. The presence of *ycf1* and its position within the IR regions plays a significant role in the genomic architecture and evolutionary dynamics of angiosperms.

Conclusion

Here, we report the in-silico analysis of the Arnaucho chili pepper along with 13 chloroplast genomes from the Capsiceae tribe, including both hot and sweet peppers, obtained from the NCBI database. We compared the 14 chloroplast genome sequences. In this comparative analysis, highly variable sites were identified, specifically non-coding regions, as well as coding regions of genes that could serve as potential markers for species identification. In this case, *rpl20*, *accD*, *ycf1*, *ycf2*, and *ndhA* were highlighted. It was also observed that this Arnaucho chili pepper exhibited patterns of codons A (adenine) and U (uracil), similar to the other types of peppers present. It was discovered in its phylogenetic reconstruction and p-distance analysis that this *Capsicum* species is closely related in evolutionary terms to *C. galapagoense*. All the genomes of the Capsiceae tribe were highly conserved, except in the IRB region at the JSB boundary. This complete cp genome of Arnaucho pepper showed different genetic characteristics regarding its SSR sequence identification compared to the other genomes. It was found to possess 36 SSR sequences, which is fewer than the other pepper genomes used in the present study. Additionally, in the other repetitive sequences, it was completely different, displaying similarities to the other types of chloroplast genomes of peppers studied. These studies may be useful for future research on the diversity or genetic improvement of this Peruvian chili pepper landrace.

Declarations

Acknowledgments

We would like to thank the Vicerrectorado de Investigación of UNTRM.

Author contributions

GC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. KRQ-H: Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. DHT-I: Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. JEB-G: Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. YAC-R: Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. SC-L: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – review & editing. CIA: Conceptualization, Data curation, Investigation, Resources, Supervision, Validation, Visualization, Writing – review & editing. PMR-G: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Funding

This research received not external funding.

Conflict of interest

The authors declare that they have no conflicts of interest in this research.

Data Availability

No datasets were generated during the current study.

Conflicts of interest

The authors declare no competing interests.

References

1. Aguilar-Meléndez A, Morrell PL, Roose ML, Kim SC. Genetic diversity and structure in semiwild and domesticated chiles (*Capsicum annuum*; Solanaceae) from Mexico. Am J Bot. 2009;96(6):1190–202
2. Aliaga, J., Portalatino, E., Obregon, K., Rodriguez, A. & Jimenez, J. (2020). Presencia del "aji nativo supano" (Capsicum chinense Jacq.) en el valle de Supe, Peru. Peruvian Agricultural Research. (1). 58-63. DOI: 10.51431/par.v1i2.584
3. Amenu, S.G., Wei, N., Wu, L. et al. Phylogenomic and comparative analyses of Coffeeae alliance (Rubiaceae): deep insights into phylogenetic relationships and plastome evolution. BMC Plant Biol 22, 88 (2022). https://doi.org/10.1186/s12870-022-03480-5
4. Amiryousefi, A., Hyvönen, J., Poczai, P. IRscope: an online program to visualize the junction sites of chloroplast genomes, Bioinformatics, Volume 34, Issue 17, September 2018, Pages 3030–3031, https://doi.org/10.1093/bioinformatics/bty220
5. Ángel Gaviria, I.J. (2023). La bioinformática como herramienta para el conocimiento de microorganismos edáficos con potencial para la producción agrícola sostenible, recuperación y conservación de suelos [Tesis de maestría, Universidad Libre].
6. Antonio, A.S., Wiedermann, L.S.M. & Veiga, F.J. (2018). The genus Capsicum: a phytochemical review of bioactive secondary metabolites. RSC Advances. 25767-25784. DOI: 10.1039/c8ra02067a
7. Arbizu, C. I., Saldaña, C. L., Ferro-Mauricio, R. D., Chávez-Galarza, J. C., Herrera, J., Contreras-Liza, S., … Maicelo, J. L. (2022). Characterization of the complete chloroplast genome of a Peruvian landrace of *Capsicum chinense* Jacq. (Solanaceae), arnaucho chili pepper. *Mitochondrial DNA Part B*, *7*(1), 156–158. https://doi.org/10.1080/23802359.2021.2014366
8. Asaf, S., Khan, A.L., Khan, A.R., Waqas, M., Kang, S., Khan, M., Lee, S. & Lee, I. (2016). Complete Chloroplast Genome of Nicotiana otophora and its Comparison with Related Species. Front Plant Sci. Sec. Evolutionary and Population Genetics. https://doi.org/10.3389/fpls.2016.00843
9. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33 2583–2585. dx.doi.org/10.1093/bioinformatics/btx198
10. Benson 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. Vol 27. Pages 573-580. https://doi.org/10.1093/nar/27.2.573
11. Birky Jr, C.W. (1995). Uniparental inheritance of mitochndrial and chloroplast genes: mechanisms and evolution. PNAS. Vol 92. 11331-11338. https://doi.org/10.1073/pnas.92.25.11331
12. Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., and Batzoglou, S. Glocal Alignment: Finding Rearrangements During Alignment, 2003. Bioinformatics, 19S1: i54-i62.

13. Caycho, E., La Torre, R. & Orjeda, G. Assembly, annotation and analysis of the chloroplast genome of the Algarrobo tree *Neltuma pallida* (subfamily: Caesalpinioideae). *BMC Plant Biol* **23**, 570 (2023). https://doi.org/10.1186/s12870-023-04581-5
14. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021)tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes.*Nucl. Acids Res.* **49**:9077–9096.
15. Chen, Q., & Zhang, D. (2019). The complete chloroplast genome sequence of the medicinal plant Nicandra physalodes (Linn.) Gaertn. (Solanaceae). Mitochondrial DNA Part B, 4(2), 3053–3054. https://doi.org/10.1080/23802359.2019.1666674
16. Chen, S., Zhao, Y., Zhang, J. Y., Zhang, J. Y., Wang, Y. P., Mou, B., … Han, Y. Z. (2021). Characterization of the complete chloroplast genome of the *Solanum tuberosum* L. cv. Shepody (Solanaceae). Mitochondrial DNA Part B, 6(8), 2342–2344. https://doi.org/10.1080/23802359.2021.1934135
17. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8), 772.
18. Dong, X., Li, X., Ding, L., Cui, F., Tang, Z., Liu, Z., Stage Extraction of Capsaicinoids and Red Pigments from Fresh Red Pepper (Capsicum) Fruits with Ethanol as Solvent, LWT - Food Science and Technology (2014), doi: 10.1016/j.lwt.2014.04.051.
19. Estrada, R., Tantalean, J.F.C., Saldaña, C.L. *et al.* Draft genome and SSR data mining of a Peruvian landrace of *Capsicum chinense,* the arnaucho chili pepper. *Genet Resour Crop Evol* (2024). https://doi.org/10.1007/s10722-024-01941-4
20. Feng, S., Zheng, K., Jiao, K., Cai, Y., Chen, C., Mao, Y., Wang, L., Zhan, X., Ying, Q. & Wang, H. (2020). Complete chloroplast genomes of four Physalis species (Solanaceae): lights into genome structure, comparative analysis, and phylogenetic relationships. BMC Plant Biol 20, 242. https://doi.org/10.1186/s12870-020-02429-w
21. Florencio, V. & Valdir, Jr. (2022). Chapter 1: Origin and Evolution of Capsicum. Chemistry and Nutritional Effects of *Capsicum*. Page: 1-14. https://doi.org/10.1039/9781839160646-00001
22. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W273-9.
23. Ge, Y., Dong, X., Wu, B., Wang, N., Chen, D., Chen, H., Zou, M., Xu, Tan, L. & Zhan, R. (2019). Evolutionary analysis of six chloroplast genomes from three *Persea americana* ecological races: Insights into sequence divergences and phylogenetic relationships. PLoS ONE 14(9): e0221827. https://doi.org/10.1371/journal.pone.0221827
24. Greiner S, Lehwark P and Bock R (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research 47: W59-W64. https://doi.org/10.1093/nar/gkz238
25. Hernandez, A.A., Pineda, A.L., Rojas, J.C. & Diaz, H.P. (2022). In vitro regeneration of arnaucho (Capsicum chinense Jacq.) from apical buds. *Manglar.* (1). 71-75. DOI: http://dx.doi.org/10.17268/manglar.2021.009
26. Ibiza, V.P.; Blanca, J.; Cañizares, J.; Nuez, F. Taxonomy and genetic diversity of domesticated Capsicum species in the Andean region. Genet. Resour. Crop Evol. 2012, 59, 1077–1088. DOI 10.1007/s10722-011-9744-z
27. Katoh, K., Standley, D. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. Vol4. Pág 772-780. doi: 10.1093/molbev/mst010
28. Kurtz, S. and Choudhuri, Jomuna V. and Ohlebusch, Enno and Schleiermacher, Chris and Stoye, Jens and Giegerich, Robert (2001) REPuter: The Manifold Applications of Repeat Analysis on a Genomic Scale, Nucleic Acids Res., Nucleic Acids Res., 29(22):4633-4642.
29. Kurt, S., Kaymaz, Y., Ateş, D. et al. Complete chloroplast genome of Lens lamottei reveals intraspecies variation among with Lens culinaris. Sci Rep 13, 14959 (2023). https://doi.org/10.1038/s41598-023-41287-y

30. Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics*30(22): 3276-3278.http://dx.doi.org/10.1093/bioinformatics/btu531
31. Letunic I and Bork P (2024) Nucleic Acids Res doi: 10.1093/nar/gkae268 Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool
32. Li, M., Qi, X.M., Lam, H. (2013). Silicon Era of Carbon-Based Life: Application of Genomics and Bioinformatics in Crop Stress Research. Mol. Sci. 14. 11444-11483. https://doi.org/10.3390/ijms140611444
33. Li, X. et al. Plant DNA barcoding: From gene to genome. Biol. Rev. Camb. Philos. Soc. 90, 157–166 (2014).
34. Li, Z., Zhang, X., Zhang, Q., & Yisilam, G. (2020). Complete chloroplast genome of Lycium ferocissimum (Solanaceae), a species native to South Africa. Mitochondrial DNA Part B, 5(1), 756–757. https://doi.org/10.1080/23802359.2020.1715301
35. Liu H, Liu X, Sun C, Li HL, Li ZX, Guo Y, Fu XQ, Liao QH, Zhang WL, Liu YQ. Chloroplast Genome Comparison and Phylogenetic Analysis of the Commercial Variety *Actinidia chinensis* 'Hongyang'. Genes (Basel). 2023 Nov 27;14(12):2136. doi: 10.3390/genes14122136. PMID: 38136958; PMCID: PMC10743354.
36. López Medina, E., López Zabaleta, A., Gil Rivero, A. E., Mostacero, J., De La Cruz, A.J. y Villena, L. (2020) Morfometría de frutos y semillas del "ají mochero" Capsicum chinense Jacq. Ciencia & Tecnología Agropecuaria, 21(3), 1-11. https://doi.org/10.21930/rcta.vol21_num3_art:1598.
37. Mansoor, S., Hamid, S., Thanh, T.T., Park, J. Suk, Y.C. (2024). Advance computational tools for multiomics data learning. Biotechnol Adv. Vol 77. https://doi.org/10.1016/j.biotechadv.2024.108447
38. McDonald MJ, Wang W-C, Huang H-D, Leu J-Y (2011) Clusters of Nucleotide Substitutions and Insertion/Deletion Mutations Are Associated with Repeat Sequences. PLoS Biol 9(6): e1000622. https://doi.org/10.1371/journal.pbio.1000622
39. Meckelmann, S., Riegel, D.W., Zonneveld, M.J., Rios, L., Pe;a, K., Ugas, R., Quinonez, L., Mueller-Seitz, E. & Petz, M. (2013). Compositional Characterization of Native Peruvian Chili Peppers (Capsicum spp.). Agric Food Chem. Vol 61. Pages: 2530-2537. DOI: 10.1021/jf304986q
40. Mehmetoğlu, E., Kaymaz, Y., Ateş, D. et al. The complete chloroplast genome of Cicer reticulatum and comparative analysis against relative Cicer species. Sci Rep 13, 17871 (2023). https://doi.org/10.1038/s41598-023-44599-1
41. Michael, D., Gurusaran, M., Santhosh, R., Khaja, Md.H., Satheesh, S.N., Suhan, S., Sivaranjan, P., Jaiswal, A. & Sekar, K. (2019). RepEx: A web server to extract sequence repeats from protein and DNA sequences. Computational Biology and Chemistry. Vol: 78. Pages: 424-430. https://doi.org/10.1016/j.compbiolchem.2018.12.015
42. Minguez-Mosquera, M.I., Jaren-Galan, M. and Garrido-FErnandez, J. (1992). Color Quality in Paprika. J. Agric. Food Cherm. Vol:40. Pages 2384-2388. https://doi.org/10.1021/jf00024a012
43. Morton, B.R. (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol.;46(4):449–59.
44. McGowan, J. (2024). RSCU calculator. Online Bioinformatics Tools. https://jamiemcgowan.ie/bioinf/rscu.html
45. Nashima K, Hosaka F, Terakami S, et al. SSR markers developed using next-generation sequencing technology in pineapple, Ananas comosus (L.) Merr. Breed Sci. 2020;70(3):415–21.
46. Park, H. S., Lee, J., Lee, S. C., Yang, T. J., & Yoon, J. B. (2016). The complete chloroplast genome sequence of *Capsicum chinense* Jacq. (Solanaceae). *Mitochondrial DNA Part B*, *1*(1), 164–165. https://doi.org/10.1080/23802359.2016.1144113
47. Perry L, Dickau R, Zarrillo S et al (1979) (2007) Fósiles de almidón y la domesticación y dispersión de chiles (*Capsicum* spp. L.) en las Américas. Ciencia 315:986–988.https://doi.org/10.1126/science.1136914

48. Ray, S. & Saty, P. (2014). Next generation sequencing technologies for next generation plant breeding. Front Plant Sci. Vol 5. https://doi.org/10.3389/fpls.2014.00367

49. Raveendar, S., Jeon, Y., Lee, J., Lee, G., Lee, K.J., Cho, G., Ma, K., Lee, S. and Chung, J. (2015). The Compmplete Chloroplast Genome Sequence of Korean Landrace “Subicho” Pepper (*Capsicum annuum* var. *annuum*). Plant breeding and Biotechnology. Vol 3. 88-94https://doi.org/10.9787/PBB.2015.3.2.08888

50. Roychowdhury, R., Prakash, S.D., Gupta, A., Parihar, P., Chandrasekhar, K., Sarker, U., Kumar, A., Pandurang, D.R. & Sudhakar, C. (2023). Multi-Omics Pipeline and Omics-Integration Approach to Decipher Plant's Abiotic Stress Tolerance Responses. Genes. Vol 14. https://doi.org/10.3390/genes14061281

51. Sharp, P.M. & Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. Vol 3. Pages: 1281-1295. doi: 10.1093/nar/15.3.1281

52. Sebastin, R., Lee, K. J., Cho, G. T., Shin, M. J., Kim, S. H., Hyun, D. Y., & Lee, J. R. (2019). The complete chloroplast genome sequence of a Bolivian wild chili pepper, *Capsicum eximium* Hunz. (Solanaceae). *Mitochondrial DNA Part B*, *4*(1), 1634–1635. https://doi.org/10.1080/23802359.2019.1601533

53. Sebastin, R., Kim, J., Jo, IH. *et al.* Comparative chloroplast genome analyses of cultivated and wild *Capsicum* species shed light on evolution and phylogeny. *BMC Plant Biol* **24**, 797 (2024). https://doi.org/10.1186/s12870-024-05513-7

54. Shirasawa, K., Asamizu, E., Fukuoka, H., Ohyama, A., Sato, S., Nakamura, Y., Tabata, S., Sasamoto, S., Wada, T., Kishida, Y., Tsuruoka, H., Fujishiro, T., Yamada, M. & Isobe, S. (2010). An interspecific linkage map of SSR and intronic polymorphism markers in tomato. Theor Appl Genet. Vol 121. Pages 731-739. https://doi.org/10.1007/s00122-010-1344-3

55. Shiragaki K, Yokoi S, Tezuka T. 2020. Phylogenetic analysis and molecular diversity of capsicum based on rDNA-ITS region. Horticulturae. 6(4):87.

56. Song, F., Zhao, W., Xu, J., Li, M. & Zhang, Y. (2022). Chloroplast Genome Evolution and Species Identification of *Styrax* (Styracaceae). BioMed Research International. https://doi.org/10.1155/2022/5364094

57. Su Q, Liu L, Zhao M, Zhang C, Zhang D, Li Y, et al. Los genomas completos del cloroplasto de diecisiete *Aegilops tauschii*: análisis comparativo del genoma e inferencia filogenética. PeerJ. 2020;8:e8678.

58. Tamura, K., Stecher, G., and Kumar, S. (2021) MEGA11: Molecular Evolutionary Genetics Analysis version 11. Molecular Biology and Evolution 38:3022-3027

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

60. Tripodi P, Rabanus-Wallace MT, Barchi L, Kale S, Esposito S, Acquadro A, Schafleitner R, van Zonneveld M, Prohens J, Diez MJ. 2021. Global range expansion history of pepper (Capsicum spp.) revealed by over 10,000 genebank accessions. Proc Natl Acad Sci USA. 118(34):e2104315118.

61. Victoria, F.C., Maia, L.C., Oliveira, A.C. (2011). In silico comparative analysis of SSR markers in plants. BMC Plat Biology. 11-15. https://doi.org/10.1186/1471-2229-11-15

62. Wang, X., Bai, S., Zhang, Z., Zheng, F., Song, L., Wen, L., Guo, M., Cheng, G., Yao, W., Gao, Y. & Li, J. (2023). Front Plant Sci. Sec. Plant Bioinformatics. Vol 13. https://doi.org/10.3389/fpls.2023.1179009

63. Write, F. (1990). The 'effective number of codons' used in a gene. Gene. Vol 87. Pages: 23-29. https://doi.org/10.1016/0378-1119(90)90491-9

64. Xiao, F., Zhao, Y., Wang, X. et al. Characterization of the chloroplast genome of Gleditsia species and comparative analysis. Sci Rep 14, 4262 (2024). https://doi.org/10.1038/s41598-024-54608-6

65. Yang, Y., Yuanye, D., Qing, L., Jinjian, L. & Xiwen, L. (2015). Complete Chloroplast Genome Sequence of Poisonous and Medicinal Plant Datura stramonium: Organizations and Implications for Genetic Engineering. PLOS ONE 10(2): e0118236. https://doi.org/10.1371/journal.pone.0118236
66. Yang, Z. & Ji, Y. Comparative and Phylogenetic Analyses of the Complete Chloroplast Genomes of Three Arcto-Tertiary Relicts: Camptotheca acuminata, Davidia involucrata, and Nyssa sinensis. Front. Plant. Sci. 8, 1536 (2017).
67. Zhang, D., Tu, J., Ding, X., Guan, W., Gong, L., Qiu, X. & Huang, Z. (2023). Analysis of the chloroplast genome and phylogenetic evolution of Bidens pilosa. BMC Genomics.(113). https://doi.org/10.1186/s12864-023-09195-7
68. Zhao, C., Sun, K., Chen, S., Liang, C., Meng, J., Tang, Y., & Song, S. (2019). Characterization the complete chloroplast genome of the tomato (Solanum lycopersicum L.) from China. Mitochondrial DNA Part B, 4(1), 1374–1376. https://doi.org/10.1080/23802359.2019.1598300
69. Zuntini, A.R., Carruthers, T., Maurin, O. *et al.* Phylogenomics and the rise of the angiosperms. *Nature* 629, 843–850 (2024). https://doi.org/10.1038/s41586-024-07324-0

Figures

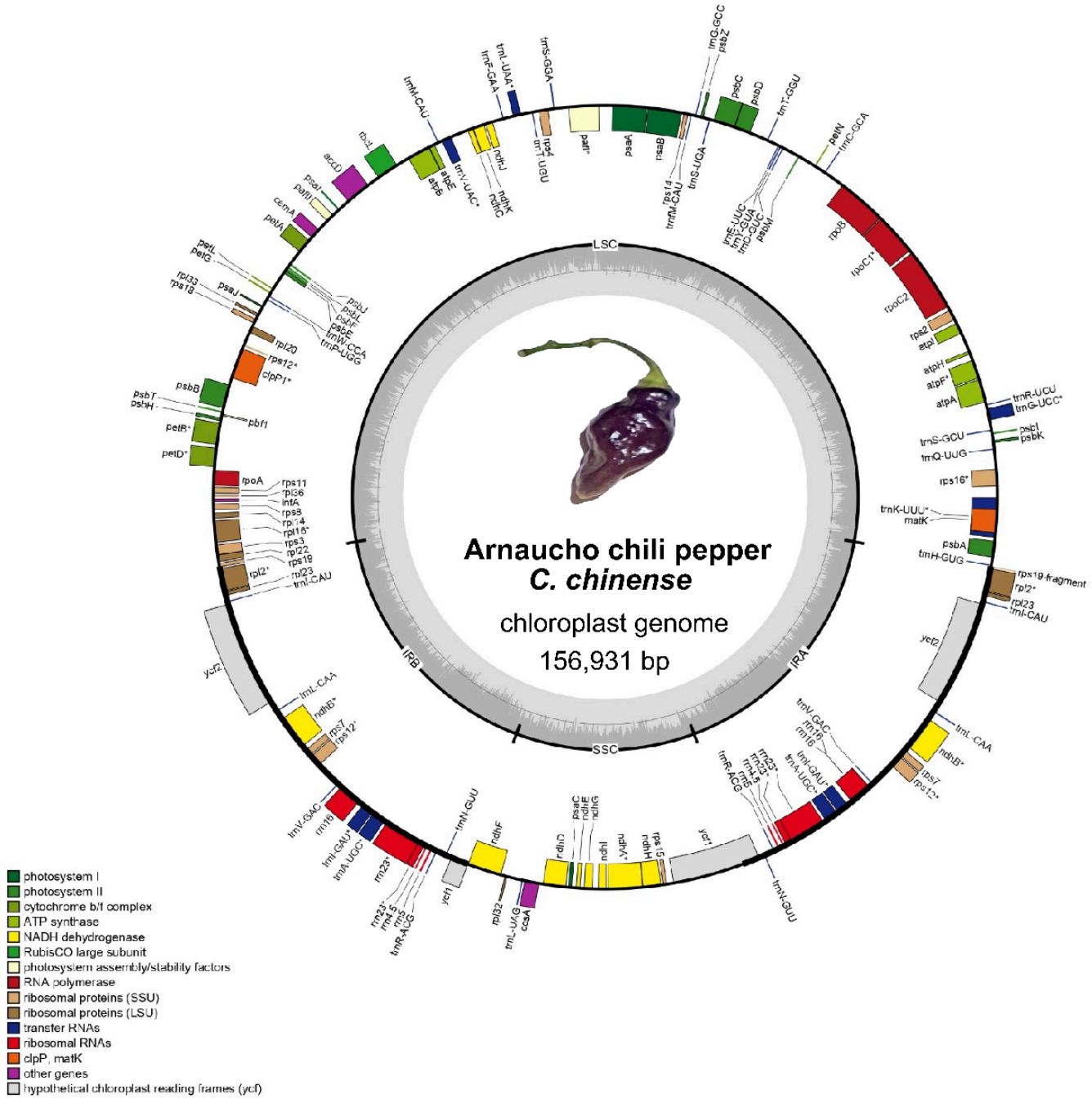

Figure 1

A circular map of the chloroplast genome of the Arnaucho chili pepper (ACP). The genes inside the circle are transcribed in a clockwise direction, while the genes outside the circle are transcribed in a counterclockwise direction. The dark gray area in the inner circle reflects the GC content of the cp genome, while the light gray area represents the AT content. In the center of the circle, the fruit of ACP is shown in the logarithmic growth phase. The genes belonging to different functional groups are shown as color-coded blocks

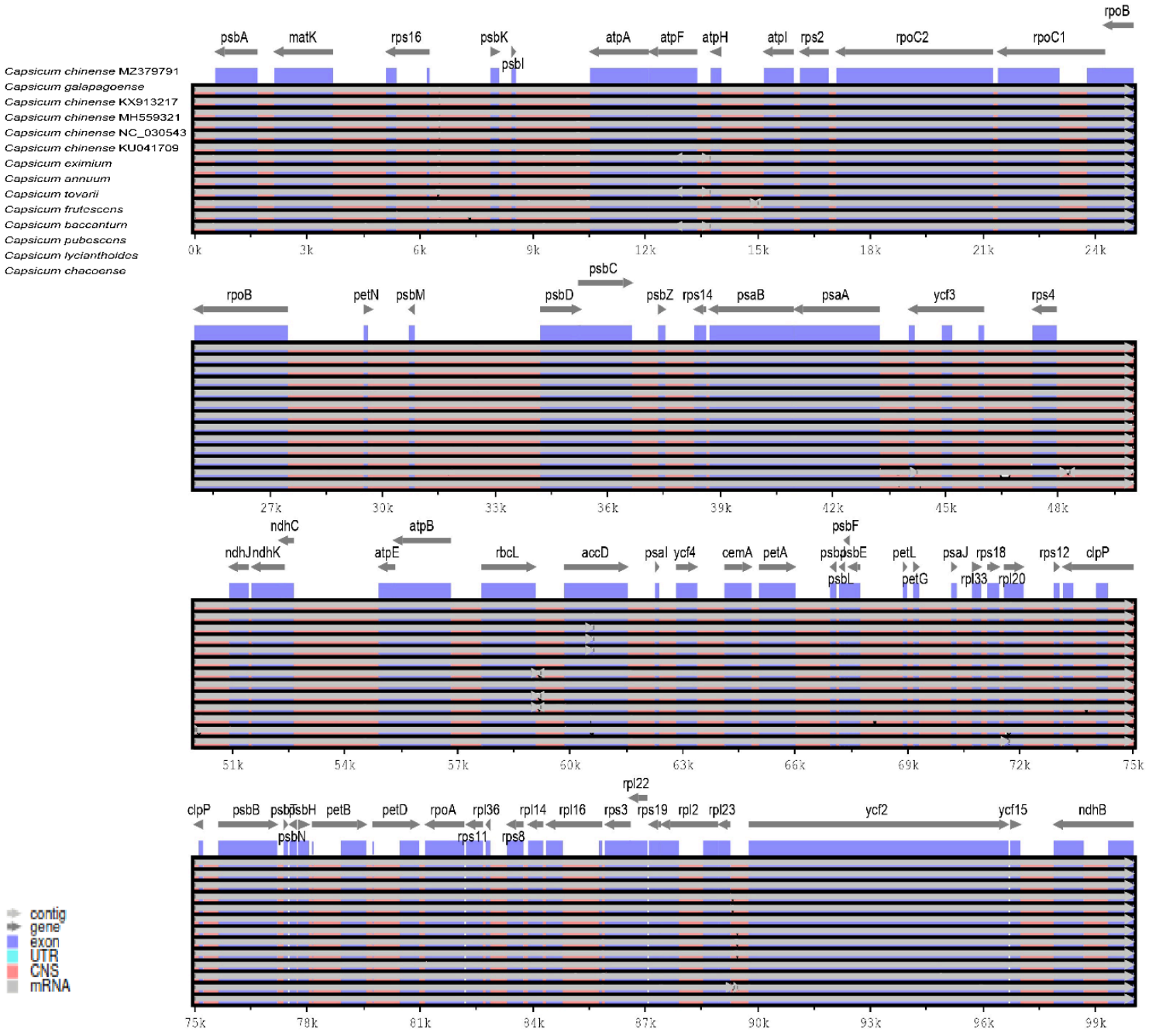

Figure 2

Graphical representation of the genomic alignment of the chloroplasts of the Capsiceae tribe. This graph shows the identity of the genomes, using the Arnaucho chili pepper as the reference genome

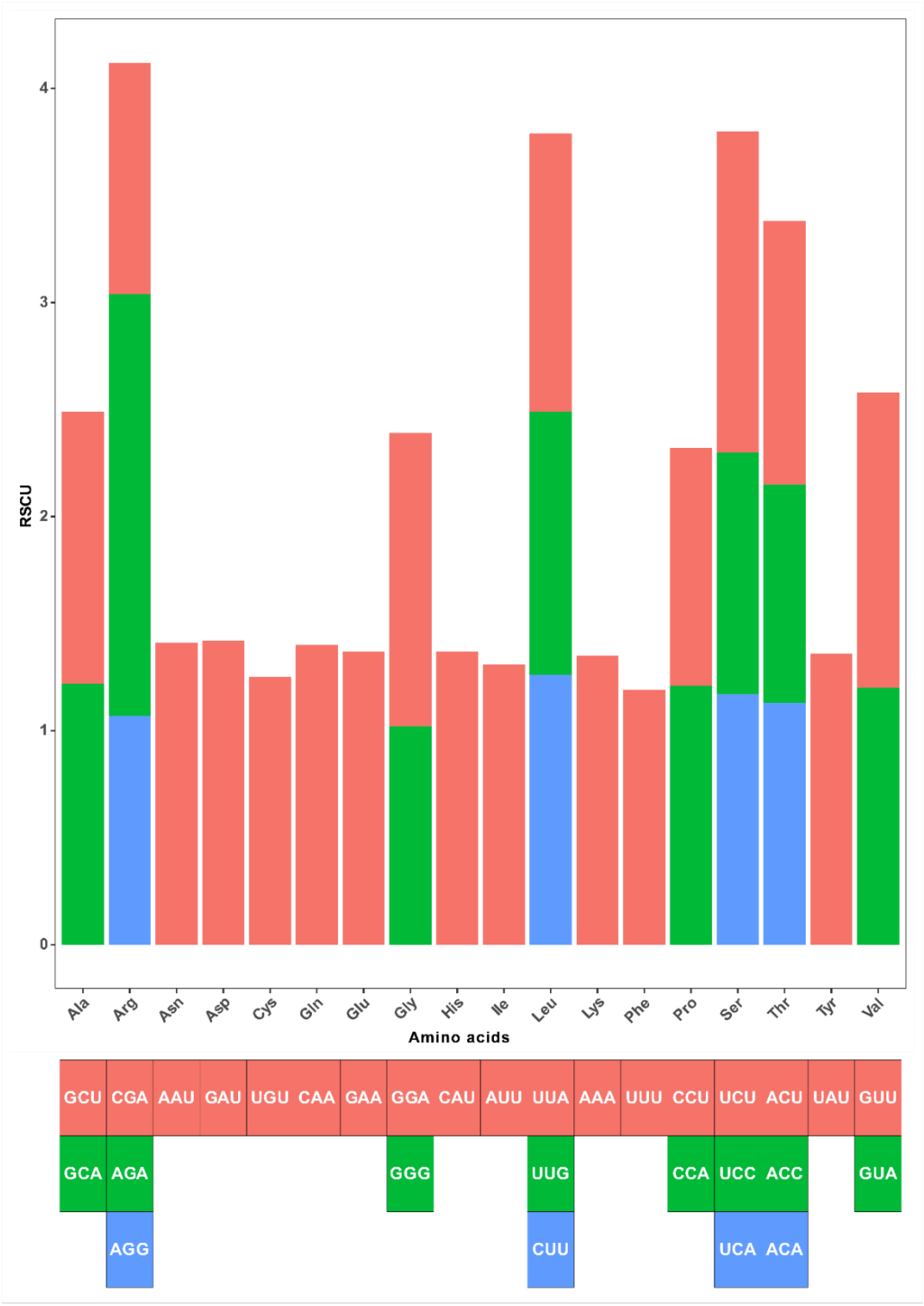

Figure 3

The RSCU values of the STOP codons and the amino acids encoded by more than one codon in the 78 genes of the chloroplast genome of Arnaucho chili pepper

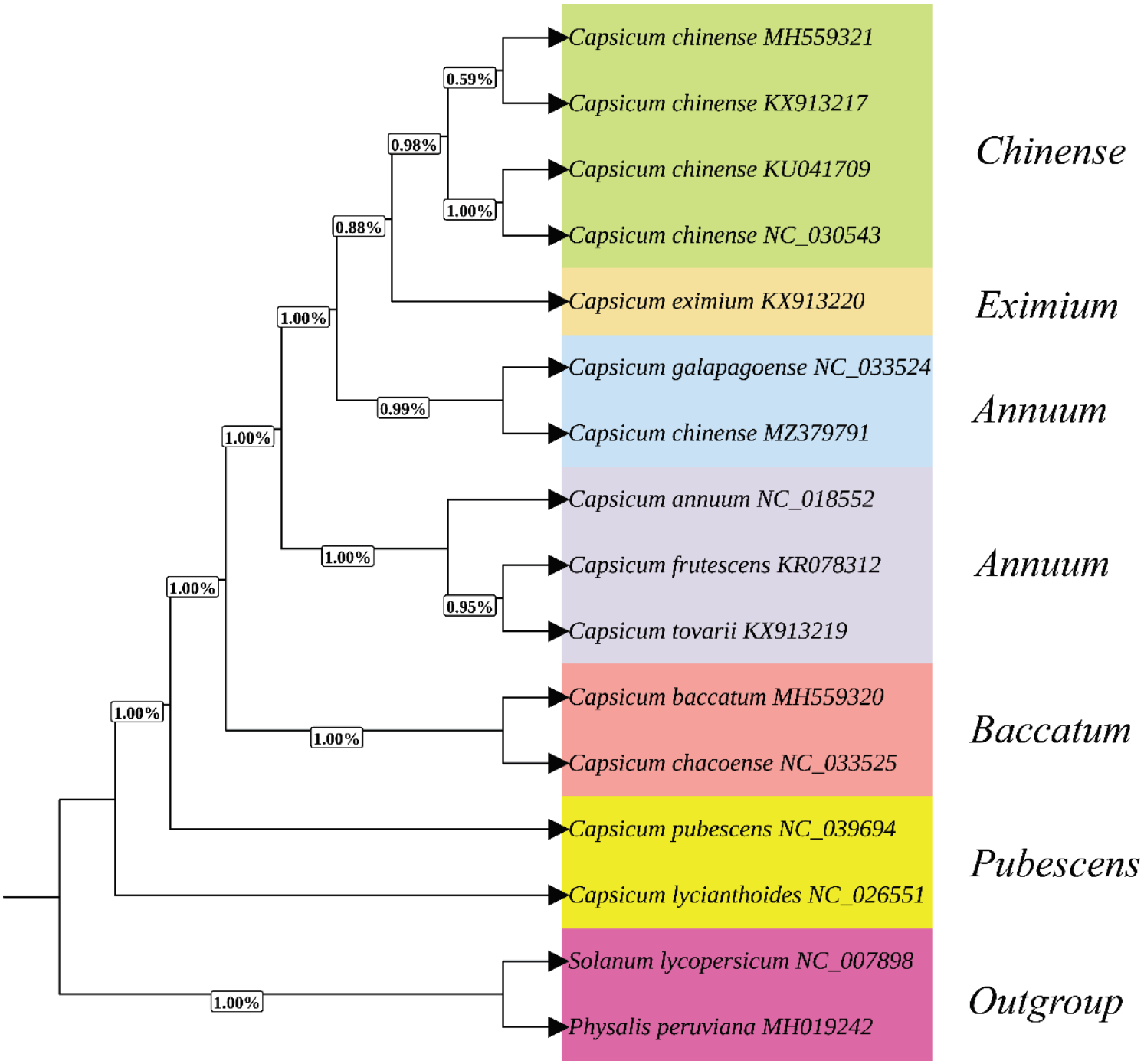

Figure 4

Phylogenetic reconstruction of the 14 chloroplast genome sequences of the Capsiceae tribe, including two genomes from an outgroup of Solanaceae. Colors highlight the clade within the *Capsicum* phylogeny.

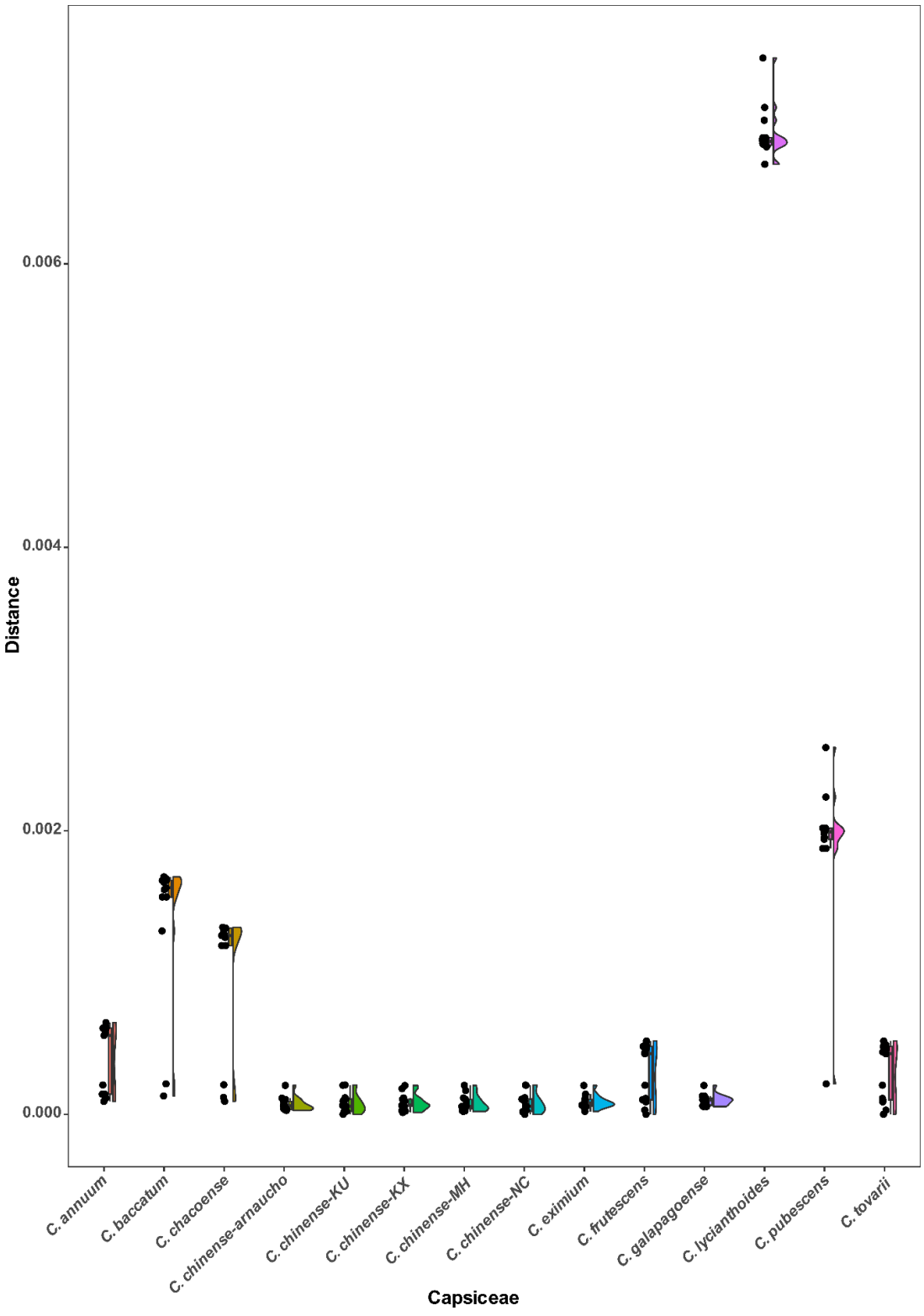

Figure 5

Estimates of Evolutionary Divergence of cp Genomes in the 14 Sequences of the Capsiceae Tribe

Analysis of repetitive sequences

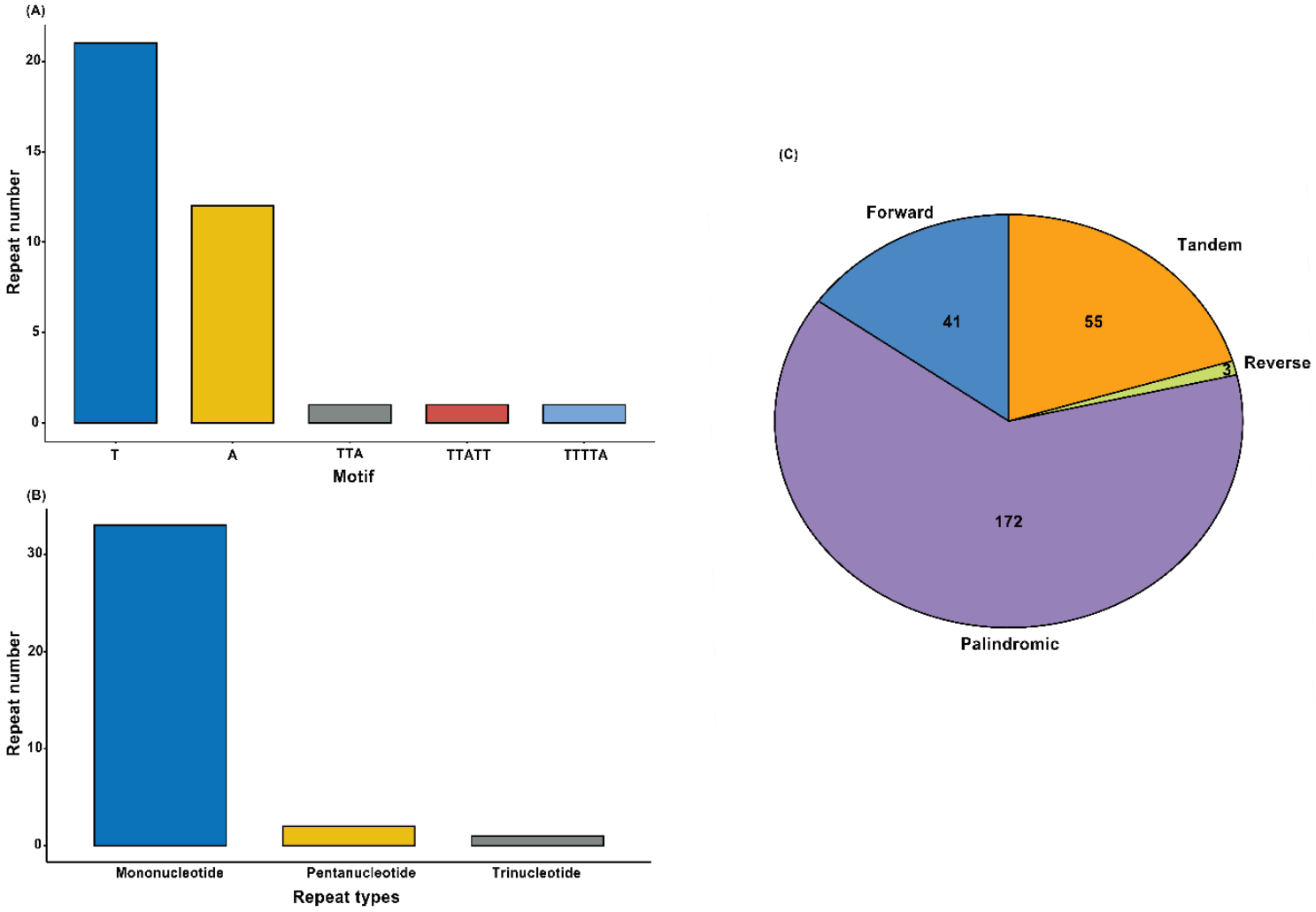

Figure 6

Frequency of simple sequence repeats (SSR) and repetitive regions in the Arnaucho chili pepper chloroplast genome. (A) The number of distinct types of SSRs. (B) The total number of SSR motifs. (C) Distribution of repetitive regions

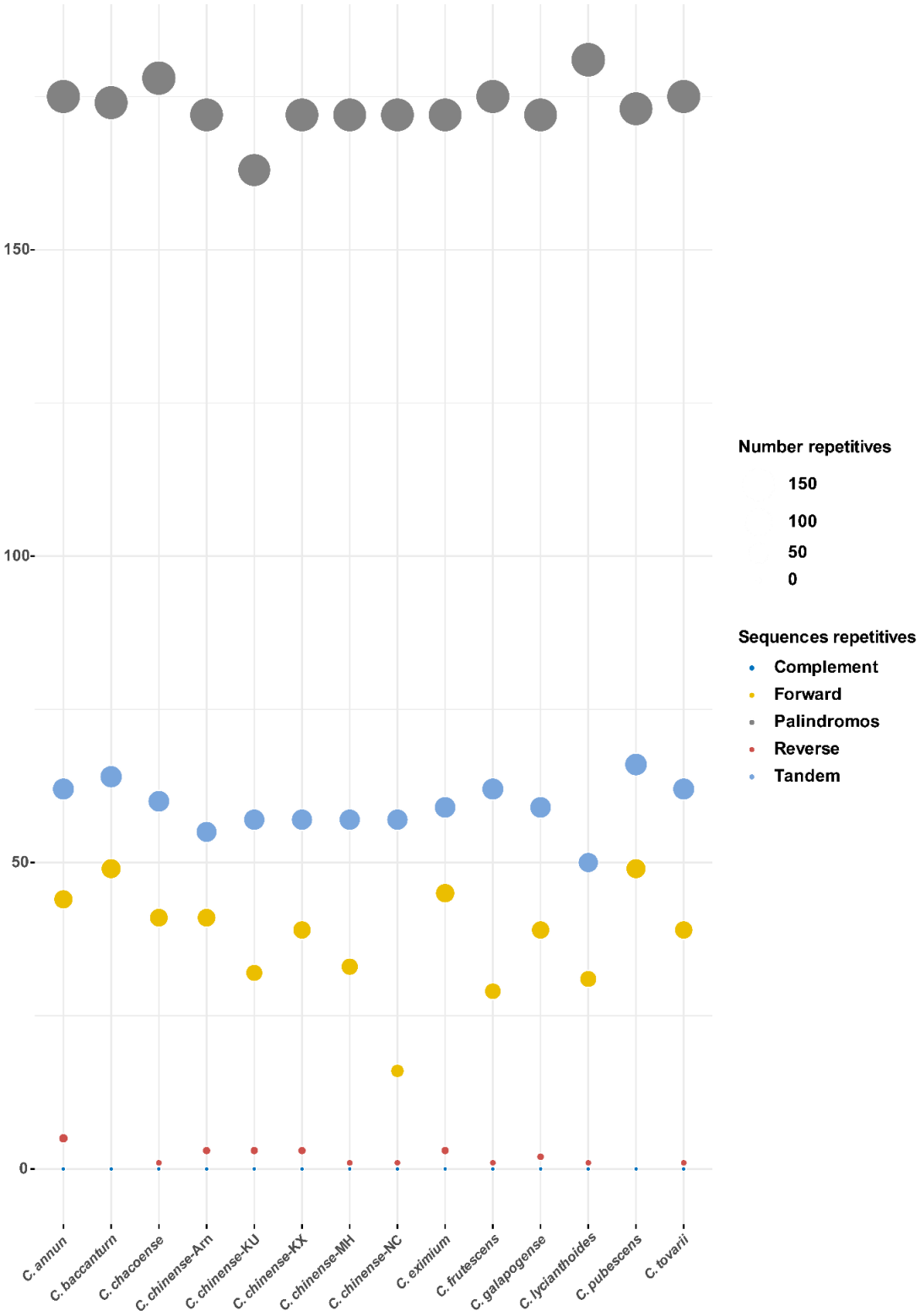

Figure 7

Comparison in the number of sequences repeated in 14 chloroplast genomes of the tribe capsiceae. The complementary repetitive sequences are represented in blue with a value of 0. The reverse repetitive sequences are represented in red with values ranging from 1 to 5, and none were found in *C. baccatum*and *C. pubescens*. The forward repetitive sequences, colored in yellow, range from 16 to 49. The tandem repetitive sequences, colored in light blue, were present in amounts of >= 50. Finally, the palindromic sequences, appearing in all Capsicum species, had a value > 150

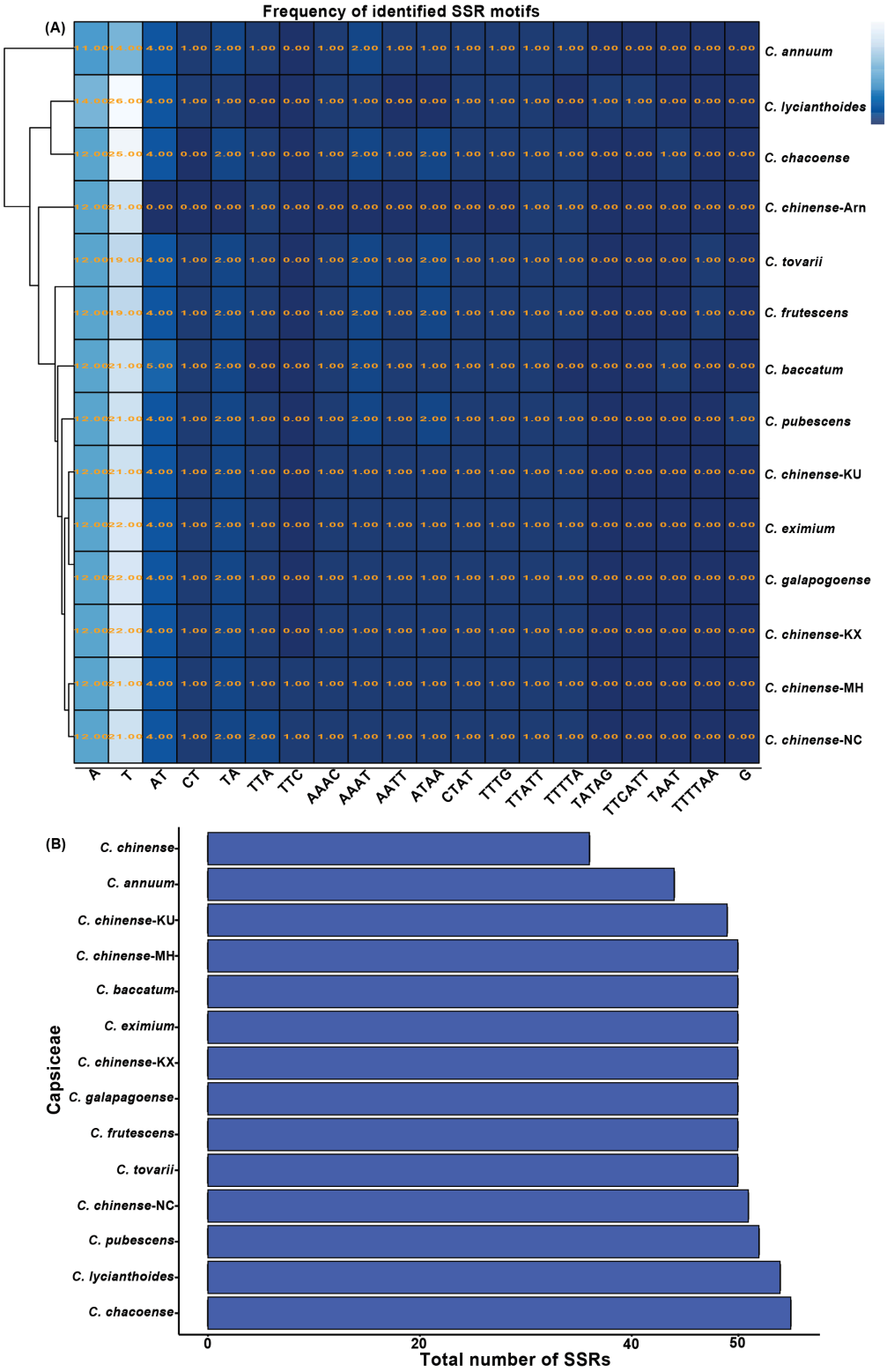

Figure 8

Comparative analysis of simple sequence repeats (SSRs). (A) Heatmap of the SSR motif frequencies identified in the Capsiceae tribe. (B) Total number of SSRs predicted in the Capsiceae tribe

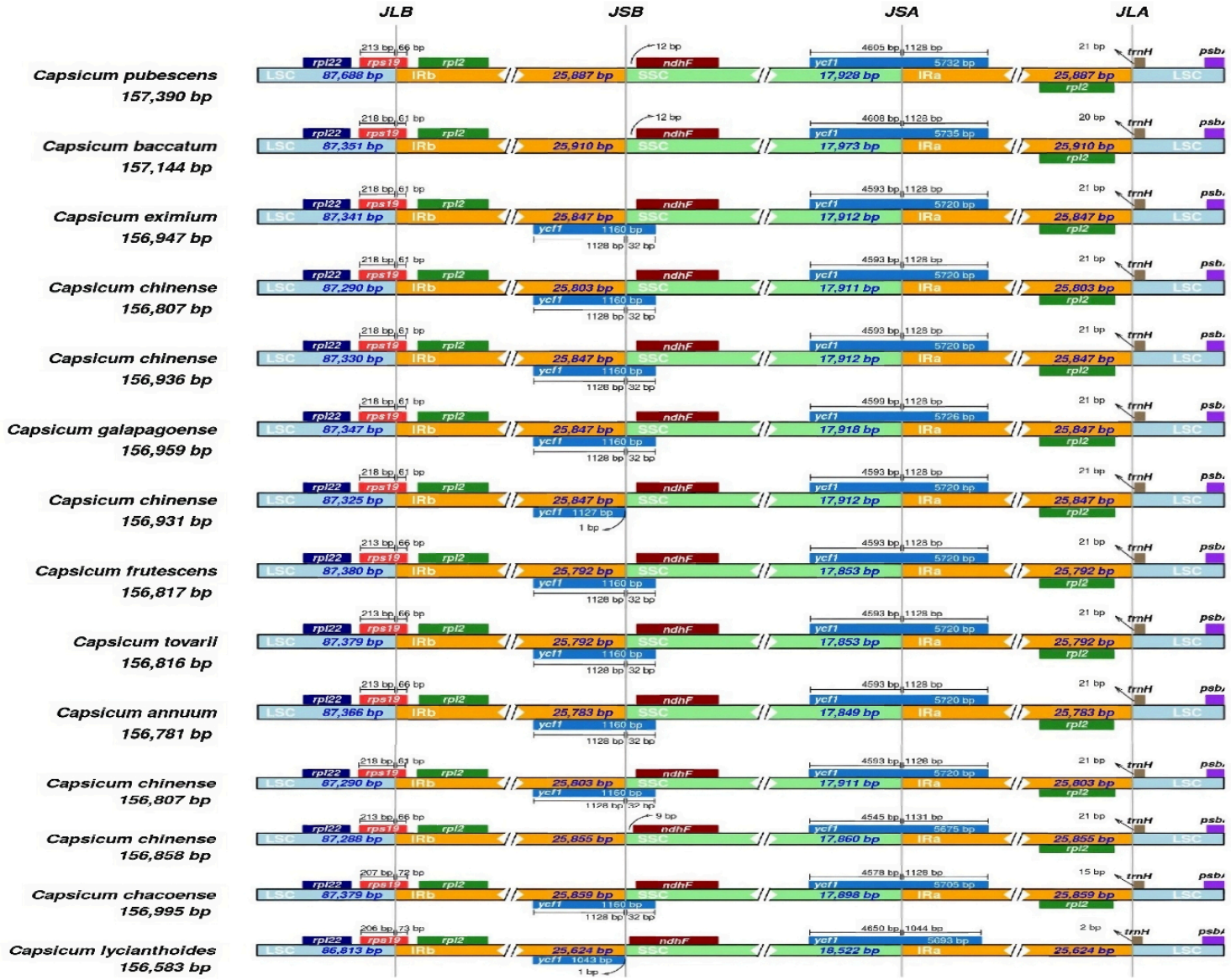

Figure 9

Comparison of boundaries between single-copy regions (LSC and SSC) and inverted repeat regions (IRB and IRA) in the genome of the Arnaucho chili pepper and 13 plastid genome sequences from the Capsiceae tribe. The genomes are represented as bars divided into each region. The boxes above and below the bars are representations of the genes. The arrows indicate the distance in base pairs (bp) from the ends of the genes to the nearest boundaries. As seen, all the boundaries present the same structure, except for the JSB boundary, where the gene *ycf1* is present in both the IRb and SSC regions in some genomes, while in others, it is only present in the IRb region, and in some cases, it is completely absent