Title: **Genome sequence for the blue-flowered Andean shrub *Iochroma cyaneum* reveals extensive discordance across the berry clade of Solanaceae**

Authors: Adrian F. Powell[1,†], Jing Zhang[1,†], Duncan A. Hauser[1], Julianne Vilela[4], Alice Hu[1], Daniel J. Gates[2,3], Lukas A. Mueller[1], Fay-Wei Li[1,5], Susan R Strickler[1], and Stacey D. Smith[6,*]

[1]Boyce Thompson Institute, Ithaca, New York, USA

[2] School of Biological Sciences, University of Nebraska, Lincoln, NE, USA

[3]Current address: Checkerspot, Inc., Alameda, California, USA

[4]Philippine Genome Center, Program for Agriculture, Livestock, Forestry and Fisheries, University of the Phillipines Los Baños, Laguna, Phillipines

[5]Plant Biology Section, Cornell University, Ithaca, New York, USA

[6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309, USA

[*]Author for correspondence: stacey.d.smith@colorado.edu

†These authors contributed equally to this work.

Abstract

The tomato family, Solanaceae, is a model clade for a wide range of applied and basic research questions. Currently, reference-quality genomes are available for over 30 species from seven genera, and these include numerous crops as well as wild species (e.g., *Jaltomata sinuosa* and *Nicotiana attenuata*). Here we present the genome of the showy-flowered Andean shrub *Iochroma cyaneum*, a woody lineage from the tomatillo subfamily Physalideae. The assembled size of the genome (2.7Gb) is more similar in size to chilipepper (2.6Gb) than to other sequenced diploid members of the berry clade of Solanaceae (e.g., potato, tomato, and *Jaltomata*). Our assembly recovers 92% of the conserved orthologous set, suggesting a nearly complete genome for this species. Most of the genomic content is repetitive (69%), with gyspy elements alone accounting for 52% of the genome. Despite the large amount of repetitive content, most of the 12 *Iochroma* chromosomes are highly syntenic with tomato. Bayesian concordance analysis provides strong support for the berry clade, including *Iochroma,* but reveals extensive discordance along the backbone, with placement of pepper and *Jaltomata* being highly variable across gene trees. The *Iochroma* genome contributes to a growing wealth of genomic resources in Solanaceae and underscores the need for expanded sampling of diverse berry genomes to dissect major morphological transitions.

Keywords: berry clade, concordance, gene family, genome assembly, Physalideae, transposable element, species tree, tomatillo

Abbreviations

LTR, Long terminal repeat; LINE, Long interspersed nuclear elements

1 | INTRODUCTION

Advances in comparative genomics rely on moving from assembling high-quality genomes from single model species to building model clades (Rogers, 2018). Model clades, as described by Donoghue and Edwards (Donoghue & Edwards, 2019), are lineages in which we sample densely across species to identify evolutionary transitions and build multilayered datasets to understand the mechanisms and drivers of those transitions. The genomic layer of clade biology has been quickly accumulated in taxa with small genomes (Feng et al., 2020; B. Y. Kim et al., 2021; Miyauchi et al., 2020), but more slowly in plants, where genomes can be as large as 149 Gb (Pellicer et al., 2010). Still, clusters of genomes have been built around plant model species and crops, where comparative evolutionary studies can result in direct applications (Ma et al., 2021; Mohd Saad et al., 2021).

One such emerging model clade is the tomato family, Solanaceae. This family comprises nearly 3000 species, roughly 40 of which have been domesticated, particularly in the fleshy-fruited subclade Solanoideae (Pickersgill, 2007; Samuels, 2015). The first published genome from this clade was potato (X. Xu et al., 2011), closely followed by tomato (Sato et al., 2012). More recently sequenced economically important species include tobacco (Sierro et al., 2014), chilipepper (S. Kim et al., 2014), eggplant (Barchi et al., 2019; Hirakawa et al., 2014a) and wolfberry (Cao et al., 2021). In addition to these crops and model organisms, many wild species have recently been sequenced, e.g. members of *Nicotiana* (S. Q. Xu et al., 2017), *Petunia* (Bombarely et al., 2016), *Solanum* (Aversano et al., 2015; Razali et al., 2017; Schmidt et al., 2017), *Capsicum* (Qin et al., 2014), and *Jaltomata* (M. Wu et al., 2018). These taxa capture wide trait variation, from fleshy to dry fruits, self-incompatible to self-compatible, and annuals to

63 perennials. Accordingly, comparative analyses have provided insights into the genomic basis for
64 a range of key traits. Studies in this family have been particularly informative with respect to
65 developmental processes (S. Kim et al., 2014), such as fruit ripening, and the evolution of
66 specialized metabolites, such as the defensive alkaloids and the colorful flavonoids and
67 carotenoids (Cardenas et al., 2015; Gebhardt, 2016).

68 Here we present a *de novo* assembly of the genome of *Iochroma cyaneum*, a blue-
69 flowered shrub native to the Andes. The genus *Iochroma* falls in the large fleshy-fruited
70 subfamily (Solanoideae) (Särkinen et al., 2013) and is related to tomatillos (*Physalis*) and
71 chilipeppers (*Capsicum*) (Deanna et al., 2019). Unlike species in these genera, *Iochroma* species
72 are woody shrubs or treelets, with some reaching up to 15m (Shaw, 1998). Moreover, while its
73 close relatives in the tomatillo tribe Physalideae are largely insect-pollinated (Knapp, 2010),
74 most species of *Iochroma* are specialized for hummingbird pollination (Smith, Hall, et al., 2008).
75 Their colorful tubular flowers arranged in large inflorescences, along with the ease of
76 hybridization among species of different colors (Smith & Baum, 2007), have made them
77 increasingly popular in the horticultural trade (Meerow et al., 2004). Given their wide range of
78 flower colors and sizes, *Iochroma* has served as a model for understanding the ecological factors
79 and genetic mechanisms that drive floral evolution (Muchhala et al., 2014; Smith, Ane, et al.,
80 2008; Smith & Rausher, 2011).

81 Comparative genomic analyses of *Iochroma* and related taxa have the potential to provide
82 new insights into the evolutionary history of Solanacaee broadly as well as the changes unique to
83 this hummingbird-pollinated lineage. For example, phylogenomic analyses may reveal
84 discordant gene histories, even in parts of the tree that were well-supported in previous
85 phylogenetic analyses with fewer markers (Gagnon et al., 2021). Moreover, the expansion of

sequenced genomes will allow us to isolate major genomic events, such as the amplification of repetitive content, rearrangements, and the gain and loss of coding genes, which may be tied to particular morphological or ecological transitions. In particular, the addition of the *Iochroma* genome will likely divide the branch between the Solanaeae (*Solanum* + *Jaltomata*) and Capsiceae (*Capsicum*+*Lycianthes*) clades, helping us to distinguish genomic variation unique to those lineages with variation that is shared due to common ancestry. In order to explore these evolutionary questions, we assembled and annotated a *de novo* genome for *Iochroma cyaneum* and applied phylogenetic and comparative analyses to estimate its relationship to other Solanaceae along with historical changes in genome content.

2 | MATERIALS AND METHODS

2.1 | Genome sequencing and assembly

Genomic DNA was prepared from fresh leaf material of *Iochroma cyaneum* (voucher: Smith 265 (WIS)) using the 2XCTAB protocol (Doyle & Doyle, 1987). We chose *I. cyaneum* because it is the type of the genus and exhibits the deep violet flowers for which the genus is named (Bentham, 1845). Although native to the northern Andes, this species is widely cultivated as an ornamental, with several commercial varieties (Meerow et al., 2004; Shaw, 1998). The sequenced accession was grown from seed from cultivated material at the Missouri Botanical Garden and originally collected from the wild by W. G. D'Arcy.

Paired-end libraries with an insert size of 400 bp were sequenced on four lanes of an Illumina Hi-Seq 2000 flow cell. Mate pair libraries of 2- and 5-kb were sequenced on two lanes.

Additionally, we sequenced a Hi-C library (Phase Genomics, Seattle, WA) on one lane of a Hi-Seq 4000 with 100x paired-end reads to assemble the contigs into larger scaffolds. All Illumina sequencing was completed at the Cornell Weill Genome Sequencing Facility and the numbers of reads are provided in Table S1. Nanopore sequencing was performed on 6 flow cells of an Oxford Nanopore Minion device to provide an additional 5,809,839 reads. Nanopore and Illumina reads were assembled with MaSurca v 3.3.2 (Zimin et al., 2013) and polished with three rounds of Pilon v1.23 (Walker et al., 2014) using Illumina reads. Hi-C data was processed using the 3D-DNA v180922 pipeline (Dudchenko et al., 2017), and the scaffolds were manually edited in Juicebox (Dudchenko et al., 2018). Gaps were filled with LR_gapcloser (G. C. Xu et al., 2019), and Pilon was used to correct errors.

2.2 | Analysis of repeat content

We examined repetitive DNA in *Iochroma* and additional Solanaceae genomes for comparison. For this purpose, we downloaded assemblies for the chilipepper *C. annuum* cv. CM334 v.1.55 (S. Kim et al., 2014), the tomato *S. lycopersicum* v.4.0 (Sato et al., 2012), *Petunia axillaris* v.1.6.2 (Bombarely et al., 2016), and *Nicotiana attenuata* r.2.0 (S. Q. Xu et al., 2017) from solgenomics.net and peppergenome.snu.ac.kr. We used LTRHarvest (Ellinghaus et al., 2008) and LTR_finder (Xu and Wang, 2007) to identify de novo putative LTR retrotransposons and LTR_retriever with default settings to filter the results and reduce false positives (Ou & Jiang, 2018). We then masked each genome using RepeatMasker v4.0.7 (Smit et al., 2013-2015) with the resulting LTR library and then used RepeatModeler v2.0.1 (Flynn et al., 2020) to identify additional repeats in the remaining unmasked regions of the genome. Known protein-coding

sequences were excluded from the RepeatModeler library using the ProtExcluder.pl script (Campbell et al., 2014). For each genome, the LTR_retriever and RepeatModeler libraries were then joined to generate a final library, which was used to mask the genome. We obtained coverage values from the RepeatMasker output, by using the fam_coverage.pl and fam_summary.pl scripts included with LTR_retriever, and inputting the estimated sizes of each genome.

2.3 | Annotation

To aid in annotation, we conducted RNA-seq on four pools of tissues (developing corollas, vegetative tissue (shoot plus root), reproductive tissue (stamen plus pistil), and seedlings) from the same accession of *I. cyaneum*. Total RNA was extracted using the Spectrum Kit (Sigma-Aldrich, St. Louis, MO) with on-column DNAse digestion (Qiagen, Valencia, CA). The corolla RNA was prepared with a TruSeq kit (Illumina, San Diego, CA) and sequenced with half of a lane of Hi-Seq2000 with 100bp paired-end reads. We also carried out 454 GS-FLX Titanium sequencing (half of plate) on normalized libraries for the corolla RNA IU at Indiana University's Center for Genomics and Bioinformatics. The remaining RNAs for the other tissues were prepared with the TruSeq kit and sequenced on a single lane of HiSeq 2500, with 100bp single reads. The 454 reads were collapsed using cd-hit v. 4.6.8 (Li & Godzik, 2006). Illumina and 454 reads were mapped to the genome assembly using Hisat2 v2.1.0) (D. Kim et al., 2015). The bam files containing mapped reads were provided as input to the BRAKER2.-2.1.5-2 pipeline (Bruna et al., 2021), which makes use of both GeneMark-ET (Lomsadze et al., 2014) and AUGUSTUS (Hoff & Stanke, 2019) for gene prediction.

Functional annotation of predicted coding genes was performed by BLASTp v2.2.31+ (Altschul et al., 1990) to the UniProt (Boutet et al., 2016) and TrEMBL (Boeckmann et al., 2003) databases using an e-value cut off of $1e^{-20}$. We also removed any predicted proteins both with few to no mapped reads (FPKM<0.01) and which had no hits with in the NCBI NR, tomato, or pepper databases. Protein domains were predicted with InterProScan v5.46-81.0 (Jones et al., 2014) and genes labeled as transposons were discarded. BUSCO v. 3 analysis (Simão et al., 2015), with the Embryophyta dataset, was used to quantify genome and annotation content and examine the completeness of the genome assembly and annotation in comparison with other published genomes. We used OrthoFinder v2.5.2 (Emms & Kelly, 2015) to identify groups of orthologous genes shared between *Iochroma*, pepper, tomato and coffee. For pepper and tomato, we used the same genome assemblies as cited above and for coffee, we used *Coffea canephora* v.1.0 (Denoeud et al., 2014). These results were used to create a Venn diagram depicting shared and unique gene clusters across taxa.

Finally, we used maximum likelihood methods to identify significantly expanded and contracted gene families in *Iochroma*. For these analyses, we expanded our sampling to include all the tips that were present in the phylogenetic analysis (see below). Again, we used OrthoFinder to identify groups of orthologous genes found in one or more of the species. We input these gene families from Orthofinder and the species tree (see below) into CAFE v.3.0 (Han et al., 2013). Before inputting, the tree was ultrametricized with penalized likelihood using the chronopl() function in the R package APE (Paradis et al., 2004). For the gene families showing significant expansion and contraction ($p<0.05$) in *Iochroma*, we conducted BLAST searches to examine their possible functions. We extracted the two longest sequences from each expanded or contracted orthogroup in *Iochroma* and ran BLAST searches using DIAMOND

blastp v0.9.30.131 (Buchfink et al., 2015). We kept the top hits for each of those sequences and retrieved the list of gene ontology (GO) terms for them with InterProScan. The resulting list of expanded or contracted *Iochroma* orthogroups and their associated GO terms was input to topGO (Alexa & Rahnenfuhrer, 2021) for enrichment analyses. We searched for enrichment in GO terms associated with biological functions and used Fisher's exact test to determine significance.

2.4 | Phylogeny estimation

We investigated the phylogenetic relationship of *I. cyaneum* to other Solanaceae using Bayesian concordance analysis (BCA) (Ane et al., 2007; Baum, 2007). This approach estimates the population or species tree with branch lengths in coalescent units using quartet methods along with the proportion of the genome that supports each clade in this tree (Larget et al., 2010). We included 7 other species of Solanaceae (*Petunia axillaris, Nicotiana attenuata*, *Solanum tuberosum*, *S. lycopersicum*, *S. melongena, Capsicum annuum*) plus *Ipomoea triloba* (Convolvulaceae) (S. Wu et al., 2018) and *Coffea canephora* (Rubiaceae) as outgroups. For the Solanaceae genomes, we used the same assembly versions and sources as listed above. For species tree estimation, we first generated posterior distributions of gene trees for the 1355 single-copy genes from the Orthofinder analysis that were present in all genomes (zero missing data). Each protein alignment was run in MrBayes v.3.2.7a (Ronquist & Huelsenbeck, 2003) for 2 million generations, sampling every 100 generations, with a mixed prior on amino acid models, an exponential prior on branch lengths with mean set to 0.001, and a gamma distribution for rate heterogeneity across sites with an estimated proportion of invariant sites. Convergence was assessed with the potential scale reduction factor (PSRF), which was near 1.0 for all model

parameters for all genes. We removed the first 5000 trees as burn-in and summarized the remaining sample from the posterior with the mbsum program in BUCKy 1.4.4 (Larget et al., 2010). We estimated the population tree and the concordance factors (CFs) in BUCKy with four MCMC chains, each of 1 million steps and the initial value for the discordance parameter, alpha, set to 1. The results of the concordance analysis were summarized as a population tree with branch lengths in coalescent units, rooted on the outgroup taxa, and concordance factors with credibility intervals for each clade.

2.5 | Synteny analysis

In order to assess patterns of synteny between *I. cyaneum* and closely related crop genomes, we first created whole genome alignments with NUCmer v3.1, part of the MUMmer software (Kurtz et al., 2004). For visualization, the alignments were then filtered to select one-to-one aligned segments with a minimum length of maximal exact matches of 2000, as well as either a minimum alignment identity of 88, in the case of *I. cyaneum* in comparison to *C. annuum*, or of 85, in the case of *I. cyaneum* to *S. lycopersicum* and *C. annuum* to *S. lycopersicum*. The coordinates of the filtered alignments were then input as links to generate plots using Circos v0.69-6 (Krzywinski et al., 2009). We used tomato as a benchmark for numbering and orienting the *Iochroma* pseudomolecules.

3 | RESULTS

3.1 | Genome assembly and annotation

The length of our de novo sequence assembly for *Iochroma cyaneum* is 2.7 Gb, making it very similar to *Capsicum annuum* (Table 1). This assembled size for *I. cyaneum* is slightly smaller than the size previously estimated from flow cytometry, ca. 3.2Gb (Gates et al., 2016). Our chromosome-level assembly (Fig. S1) was quite similar to *C. annuum*, with 84% of the assembly anchored, and our sequencing strategy resulted in a lower percentage of N bases and gaps (Table 1). Although the genomes of *Iochroma* and *Capsicum* are over three times the size of those in sequenced *Solanum* species (Bolger et al., 2014; Hirakawa et al., 2014b; Sato et al., 2012), we recovered similar numbers of annotated genes (Table 1). Our annotation for *Iochroma* includes 92% of the highly conserved benchmarking universal single-copy orthologs (BUSCOs). Overall, the BUSCO analysis showed few fragmented or missing BUSCOs (Figure 1), suggesting that the quality of the genome is on par with those of related economically important plants. In addition to these highly conserved orthologous genes, we found a large number of unique gene clusters in *I. cyaneum*, nearly twice those found in tomato or pepper (Figure 1).

Our CAFE analyses revealed a strong bias toward gene family expansion in *Iochroma*. A total of 1959 gene families had a significant change in size along the *Iochroma* branch ($p<0.05$) with 654 contracted and 1305 expanded (Supplemental Table S2). The contracted families were spread across a range of biological processes, with the most significant enrichment in ribonucleoprotein complex assembly ($p=0.0043$, Supplemental Figure S2). By contrast, the most highly enriched GO terms for the expanded gene families were all related to pollen recognition

(p=0.00037, Supplemental Figure S3). We used BLAST searches to determine the identity of the nine expanded families with this GO term, and all appear to be G-type lectin S-receptor-like serine/threonine-protein kinases (Supplemental Table S3).

	Iochroma cyaneum	***Capsicum annuum***	***Solanum lycopersicum***
Genome assembly total length (Mb)	2716.02	2633.68	782.52
Percentage of assembly assigned to chromosomes	84.13	86.00	98.77
Number of contigs	37,881	117,244	448
Contig N50 (kb)	212.94	55.87	6007.83
Longest contig (kb)	3996.25	608.96	26291.69
Number of N bases (Mb)	0.64	78.12	0.04
Number of gaps	19176	217286	435
Number of genes	38,625	34,903	34,075
Repeat percentage of genome (%)	69.35	72.26	58.30

TABLE 1 Summary statistics for *Iochroma cyaneum* genome assembly compared to closely related Solanaceae. Values for assembly length, number of N bases, and number of gaps based on currently available assemblies on SolGenomics.net (SL4.0 for tomato and v.1.55 for pepper) calculated with assembly-stats 0.1.4 (Trizna, 2020). Contig statistics were calculated with the same tool after splitting the assemblies at Ns. Remaining values estimated during the comparative repeat analyses (Figure 3) or, for annotation information, gathered from the literature (Hosmani et al., 2019; S. Kim et al., 2014).

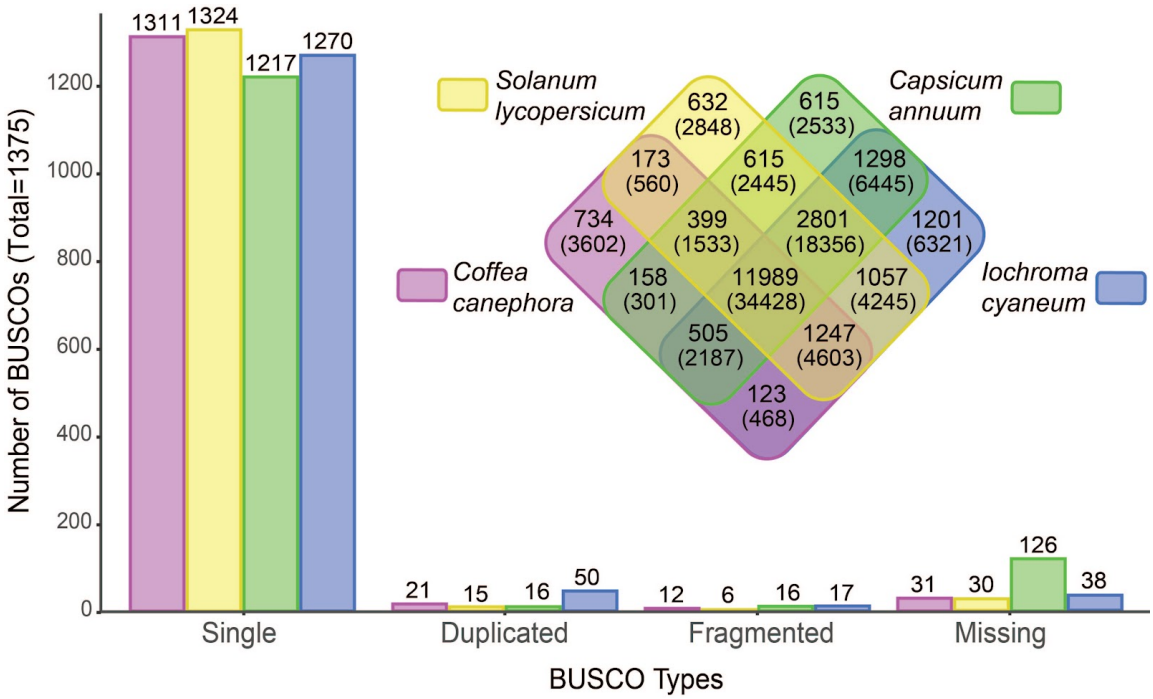


FIGURE 1 Comparison of *I. cyaneum* annotation to related crop genomes. Bar graph shows the results of the BUSCO analysis, with coffee, tomato, pepper, and *I. cyaneum*, left to right, for each BUSCO type. The numbers of genes in each category are shown at the top of each bar. Inset is a Venn diagram showing the results of the orthogroup analysis, with unique and shared clusters shown for each species. The total numbers of genes in each orthogroup are shown in parentheses.

3.2 | Phylogeny

Our phylogenetic analysis recovered the core relationships among lineages of Solanaceae that have been estimated in previous studies (Olmstead et al., 2008; Särkinen et al., 2013). *Nicotiana* is sister to the large fleshy-fruited clade containing tomato, potato, eggplant, *Jaltomata*, pepper and *Iochroma* (CF=0.72; Figure 2A). Together, they form the X=12 clade, united by the base chromosome number of 12 (Olmstead & Palmer, 1992). We find strong agreement across the 1355 genes for all the relationships within *Solanum* (CF=0.91-0.99), but less so among the other

fleshy-fruited species. For example, the estimated proportion of the genome for which the true tree places *Capsicum* sister to *Iochroma* is 0.43 and there is even less agreement regarding the placement of *Jaltomata*. Indeed, the population tree shown in Figure 2 varies from the primary concordance tree in *Jaltomata*'s position, putting it instead sister to *Capsicum* + *Iochroma* with a CF of 0.32, with an overlapping credibility interval (0.287-0.353) (Supplemental Table S4). We also estimate a sizeable proportion (23%) of the genome supporting a *Jaltomata*+*Capsicum* relationship (Figure 2B), and 19% placing *Capsicum* closer to *Solanum* than to *Iochroma* (Supplemental Table S4). Overall, these analyses point to significant discordance along the backbone of the berry clade, with large numbers of loci supporting alternate relationships to those in the population tree.

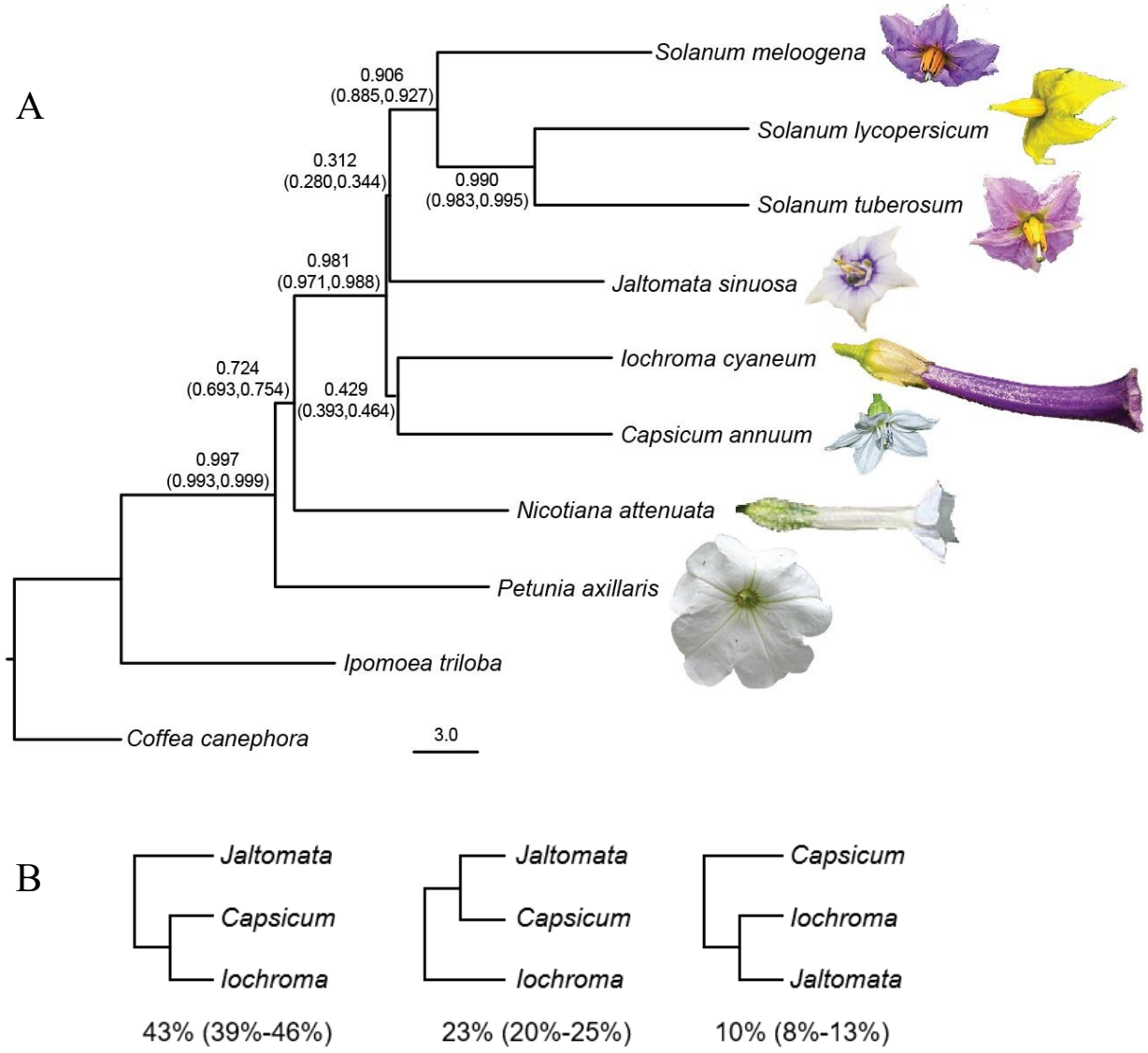
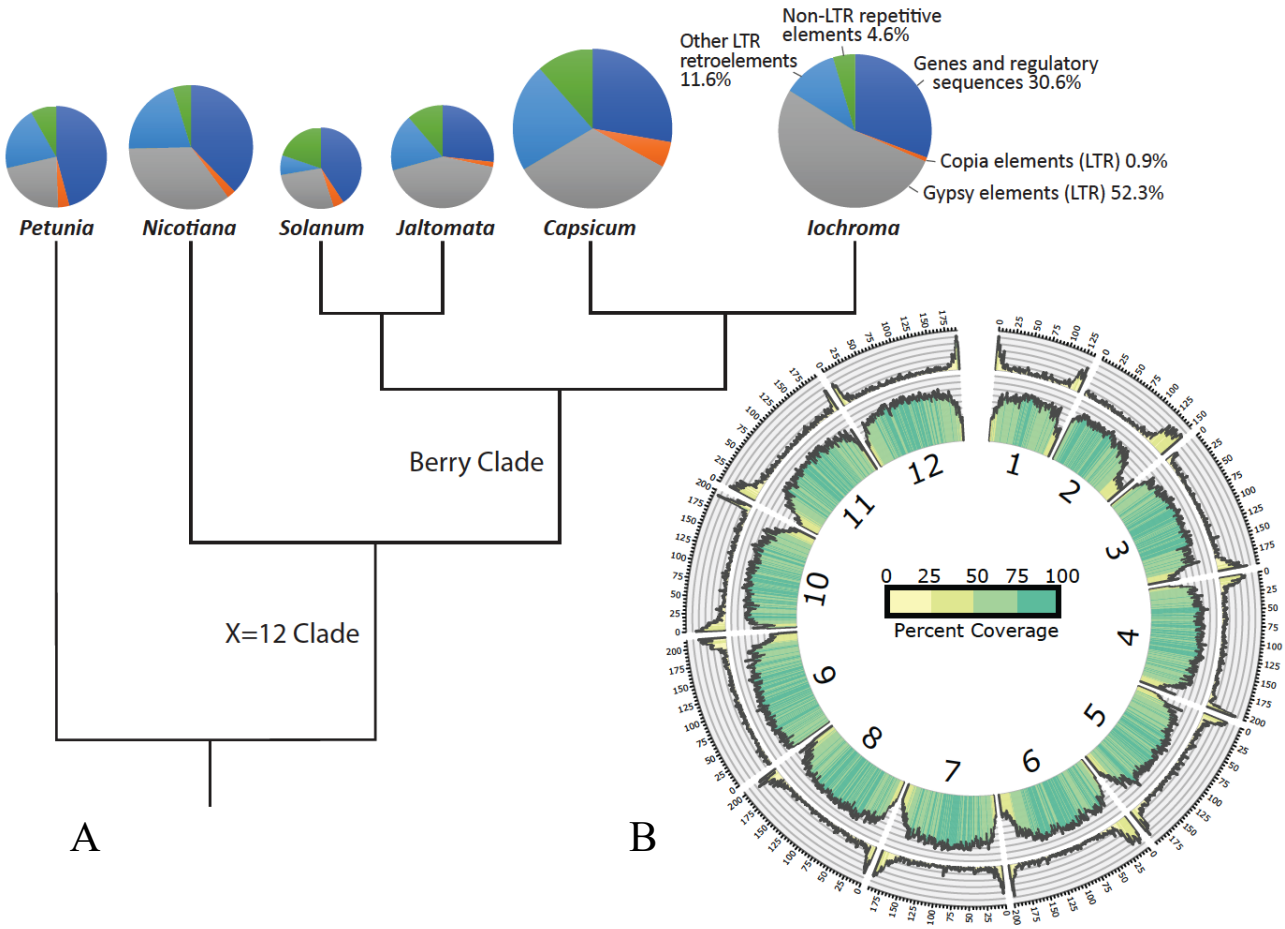


FIGURE 2 Phylogenetic position of *Iochroma*. (A) Population tree for Solanaceae estimated with BUCKy. Branch lengths are in coalescent units, and branches are annotated with the estimated genome-wide concordance factors (with credibility intervals in parentheses). Each concordance factor corresponds to the proportion of the genome estimated to have the clade in its history. Photos are from wikimedia commons with the exception of *Jaltomata sinousa* (image from Thomas Mione, CCSU). (B) Genome-wide variation in the relationships among *Jaltomata*, *Iochroma*, and *Capsicum*. Concordance factors (and their credibility intervals) are shown as percentages.

3.3 | Repetitive content in *Iochroma*

Our analyses show that the *Iochroma* genome comprises largely repetitive content as in other Solanaceae, and indeed in most plant genomes (Feschotte et al., 2002). Only 31% of the *Iochroma* genome is non-repetitive, which is slightly more than *Capsicum* and *Jaltomata* but less than the other genomes analyzed (Figure 3A, Supplemental Table S5). Despite being closely related and sharing similar percentages of repetitive DNA, the composition of the repeats varies markedly between *Iochroma* and *Capsicum*. In *Iochroma*, *gyspy* elements account for the majority of the repetitive content (75%) and over half (52%) of the entire genome. The other types of elements have contracted in *Iochroma*, which has a smaller proportion of *copia* elements among its LTR repeats than any other Solanaceae examined (Supplemental Table S5). In this context, it is worth noting that all the lineages have a significant fraction of repetitive elements that cannot be classified, either within interspersed repeats or as a type of LTR specifically. Nonetheless, as the same pipeline was applied to all taxa, the estimated variation in the fraction of each element in the genome points to substantial macroevolutionary shifts in the composition of the repetitive DNA.

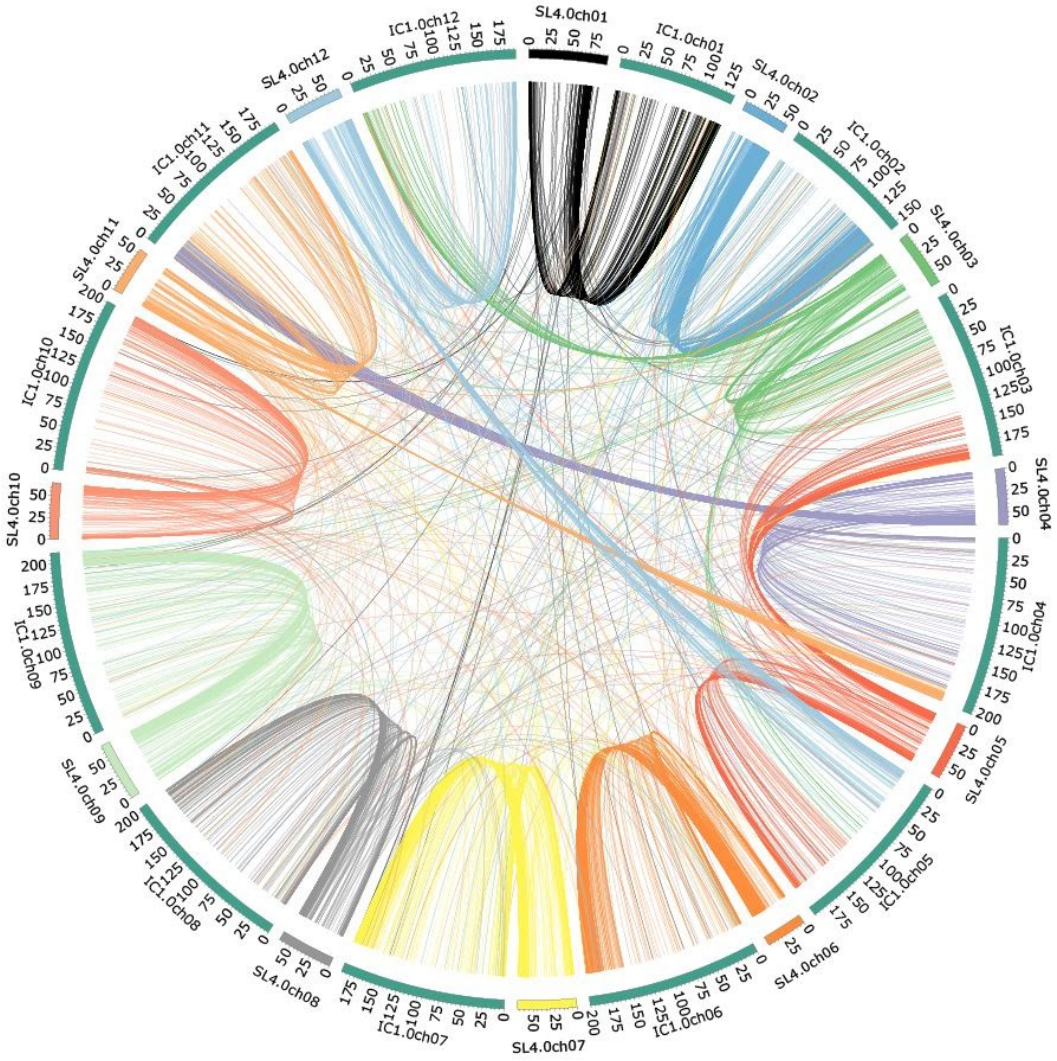
We also examined how this repetitive content was distributed along chromosomes within the *I. cyaneum* genome. We found that the non-repetitive genic regions are clustered at the very ends of the chromosomes while the centromeric regions tended to be less gene rich and more repetitive (Figure 3B). While most chromosomes have genic regions at either end, two of them (chromosomes 2 and 9) have only a single cluster at one end. This chromosomal organization (with repetitive DNA most dense at the center and coding regions at the distal ends) is common for plant genomes and was also observed in *Capsicum* (S. Kim et al., 2014).



317 **FIGURE 3** Repetitive content in *Iochroma* and related Solanaceae. (A) Phylogenetic
318 relationships from Figure 2. The pie charts for each species are proportional to genome size.
319 The other LTR retroelements category includes caulimovirus, ERK and unknown
320 retroelements and the non-LTR elements category includes LINEs, DNA elements, simple
321 and low complexity repeats, and other unclassified repetitive elements (see Supplementary
322 Table S2). (B) The distribution of repetitive content across *Iochroma* chromosomes. The
323 inside ring shows the percentage of repetitive content in each 1 Mb window and the outside
324 ring shows the percentage of annotated genic content in that window, where the gray lines
325 denote 10% increments. Each genomic window is colored to show the percent coverage of
326 repetitive content in that window, as indicated by the legend in the center. Chromosomes are
327 numbered and ordered following patterns of synteny with tomato (Figure 4). The length of
328 each chromosome is shown with the outermost ring in units of Mb.

3.4 | Collinearity between *Iochroma* and other Solanaceae

Despite the large difference in genome size between *Iochroma* and tomato, we found strong synteny for much of the genome. Most *Iochroma* chromosomes (1, 2, 4, 6-10) were easily aligned to tomato, having only small structural arrangements between the two taxa. For example, the content of *Iochroma* chromosome 9 closely matches that of tomato chromosome 9, although a few areas that match more highly to sectors of tomato chromosomes 1 and 11 (Figure 4). We did, however, observe some connections that indicate major rearrangements between the two taxa. In one clear case, the roughly 20Mb at 3' end of tomato chromosome 4 is highly syntenic with the 5' end of *Iochroma* chromosome 11, suggesting a translocation event (Figure 4). This relationship between chromosome 4 and 11 is apparent in our synteny analysis of *Iochroma* and pepper (Figure S4) but not pepper and tomato (Figure S5), which is consistent with a translocation event specific to the *Iochroma* branch of the phylogeny. In fact, visual comparison of the two synteny maps (tomato vs. *Iochroma*, Figure 4, and tomato vs. pepper, Figure S5) points to no major shared re-arrangements in *Iochroma* and *Capsicum*, suggesting that instead most of the translocations and inversions are lineage-specific. This result is consistent with the likely short duration of shared history of the two genera (Figure 2).



346

347 **FIGURE 4** Patterns of synteny between tomato and *Iochroma*. Tomato and *Iochroma cyaneum*
348 chromosomes are shown with lines connecting syntenic segments. Line coloring follows tomato. The
349 length of each chromosome is marked in 25Mb increments.

350

351 **4 | DISCUSSION**

352

353 The family Solanaceae has witnessed an explosion in whole genome sequencing accompanied by
354 efforts to expand beyond crop species into wild relatives (Bolger et al., 2014; Cao et al., 2021;
355 M. Wu et al., 2018). Analyses of these new genomes have solidified aspects of the family's

evolutionary history, such as the whole genome triplication at the base of the family (Bombarely et al., 2016; Cao et al., 2021; Sato et al., 2012), while also revealing the complexities of the phylogenetic relationships and genomic rearrangements (Barchi et al., 2019). As the first member of the tomatillo subfamily (Physalideae) with a chromosome-level assembly, our analysis of the *Iochroma* genome brings new insights regarding the radiation of the berry clade and the accompanying changes in genome size, content and organization.

4.1 | Discordance along the berry clade backbone

Phylogenetic analyses including *Iochroma* together with seven other Solanaceae points to significant discordance within the berry-fruited clade Solanoideae. This clade includes the pepper and its allies (Capsiceae), tomatillo and its allies (Physalideae) and the large genus *Solanum* and its sister genus *Jaltomata* (Solaneae). Recent family-level analyses with plastid and nuclear markers have shown strong support for the dominant relationship, with *Capsicum* more closely related to *Physalis* and *Jaltomata* sister to *Solanum* (Olmstead et al., 2008; Särkinen et al., 2013). Nevertheless, alternative relationships have often appeared in phylogenetic analyses (Bohs & Olmstead, 1997; Olmstead et al., 1999; Smith & Baum, 2006) and previous phylogenomic analyses suggest extensive discordance involving *Capsicum* and *Jaltomata* (M. Wu et al., 2018). Our Bayesian concordance analysis expands the scope of this discordance, as the relationship of *Iochroma* to these two taxa is also highly variable across the genome. Following previous family-level studies (Olmstead et al., 2008; Särkinen et al., 2013), we expected *Iochroma* to be most closely related to *Capsicum* and indeed, 43% of the genes in the genome are estimated to follow this dominant history (Figure 2A). However, many genes show

alternative resolutions, i.e., with *Capsicum* sister to *Jaltomata* (22%) or *Jaltomata* sister to *Iochroma* (10%) (Figure 2B). Meanwhile, the position of *Jaltomata* is nearly evenly split across gene trees between appearing as sister to *Solanum* (31%) versus sister to *Capsicum+Iochroma* (32%). These patterns contrast with other nodes in the tree (e.g., the common ancestor of *Solanum*, the common ancestor of Solanaceae), where nearly all genes share the same underlying history. The high discordance along the backbone of the berry clade may reflect a range of evolutionary processes, including hybridization and introgression or incomplete lineage sorting due to rapid radiation and/or large population sizes (Maddison, 1997). In the case of *Iochroma*, the large difference between the dominant history (43% for *Capsicum* sister) and the minor histories (22%, 10%) is most consistent with incomplete lineage sorting (Baum, 2007). Expanding the phylogenomic analysis to include other major lineages of the large and diverse berry clade (ca. 2000 species) would be valuable to distinguish among these possible causes.

4.2 | Gene family evolution in *Iochroma*

Although quite similar in total genome size, our annotation pipeline retrieved more gene models in *Iochroma* than were estimated in pepper (38.6K vs 34.9K, Table 1), and we estimate a slightly higher proportion of non-repetitive (including genic) content in *Iochroma* (30.6% vs. 27.7%). Consistent with the possibility of gene family expansion along the *Iochroma* lineage, the orthogroup analysis recovered a larger number of unique orthogroups compared to pepper and more genes in those orthogroups (Figure 1). Using maximum likelihood birth-death models, we estimated significant expansions in 1305 gene families (Supplemental Table S2), and we found that these families were enriched for function in pollen recognition (Figure S3). BLAST searches

suggest that these orthogroups that are significantly expanded in *Iochroma* and involved in pollen recognition are G-type lectin S-receptor-like serine/threonine-protein kinases. Receptor kinases are known to play an important role in sporophytic self-incompatibility in the Brassiceae, but they have not been documented to be involved in pollen recognition in species with gametophytic self-incompatibility, like Solanaceae (Kachroo et al., 2001; McCubbin & Kao, 2000). Beyond pollination, these G-type lectin receptor-like kinases (LecRLKs) are known to be involved in other aspects of signalling, in particular, mediating responses to insect attacks (Gilardoni et al., 2011). Plant-insect interactions have emerged as major drivers of genome evolution, especially in Solanaceae (De-la-Cruz et al., 2021; Fan et al., 2020), and our findings from *Iochroma* suggest that LecRLKs merit additional investigation as mediators of these interactions (Sun et al., 2020).

4.3 | Diversity and distribution of repetitive DNA

With a genome estimated at 3.2Gb with flow cytometry (Gates et al., 2016) and 2.7Gb in our reference assembly (Table 1), *Iochroma* presents the largest diploid genome sequenced in the Solanaceae thus far, and is most similar in size to *Capsicum*. The large size of the pepper genome compared to tomato was attributed to the expansion of repetitive content, and in particular, LTR retroelements (S. Kim et al., 2014). Using a single pipeline for six Solanaceae species, we estimated that the proportion of the genome occupied by LTRs in *Iochroma* is even higher than in pepper and roughly 1.5 times that in tomato (Supplemental Table S5). We also uncovered a high turn-over in the type of LTR retrotransposon in *Iochroma*, which has much higher proportion of *gypsy* elements compared to pepper (81% versus 55%) and a

correspondingly smaller proportion of the other classes of retroelements (Figure 3, Supplemental Table S5). Thus, while the genomes of these species are both composed of over 60% LTR retrotransposons, the individual classes of element have shifted dramatically in frequency, possibly due to rounds of TE expansion and contraction (i.e., the genomic 'accordion', Kapusta et al., 2017). Although LTR retrotransposons, like other transposable elements (TEs), seem to be largely inactive (Feschotte et al., 2002), lineage specific amplification and contractions are often uncovered in comparative genomic analyses in plants (e.g., Lee et al., 2017; Zhang et al., 2019). Whole genome duplications and hybridization events are hypothesized to trigger TE proliferation (Wendel et al., 2016), offering an intriguing area for future research given the apparent frequency of hybridization in Iochrominae (Smith & Baum, 2006) and possibly more broadly in the tomatillo clade (Zamora-Tavares et al., 2016).

As in many plant genomes, we also found that the repetitive content in the *Iochroma* genome is found in the centers of the chromosomes, with genic regions clustered at the tips (Figure 3B). This organization is common to plants with metacentric chromosomes, and the repetitive content plays a key role in coordinating chromosome movement during meiosis and mitosis (Nagaki et al., 2003; Zhong et al., 2002). All twelve chromosomes of *I. cyaneum* are metacentric, and such highly symmetric karyotypes are typical in the genus (Deanna et al., 2018). Tomato and pepper share this karyotype (mostly or all metacentric; Chiarini et al., 2018) and in turn, this chromosomal organization, with an expanse of repetitive content at the center and gene-rich content only near the ends (Jouffroy et al., 2016; S. Kim et al., 2014).

Despite their similarity in genome organization, pattern of synteny between these three taxa suggest several major rearrangements. The comparison of tomato and *Iochroma* revealed regions of up to 50Mb with disrupted synteny, likely due to translocations, towards the ends of

chromosomes 4, 5, 11 and 12 (Figure 4). Given that *Iochroma* is more closely related to pepper than to tomato, we expected fewer rearrangements between them, but instead observed less synteny than with tomato (Supplemental Figure S4). These results suggest that genomic events such as large translocations, inversions, and deletions, are frequent at this ca. 20-million year intergeneric scale (Barchi et al., 2019) and that a much more dense taxon sampling will be needed to infer the order and timing of any particular event. The addition of a high-quality reference genome for *Physalis* (Lemmon et al., 2018) will aid in determining which of the rearrangements that appear distinct to *Iochroma* are in fact shared more widely across the tomatillo clade. Karyotypic analyses across Physalideae point to several shifts in chromosome size, symmetry, and number that can help to guide taxon sampling (Deanna et al., 2018; Rodriguez et al., 2020). With more targeted sampling across the berry clade together with the development of new comparative genomic tools (e.g., GENESPACE, Lovell et al., 2018), we may look toward building a berry core-genome that captures the shared elements in fleshy-fruited common ancestor as well as a pan-genome that spans the genomic diversity of the clade.

5 | CONCLUSIONS

With clusters of genomes emerging around crop species of Solanaceae, our challenge now is to expand in terms of phylogenetic diversity, using wild species to span the connections between these clusters. The berry clade of Solanaceae comprises roughly 50 genera (Hunziker, 2001), but 20 genomes sequenced thus far include only 5 of these. As a member of the tomatillo clade, the addition of the *Iochroma* splits the evolutionary path between pepper and tomato, with slightly closer affinity to pepper. Nevertheless, our phylogenetic analyses reinforce and expand the

findings of M. Wu et al. (2018), namely that the relationships among berry clade genera are highly discordant across the genome. This discordance has important implications for downstream applications of these comparative genomics resources. For example, the genes that underlie traits of interest, such as fruit characteristics or secondary metabolites, may not follow the inferred species tree, potentially leading to incorrect inferences about the number and timing of evolutionary transitions (Hahn & Nakhleh, 2016). Moreover, the disagreement about relationships means there is no clear sister group for genomic comparison with crop-containing genera (*Solanum*, *Capsicum*). Instead, functional comparative studies will need to make use of the suite of sequenced berry clade genomes to reconstruct gene histories and dissect the origins of mutations with functional consequences (Martin & Orgogozo, 2013). Adding genomic resources for other genera is unlikely to resolve the deeply discordant backbone of the berry species tree, but will allow us to build a more complete picture of the evolutionary diversification of this economically important clade of plants.

DATA AVAILABILITY

Raw sequencing reads used in the assembly of the genome are available from the NCBI database under BioProject PRJNA777841. The completed genome assembly and annotation files are available on the Sol Genomics Network (https://solgenomics.net) website.

ACKNOWLEDGEMENTS

Funding for this work was provided by the National Science Foundation (NSF-DEB 1413855 and 1355518 to SDS) and start-up funds from the University of Nebraska-Lincoln to SDS. SDS also thanks Lynn Bohs for providing seed of this accession of *I. cyaneum*.

AUTHOR CONTRIBUTIONS

This research was conceived by SDS and designed by SDS, SRS, LAM, and DJG. SDS provided the materials, and DAH and FWL completed the nanopore sequencing. Preliminary analyses were carried out by JAV, AH, and SRS. Final analyses were completed by AFP, JZ, and SDS. SDS drafted the manuscript and all authors have reviewed and approved the submission.

CONFLICT OF INTEREST DISCLOSURE

The authors declare that they have no conflict of interest.

ETHICAL STANDARD

Research conducted for this manuscript complies with the ethical rules applicable for this journal, as stated in the Instructors for Authors.

REFERENCES

Alexa, A., & Rahnenfuhrer, J. (2021). topGO: Enrichment Analysis for Gene Ontology. R package version 2.45.0.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology, 215*, 403-410.

Ane, C., Larget, B., Baum, D. A., Smith, S. D., & Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution, 24*, 412-426.

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., . . . Carputo, D. (2015). The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell, 27*, 954-968.

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A., . . . Rotino, G. L. (2019). A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports, 9*.

Baum, D. A. (2007). Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon, 56*, 417-426.

Bentham, G. (1845). *Iochroma tubulosa*. *Edward's Botanical Register, 31*, t. 20.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., . . . Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research, 31*, 365-370.

Bohs, L., & Olmstead, R. G. (1997). Phylogenetic relationships in *Solanum* (Solanaceae) based on ndhF sequences. *Systematic Botany, 22*, 5-17.

Bolger, A., Scossa, F., Bolger, M. E., Lanz, C., Maumus, F., Tohge, T., . . . Fernie, A. R. (2014). The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature Genetics, 46*, 1034-1038.

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C. S., . . . Kuhlemeier, C. (2016). Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nature Plants, 2*.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., . . . Xenarios, I. (2016). UniProtKB/Swiss-Prot, the manually annotated section of the UniProt

KnowledgeBase: How to use the entry view. *High-Throughput Next Generation Sequencing: Methods and Application, 1374*, 23-54.

Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform, 3*, lqaa108.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods, 12*, 59-60.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., . . . Yandell, M. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology, 164*, 513-524.

Cao, Y. L., Li, Y. L., Fan, Y. F., Li, Z., Yoshida, K., Wang, J. Y., . . . Liu, Z. J. (2021). Wolfberry genomes and the evolution of Lycium (Solanaceae). *Commun Biol, 4*, 671.

Cardenas, P. D., Sonawane, P. D., Heinig, U., Bocobza, S. E., Burdman, S., & Aharoni, A. (2015). The bitter side of the nightshades: Genomics drives discovery in Solanaceae steroidal alkaloid metabolism. *Phytochemistry, 113*, 24-32.

Chiarini, F., Sazatornil, F., & Bernardello, G. (2018). Data reassessment in a phylogenetic context gives insight into chromosome evolution in the giant genus Solanum (Solanaceae). *Systematics and Biodiversity, 16*, 397-416.

De-la-Cruz, I. M., Hallab, A., Olivares-Pinto, U., Tapia-Lopez, R., Velazquez-Marquez, S., Pinero, D., . . . Nunez-Farfan, J. (2021). Genomic signatures of the evolution of defence against its natural enemies in the poisonous and medicinal plant *Datura stramonium* (Solanaceae). *Sci Rep, 11*, 882.

561 Deanna, R., Larter, M. D., Barboza, G. E., & Smith, S. D. (2019). Repeated evolution of a
562 morphological novelty: a phylogenetic analysis of the inflated fruiting calyx in the
563 Physalideae tribe (Solanaceae). *American Journal of Botany*, in press.
564 Deanna, R., Smith, S. D., Sarkinen, T., & Chiarini, F. (2018). Patterns of chromosomal evolution
565 in the florally diverse Andean Glade Iochrominae (Solanaceae). *Perspectives in Plant*
566 *Ecology Evolution and Systematics, 35*, 31-43.
567 Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., . . .
568 Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution
569 of caffeine biosynthesis. *Science, 345*, 1181-1184.
570 Donoghue, M. J., & Edwards, E. J. (2019). Model clades are vital for comparative biology, and
571 ascertainment bias is not a problem in practice: a response to Beaulieu and O'Meara
572 (2018). *American Journal of Botany, 106*, 327-330.
573 Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure from small quantities of
574 fresh leaf tissues. *Phytochemical Bulletin, 19*, 11-15.
575 Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden,
576 E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields
577 chromosome-length scaffolds. *Science, 356*, 92-95.
578 Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., . . .
579 Aiden, E. L. (2018). The Juicebox Assembly Tools module facilitates de novo assembly
580 of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*,
581 https://doi.org/10.1101/254797.
582 Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software
583 for de novo detection of LTR retrotransposons. *Bmc Bioinformatics, 9*, 18.

584 Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome
585 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology, 16*,
586 157.
587 Fan, P., Wang, P., Lou, Y. R., Leong, B. J., Moore, B. M., Schenck, C. A., . . . Last, R. L.
588 (2020). Evolution of a plant gene cluster in Solanaceae and emergence of metabolic
589 diversity. *Elife, 9*.
590 Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., . . . Zhang, G. (2020).
591 Dense sampling of bird diversity increases power of comparative genomics. *Nature, 587*,
592 252-257.
593 Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics
594 meets genomics. *Nature Reviews Genetics, 3*, 329-341.
595 Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F.
596 (2020). RepeatModeler2 for automated genomic discovery of transposable element
597 families. *Proc Natl Acad Sci U S A, 117*, 9451-9457.
598 Gagnon, E., Hilgenhof, R., Orejuela, A., McDonnell, A., Sablok, G., Aubriot, X., . . . Sarkinen,
599 T. (2021). Phylogenomic data reveal hard polytomies across the backbone of the large
600 genus *Solanum* (Solanaceae). *bioRxiv*, https://doi.org/10.1101/2021.1103.1125.436973.
601 Gates, D. J., Strickler, S. R., Mueller, L. A., Olson, B. J. S. C., & Smith, S. D. (2016).
602 Diversification of R2R3-MYB transcription factors in the tomato family Solanaceae.
603 *Journal of Molecular Evolution, 83*, 26-37.
604 Gebhardt, C. (2016). The historical role of species from the Solanaceae plant family in genetic
605 research. *Theoretical and Applied Genetics, 129*, 2281-2294.

606 Gilardoni, P. A., Hettenhausen, C., Baldwin, I. T., & Bonaventure, G. (2011). *Nicotiana*
607 *attenuata* LECTIN RECEPTOR KINASE1 suppresses the insect-mediated inhibition of
608 induced defense responses during *Manduca sexta* herbivory. *Plant Cell, 23*, 3512-3532.
609 Hahn, M. W., & Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution,*
610 *70*, 7-17.
611 Han, M. V., Thomas, G. W., Lugo-Martinez, J., & Hahn, M. W. (2013). Estimating gene gain
612 and loss rates in the presence of error in genome assembly and annotation using CAFE 3.
613 *Molecular Biology and Evolution, 30*, 1987-1997.
614 Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., . . . Fukuoka,
615 H. (2014a). Draft Genome Sequence of Eggplant (Solanum melongena L.): the
616 Representative Solanum Species Indigenous to the Old World. *DNA Research, 21*, 649-
617 660.
618 Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., . . . Fukuoka,
619 H. (2014b). Draft genome sequence of eggplant (*Solanum melongen*a L.): the
620 representative solanum species indigenous to the old world. *DNA Research, 21*, 649-660.
621 Hoff, K. J., & Stanke, M. (2019). Predicting genes in single genomes with AUGUSTUS. *Curr*
622 *Protoc Bioinformatics, 65*, e57.
623 Hosmani, P. S., Flores-Gonzalez, M., Van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E.,
624 . . . Saha, S. (2019). *An improved de novo assembly and annotation of the tomato*
625 *reference genome using single-molecule sequencing, Hi-C proximity ligation and optical*
626 *maps*. bioRxiv. https://doi.org/10.1101/767764.
627 Hunziker, A. T. (2001). *The genera of Solanaceae*. Konigstein, Germany: A. R. G. Ganter
628 Verlag.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W. Z., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*, 1236-1240.

Jouffroy, O., Saha, S., Mueller, L., Quesneville, H., & Maumus, F. (2016). Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *Bmc Genomics, 17*.

Kachroo, A., Schopfer, C. R., Nasrallah, M. E., & Nasrallah, J. B. (2001). Allele-specific receptor-ligand interactions in *Brassica* self-incompatibility. *Science, 293*, 1824-1826.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A, 114*, E1460-E1469.

Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., . . . Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *Elife, 10*.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Methods, 12*, 357-360.

Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., . . . Choi, D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature Genetics, 46*, 270-+.

Knapp, S. (2010). On 'various contrivances': pollination, phylogeny and flower form in the Solanaceae. *Philosophical Transactions of the Royal Society B-Biological Sciences, 365*, 449-460.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research, 19*, 1639-1645.

652 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S.
653 L. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5*,
654 R12.
655 Larget, B. R., Kotha, S. K., Dewey, C. N., & Ane, C. (2010). BUCKy: Gene tree/species tree
656 reconciliation with Bayesian concordance analysis. *Bioinformatics, 26*, 2910-2911.
657 Lee, J., Waminal, N. E., Choi, H. I., Perumal, S., Lee, S. C., Nguyen, V. B., . . . Yang, T. J.
658 (2017). Rapid amplification of four retrotransposon families promoted speciation and
659 genome size expansion in the genus *Panax*. *Sci Rep, 7*, 9045.
660 Lemmon, Z. H., Reem, N. T., Dalrymple, J., Soyk, S., Swartwood, K. E., Rodriguez-Leal, D., . .
661 . Lippman, Z. B. (2018). Rapid improvement of domestication traits in an orphan crop by
662 genome editing. *Nature Plants, 4*, 766-770.
663 Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of
664 protein or nucleotide sequences. *Bioinformatics, 22*, 1658-1659.
665 Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads
666 into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research, 42*,
667 e119.
668 Lovell, J. T., Jenkins, J., Lowry, D. B., Mamidi, S., Sreedasyam, A., Weng, X., . . . Juenger, T.
669 E. (2018). The genomic landscape of molecular responses to natural drought stress in
670 Panicum hallii. *Nature Communications, 9*, 5213.
671 Ma, Z., Zhang, Y., Wu, L., Zhang, G., Sun, Z., Li, Z., . . . Wang, X. (2021). High-quality
672 genome assembly and resequencing of modern cotton cultivars provide resources for crop
673 improvement. *Nature Genetics, 53*, 1385-1391.
674 Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology, 46*, 523-536.

675 Martin, A., & Orgogozo, V. (2013). The loci of repeated evolution: a catalog of genetic hotspots
676 of phenotypic variation. *Evolution, 67*, 1235-1250.
677 McCubbin, A. G., & Kao, T. (2000). Molecular recognition and response in pollen and pistil
678 interactions. *Annu Rev Cell Dev Biol, 16*, 333-364.
679 Meerow, A. W., Schoellhorn, R. J., & Kartuz, M. (2004). Four cultivars of *Iochroma*.
680 *Hortscience, 39*, 194-197.
681 Miyauchi, S., Kiss, E., Kuo, A., Drula, E., Kohler, A., Sanchez-Garcia, M., . . . Martin, F. M.
682 (2020). Large-scale genome sequencing of mycorrhizal fungi provides insights into the
683 early evolution of symbiotic traits. *Nature Communications, 11*, 5125.
684 Mohd Saad, N. S., Severn-Ellis, A. A., Pradhan, A., Edwards, D., & Batley, J. (2021). Genomics
685 armed with diversity leads the way in Brassica improvement in a changing global
686 environment. *Front Genet, 12*, 600789.
687 Muchhala, N., Johnsen, S., & Smith, S. D. (2014). Competition for hummingbird pollination
688 shapes flower color variation in Andean Solanaceae. *Evolution, 68*, 2275-2286.
689 Nagaki, K., Talbert, P. B., Zhong, C. X., Dawe, R. K., Henikoff, S., & Jiang, J. (2003).
690 Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key
691 functional DNA element of Arabidopsis thaliana centromeres. *Genetics, 163*, 1221-1225.
692 Olmstead, R. G., Bohs, L., Migid, H. A., Santiago-Valentin, E., Garcia, V. F., & Collier, S. M.
693 (2008). A molecular phylogeny of the Solanaceae. *Taxon, 57*, 1159-1181.
694 Olmstead, R. G., & Palmer, J. D. (1992). A chloroplast DNA phylogeny of the Solanaceae:
695 subfamilial relationships and character evolution. *Annals of the Missouri Botanical*
696 *Garden, 79*, 346-360.

Olmstead, R. G., Sweere, J. A., Spangler, R. E., Bohs, L., & Palmer, J. D. (1999). Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. In M. Nee, D. E. Symon, R. N. Lester, & J. P. Jessop (Eds.), *Solanaceae IV* (pp. 111-137). Kew: Royal Botanical Gardens.

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology, 176*, 1410-1422.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics, 20*, 289-290.

Pellicer, J., Garcia, S., Canela, M. A., Garnatje, T., Korobkov, A. A., Twibell, J. D., & Valles, J. (2010). Genome size dynamics in Artemisia L. (Asteraceae): following the track of polyploidy. *Plant Biol (Stuttg), 12*, 820-830.

Pickersgill, B. (2007). Domestication of plants in the Americas: insights from Mendelian and molecular genetics. *Ann Bot, 100*, 925-940.

Qin, C., Yu, C. S., Shen, Y. O., Fang, X. D., Chen, L., Min, J. M., . . . Zhang, Z. M. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 5135-5140.

Razali, R., Bougouffa, S., Morton, M. J. L., Lightfoot, D. J., Alam, I., Essack, M., . . . Negrão, S. (2017). The genome sequence of the wild tomato Solanum pimpinellifolium provides insights into salinity tolerance. *bioRxiv*.

718 Rodriguez, J., Deanna, R., & Chiarini, F. (2020). Karyotype asymmetry shapes diversity within
719 the physaloids (Physalidinae, Physalideae, Solanaceae). *Systematics and Biodiversity, 19*,
720 168-185.
721 Rogers, J. (2018). Adding resolution and dimensionality to comparative genomics: moving from
722 reference genomes to clade genomics. *Genome Biology, 19*, 115.
723 Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under
724 mixed models. *Bioinformatics, 19*, 1572-1574.
725 Samuels, J. (2015). Biodiversity of food species of the Solanaceae family: A preliminary
726 taxonomic inventory of the subfamily Solanoideae. *Resources, 4*, 277-322.
727 Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for
728 evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *Bmc*
729 *Evolutionary Biology, 13*.
730 Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., . . . Consortium, T. G.
731 (2012). The tomato genome sequence provides insights into fleshy fruit evolution.
732 *Nature, 485*, 635-641.
733 Schmidt, M. H. W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., . . .
734 Usadel, B. (2017). De novo assembly of a new Solanum pennellii accession using
735 nanopore sequencing. *Plant Cell, 29*, 2336-2348.
736 Shaw, J. M. H. (1998). Iochroma - a review. *New Plantsman, 5*, 153-191.
737 Sierro, N., Battey, J. N. D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., . . . Ivanov, N. V.
738 (2014). The tobacco genome sequence and its comparison with those of tomato and
739 potato. *Nature Communications, 5*.

740 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015).
741 BUSCO: assessing genome assembly and annotation completeness with single-copy
742 orthologs. *Bioinformatics, 31*, 3210-3212.
743 Smit, A. F. A., Hubley, R., & Green, P. (2013-2015). RepeatMasker Open-4.0. Retrieved from
744 http://www.repeatmasker.org
745 Smith, S. D., Ane, C., & Baum, D. A. (2008). The role of pollinator shifts in the floral
746 diversification of *Iochroma* (Solanaceae). *Evolution, 62*, 793-806.
747 Smith, S. D., & Baum, D. A. (2006). Phylogenetics of the florally diverse Andean clade
748 Iochrominae (Solanaceae). *American Journal of Botany, 93*, 1140-1153.
749 Smith, S. D., & Baum, D. A. (2007). Systematics of Iochrominae (Solanaceae): Patterns in floral
750 diversity and interspecific crossability. *Acta Horticulturae, 745*, 241-254.
751 Smith, S. D., Hall, S. J., Izquierdo, P. R., & Baum, D. A. (2008). Comparative pollination
752 biology of sympatric and allopatric Andean *Iochroma* (Solanaceae). *Annals of the*
753 *Missouri Botanical Garden, 95*, 600-617.
754 Smith, S. D., & Rausher, M. D. (2011). Gene loss and parallel evolution contribute to species
755 difference in flower color. *Molecular Biology and Evolution, 28*, 2799-2810.
756 Sun, Y., Qiao, Z., Muchero, W., & Chen, J. G. (2020). Lectin receptor-like kinases: The sensor
757 and mediator at the plant cell surface. *Frontiers in Plant Science, 11*, 596301.
758 Trizna, M. (2020). assembly-stats 0.1.4 doi: 10.5281/zenodo.3968774.
759 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Earl, A. M.
760 (2014). Pilon: an integrated tool for comprehensive microbial variant detection and
761 genome assembly improvement. *Plos One, 9*, e112963.

762 Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome
763 architecture. *Genome Biology, 17*, 37.
764 Wu, M., Kostyun, J., & Moyle, L. C. (2018). Genome sequence of *Jaltomata* addresses rapid
765 reproductive trait evolution and enhances comparative genomics in the hyper-diverse
766 Solanaceae. *bioRxiv*.
767 Wu, S., Lau, K. H., Cao, Q., Hamilton, J. P., Sun, H., Zhou, C., . . . Fei, Z. (2018). Genome
768 sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for
769 genetic improvement. *Nature Communications, 9*, 4580.
770 Xu, G. C., Xu, T. J., Zhu, R., Zhang, Y., Li, S. Q., Wang, H. W., & Li, J. T. (2019).
771 LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome
772 assembly. *Gigascience, 8*.
773 Xu, S. Q., Brockmoller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z. H., . . . Baldwin,
774 I. T. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis.
775 *Proceedings of the National Academy of Sciences of the United States of America, 114*,
776 6133-6138.
777 Xu, X., Pan, S. K., Cheng, S. F., Zhang, B., Mu, D. S., Ni, P. X., . . . Consortiu, P. G. S. (2011).
778 Genome sequence and analysis of the tuber crop potato. *Nature, 475*, 189-U194.
779 Zamora-Tavares, M. D. P., Martinez, M., Magallon, S., Guzman-Davalos, L., & Vargas-Ponce,
780 O. (2016). Physalis and physaloids: A recent and complex evolutionary history.
781 *Molecular Phylogenetics and Evolution, 100*, 41-50.
782 Zhang, T., Qiao, Q., Novikova, P. Y., Wang, Q., Yue, J., Guan, Y., . . . Qiong. (2019). Genome
783 of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation
784 to high altitude. *Proc Natl Acad Sci U S A, 116*, 7137-7146.

785 Zhong, C. X., Marshall, J. B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., . . . Dawe, R. K.
786 (2002). Centromeric retroelements and satellites interact with maize kinetochore protein
787 CENH3. *Plant Cell, 14*, 2825-2836.
788 Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The
789 MaSuRCA genome assembler. *Bioinformatics, 29*, 2669-2677.
790

Table S1 Number of reads from each of the sequenced libraries

Library type	Number of lanes / cells	Total number of reads
Paired-end with 400bp inserts (first round)	2	768,000,000
Paired-end with 400bp inserts (second round)	2	738,216,828
Mate pair with 2kb inserts	1	396,495,640
Mate pair with 5kb inserts	1	406,371,714
Nanopore	6	5,809,839
Hi-C	2	644,055,990

Links to Supplemental Tables S2 and S3

TABLE S2 Change in size of orthogroups in *Iochroma*. Increases or decreases in size (estimated) by CAFE along with p-values for the significance of that change.

TABLE S3 Enriched gene ontology (GO) terms among expanded gene families. The GO enrichment analysis indicated significant enrichment associated with GO:0048544. The nine significantly expanded gene families with this GO term are listed along with the estimated increase in size along the *Iochroma* branch from CAFE and the top BLAST hits for the longest transcripts in that orthogroup from *Iochroma* (IC) and *Capscium* (CA).

Table S4 Concordance factors (CF) for splits not present in the population tree, greater than 0.05. Credibility intervals for CFs are shown below estimates in parentheses. Taxa are abbreviated to the first letter of the genus name and first three letters of species name (e.g. Smel=*Solanum melongena*).

Split	**Sample-wide CF**	**Genome-wide CF**	**SD of mean sample-wide CF**
{Smel,Slyc,Natt,Itri,Stub,Paxi,Ccan|Icya,Jsin,Cann}	0.320 (0.299,0.341)	0.320 (0.287,0.353)	0.005
{Smel,Icya,Slyc,Natt,Itri,Stub,Paxi,Ccan|Jsin,Cann}	0.225 (0.206,0.244)	0.225 (0.196,0.255)	0.004
{Smel,Slyc,Jsin,Stub,Cann|Icya,Natt,Itri,Paxi,Ccan}	0.188 (0.170,0.206)	0.188 (0.161,0.216)	0.003
{Smel,Icya,Slyc,Jsin,Itri,Stub,Cann,Ccan|Natt,Paxi}	0.188 (0.171,0.205)	0.188 (0.162,0.215)	0.003
{Smel,Icya,Slyc,Jsin,Stub|Natt,Itri,Cann,Paxi,Ccan}	0.144 (0.127,0.162)	0.144 (0.119,0.170)	0.003
{Smel,Slyc,Natt,Itri,Stub,Cann,Paxi,Ccan|Icya,Jsin}	0.101 (0.083,0.118)	0.101 (0.078,0.125)	0.005
{Smel,Icya,Slyc,Stub,Cann|Jsin,Natt,Itri,Paxi,Ccan}	0.086 (0.070,0.102)	0.086 (0.065,0.108)	0.004
{Smel,Icya,Slyc,Jsin,Stub,Cann,Paxi|Natt,Itri,Ccan}	0.080 (0.066,0.095)	0.080 (0.060,0.102)	0.003
{Smel,Slyc,Stub,Cann|Icya,Jsin,Natt,Itri,Paxi,Ccan}	0.057 (0.046,0.069)	0.057 (0.041,0.075)	0

Table S5 Repeat analysis across Solanaceae. Major classes of repeats identified by RepeatModeler are interspersed repeats (including long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), DNA repeats), simple repeats and low complexity repeats. These are shown in bold. Percentages for interspersed repeats and the two types of non-interspersed repeats are out of 100% of the genome. The LTRs are broken down in their subcategories (*copia*, *gypsy*, etc), with the percentages (out of all LTRs) shown.

	Iochroma cyaneum	*Capsicum annuum*	*Jaltomata sinuosa*	*Solanum lycopersicum*	*Nicotiana attenuata*	*Petunia axillaris*
Interspersed Repeats:	**68.69%**	**71.55%**	**72.64%**	**57.90%**	**61.73%**	**53.08%**
LTR	64.73%	60.79%	62.04%	39.49%	57.66%	45.98%
LINE	0.43%	0.48%	0.76%	0.99%	0.28%	0.74%
DNA	0.22%	1.25%	0.81%	0.84%	0.45%	1.48%
Unclassified	3.31%	9.03%	9.04%	16.58%	3.34%	4.88%
Other (non interspersed) repeats:						
Simple Repeats	**0.52%**	**0.57%**	**0.53%**	**1.16%**	**0.53%**	**0.90%**
Low complexity	**0.14%**	**0.14%**	**0.15%**	**0.24%**	**0.12%**	**0.18%**
Total repetitive content	**69.35%**	**72.26%**	**73.32%**	**59.30%**	**62.38%**	**54.16%**
LTR Subcategories						
copia	1.35%	8.57%	2.61%	10.75%	3.87%	7.88%
gypsy	80.73%	55.20%	68.39%	69.40%	60.37%	47.71%
Unknown	17.82%	35.91%	28.67%	19.12%	35.69%	44.24%
Caulimovirus	0.08%	0.32%	0.33%	0.73%	0.06%	0.15%
ERVK	0.00%	0.00%	0.00%	0.00%	0.00%	0.008%
Pao	0.00%	0.00%	0.001%	0.00%	0.00%	0.00%

Figure S1 Hi-C heatmap showing 12 assembled chromosomes.

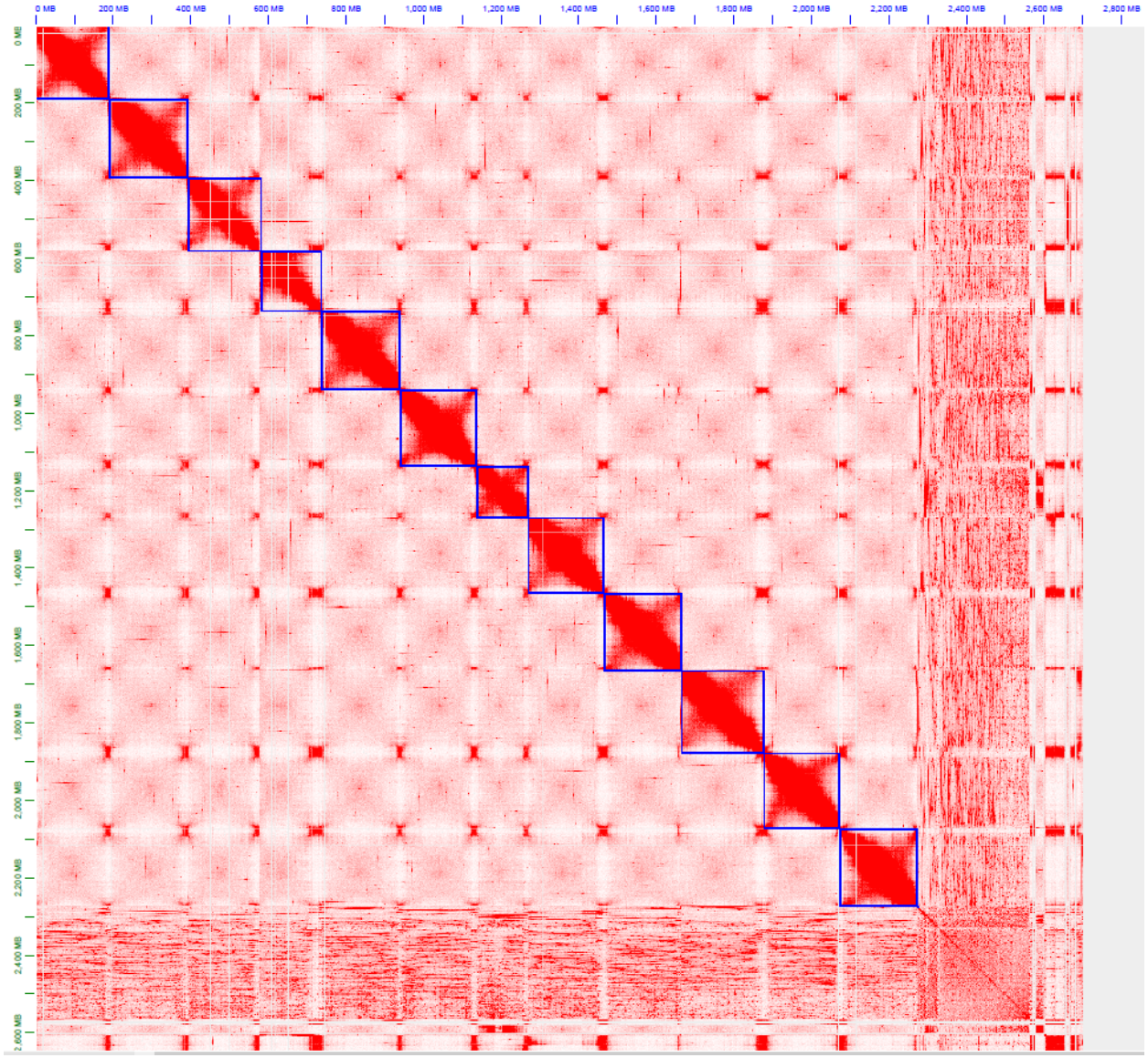


Figure S2 Directed subgraph for the top five gene ontology (GO) terms from enrichment analysis of the contracted gene families identified by CAFE. The level of significance for each GO term is indicated by the color, with yellow being non-significant, orange being moderately significant, and red being highly significant. The top five terms are shown in the red boxes. A descriptor and p-value are given for each GO term along with the level of enrichment (here the number of gene families in the contracted set compared to the entire set of gene families in *Iochroma*).

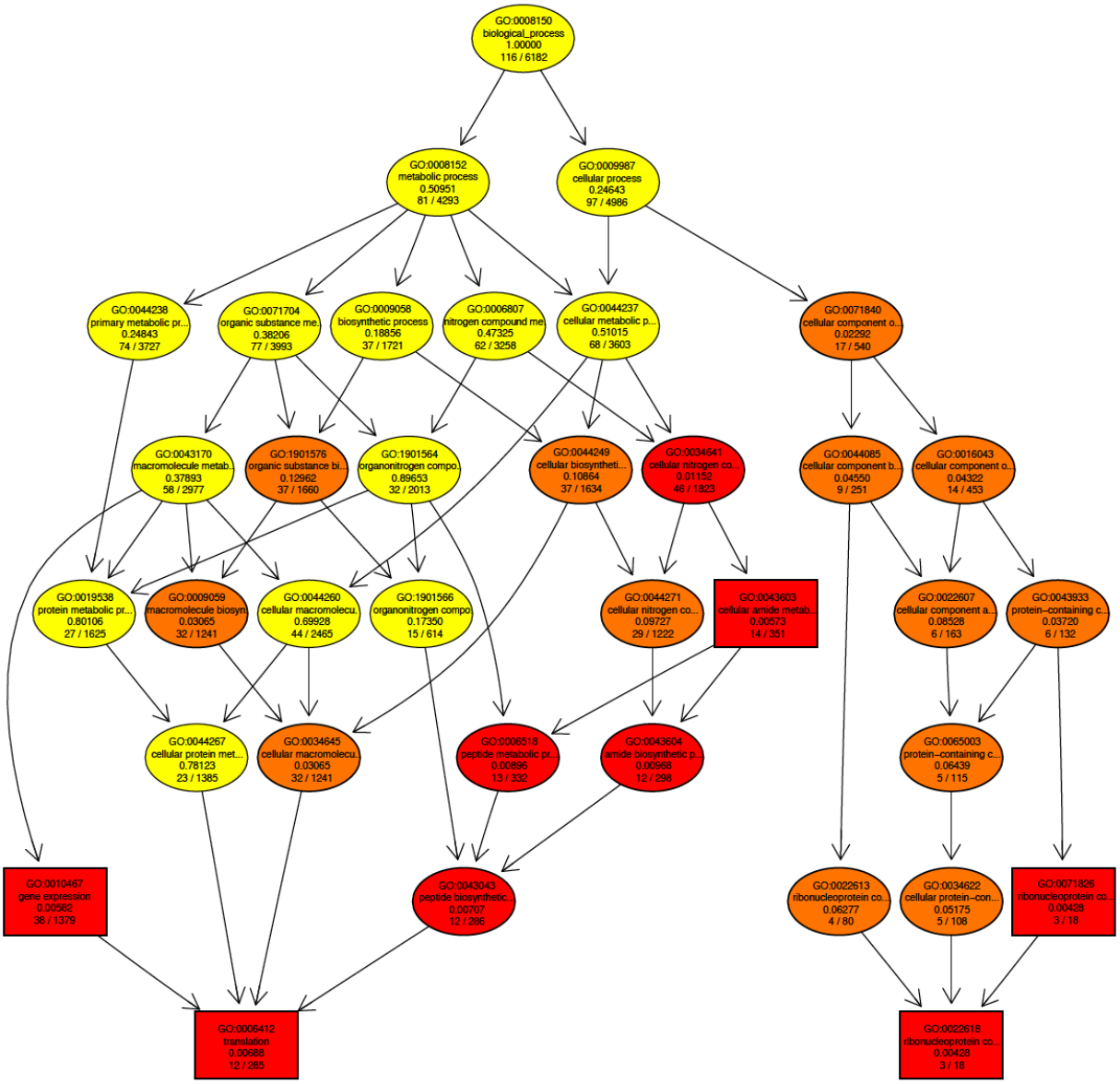


Figure S3 Directed subgraph for the top five gene ontology (GO) terms from enrichment analysis of the expanded gene families identified by CAFE. The level of significance for each GO term is indicated by the color, with yellow being non-significant, orange being moderately significant, and red being highly significant. The top five terms are shown in the red boxes. A descriptor and p-value are given for each GO term along with the level of enrichment (here the number of gene families in the expanded set compared to the entire set of gene families in *Iochroma*).

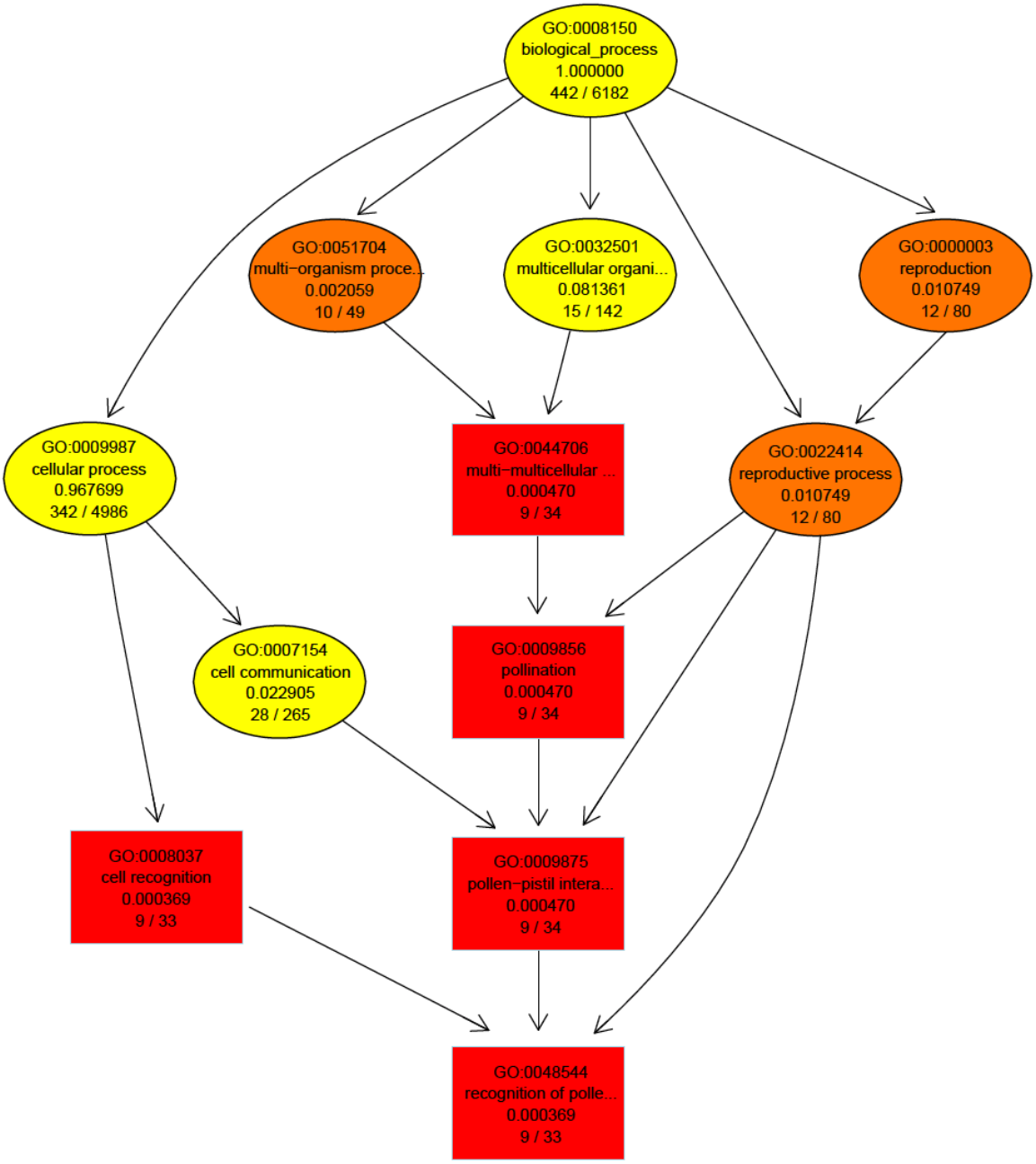


Figure S4 Patterns of synteny between *Iochroma* and pepper. Pepper chromosomes (*C. annuum* assembly v. 1.55) and *Iochroma cyaneum* (v. 1.0) are shown with lines connecting syntenic segments. Line coloring follows pepper.

Figure S5 Patterns of synteny between tomato and pepper. Tomato chromosomes (*S. lycopersicum* assembly v.4.0) and pepper chromosomes (*C. annuum* assembly v. 1.55) are shown with lines connecting syntenic segments. Line coloring follows tomato.