Exploring Heliconia chloroplast genomic features through assembly and analysis of four complete chloroplast genomes

Xin Cheng
University of Chinese Academy of Sciences
Ting Yang
BGI Research
Chengcheng Shi
BGI Research
Xin Liu (liuxin@genomics.cn)
University of Chinese Academy of Sciences

Research Article

Keywords: Heliconiaceae, Heliconia, chloroplast genome, phylogeny

Posted Date: January 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3849310/v1

Additional Declarations: No competing interests reported.

Abstract

Background

In the field of *Heliconia* phylogeny, the analysis has traditionally relied on the use of partially conserved chloroplast and nuclear genes, which serve as important markers for studying coevolution. However, the lack of complete chloroplast genomes for *Heliconia* species has posed a challenge in achieving a more comprehensive understanding of *Heliconia* chloroplast genomes and developing specific molecular markers for conducting in-depth phylogenetic studies within the genus.

Results

In this study, we performed sequencing and assembly of the complete chloroplast genomes of four representative *Heliconia* species of the Zingiberales order: *Heliconia bihai*, *Heliconia caribaea*, *Heliconia orthotricha*, and *Heliconia tortuosa*. The chloroplast genomes of these *Heliconia* species exhibited the typical quadripartite structure and ranged in length from 161,680 bp to 161,913 bp, all containing 86 protein-coding genes. Comparative analysis between the *Heliconia* chloroplast genomes and those of Zingiberales species revealed a high overall similarity in chloroplast genome structure. However, we observed significant variability in the single-copy (SC) regions and noticed a high degree of A/T base preference. Additionally, there were variable amplifications in the inverted repeat (IR) regions. While no genes with high nucleotide diversity were identified, three positively selected genes in *Heliconiaceae*, including *ndhD*, *rpl2* and *ycf2*, were discovered when compared to other Zingiberales plants. Moreover, phylogenetic analysis provided strong support for the formation of a monophyletic clade consisting of *Heliconiaceae* species. This clade was nested within the tribe *Heliconiaceae* of the Zingiberales order, with high bootstrap support, reinforcing their evolutionary relatedness.

Conclusions

The results of this study have offered insights into the chloroplast genomes of *Heliconia*, and the dataset produced by our research serves as a valuable resource for subsequent studies on the *Heliconia* evolutionary trajectory.

Background

Heliconia, a genus belonging to the *Heliconiaceae* family, is a unique group of flowering plants comprising nearly 200 species [1, 2]. These plants are primarily found in tropical America and certain islands in the western Pacific [3]. The inflorescences of *Heliconia* consist of upright or drooping cone-like structures, from which emerge the vibrant and eye-catching bracts. The colorful portions that catch our attention are waxy bracts, while the slender sections within these bracts house the true flowers of the *Heliconia* plant. *Heliconiaceae* holds great significance not only in the international fresh-cut flower

market, where its exotic floral bracts are highly sought after [2], but also in the realm of studying coevolution between plants and animals [4]. A well-known previous study investigated the relationship between hummingbirds and their *Heliconia* food plants across the Lesser Antilles islands [5], which offered valuable insights into the intricate relationships between flowering plants and their avian pollinators. By delving into the intricate evolutionary processes within the *Heliconiaceae* family, we can contribute to a better understanding of the broader phenomenon of coevolution. In this aspect, previous studies on *Heliconia* species evolution can date back to the 1980s when the cladistic morphological analyses contributed to understanding on *Heliconia* infrageneric taxonomic systems [6–9]. The taxonomic and morphological aspects of the genus, as well as its ecological significance in tropical forests, have garnered considerable interest.

Genetic makers from plastid and nuclear genomes were used to analyze *Heliconia* evolution, revealing that the diversity of *Heliconia* originated in the Late Eocene (39 million years ago) and experienced rapid diversification during the Early Miocene[1]. However, studies focusing on the molecular diversity of *Heliconia* are relatively infrequent, majorly utilizing genetic makers to study species level and population-level evolution [10–15]. Previous studies utilized Amplified Fragment Length Polymorphism (AFLP) makers to study cultivated *Heliconia* species[16] and the genetic diversity of *H. bihai* populations [17]. Random Amplified Polymorphic DNA (RAPD) markers were also applied to study evolutionary relationship among *Heliconia* species, revealing the monophyletic nature of the *Heliconia* genus [15]. Furthermore, In the larger group of Zingiberales order, to which Heliconia belongs, more genetic makers or representative whole chloroplast genomes were utilized to depict the evolutionary process of Zingiberales species, indicating *Heliconia* as the sister group to the remaining families in Zingiberales [18, 19]. To attain a comprehensive understanding of the correlation between morphological and molecular diversity in *Heliconiaceae*, it is essential to gather more extensive molecular data and conduct comparative analyses with other species within the Zingiberales order. This will be instrumental in elucidating the evolutionary patterns and relationships among *Heliconia* species and its coevolutionary dynamics with hummingbirds.

In this study, we assembled the chloroplast genomes of four representative *Heliconia* species, including *Heliconia bihai* [20], *Heliconia caribaea* [21], *Heliconia orthotricha* [22], and *Heliconia tortuosa* [23] (Pic. S1). We conducted a comprehensive examination of the complete chloroplast genome structures of these species, undertaking a detailed analysis and comparison of their structural and genomic features with those of other Zingiberales species. With the chloroplast genomes, our research aims to contribute to a deeper understanding of the phylogenetic relationships within the *Heliconiaceae* family.

Materials and methods

Plant materials and DNA sequencing

Four representative *Heliconia* species, namely *Heliconia bihai*, *Heliconia caribaea*, *Heliconia orthotricha*, and *Heliconia tortuosa*, were selected for our study. *H. bihai*, *H. caribaea*, and *H. tortuosa* samples were

collected from tropical America by Kress lab, while *H. orthotricha* was obtained from the Guangdong Flower Market. Fresh leaves were carefully collected and immediately snap-frozen in liquid nitrogen. The samples were then stored at -80 ℃ until DNA extraction. DNA extraction was performed using the modified CTAB method [24]. Subsequently, the DNA samples were sequenced on BGISEQ-500 platforms (MGI, Shenzhen, China) using the whole genome strategy at BGI Research Qingdao lab, following the manufacturer instructions [25] .

Chloroplast genome assembly and annotation

The *de novo* assembly of four chloroplast genomes was performed using NOVOplasty (version 4.3.3) [26] with parameters of “Genome Range: 150,000-190,000; *K*-mer: 31; Seed Input: *Heliconia collinsiana*; Combined reads: All clean reads”. For the homology-based assembly of the chloroplast genomes, MITObim version 1.9.1 (relies on MIRA 4.0.2) (https://github.com/chrishah/MITObim) was utilized with parameters of “Read Pool: Extracted all clean reads with a depth of 20X; -quick Heliconia collinsiana” [27]. The resulting assemblies from both methods were then aligned and refined against the reference chloroplast genome of *Heliconia collinsiana* (NC_020362.1). Finally, we conducted manual curation to derive circular sequences. To visualize the chloroplast genome maps, the online program OGDRAW v1.3.1 [28] (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) was employed.

Chloroplast genome analysis and statistics

The identification of simple sequence repeats (SSRs) was performed using the online MISA-web tool [29, 30]. The minimum number of repeats was set to 10, 5, 4, 3, 3, and 3 for mononucleotide (mono-), dinucleotide (din-), trinucleotide (tri-), tetranucleotide (tetra-), pentanucleotide (penta-), and hexanucleotide (hexan-) SSRs, respectively [31]. Tandem repeat sequences were detected using Tandem Repeats Finder with default parameters [32]. The parameters used were 2, 7, and 7 for weights of match, mismatch, and indels, respectively. The detection parameters were set to 80 for the matching probability (Pm), 10 for the indel probability (Pi), a minimum alignment score of 50, and a maximum period size of 500. Long repeat sequences were analyzed using REPuter [33]. The analysis identified forward (F), reverse (R), complement (C), and palindromic (P) repeats with default parameters. The parameters used were, ‘-f’ to compute maximal forward repeats, ‘-p’ to compute maximal palindromes, ‘-h’ to search for repeats up to the given hamming distance, and ‘-l’ to specify the desired length of repeats. Codon usage was analyzed using MEGA11 [34], and the relative synonymous codon usage (RSCU) and amino acid frequencies were calculated with default settings. Additionally, the GC content of the three positions was analyzed using CUSP in the EMBOSS program [35].

Comparative analysis of the chloroplast genomes

Signals of natural selection were evaluated for all protein coding genes. The non-synonymous (*K*a) and synonymous (*K*s) substitution ratio (*K*a/*K*s) of each gene was calculated in the background of different species in Zingiberales. The protein sequences of protein coding genes in each pair of the species were aligned using MAFFT (v7.407) [36]. Subsequently, the coding DNA sequences (CDS) were converted into codon alignments based on the protein sequence alignment using the Perl script pal2nal (v14) [37]. The

KaKs calculator (v2.0) [38], utilizing its model-averaging method, was employed to compute the values for *K*a (non-synonymous substitutions), *K*s (synonymous substitutions), and the *K*a/*K*s ratio.

The pairwise alignments and sequence divergence analysis were conducted for *H. bihai*, *H. caribaea*, *H. orthotricha*, and *H. tortuosa*, along with seven additional Zingiberales species, namely *Canna indica* (MK561603), *Costus pulverulentus* (KF601573), *Musa acuminata* (NC_058940), *Orchidantha fimbriata* (KF601569.1), *Thaumatococcus daniellii* (KF601575.1), *Ravenala madagascariensis* (NC_022927.1), and *Zingiber officinale* (NC_044775). The alignments and sequence comparisons were performed using the mVISTA tool with LAGAN and Shuffle-LAGAN modes [39]. The analysis was carried out to assess the contraction and extension of the inverted repeat (IR) borders across the four major regions (LSC/IRa/SSC/IRb) in the chloroplast genome sequences of all eleven species. This assessment was carried out using the web tool IRSCOPE [40].

Phylogenetic analysis

We obtained 22 chloroplast genomes from the NCBI database, including *Oryza sativa*(NC_031333.1), *Canna indica* (MK561603.1), *Costus pulverulentus* (KF601573.1), *Heliconia acuminata* (MH603423.1), *Heliconia collinsiana* (NC_020362.1), *Heliconia meridensis* (MH603426.1), *Heliconia nutans* (MH603425.1), *Orchidantha fimbriata* (KF601569.1), *Thaumatococcus daniellii* (KF601575.1), *Ensete glaucum* (LC610748.1), *Musa acuminata* (NC_058940.1), *Musa balbisiana* (NC_028439.1), *Ravenala madagascariensis* (NC_022927.1), *Amomum compactum* (NC_036992.1), *Amomum krervanh* (NC_036935.1), *Curcuma roscoeana* (MT395652.1), *Kaempferia elegans* (NC_040852.1), *Lanxangia tsao-ko* (MK937808.1), *Roscoea schneideriana* (MZ569051.1), *Wurfbainia compacta* (MG000589.1), *Zingiber officinale* (NC_044775.1), and *Zingiber spectabile* (NC_020363.1). In addition to the seven species from *Heliconiaceae* family, we included 18 additional species and used the monocotyledonous plant rice (*Oryza sativa*) as an outgroup. To align the chloroplast genome single-copy sequences, we employed the MAFFT software [36], and we then used Gblocks (Version 0.91b, http://molevol.cmima.csic.es/castresana/Gblocks_server.html) for extracting conserved sites from the multiple sequence alignment. Subsequently, we extracted the Fourfold Degenerate Third Codon Transversion (4dtv) sites for the construction of the phylogenetic tree. Maximum likelihood (ML) analysis was performed using the RAxML program [41] with the parameter '-N 1000 -m GTRGAMMAI -f a -x 123 -p 123 -k -O -o Oryza_sativa' as the nucleotide substitution model. MEGA11 [34] was used with default parameters to construct the Neighbor-Joining evolutionary tree. To visualize the phylogenetic relationships, we utilized the iTOL online tool (https://itol.embl.de/) [42].

For the analysis of shared genes among the 26 species, we generated a high-quality alignment file using the MAFFT [36] with default parameters. These alignment files, along with the chloroplast genome sequences, were used as input files for codeml. In the initial run, the ctl file parameters were set to 'runmode = 0, CodonFreq = 2, and model = 0'. In the second run, the parameters were adjusted to 'mode = 2', focusing on the *Heliconiaceae* family as the foreground branch, allowing for the calculation of different evolutionary rates [36]. The DnaSP v5 software [43] was employed to compare the aligned sequences, calculate nucleic acid diversity, and obtain the value of π.

Results

Assemble the Heliconia chloroplast genomes

Using the generated sequencing data, we successfully assembled the chloroplast genomes of four *Heliconia* species, including *H. bihai*, *H. caribaea*, *H. orthotricha*, and *H. tortuosa* (Table S1). The chloroplast genomes of these four *Heliconia* species exhibited significant similarity. The sizes of the chloroplast genomes were as follows, 161,745 bp for *H. bihai*, 161,908 bp for *H. caribaea*, 161,689 bp for *H. orthotricha*, and 161,672 bp for *H. tortuosa*. A total of 132 genes were identified in these chloroplast genomes, comprising 86 coding sequences (CDS), 8 ribosomal RNAs (rRNAs), and 38 transfer RNAs (tRNAs) (Fig. 1a, Table 1, S2). Among these genes, 18 were identified as splitting genes in *H. bihai*, *H. orthotricha*, and *H. tortuosa*, with 16 of them containing a single intron each, while two genes (*clpP* and *ycf3*) had two introns each. Notably, *H. caribaea* had 19 splitting genes, including the unique presence of *accD*, which sets it apart from other species in terms of splitting genes (Table S3). The chloroplast genomes of these *Heliconia* species displayed a quadripartite structure, similar to the majority of angiosperms. This structure consisted of a large single-copy (LSC) region (89,772 bp for *H. bihai*, 89,861 bp for *H. caribaea*, 89,734 bp for *H. orthotricha*, and 89,775 bp for *H. tortuosa*), a small single-copy (SSC) region (18,757 bp for *H. bihai*, 18,779 bp for *H. caribaea*, 18,704 bp for *H. orthotricha*, and 18,656 bp for *H. tortuosa*), and two inverted repeat (IR) regions (26,608 bp for *H. bihai*, 26,634 bp for *H. caribaea*, 26,617 bp for *H. orthotricha*, and 26,629 bp for *H. tortuosa*) (Fig. S1). The GC content in the LSC, SSC, and IR regions of all four chloroplast genomes was 35.4%, 31.3%, and 42.8%, respectively (Table 1), reflecting a notable bias toward the usage of A/T bases in the *Heliconia* chloroplast genome.

Table 1
General characteristics of four *Heliconia* chloroplast genomes.

Charateristics and parameters	*Heliconia bihai*	*Heliconia caribaea*	*Heliconia orthotricha*	*Heliconia tuotorsa*
Total cp genome size (bp)	161,745	161,908	161,689	161,672
LSC length (bp)	89,772	89,861	89775	89,734
SSC length (bp)	18,757	18,779	18656	18,704
IR length (bp)	26,608	26,634	26629	26,617
Total number of genes	132	132	132	132
CDS genes	86	86	86	86
rRNAs genes	8	8	8	8
tRNAs genes	38	38	38	38
Total GC content (%)	37.36	37.34	37.36	37.36
GC content for LSC (%)	35.39	35.36	35.39	35.38
GC content for SSC (%)	31.29	31.27	31.29	31.34
GC content for IR (%)	42.82	42.83	42.82	42.82
Coding GC (%)	38.17	38.17	38.18	38.13
1st letter GC (%)	45.74	45.75	45.82	45.68
2nd letter GC (%)	38.41	38.41	38.47	38.40
3rd letter GC (%)	30.37	30.35	30.24	30.30

Heliconia chloroplast repeat sequence features

Much like the role of mitochondrial genomes in vertebrate genetics, chloroplast genomes serve as a common tool for resolving phylogenetic and evolutionary debates [44]. Three main types of repeat sequences were found in organelle genomes, including simple sequence repeats (SSRs) [45], tandem repeats (TRs), and dispersed repeats (DRs). Among these, SSRs exhibited high variability within a species, making them valuable markers for population genetics and phylogenetic analyses [46]. In the case of *Heliconia* chloroplasts, we observed similarities but not complete consistency in repeat sequences. Focusing on SSRs, we found minimal variation in their numbers among the four *Heliconia* genomes, with 73 in *H. bihai* and *H. caribaea*, 71 in *H. tortuosa*, and 68 in *H. orthotricha*. Despite the similarity in the number of encoded genes, notable differences in SSR types were observed as well. Specifically, *H. bihai* and *H. caribaea* featured monomeric, dinucleotide, trinucleotide, tetranucleotide, and pentanucleotide SSR types, while *H. tortuosa* and *H. orthotricha* additionally included the hexanucleotide

SSR type in the SSC region (Table S4). Most SSRs were concentrated in the LSC regions, with only one SSR located within coding genes across all four *Heliconia* species. By comparing the chloroplast genome data of other species within the Zingiberales order that have been sequenced to date (Fig. 1b), we found that the presence of both ACT and AATC types of SSRs in the genome could potentially serve as an indicator for classifying a species as belonging to the *Heliconia* genus (Table S5). Shifting our focus to tandem repeats (TRs), our detailed analysis revealed that most repeat units were predominantly composed of A or T, with the longest repeat sequence spanning approximately 120 base pairs (Table S6). Transitioning to dispersed repeats (DRs), *H. bihai* exhibited two types (forward repeat and reverse repeat,), while *H. caribaea* and *H. tortuosa* showed three types (forward repeat, reverse repeat, and palindromic repeat). *H. orthotricha*, on the other hand, possessed all four types of dispersed repeats (forward repeat, reverse repeat, complemented repeat and palindromic repeat) but had a comparatively lower quantity of DRs. Overall, we identified repeat features in the *Heliconia* chloroplasts, which might be used as genetic markers for distinguishing *Heliconia* species among themselves and from other species.

Heliconia chloroplast codon usage features

Codon usage bias refers to the uneven utilization of different codons that encode the same amino acid within a genome. Research on codon usage bias contributes to our understanding of genome evolution, gene expression regulation, and the adaptability of organisms to environmental changes [47]. In our analysis of the 86 CDS in chloroplast genomes, we computed the frequency of codon usage and relative synonymous codon usage (RSCU) (Fig. 1c and Table S7). The CDS in these chloroplast genomes encode 20 amino acids using 64 codons, including the termination codon. Among these 64 codons, 30 of them exhibit an RSCU value greater than 1, with 29 of them ending with an A or T bases. This observation indicates a preference for A or T endings in the codons of the *Heliconia* chloroplast genomes, which is consistent with the previously mentioned decrease in GC content at the third position of codons (30.3%) compared to the first (45.7%) and second (37.4%) positions. Regarding the codon usage bias among the four chloroplast genomes, there are six codons each for arginine (Arg), leucine (Leu), and serine (Ser), while only one codon each is present for methionine (Met) and tryptophan (Trp). Within the spectrum of amino acids, Isoleucine (Ile) stands out as the most frequently occurring amino acid, predominantly encoded by the ATT codon with a frequency of 41%. Conversely, cysteine (Cys) is the least common amino acid, with the TGC codon having the lowest frequency at 3%, across four chloroplast genomes. Except for methionine (Met) and tryptophan (Trp), nearly all amino acids are encoded by 2–6 synonymous codons. Our analysis reveals a codon usage bias favoring A or T endings in the codons of *Heliconia* chloroplast genomes, highlighting the significance of studying codon usage patterns in understanding genome evolution and gene expression regulation.

Positively selected genes in Heliconia chloroplast genomes

In addition to repeat features, we also investigated the chloroplast genes to reveal possible genes associated with the visual diversity of *Heliconia* and its thriving presence in tropical forest ecosystems. Our selective pressure analysis reveals insights into the chloroplast genes under selection and nucleotide

diversity within specific genes in the chloroplast genomes. During the positive selection analysis of the genes used in constructing the phylogenetic tree, we observed that *Heliconia*, as a foreground branch, did not undergo significant positive selection. However, within the *Heliconiaceae* family, three genes (*ndhD*, *rpl2*, and *ycf2*) showed a trend of positive selection (*K*a/*K*s > 1) (Table S8). This suggests their potential biological roles in the evolution of these plants. Additionally, focusing on protein-coding genes, we analyzed nucleotide diversity in a total of 12 species from the Zingiberales order (Table S9). Among these genes, the *ndhD* gene exhibited notably high nucleotide diversity, with a PAI (per-site average information) value exceeding 0.2. Several other genes, including *ccsA, cemA, infA, matK, ndhD, rpl, rpo, rps* also displayed PAI values greater than 0.05. However, among the *Heliconia* species, we did not observe coding genes with high nucleotide diversity (Table S10). By considering selective pressures and nucleotide diversity in specific chloroplast genes, our study provides valuable insights into the genetic dynamics and adaptive processes within *Heliconia* and related species.

Structural conservation and variations in Heliconia chloroplast genomes

In our comparative analysis of *Heliconia* chloroplast genomes alongside five closely related species (*Canna indica, Costus pulverulentus*, *Musa acuminata, Ravenala madagascariensis, Zingiber officinale*), we observed a significant degree of structural conservation in the overall chloroplast architecture. Specific structural variations were identified at distinct boundaries, including LSC/IRb, IRb/SSC, SSC/IRa, and IRa/LSC (Fig. 2). These boundary regions in the four *Heliconia* species remained consistent yet exhibited unique features, setting them apart from other plants in Zingiberale. Noteworthy is the absence of the *rps19* gene in the IR region of *Heliconia* chloroplasts, distinguishing it from other Zingiberales plants where the IR region includes the *rps19* gene. Furthermore, an elongated separation of approximately 150 base pairs at the boundary between the inverted repeat B (IRb) and the small single-copy region (SSC) in the *Heliconia* chloroplast genomes for the *ndhF* gene was noted. This contrasts with other species, where the distance typically falls within the range of approximately 10 to 60 base pairs.

To further explore functional sequence variations within highly conserved and maternally inherited chloroplast genomes, which can serve as valuable genetic markers for species differentiation [48], we conducted a comparative analysis of *Heliconia* chloroplast genomes to those of related species. We compared the chloroplast genomes of *Canna indica, Costus pulverulentus, Musa acuminata, Orchidantha fimbriata, Thaumatococcus daniellii, Ravenala madagascariensis*, and *Zingiber officinale* from the Zingiberales order with the reference *H. bihai* chloroplast genome (Fig. 3). The result illustrates that the *Heliconia* chloroplast genomes exhibits no significant differences in the exonic regions. However, in the conserved noncoding sequences (CNS) region, notable genetic diversity and variation are evident, primarily occurring in the LSC and SSC regions. The contraction of the inverted repeat (IR) region results in a slightly smaller chloroplast genome size in *Heliconia* compared to other species within the Zingiberales order. Our analysis of *Heliconia* chloroplast genomes and related species reveals functional sequence variations in CNS, serving as genetic markers for species differentiation.

Phylogeny of Heliconia species revealed by chloroplast genomes

The chloroplast genomes, with their maternal inheritance and relatively low mutation rate, are widely used to investigate phylogenetic relationships among green plants [44]. Analyzing the entire chloroplast genome yields more reliable results, providing substantial insights into the genetic evolution of plant species. In our study, we carefully selected 26 diverse plant species, representing major clades of Zingiberales plants (Fig. 4), and including representatives from different families such as *Cannaceae, Costaceae*, and *Heliconiaceae*, *Musaceae, Marantaceae*, Lowiaceae, along with *Strelitziaceae* and *Zingiberaceae*. Fourfold Degenerate Third Codon Transversion (4dtv) site mutations from single-copy genes were employed in constructing evolutionary trees using two different methods. In the maximum likelihood (ML) tree, Zingiberales diverge from three distinct terminal nodes. *Heliconiaceae* plants formed a distinct branch, and emerge as the sister clade to *Musaceae*, *Strelitziaceae*, and *Lowiaceae.* While in the Neighbor-Joining tree, Zingiberales diverge from two distinct terminal nodes. *Musaceae* emerged as sister branches to *Heliconiaceae* and *Strelitziaceae*, forming a distinct clade (Fig. S2). Due to incomplete genomic data for certain species, additional data support is required for the construction of the phylogenetic tree of the chloroplast genomes of plants in the Zingiberales. Our results highlight the intricate relationships within the Zingiberales order and underscore the potential influence of the chosen methodology on the inferred evolutionary relationships among the studied species.

Discussion

The high conservation of chloroplast genomes in terrestrial plants encompasses their structure, length, and gene content. In our study, we successfully assembled complete chloroplast genomes of *Heliconia* plants, closely resembling the reported structure of *Heliconia collinsiana*. However, the analysis of repetitive sequences, specifically SSRs, revealed distinguishable patterns not only among different *Heliconia* species but also across genera within the Zingiberales order. Furthermore, the low GC content observed in both codon positions and repetitive sequences suggests a strong preference for A/T bases in *Heliconia*. Compared to other Zingiberales species, the chloroplast genome of *Heliconia* is slightly shorter, attributed to a reduction in the length of the IR region and an expansion at the SSC region boundary. Leveraging the assembled complete chloroplast genomes of *Heliconia*, we conducted phylogenetic analyses to determine the relationships among closely related species such as *Musa acuminata*.

Our study highlighted specific genes within the chloroplast genome of *Heliconia* that exhibit notable nucleotide diversity, possibly playing crucial roles in chloroplast functionality. Genes such as *ccsA, cemA, infA, matK, ndhD, rpl, rpo, rps* encode proteins involved in various biological processes. For instance, *ccsA* encodes a crucial component in the synthesis of cytochrome c within the chloroplast [49]. *cemA* encodes a subunit of chloroplast ATP synthase involved in energy production during photosynthesis [50], whereas *infA* encodes a protein crucial for tRNA processing, contributing to chloroplast protein synthesis [51]. Gene *matK* encodes a splicing enzyme that facilitates RNA splicing [52], while *ndhD* is part of the chloroplast NADH dehydrogenase-like (NDH) complex, which plays a crucial role in photosynthesis, particularly in electron transport interactions with photosystem I (PSI) [53, 54]. Additionally, ribosomal proteins encoded by *rps18*, *rps3*, *rpl22*, and *rps15* are directly engaged in protein synthesis [55, 56]. These genes are vital for plant growth, development, and metabolic processes, supporting chloroplast structure

and function [57]. While the current chloroplast genome offers valuable genetic resources for understanding the diverse appearances of *Heliconia* species, future research focusing on complete nuclear genomes will significantly enhance our understanding and applications related to *Heliconia*.

Abbreviations

IR Inverted repeat regions

LSC Large single-copy region

SSC Small single-copy region

rRNAs Ribosomal RNAs

tRNAs Transfer RNAs

PE Paired-end

BI Bayesian inference

ML Maximum likelihood

CDS Protein-coding genes

JLB Junction between LSC and IRb

JSB Junction between SSC and IRb

JSA Junction between SSC and IRa

JLA Junction between LSC and IRa

CNS Conserved non-coding sequence

Declarations

Ethics approval and consent to participate

We confirm that the collection of plant material and experimental research followed all local and national guidelines and legislation.

Consent for publication

All necessary consents for publication have been obtained.

Availability of data and materials

The complete chloroplast genomes generated during the current study were deposited in NCBI database(PP093761, PP093760, PP093759, PP093762). The other accession numbers for the remaining datasets analyzed in this study are listed in the Table S11.

Competing interests

Funding

This work was supported by the National Key Research and Development Program of China, grant number 2021YFD2200502.

Authors' contributions

X.C., T.Y. and C.S. performed the data analysis. X.C. and X.L. wrote the manuscript.

Acknowledgements

We sincerely thank Dr. John Kress from Academic Department of National Museum of Natural History for help in providing samples.

References

1. Iles WJD, Sass C, Lagomarsino L, Benson-Martin G, Driscoll H, Specht CD. The phylogeny of Heliconia (Heliconiaceae) and the evolution of floral presentation. Mol Phylogenet Evol 2017, 117:150–167U https://linkinghub.elsevier.com/retrieve/pii/S1055790316303906.
2. Linares A, Gallardo-López F, Villarreal M, Landeros-Sánchez C, López-Romero G. Global vision of heliconias research as a cut flowers: a review. Ornam Hortic. 2020;26:633–46.
3. Heliconia L. [https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:331205–2.
4. Altshuler DL, Clark CJ. Darwin's Hummingbirds. Science. 2003;300(5619):588–589U. https://www.science.org/doi/510.1126/science.1084477.
5. Temeles EJ, Kress WJ. Adaptation in a Plant-Hummingbird Association. Science. 2003;300(5619):630–633U. https://www.science.org/doi/610.1126/science.1080003.
6. Lennart. Revision of Heliconia subgen. Taeniostrobus and subgen. Heliconia (Musaceae-Heliconioideae). Opera Bot. 1992;11:5–98.
7. Kress WJ. Systematics of Central American Heliconia (Heliconiaceae) with pendant inflorescences. J Arnold Arboretum. 1984;65:429–532.
8. Kress WJ, Betancur B, Echeverry. Heliconias: llamaradas de la selva colombiana. Cristina Uribe Ediciones; 1999.
9. Kress WJ, Prince LM, Hahn WJ, Zimmer EA. Unraveling the Evolutionary Radiation of the Families of the Zingiberales Using Morphological and Molecular Evidence. *Systematic Biology* 2001, 50(6):926–944%L 921%U http://academic.oup.com/sysbio/article/950/926/926/1628904.

10. Côrtes MC, Uriarte M, Lemes MR, Gribel R, John Kress W, Smouse PE, Bruna EM. Low plant density enhances gene dispersal in the Amazonian understory herb < i > Heliconia acuminata. Mol Ecol. 2013;22(22):5716–5729L. https://onlinelibrary.wiley.com/doi/5710.1111/mec.12495. 5711%U.
11. Stein K, Rosche C, Hirsch H, Kindermann A, Köhler J, Hensen I. The influence of forest fragmentation on clonal diversity and genetic structure in < i > Heliconia angusta, an endemic understorey herb of the Brazilian Atlantic rain forest. *Journal of Tropical Ecology* 2014, 30(3):199–208%L 194%U https://www.cambridge.org/core/product/identifier/S0266467414000030/type/journal_article.
12. Suárez-Montes P, Fornoni J, Núñez-Farfán J: Conservation Genetics of the Endemic Mexican Heliconia uxpanapensis in the Los Tuxtlas Tropical Rain Forest: Conservation Genetics in Heliconia. 2011, 43(1):114–121%L 113%U https://onlinelibrary.wiley.com/doi/110.1111/j.1744-7429.2010.00657.x.
13. Westerband AC, Horvitz CC, 1293%U. https://bsapubs.onlinelibrary.wiley.com/doi/1210.3732/ajb.1500041.
14. Isaza L, Marulanda ML, López AM. Genetic diversity and molecular characterization of several Heliconia species in Colombia. Genet Mol Res. 2012;11(4):4552–4563U. http://www.funpecrp.com.br/gmr/year2012/vol4511-4554/pdf/gmr1973.pdf.
15. Marouelli LP, Inglis PW, Ferreira MA, Buso GSC. Genetic relationships among Heliconia (Heliconiaceae) species based on RAPD markers. Genet Mol Res. 2010;9(3):1377–1387U. http://www.funpecrp.com.br/gmr/year2010/vol1379-1373/pdf/gmr1847.pdf.
16. Isaza L, Marulanda ML, López AM. Genetic diversity and molecular characterization of several Heliconia species in Colombia. Genet Mol Res. 2012;11(4):4552–63.
17. Martén-Rodríguez S, John Kress W, Temeles EJ, Meléndez-Ackerman E. Plant–pollinator interactions and floral convergence in two species of Heliconia from the Caribbean Islands. Oecologia. 2011;167(4):1075–1083U. http://link.springer.com/1010.1007/s00442-00011-02043-00448.
18. Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW. Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* 2013, 29(1):65–87%L 61%U https://onlinelibrary.wiley.com/doi/10.1111/j.1096-0031.2012.00418.x.
19. Barrett CF, Specht CD, Leebens-Mack J, Stevenson DW, Zomlefer WB, Davis JI, 112%U. https://academic.oup.com/aob/article-lookup/doi/110.1093/aob/mct1264.
20. Heliconia bihai. [https://www.monaconatureencyclopedia.com/heliconia-bihai/?lang=en].
21. Heliconia caribaea purpurea. [https://www.shaileshnursery.com/portfolio-items/heliconia-caribaea-purpurea/].
22. Heliconia orthotricha. [https://www.monaconatureencyclopedia.com/heliconia-orthotricha/?lang=en].
23. Heliconia tortuosa. [https://commons.wikimedia.org/wiki/Category:Heliconia_tortuosa].
24. Aboul-Maaty NA-F, Oraby HA-S. Extraction of high-quality genomic DNA from different plant orders applying a modified CTAB-based method. Bull Natl Res Centre. 2019;43(1):25.

25. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, Qu S, Mei X, Chen H, Yu T, et al. A reference human genome dataset of the BGISEQ-500 sequencer. Gigascience. 2017;6(5):1–9.
26. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: *de novo assembly of organelle genomes from whole genome data. Nucleic Acids Research 2016:gkw955%L 952%U* https://academic.oup.com/nar/article-lookup/doi/910.1093/nar/gkw*1955.*
27. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res. 2013;41(13):e129–9.
28. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res 2019, 47(W1):W59-W64%L 52%U https://academic.oup.com/nar/article/47/W51/W59/5428289.
29. Thiel T, Michalek W, Varshney R, Graner A: Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). 2003, 106(3):411–422%L 411%U http://link.springer.com/410.1007/s00122-00002-01031-00120.
30. Beier S, Thiel T, Münch T, Scholz U, Mascher M, 2583%U. https://academic.oup.com/bioinformatics/article/2533/2516/2583/3111841.
31. Martin G, Baurens F-C, Cardi C, Aury J-M, D'Hont A. The Complete Chloroplast Genome of Banana (Musa acuminata, Zingiberales): Insight into Plastid Monocotyledon Evolution. PLoS ONE 2013, 8(6):e67350%L 67353%U https://dx.plos.org/67310.61371/journal.pone.0067350.
32. Benson G, 572%U. https://academic.oup.com/nar/article-lookup/doi/510.1093/nar/1027.1092.1573.
33. Kurtz S, 4632%U. https://academic.oup.com/nar/article-lookup/doi/4610.1093/nar/4629.4622.4633.
34. Kumar S, Nei M, Dudley J, Tamura K, 292%U. https://academic.oup.com/bib/article-lookup/doi/210.1093/bib/bbn1017.
35. Rice P, Longden I, Bleasby A, 271%U. https://linkinghub.elsevier.com/retrieve/pii/S0168952500020242.
36. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30(4):772–80.
37. Suyama M, Harrington E, Bork P, Torrents D. Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. PLoS Comput Biol. 2006;2(6):e76.
38. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteom Bioinf. 2010;8(1):77–80.
39. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. Glocal alignment: finding rearrangements during alignment. Bioinformatics 2003, 19(suppl_1):i54-i62%L 53%U https://academic.oup.com/bioinformatics/article/19/suppl_51/i54/227687.

40. Amiryousefi A, Hyvönen J, Poczai P, 3033%U. https://academic.oup.com/bioinformatics/article/3034/3017/3030/4961430.
41. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
42. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021, 49(W1):W293-W296%L 292%U https://academic.oup.com/nar/article/249/W291/W293/6246398.
43. Librado P, Rozas J, 1453%U. https://academic.oup.com/bioinformatics/article/1425/1411/1451/332507.
44. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17(1):134.
45. Song S-L, Lim P-E, Phang S-M, Lee W-W, Hong DD, Prathep A. Development of chloroplast simple sequence repeats (cpSSRs) for the intraspecific study of Gracilaria tenuistipitata (Gracilariales, Rhodophyta) from different populations. BMC Res Notes. 2014;7(1):77.
46. Fan H, Chu J-Y. A Brief Review of Short Tandem Repeat Mutation. Genom Proteom Bioinform. 2007;5(1):7–14U. https://linkinghub.elsevier.com/retrieve/pii/S1672022907600096.
47. Parvathy ST, Udayasuriyan V, Bhadana V, 534%U. https://link.springer.com/510.1007/s11033-11021-06749-11034.
48. Wysocki WP, Clark LG, Attigala L, Ruiz-Sanchez E, Duvall MR. Evolution of the bamboos (Bambusoideae; Poaceae): a full plastome phylogenomic analysis. BMC Evol Biol. 2015;15(1):50U. 10.1186/s12862-12015-10321-12865. https://bmcevolbiol.biomedcentral.com/articles/.
49. Xie Z, Merchant S, 4632%U. https://linkinghub.elsevier.com/retrieve/pii/S0021925818824929.
50. Sonoda M, Katoh H, Katoh A, Ohkawa H, Vermaas W, Ogawa T. Structure and Function of Cema Homologue (PXCA) in Cyanobacteria. In: *The Chloroplast: From Molecular Biology to Biotechnology.* Edited by Argyroudi-Akoyunoglou JH, Senger H. Dordrecht: Springer Netherlands; 1999: 149–154.
51. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, et al. Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. Plant Cell. 2001;13(3):645–58.
52. Barthet MM, Hilu KW: Expression of < i > matK: functional and evolutionary implications. 2007, 94(8):1402–1412%L 1403%U https://bsapubs.onlinelibrary.wiley.com/doi/1410.3732/ajb.1494.1408.1402.
53. Peng L, Yamamoto H, Shikanai T. Structure and biogenesis of the chloroplast NAD(P)H dehydrogenase complex. Biochim et Biophys Acta (BBA) - Bioenergetics. 2011;1807(8):945–953U. https://linkinghub.elsevier.com/retrieve/pii/S0005272810007231.
54. Shen L, Tang K, Wang W, Wang C, Wu H, Mao Z, An S, Chang S, Kuang T, Shen J-R, et al. Architecture of the chloroplast PSI–NDH supercomplex in Hordeum vulgare. Nature. 2022;601(7894):649–654L. https://www.nature.com/articles/s41586-41021-04277-41586. 641%U.

55. Asakura Y, Barkan A, 1651%U. https://academic.oup.com/plphys/article/1142/1654/1656/6106473.
56. Ostheimer GJ. Group II intron splicing factors derived by diversification of an ancient RNA-binding domain. *The EMBO Journal* 2003, 22(15):3919–3929%L 3911%U http://emboj.embopress.org/cgi/doi/3910.1093/emboj/cdg3372.
57. Shikanai T, Shimizu K, Ueda K, Nishimura Y, Kuroiwa T, Hashimoto T. The Chloroplast clpP Gene, Encoding a Proteolytic Subunit of ATP-Dependent Protease, is Indispensable for Chloroplast Development in Tobacco. *Plant and Cell Physiology* 2001, 42(3):264–273%L 262%U http://academic.oup.com/pcp/article/242/263/264/1859197/The-Chloroplast-clpP-Gene-Encoding-a-Proteolytic.

Figures

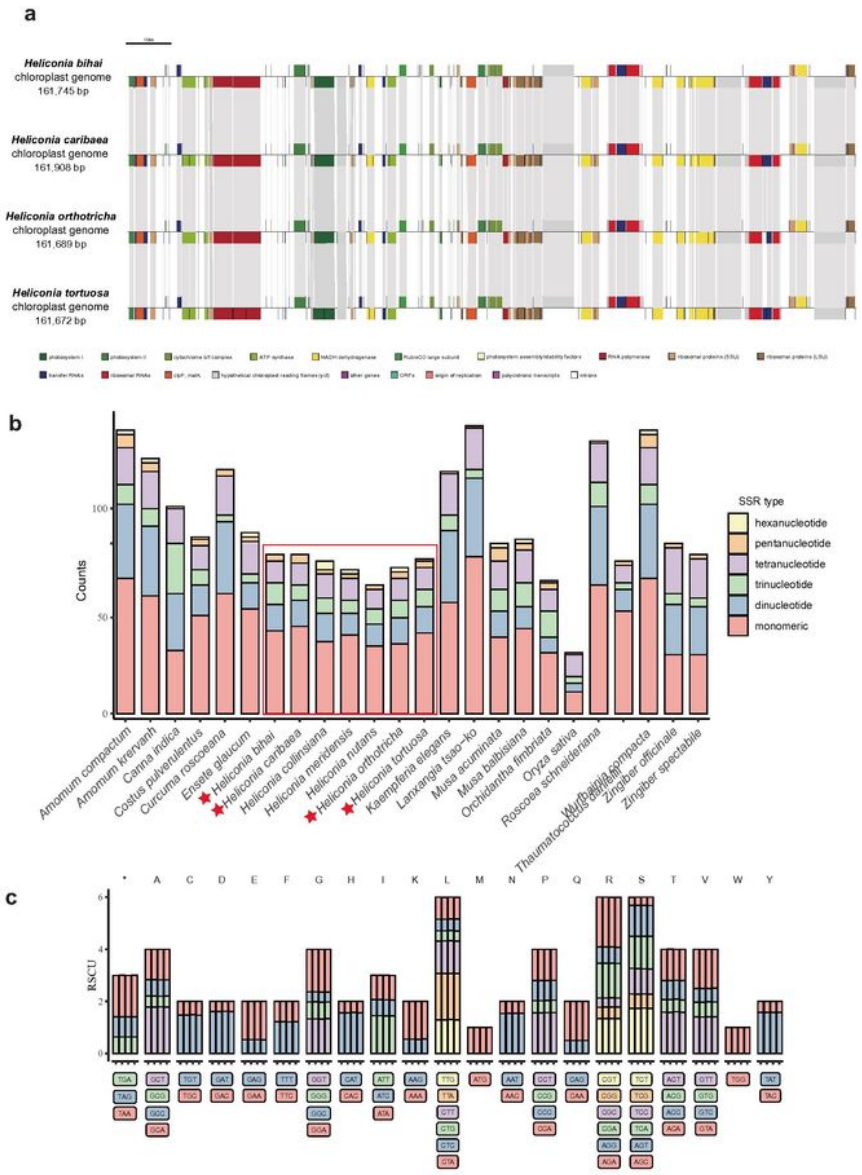

Figure 1 | Chloroplast genomes of four *Heliconia* species. a. Genetic features. Genes were shown along the four chloroplast genomes shown in linear forms. **b**. Simple sequence repeats (SSRs) compositions. Types and numbers of SSRs in the four chloroplast genomes comparing to those of related species. **c**. Codon usage patterns. Codon usage patterns of the four chloroplast genomes are shown in order.

Figure 1

See image above for figure legend.

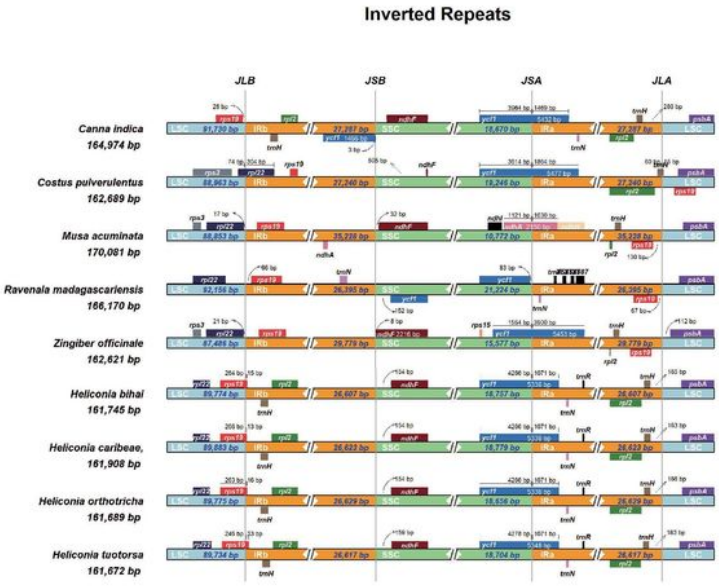

Figure 2 | Structural variations in the chloroplast genomes. Chloroplast genomes of eight Zingiberales species are compared to indicate the major chloroplast genome regions including LSC, SSC and IR regions. Genes transcribed forward are shown above the lines, whereas genes transcribed reversely are shown below the lines. Gene lengths in the corresponding regions are displayed above the boxes of gene names. JLB (LSC/IRb), JSB (IRb/SSC), JSA (SSC/IRa) and JLA (Ira/LSC) denoted the junction sites between each corresponding two regions.

Figure 2

See image above for figure legend.

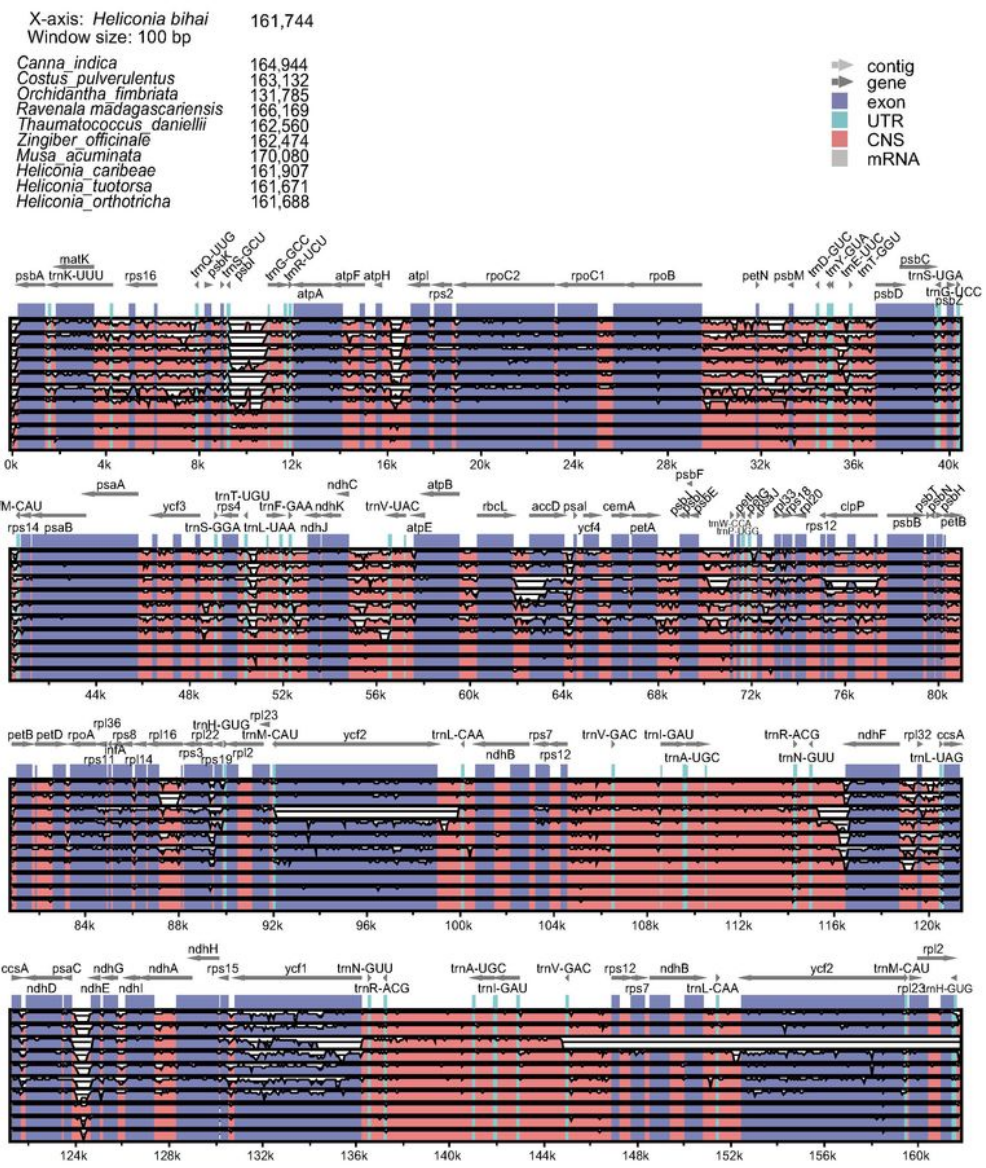

Figure 3 | Comparing the four *Heliconia* chloroplast genomes to these of the other Zingiberales species. Chloroplast genomes are shown with genes indicated, and the vertical scale indicates the percentage of identity, ranging from 50% to 100%.

Figure 3

See image above for figure legend.

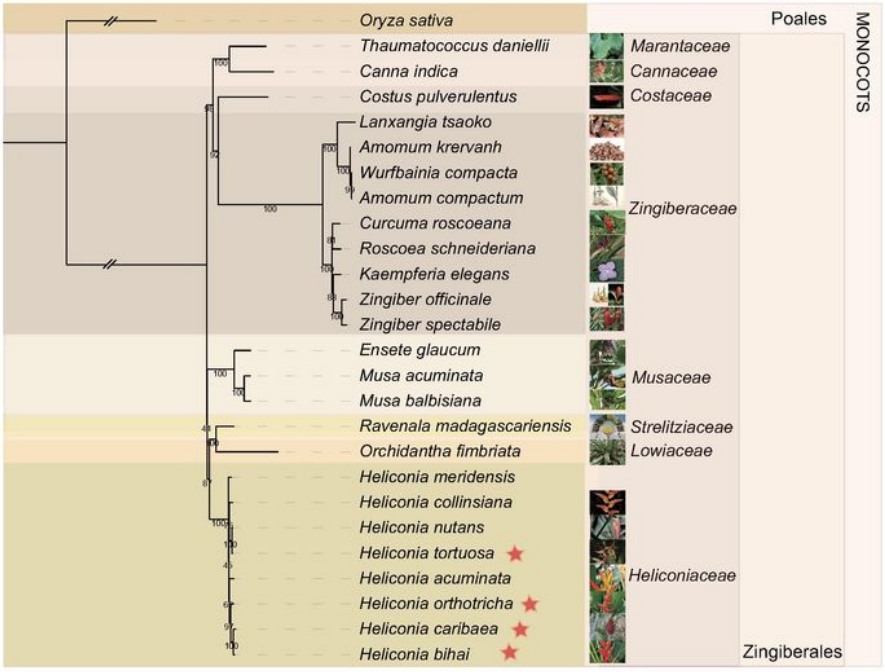

Figure 4 | Phylogenetic tree of *Heliconia* and related species. Maximum likelihood (ML) phylogenetic tree was constructed for 26 species from Zingiberales order, and rice (*Oryza sativa*) as an outgroup. The bootstrap values are shown for each branch.

Figure 4

See image above for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable.xlsx
- Supplementaryfig1.pdf
- Supplementaryfig2.pdf
- Supplementarypic1.pdf