

RESEARCH ARTICLE OPEN ACCESS

Three Novel Genomes Broaden the Wild Side of the *Capsicum* Pangenome

Christina Papastolopoulou^{1,2}  | Ronald Nieuwenhuis²  | Sven Warris²  | Linda V. Bakker²  | Jan van Haarst² | Jan Cordewener² | Thamara Hesselink² | Hetty van den Broeck²  | Willem van Dooijeweert³ | Hans de Jong^{3,4}  | Julapark Chunwongse⁵ | Sara Diaz Trivino²  | Elio Schijlen²  | Dick de Ridder¹  | Sandra Smit¹  | Sander A. Peters^{1,2} 

¹Bioinformatics Group, Wageningen University & Research, Wageningen, the Netherlands | ²Wageningen Plant Research, Cluster Applied Bioinformatics, Wageningen University & Research, Wageningen, the Netherlands | ³Centre for Genetic Resources, Wageningen University & Research, Wageningen, the Netherlands | ⁴Laboratory of Genetics, Wageningen University & Research, Wageningen, the Netherlands | ⁵Center for Agricultural Biotechnology, Kasetsart University, Nakhon Pathom, Thailand

Received: 6 June 2025 | **Revised:** 27 October 2025 | **Accepted:** 3 November 2025

Keywords: *Capsicum annuum* | *Capsicum chacoense* | *Capsicum galapagoense* | pangenomics | pepper | wild species

ABSTRACT

This study presents three genome assemblies within the *Capsicum* genus, enabling comprehensive comparative analyses for the *Annuum* and *Baccatum* complexes within the genus. We produced highly continuous assemblies of the nuclear genomes and complete chloroplast assemblies. Subsequent genome annotation identified 34,580 genes in nonpungent *C. annuum* cv. ECW, and 32,704 and 33,994 genes in pungent *C. chacoense* and *C. galapagoense*, respectively. These assemblies, including the first complete genomes for *C. chacoense* and *C. galapagoense*, provide additional genomic resolution within the *Capsicum* genus. The novel genomes were analyzed within a pangenomic framework, integrating 16 *Capsicum* genomes across the *Annuum*, *Baccatum*, and *Pubescens* complexes. Homology grouping was used to identify core, accessory and unique genes and showed a wide spectrum of genetic diversity, particularly in homology groups exclusive to *C. chacoense* and *C. galapagoense*. Out of 79,267 homology groups identified, 13% were core groups, present in all accessions, corresponding to approximately 30% of core genes per genome. Comparative analyses revealed distinct species and genus-specific genomic characteristics. Additionally, we used the graph pangenome to illustrate locus-level exploration by examining the *Pun1* locus associated with capsaicinoid biosynthesis, identifying multiple *Pun1*-like genes including their genomic position and homology information. The integration of these new resources into a dynamic *Capsicum* pangenome framework provides a versatile platform for extracting genetic information relevant to both fundamental research and breeding applications.

1 | Introduction

Plant comparative genomics, evolutionary studies, and genomics-assisted precision breeding benefit from the rapidly growing availability of sequencing data and the development of high-quality genome assemblies. However, these research fields have also highlighted limitations of comparisons against a single reference genome, as plant genomes can contain up to 90%

of repetitive sequences and have complex structural variation (SV), even within the species boundary. Pangenomic analyses, comparing all constituent genomes at once rather than each individually against a reference, have therefore recently become mainstream in plant genomics (Schreiber et al. 2024).

Capsicum is an important food crop, with an annual worldwide production and area harvested amounting to close to 40 million

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Plant Direct* published by American Society of Plant Biologists and the Society for Experimental Biology and John Wiley & Sons Ltd.

tonnes and over 2 million hectares, respectively (<https://www.fao.org/faostat/en/#data/QCL>). The *Capsicum* genus belongs to the Solanaceae family, including (among others) tomato, potato, eggplant, tobacco, and petunia. *Capsicum* is well known for its nutritional, medicinal, ornamental, and pharmacological values and is renowned for its flavoring agent capsaicin, the characteristic ingredient of the spice. Its high vitamin A and C content and richness in iron, potassium and magnesium make *Capsicum* an important nutritional crop. Antioxidant, anti-mutagenic, anti-carcinogenic, and immunosuppressive activities have been attributed to capsaicin as well (Saxena et al. 2016). In the *Capsicum* genus, currently 42 *Capsicum* spp. have been identified, which have been divided into two groups based on their chromosome numbers, $2n=24$ and $2n=26$, respectively—except for *Capsicum annuum* var. *glabriusculum* that has been profiled as a tetraploid ($2n=4x=48$) (Back et al. 2024; Jarret et al. 2019; Mongkolporn and Taylor 2011; Pickersgill 1991). *Capsicum* genomes are the second largest of the Solanaceae family, with genome sizes between 3–3.9 Gb, of which approximately 80% is comprised of repetitive elements. Major genomic rearrangements, including extensive chromosomal translocations and genomic rearrangements, contribute to a rich genetic diversity in the *Capsicum* species (Sanatombi 2024).

Several *Capsicum* genomes have been assembled over the past decade. Since the release of the first draft genome of *C. annuum* CM334 (Kim et al. 2014), followed by draft genomes of *C. chinense* and *C. baccatum* (Kim et al. 2017), advances in sequencing technologies have progressively improved assembly quality, leading to chromosome-scale and even telomere-to-telomere genome assemblies (Hulse-Kemp et al. 2018; Qin et al. 2014; Liu et al. 2023; Chen et al. 2024). At present, genome assemblies are available for five *Capsicum* species. However, the majority represent domesticated accessions, with only limited coverage of wild relatives. This imbalance between cultivated and wild genome availability restricts our ability to capture the full genomic diversity of the genus and to leverage the allelic richness of wild species. Therefore, additional high-quality genome assemblies spanning diverse wild *Capsicum* accessions are needed. Integrating these new genomes into a pangenome, a comprehensive genomic framework that incorporates variation across multiple individuals or species, can further enhance our ability to explore genetic diversity at both the species and genus levels.

Phylogenetic relationships within the *Capsicum* genus were studied by Moscone et al. (2007), who proposed the *Annuum* (*C. annuum*, *C. chacoense*, *C. chinense*, *C. frutescens*, and *C. galapagoense*), *Baccatum* (*C. baccatum* and *C. praetermissum*), and *Pubescens* (*C. pubescens*, *C. eximium* and *C. cardenasii* and *C. tovarii*) complexes, respectively. Among these species, *C. annuum* and the closely related *C. chinense* and *C. frutescens* are used for food production, with *C. annuum* being the most widely used crop. *C. baccatum* and *C. pubescens* are grown for commercial purposes as well, although to a limited extent. Nonetheless, these closely related species are becoming increasingly important for breeding, as they harbor beneficial traits that can be used for crop improvement.

To benefit from the allelic richness and to advance (introgression) breeding for *Capsicum* crop improvement, assessment of

genetic diversity and insight in chromosome topology in the *Capsicum* genus is essential. Disclosing the genetic diversity and chromosome synteny and collinearity is not only critical for precision breeding and selection of compatible breeding parents for interspecific hybridization breeding but is also required to achieve a deeper understanding of the genetic basis of complex traits. However, while genetic diversity has been studied in terms of SNPs and small insertions and deletions (InDels), SVs, are still largely undisclosed. In this respect, de novo assembly comparisons are preferred over biased read mapping to capture large SVs (Khan et al. 2020). Other approaches to study genetic diversity, synteny and collinearity involve the use of genetic and physical marker information to position and order alleles. However, genetic markers often have discrepant physical and genetic positions in the genome, complicating the precise delineation of genetic loci. Furthermore, most genetic maps for *Capsicum* have been constructed using intra- and interspecific populations that have very few genetic markers in common, and sometimes chromosomes are not congruent with intraspecific maps (Lefebvre et al. 2002; Mongkolporn and Taylor 2011).

The complex genomic diversity within and among *Capsicum* species poses significant challenges for introgressive hybridization breeding. The observed incongruency between linkage maps, such as that between *C. frutescens* and *C. annuum*, highlights the difficulties in predicting genetic outcomes when translocations occur. This genetic variability can lead to unexpected traits and incompatibilities, making it hard for breeders to develop consistent and desirable varieties. Moreover, while recent pangenome studies have identified valuable genetic variations, they still rely on a single reference genome as an anchor for comparative analyses or are limited to gene-based analysis (Lee et al. 2022; Liu et al. 2023; Ou et al. 2018), which can limit the discovery of novel traits and hinder the full exploration of genomic diversity. A reference-free approach could allow for a more comprehensive understanding of the genetic landscape, enabling breeders to tap into the full potential of the *Capsicum* clade.

To advance our understanding of genomic diversity in the *Capsicum* genus, we present three high-quality de novo genome sequences, including two wild *Capsicum* species (*C. chacoense* and *C. galapagoense*), complete with detailed structural and functional annotations. We present a pangenome at the genus level, utilizing PanTools, an alignment-free graph-based approach that effectively eliminates reference bias, ensuring a more accurate representation of genetic variation (Jonkheer et al. 2022; Sheikhezadeh Anari et al. 2018; Sheikhezadeh et al. 2016). The pangenome incorporates structural, functional, and phenotypic annotations, allowing researchers to extract valuable biological insights during analysis. Through this comprehensive framework, we analyze genetic diversity across domesticated and wild *Capsicum* species, as well as show examples of using the pangenome to explore loci of interest across the *Capsicum* genomes. This pangenome facilitates the understanding and mining of *Capsicum* diversification but also lays a foundation for advanced breeding initiatives, with future applications in enhancing food security and improving the livelihoods of communities worldwide.

2.1 | Three De Novo *Capsicum* Genome Assemblies

To enhance our understanding of the genetic diversity within the *Capsicum* genus, we aimed to assemble the genomes of the following three distinct pepper lines: the nonpungent *C. annuum* cv. Early Cal Wonder (ECW) (CGN16883) and the wild and pungent *C. chacoense* (CGN21477, PI260429) and *C. galapagoense* (CGN22208, AC1501). The goal was to generate genomes that enrich the Annuum and Baccatum complexes within the *Capsicum* genus with both cultivated and wild species. These new genomes complement previously established ones (referenced in Chen et al. 2024; Hulse-Kemp et al. 2018; Kim et al. 2014, 2017; Lee et al. 2022; Liao et al. 2022; Liu et al. 2023; Qin et al. 2014; Shirasawa et al. 2023), providing a more comprehensive resource for comparative genomics and evolutionary studies in *Capsicum*.

For de novo genome assembly, we combined multiple (sequencing) technologies. This included large-insert sequencing with PacBio Sequel technology, Illumina-based 10x Genomics linked-reads, and Bionano optical mapping technology (Table S1). These data were integrated to assemble contigs and scaffold them into near chromosome-length sequences. For *C. chacoense* and *C. galapagoense*, this approach produced large scaffolds, with alignment to a chromosome-level *Capsicum* reference indicating that one to three large scaffolds corresponded to each chromosome (Figure S1). For *C. annuum* ECW, we further anchored the scaffolds using the genetic map from Hulse-Kemp et al. 2016, resulting in a near chromosome-level assembly. Ultimately, we obtained genomes with total lengths ranging from 3.04 to 3.13 Gb and an N50 final scaffold length from 213 to 256 Mb (Figure 1, Table S2).

Completeness of the reconstructed genome sequence was assessed by determining the number of BUSCO (Benchmarking Universal Single-Copy Orthologs) genes. We find at least 97.9% out of a total of 5950 complete single-copy genes in the Solanales odb10 data set (Manni, Berkeley, Seppey, and Zdobnov 2021; Manni, Berkeley, Seppey, Simão, et al. 2021), indicating a high-quality recovery of the *C. annuum*, *C. chacoense* and *C. galapagoense* genomes (Figure 1, Table S2). We assessed the

completeness of the chromosomes in *C. annuum* ECW with 5303 markers aligning confidently to the 12 assembled chromosomes, indicating the successful scaffolding and reconstruction of the genome. In chromosomes 1 and 8, the alignment of the markers was inconsistent with the chromosome reconstruction due to a known translocation in the interspecific marker set, while in chromosome 2 a part of the chromosome is not covered by markers (Figure S2). Next to that, we identified nine out of the 24 telomeres in the assembly (Figure S3). All three genomes displayed a repetitive sequence content of 80%, which is similar to previously published *Capsicum* genomes (Chen et al. 2024; Kim et al. 2021, 2017; Kim et al. 2014; Lee et al. 2022; Liao et al. 2022) (Table S3).

In addition to the nuclear genome assemblies, the complete chloroplast genome assemblies (a single contig for the 3 *Capsicum* species, measuring 156 Kb) demonstrated a high consistency with previously published chloroplast genomes (Figures S4, S5, S6). This reinforces the reliability of the genomic data obtained and highlights the stability of chloroplast genome structure across these species.

Next, we combined ab initio predicted gene models with evidence-based predictions, annotating 34,580 protein-coding genes for *C. annuum*, 32,704 for *C. chacoense*, and 33,994 for *C. galapagoense*, with 98.3% of protein BUSCOs or more (Figure 1, Table S2). This result indicates that we have achieved a high level of structural annotation completeness.

2.2 | Comprehensive *Capsicum* Pangenome Analysis Reveals High Genetic Diversity and Evolutionary Relationships Across Domesticated and Wild Species

In total, 16 annotated genome assemblies from domesticated and wild *Capsicum* species were selected to construct a *Capsicum*-clade pangenome representing the Annuum, Baccatum, and Pubescens complexes. Specifically, nine *C. annuum* domesticated cultivars were included, along with *C. chinense* PI159236, *C. baccatum* var. *pendulum* PI632928, and the wild *C. annuum* var. *glabriusculum*, *C. chacoense* CGN21477, *C. galapagoense* CGN22208, *C. baccatum* var. *baccatum* PBC81, and *C. pubescens*

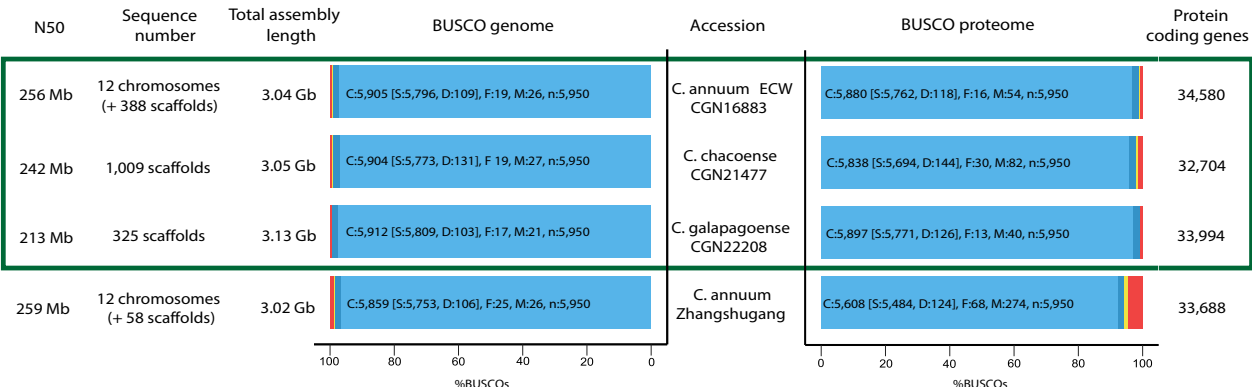


FIGURE 1 | Genome assembly and annotation metrics for the three new *Capsicum* genomes. Genome assembly and structural annotation statistics including BUSCO results for three de novo *Capsicum* genomes, compared to the public *C. annuum* Zhangshugang assembly. The BUSCO score was calculated using the Solanales_odb10 set.

Grif 1614. These assemblies were selected based on their high quality and phylogenetic representativeness across the three *Capsicum* clades, ensuring coverage of both domesticated and wild diversity while maintaining consistency with widely used reference genomes (Table S4).

To explore the evolutionary relationship of the newly introduced genomes to the rest of the published *Capsicum* genomes, we estimated genomic distances using two methods. The core-gene approach focuses on the conserved (core) genes shared among the genomes, providing insights into their evolutionary history, while the *k*-mer mash distance approach uses sequence-based comparisons to estimate genomic distances. These methods were chosen to ensure a comprehensive analysis from both gene-based and full genome-based perspectives. The resulting phylogenies were depicted using a bootstrap core SNP tree for the core-gene phylogeny and a clustered heatmap for the *k*-mer mash distance genome relationships (Figures 2 and S7). Both methods confirm the existing phylogenies and place *C. annuum* and *C. galapagoense* within the Annuum clade, while *C.*

chacoense is placed in the Baccatum clade, which is consistent with previously published results (Carrizo García et al. 2016; D'Agostino et al. 2018; Liu et al. 2023).

Homology grouping enabled us to study the diversity in genes captured in this pangenome. The 545,775 genes for 16 *Capsicum* accessions were assigned to 79,267 homology groups (HGs), which can be classified based on presence/absence variation in the accession (Figure 3, Table S5). There are 10,368 core groups (13.1%) present in all accessions, corresponding to 29.3%–37.6% of core genes per accession (Figure S8). GO enrichment analysis of these core genes revealed their involvement in essential biological processes including transport, photosynthesis, DNA/RNA processes, protein modification/metabolism, and development and defense (Table S6). In contrast to the highly reliable core genes, there are 21,444 unique HGs (only found in one genome), which contain variation in specific processes such as metabolic pathways for specific compounds, developmental processes specific to reproduction and plant defense (Table S7). However, these putative genes

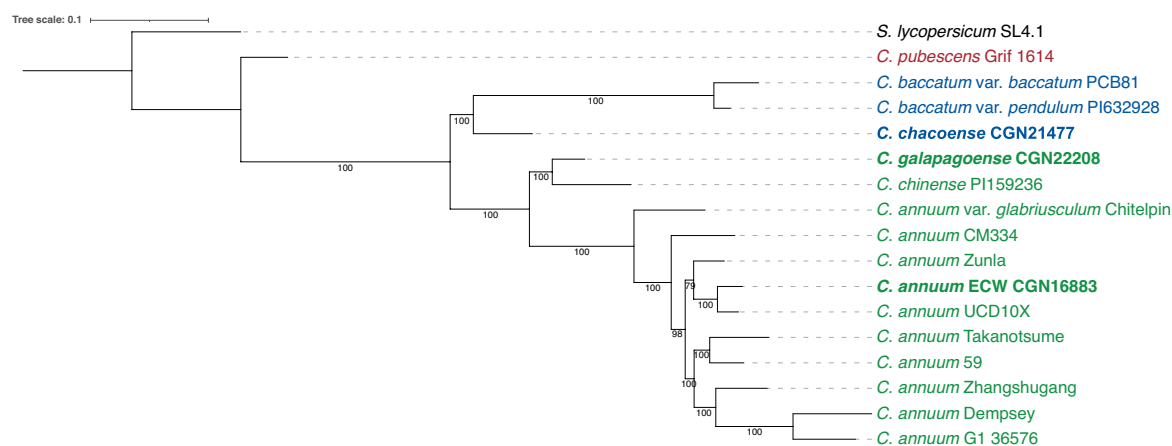


FIGURE 2 | Core phylogeny of *Capsicum* proteomes. A maximum likelihood (ML) core SNP tree was constructed based on 6023 single-copy ortholog groups. Bootstrap values were calculated based on 1000 iterations, with *Solanum lycopersicum* SL4.1 serving as the outgroup. The following three accessions generated in this study are highlighted in bold, with their respective clades distinguished by color: Annuum (green), Baccatum (blue), and Pubescens (red).

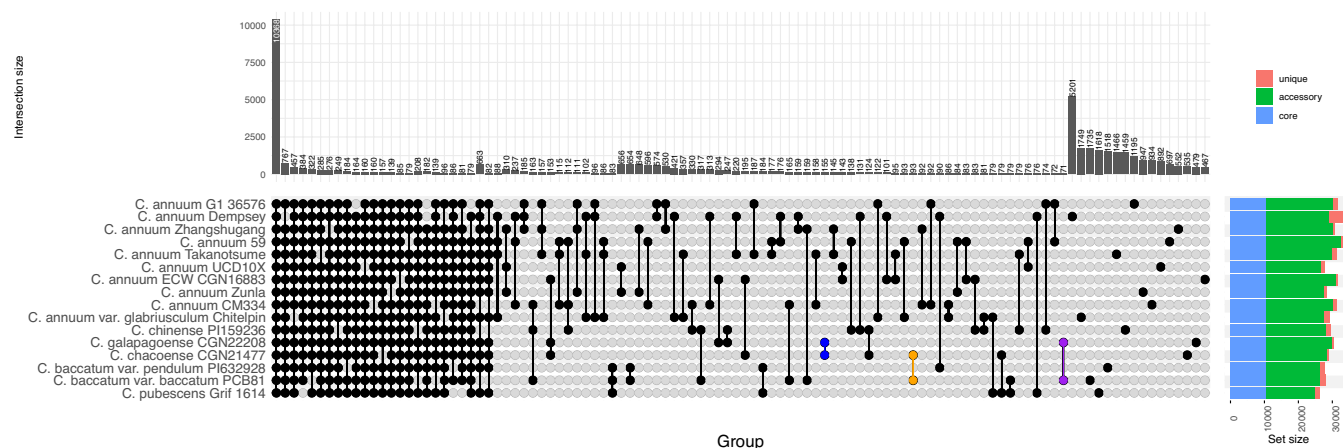


FIGURE 3 | Core, accessory, and unique homology groups in the *Capsicum* pangenome. Upset plot showing the most populated homology groups in the pangenome. The bar plot shows the distribution of the three homology group categories per genome. The unique homology groups between the wild *C. chacoense* and *C. galapagoense*, *C. chacoense* and *C. baccatum* var. *baccatum* and *C. galapagoense* and *C. baccatum* var. *baccatum* are highlighted in blue, orange, and purple, respectively.

(corresponding to 1.4%–14.6% per accession) should be treated with care, and the inspection of supporting evidence like function, expression, or experimental validation is highly recommended. More reliable genetic variation is found in the 47,455 (59.9%) accessory groups, occurring in a subset of accessions (56.1%–66.3% of genes per accession). GO analysis of these accessory groups showed an enrichment for GO terms related to defense response, specialized developmental processes, and energy metabolism (Table S8).

These genes were further subdivided based on their presence/absence pattern in wild and domesticated accessions (as defined in Table S4). Groups specific to domesticated genomes showed enrichment for terms related to growth and yield, adaptation to cultivation conditions, stress tolerance, plant architecture, and changes in reproductive development (Table S9). In contrast, the 5008 groups specific to wild species showed enrichment for more fundamental survival-oriented processes with emphasis on defense response and energy management (Table S10). Interestingly, 510 accessory groups of genes are only found in a combination of the four wild accessions, of which 155 apparently are exclusive to both *C. chacoense* and *C. galapagoense*, 93 to *C. chacoense* and *C. baccatum* var. *baccatum*, and 71 to *C. galapagoense* and *C. baccatum* var. *baccatum* genomes (Figure 3). Notably, *C. chacoense* and *C. galapagoense* contain 1169 unique and accessory groups specific to themselves, with GO terms enriched for binding activities and defense response (Table S11). Considering that domestication in general selects for traits beneficial to humans (e.g., growth, development, reproduction), while wild species maintain a broader range for traits related to survival (e.g., defense and stress tolerance), the observed differences in enrichment therefore possibly reflect selection of human-preferred traits historically targeted by *Capsicum* domestication. These results highlight unique genetic features of the novel genomes, strengthening the wild and domesticated parts of the pangenome with potentially interesting genes/alleles for breeding.

We used HGs as a statistic to assess the saturation of our pangenome. By progressively adding genomes and subsequently analyzing the number of core, accessory, and unique genes we can use the HGs as a proxy for “new” genes. We plotted the number of core HGs (present in all genomes so far) as a function of genome number. The declining and stabilizing core curve is clearly plateauing (Figure S8), indicating the level of saturation for the core genes of the *Capsicum* genus pangenome.

2.3 | Chromosome Structural Variations Reveal Evolutionary Divergence in *Capsicum*

Previous comparative sequence alignment studies revealed species- and genus-specific collinearity breaks in the *Solanaceae* clade. These synteny breaks between tomato, pepper, and potato were validated through comparative FISH analyses, ruling out the possibility that collinearity disruptions were due to assembly errors (Peters et al. 2012).

In this study, we analyzed the chromosomes structure of *C. galapagoense*, *C. chacoense*, *C. annuum* ECW, *C. pubescens*, and *C. baccatum* genomes, using comparative sequence alignment

with the *Solanum lycopersicum* cv. Heinz reference genome (Su et al. 2021). Dot plot comparison (Figures S9–S12) revealed similar topologies among these five *Capsicum* genomes, aligning with previously observed patterns in chromosomes 2L, 6S, 10L, and 11L of *C. annuum*. These findings support the existence of genus-specific synteny breaks between *Capsicum* accessions in this study and *S. lycopersicum*, while these structural rearrangements apparently are shared between *Capsicum* and potato (Peters et al. 2012). This indicates that the observed synteny breaks in chromosome 2L are specific to the tomato clade and arose after the split of tomato from the common ancestor of *Capsicum* and potato (Peters et al. 2012). Beyond these clade-specific patterns, we observed additional species-specific SVs. *C. chacoense* exhibited two minor inversions in chromosome 2L compared to *C. galapagoense*, *C. annuum* ECW, *C. pubescens*, and *C. baccatum*, which apparently are specific to *C. chacoense*, while *C. pubescens* exhibited additional SVs in chromosome 11L. Overall, *C. chacoense* and *C. baccatum* showed more frequent SV than *C. galapagoense* (Figures S9, S10, S11, S12), consistent with its more distant phylogenetic relationship to *C. annuum* (Figure 2). These findings underscore both clade-specific synteny breaks and species-specific SVs in the *Capsicum* genus. Moreover, the development of a comprehensive *Capsicum* pangenome facilitates more detailed analysis and comparison of these SVs across species.

2.4 | Copy Number Variation in the *Capsicum* pangenome for the *Pun1* locus

Our comprehensive *Capsicum* data set captures rich, previously unexplored genomic variation across three clades, spanning both wild and domesticated accessions, including both pungent and nonpungent varieties (Table S4). The use of this pangenome allows for efficient comparison and analysis of variation in key genes and loci across different genomes, offering insights into genetic diversity at a broader scale. In this section, we describe a use case on pungency.

Pungency is a complex polygenic trait that is regulated by multiple quantitative trait loci involving a pathway of at least nine essential biosynthetic genes. As this trait is of major interest in pepper breeding programs, we selected the *Pun1* locus as a showcase to examine genetic diversity across both pungent and nonpungent accessions (Chen et al. 2024; Stewart et al. 2007; Stewart et al. 2005). Several studies established the *Pun1* locus position on chromosome 2 (Chen et al. 2024; Stewart et al. 2005).

Looking across genomes, chromosome 2 shows largely conserved synteny, with minor variations, across 15 of the 16 genomes, except for *C. galapagoense* (Figures 4, Figure S13). In *C. galapagoense* scaffolds 6 and 12 align partially to chromosome 2, but partially to other chromosomes (Figure S14), suggesting a chromosomal rearrangement. Verifying the situation, consistent long-read coverage across both breakpoints supports the accuracy of the *C. galapagoense* assembly, indicating that these represent a genuine rearrangement rather than an assembly artifact (Figure S15).

Next, zooming in to the *Pun1* locus, initial examination of SVs in our pangenome revealed that all accessions show collinear

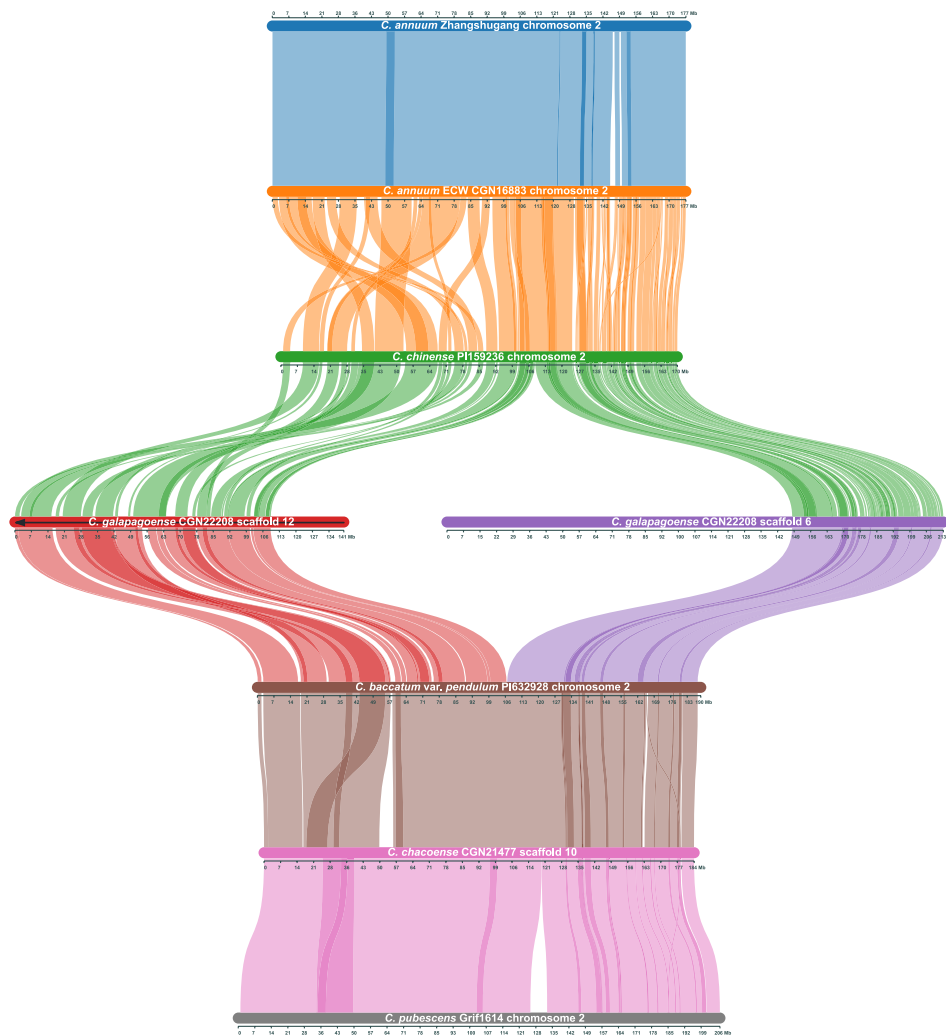


FIGURE 4 | Synteny analysis of chromosome 2 across seven *Capsicum* assemblies. Comparison of chromosome 2 syntenic relationships between *C. annuum* cv. Zhangshugang (reference) and six *Capsicum* genomes: *C. annuum* cv. ECW, *C. chinense*, *C. galapagoense*, *C. baccatum* var. *pendulum*, *C. chacoense*, and *C. pubescens*. The analysis reveals high collinearity in chromosome 2 structure across all examined species, except for *C. galapagoense*, where two scaffolds partially align to chromosome 2 of the others; this scaffold structure is supported by long-read data.

organization, without significant structural polymorphisms (Figures 5 and S16). Nevertheless, a combination of the homology grouping and a BLASTp search using the *C. annuum* CM334 *Pun1* protein sequence as a query revealed up to nine *Pun1*-like genes, clustering into nine distinct HGs, with all 16 genomes containing a copy in the same HG as the query sequence. A maximum likelihood phylogenetic tree confirmed these HGs and highlighted sequence diversity among the copies. Based on sequence similarity to previous studies, we identified two major HGs: group A, homologous to capsaicin synthase 1 (CS1), and group B, homologous to capsaicin synthase 2 (CS2). Among our de novo genomes, *C. annuum* ECW and *C. galapagoense* contained eight and nine copies, respectively, while *C. chacoense* had four copies (Figures 5, S16, S17). However, further analysis is needed to determine which *Pun1*-like copies are functionally active to link these findings to pungency (Chen et al. 2024).

Nonetheless, the above results are illustrative for a graph pangenome-based macro- and micro-synteny analyses, providing critical insights into genomic complexity.

3 | Discussion

Our study significantly enhances the genomic resources available for the *Capsicum* genus by providing high-quality genome assemblies for three key species: the domesticated *C. annuum* and two wild relatives, *C. chacoense* and *C. galapagoense*. The inclusion of these wild species is significant as they represent two major evolutionary clades within *Capsicum*, harboring traits with high potential for pepper breeding and adaptive studies, traits that were previously inaccessible because of the lack of genomic resources. By leveraging advanced sequencing technologies, we generated a near chromosome-level and two highly contiguous genome assemblies, as reflected by high N50 contig lengths and the BUSCO results. These assemblies and their detailed annotations represent a robust foundation for comparative genomic analyses and advance our understanding of genetic diversity within the *Capsicum* genus opening new avenues for breeding applications and phylogenetic research, providing crucial insights into the evolutionary relationships and trait variability across *Capsicum* species. To further explore genomic diversity, we

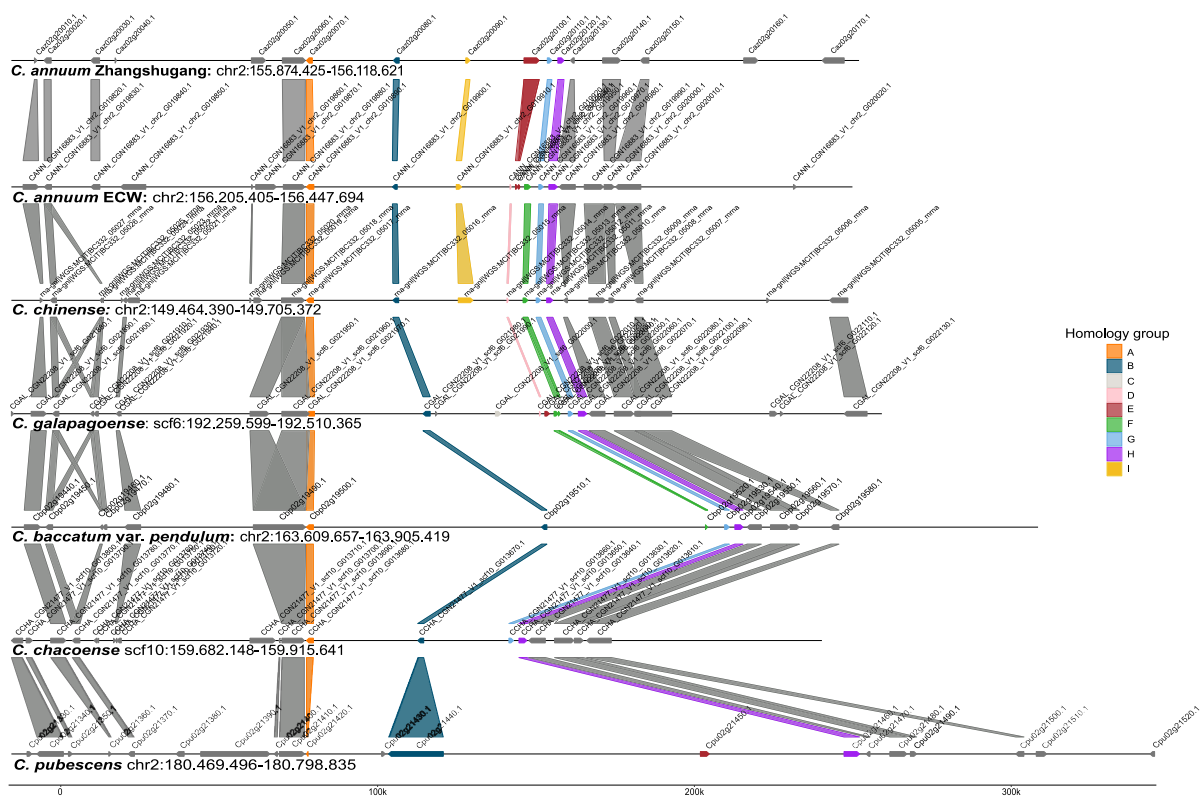


FIGURE 5 | Synteny plot of the *Pun1* region for a subset of the *Capsicum* genomes. mRNAs grouped in the same homology group are connected and mRNAs that had a similarity hit to the *Pun1* gene of *C. annuum* CM334 are marked with the same color per homology group.

constructed a *Capsicum*-clade pangenome, incorporating 16 genome assemblies along with their coding genes and their functions spanning both domesticated and wild species. This dataset includes nine *C. annuum* cultivars alongside a diverse array of accessions. By integrating these additional genomes, we captured significant genomic diversity and complexity. Compared to previous studies, such as Lee et al. (2022), which focused primarily on a core *C. annuum* collection with 11 additional accessions, our approach resulted in a substantially larger number of HGs and a significant reduction in core groups—from 20,000 to just 10,000 (Figure 2). Furthermore, the number of core HGs for the different genome combinations within the *Capsicum* pan-proteome clearly levels off, while the dispensable (unique and accessory) still increases (Figure S8), suggesting that the genus-level *Capsicum* pangenome is fully capturing the core genes of the genus but more genomes are needed to capture all the variation. These results provide deeper insights into allelic diversity at the genus level, offering a valuable resource for identifying trait-associated genes and informing breeding strategies. Furthermore, the resources presented here will be supportive of future evolution and genetic diversity studies as well as the development of advanced breeding applications.

The use of the pangenome enables us to efficiently analyze and compare variations in key genes in loci of interest across the genomes, offering valuable insights into genetic diversity on a broader scale. For example, pungency, a complex polygenic trait regulated by multiple quantitative trait loci and involving at least nine essential biosynthetic genes, is of key interest in pepper breeding programs. Here, we focused on

the *Pun1* locus, a critical determinant of pungency, and examined its genetic diversity across both pungent and nonpungent accessions (Chen et al. 2024; Stewart et al. 2007; Stewart et al. 2005). We used the *Pun1* protein from accession CM334 as a query to identify homologs across a range of accessions and the pangenome to extract the relevant genomic information. This example highlights the interconnection between pangenomes, synteny analysis, and gene model correction. While we identified highly similar proteins in 16 accessions, we revealed copy number variation and distinct differences between *Pun1* gene copies. Our attention was drawn to the *Pun1* gene in *C. chacoense*, which at first had an unusually large size. Only after inspection of aligned full-length transcriptome data could the *Pun1* gene annotation for *C. chacoense* be corrected. This example illustrates that pangenomic approaches do not only provide insight into gene arrangements and CNV but also may draw attention to discrepancies in gene annotations and/or genomic sequences that need closer inspection and possibly correction.

However, the pungency trait is more complex by essentially being a quantitative trait that is affected by both genetic and environmental factors. Besides *Pun1*, several capsaicinoid biosynthesis pathway genes have been identified, and multiple QTLs have been associated with the localization of capsaicin in placental and pericarp tissue. Nonetheless, given the transcriptome data that was included in recent publications (Back et al. 2024; Chen et al. 2024), identification of transcriptionally active *Pun1* copies may provide more context for future presence-absence variation (PAV) studies and linking the expressions of gene copies to the pungency trait.

Structural variation analysis can leverage complementary methods, each with unique strengths and limitations. Sequence-based approaches, such as whole-genome sequencing, provide a comprehensive and linear representation of SV topologies, capturing the detailed arrangement and structure of genomic segments. In contrast, BAC-FISH (bacterial artificial chromosome-fluorescence in situ hybridization) analysis previously offered direct physical evidence of SVs in Solanaceae spp. including *Capsicum* (Peters et al. 2012) by visualizing the spatial arrangement of specific markers. While BAC-FISH effectively confirmed the presence and location of SVs with high confidence, it lacked the resolution to elucidate the complete topology of the SV segments located between markers.

Conversely, the sequence comparison in this study provides additional indirect evidence of genus and species-specific SVs, although it excels at reconstructing their detailed structural organization. Together, these methods provide complementary insights, with BAC-FISH validating SV presence and sequence analysis elucidating its topology in finer detail. In general, the detection of SV is highly valuable, because it provides both opportunities for enhancing genetic diversity and challenges in managing genomic integrity during introgressive hybridization breeding. Previous data suggest that long-distance intrachromosomal rearrangements and local gene rearrangements have frequently evolved during speciation in the Solanaceae family, with small changes being more prevalent than large-scale differences (Peters et al. 2012).

This disruption of gene adjacency and synteny, often due to gene and LTR retrotransposon insertions (Peters et al. 2012), has significant implications for hybridization efforts. A study of 12 *C. annuum* accessions, including sweet, hot, and blocky-type peppers, revealed extensive SVs associated with important traits such as virus resistance, pungency, and fruit morphology (Lee et al. 2022). Notably, transposable elements (TEs) in peppers are highly correlated with SVs and have played a significant role in shaping *Capsicum* genomic diversity (Lee et al. 2022; Peters et al. 2012).

While SVs can enhance genetic diversity and the adaptive potential of hybrids, they can also introduce crossing barriers and linkage drag, which limits the use of germplasm for introgressive hybridization. Understanding the effects of SVs is thus imperative as they can influence breeding efforts. For instance, the suppression of recombination in interspecific hybrids between L-type tomato and S-type *Solanum lycopersicoides* or *S. sitiens* chromosome 10 is explained by a large paracentric inversion (Chetelat et al. 1997; Pertuzé et al. 2002). In the context of *Capsicum*, the observed large SVs between our newly assembled genomes and the cultivated *C. annuum* genomes (Figures S1 and S9) may explain the compatible and incompatible crossing outcomes between *C. annuum*, *C. frutescens*, and *C. chinense* (Hazarika et al. 2023). Moreover, we achieved a significant advancement enabled by the newly provided genomic resources, in particular the ability to identify species- and genus-specific SVs, which can uncover unique genetic adaptations and further inform targeted breeding strategies (Figures S9–S12). Such insights pave the way for leveraging species-specific traits in breeding programs, enhancing the precision and effectiveness of introgressive hybridization.

In summary, insights from our pangenome analysis, including the three new *Capsicum* genomes, provide valuable resources for elucidating genome content and organization. This understanding aids in evaluating collinearity, synteny, and linkage at both the chromosome and gene levels.

4 | Materials and Methods

4.1 | Plant Material

Seeds of three different *Capsicum* lines (*C. annuum* Early Cal Wonder [CGN16883], *C. chacoense* PI260429, *C. galapagoense* AC150) were obtained from Centre for Genetic Resources, the Netherlands (CGN) and were sown, germinated, and grown under standard greenhouse conditions at Wageningen University and Research.

4.2 | DNA Extraction and PacBio Sequencing

Plant leaves were harvested, snap frozen in liquid nitrogen and stored at -80°C . DNA isolation was done using 2–3 g per extraction following standard CTAB-based protocols. DNA quantity and purity was analyzed by Qubit (Invitrogen), OD values (NanoDrop), and Fragment Analyzer (Agilent). Obtained gDNA was used for two different SMRTbell library protocols (normal and express kit) without initial DNA shearing. Of each line, 15 μg of DNA was used for library preparation using the SMRTbell template prep kit 1.0 according to the manufacturer's protocol (PacBio; Procedure and Checklist Preparing > 30-kb SMRTbell libraries using Megaruptor Shearing and BluePippin Size selection on PacBio RSII and Sequel systems). Resulting libraries were used for SMRTbell-polymerase complexing using the Sequel Binding kit 2.0/2.1, sequencing primer v3/v4 and sequencing on multiple SMRT cells on a PacBio Sequel instrument. All sequencing of these “normal” SMRTbell libraries was done with approximately 5 pM on plate loading, using sequencing kit v2.1, 120-min immobilization time, and 600-min movie time per SMRT cell. In parallel, SMRTbell libraries were made using the SMRTbell Express Template Prep Kit (PacBio) following the Procedure and Checklist Preparing Greater than 15 Kb Libraries Using SMRTbell Express Template Preparation kit 2. Express SMRTbell libraries were subjected to DNA polymerase complexing using the Sequel binding kit v2.1/3.0, Sequel Polymerase 3.0, sequencing primer v4 and sequencing on multiple SMRT cells on a PacBio Sequel instrument. All sequencing of these “express” SMRTbell libraries was done with approximately 10–20 pM on plate loading, using sequencing kit 2.1/3.0, 120-min immobilization time and 600-min movie time per SMRT. We achieved a sequence coverage of 54 $\times$, 52 $\times$, and 75 $\times$ for the *C. annuum*, *C. chacoense*, and *C. galapagoense* reference genomes, respectively, given an expected size of 3.5 Gb for each genome (Table S1; Moscone et al. 2007).

4.3 | *Capsicum* Genome Sequencing, 10x Genomics Linked-Reads

Frozen leaf material of all three *Capsicum* lines was used for nuclei and ultrahigh molecular weight DNA extraction following

Bionano Prep Plant Tissue DNA Isolation Base Protocol. Of each line, one nanogram of ultrahigh molecular weight DNA was used for preparation of a linked-read library using Chromium genome library, gel bead and multiplexing kit, and a Chromium controller, following the Chromium genome reagent kits user. Linked-read libraries were used for 2x150 nt paired-end sequencing on Illumina HiSeq2500 and NovaSeq 6000 instruments.

4.4 | *Capsicum* Genome Optical Mapping

Of all three *Capsicum* lines, UHMW DNA was used for Bionano Genomics optical genome mapping. DNA was processed using the Bionano Genomics DLS Kit following Bionano-Prep-Direct-Label-and-Stain-DLS-Protocol rev D. Labeled DNA was loaded on multiple Saphyr v2 chips using separate flowcells for the individual lines and analyzed on a Bionano Genomics Saphyr instrument. Genome mapping data was further processed for molecule quality reporting, de novo assembly and hybrid scaffolding using Bionano Access v1.3 and Bionano Solve v3.3. Sequence-specific labeling of approximately 700-ng genomic DNA and subsequent backbone staining and DNA quantification for Bionano mapping was done using a direct label enzyme (DLE-1) according to the manufacturer protocol 30206F Bionano Prep Direct Label and Stain Protocol (<https://bionanogenomics.com/wp-content/uploads/2018/04/30206-BionanoPrep-Direct-Label-and-Stain-DLS-Protocol.pdf>). Chip loading and real-time analysis were carried out on a Bionano Genomics Saphyr analyzer according to the manufacturer system guide protocol 30143C (<https://bionanogenomics.com/wp-content/uploads/2017/10/30143Saphyr-System-User-Guide.pdf>). Molecule quality hybrid scaffold report was carried out using the Bionano Solve analysis pipeline (<https://bionanogenomics.com/support-page/data-analysis-documentation/>).

4.5 | *Capsicum* PacBio Isoseq

Different plant tissues (roots, flowers, fruits, and leaves) were collected, snap frozen in liquid nitrogen, and stored at -80°C . Of each sampled tissue, 50–70 mg was used for RNA extraction by Direct-zol RNA miniprep (Zymo Research) with minor modifications. RNA quantity and purity were analyzed by Qubit (Invitrogen), OD values (NanoDrop), and Bioanalyzer (Agilent). Of each line, RNA of fruit, flower and root was equally pooled to 2 μg , which was used for subsequent Iso-Seq library preparation according to the manufacturer's protocol (PacBio; Procedure Checklist Iso-Seq Template Preparation for Sequel Systems). RNA of leaves was used separately for Iso-Seq library preparation following the same procedures. Each RNA sample was split into two separate cDNA reactions to minimize the risk of potential length-based PCR bias during the full-length cDNA synthesis. First-strand cDNA synthesis employed the Clontech SMARTer PCR cDNA synthesis kit. Next, cDNA reactions were used for large-scale PCR amplification with 16 parallel reactions. PCR amplification of the cDNA was performed with 16 and 20 cycles for mixed (roots, flowers, and fruits) and leaf cDNA templates, respectively. Per template, parallel PCR reactions were pooled and purified, and two different Ampure XP elution fractions were collected and quantified by Qubit and

BioAnalyzer. Collected amplified cDNA fractions were pooled equimolar and approximately 5 pmol amplified cDNA was used for SMRTbell template preparation (SMRTbell template prep kit 1.0). SMRTbell-polymerase complexes were made using Sequel binding kit 3.0, Sequel polymerase 3.0, and Sequel sequencing primer kit v4. SMRTbell complexes were loaded at approximately 7–8 pM for sequencing on a PacBio Sequel system using multiple SMRT cells with 120-min immobilization, 240-min pre-extension, and 1200-min movie time per SMRT cell. Resulting subreads were used for circular consensus (CCS) read construction with PacBio SMRT link v12. Subsequently, CCS reads were clustered into high-quality isoforms, which were further collapsed to unique isoforms using the PacBio isoseq v 4.0.0 Bioconda package. Protein coding regions were predicted in the transcripts using GeneMarkS-T (Gabriel et al. 2023).

4.6 | Genome De Novo Assembly and Scaffolding

Three *Capsicum* lines were de novo assembled using the Flye assembler (Kolmogorov et al. 2019). For *C. annuum* and *C. chacoense*, Flye version 2.4 was used, while for *C. galapagoense*, the assembly was performed using commit 786ec5c8c271f1712d3d6bb7420210026afddc22, between versions 2.4.1 and 2.4.2, to manage the size of assembled contigs during consensus and polishing steps. The resulting assemblies were then polished using Racon v1.3.3 and Arrow v2.3.3 using PacBio CLR data (github.com/lbcbsci/racon). Scaffolding was then done using Arcs v1.0.5 and LINKS v1.8.6 with 10x Genomics linked-read data (Warren et al. 2015; Yeo et al. 2017).

Extended scaffolding was done using Bionano Genomics Solve v3.3.1 based on optical mapping data from DLE1 labeled DNA. A de novo assembly and a hybrid scaffold were constructed with Bionano Genomics Solve v3.3.2.1, using molecules larger than 120 kb. After detecting conflict regions with the Bionano Genomics Access Suite (v1.3.0), we manually inspected the conflict regions using 10x linked-reads mapped to the super-scaffolds. Mapping and visualization of the 10x derived scaffolds was done with Longranger WGA v2.2.2 (Marks et al. 2019) and Loupe v2.1.1, respectively. Gap filling was performed using pb-jelly v15.8.24 on both the scaffolded and nonscaffolded parts (English et al. 2012). A final round of polishing utilizing Pilon v1.23 and 10x genomics linked Illumina PE reads was performed after gap filling based on a mapping derived from Longranger align v2.2.2 (<https://support.10xgenomics.com/genomeexome/software/downloads/latest>).

Scaffolding using Allmaps v0.8.12 was done on *C. annuum* for chromosome anchoring using markers from Hulse-Kemp et al. 2016. This interspecific map contains a known translocation between chromosome 1 and chromosome 8, which was taken into account for this during scaffolding.

The resulting de novo assemblies underwent optimization and filtration to eliminate taxonomic and organellar contamination. Specifically, contigs and scaffolds longer than 10 kb, exhibiting a GC percentage between 25 and 45 and demonstrating a Solanaceae taxonomic hit (attained through Blobtools v1.1.1) along with PacBio CLR read coverage ranging from 10x to 300x for the same accession, were retained (Laetsch and Blaxter 2017).

Addressing taxonomic and organellar contamination as well as any residual adapter sequences within the genomic sequence, as identified by FCS-GX v0.2.3 (database build 09/07/2023) and FCS-adaptor, was also part of the process, involving their removal or hard masking (Astashyn et al. 2023). Furthermore, contigs and scaffolds flagged as exhibiting mitochondrial or chloroplast contamination, as determined by MitoHiFi v3.2 (blast hits against publicly available *C. annuum* organelles that were larger than 30% of the length of the respective public organelle), were excluded from the genomic assembly (Uliano-Silva et al. 2023). The completeness and redundancy of the genome assemblies were assessed with BUSCO v5.4.7 (Manni, Berkeley, Seppey, and Zdobnov 2021; Manni, Berkeley, Seppey, Simão, and Zdobnov 2021). Telomeric sequences were identified using the plant telomeric repeat sequence TTTAGGG with tidk v0.2.65 (Brown et al. 2025).

4.7 | Chloroplast Assembly and Annotation

To assemble the chloroplast sequence of each of the three *Capsicum* species, PacBio CLR from each line was aligned to the two public chloroplast assemblies available in NCBI. From these mappings, primarily mapped reads were used as input to the tool ptGAUL v1.0.5 with 250× coverage (Zhou et al. 2023). The resulting chloroplast assemblies were annotated using the CPGAVAS2 web server and their orientation was corrected according to the commonly accepted order of long single copy (LSC) region, first inverted repeat (IR), short single copy (SSC) region, and the second inverted repeat (Shi et al. 2019).

4.8 | Repeat Masking and Structural and Functional Annotation

RepeatModeler with the -LTRstruct parameter was used to construct a de novo library of repetitive elements specific to the three *Capsicum* nuclear genomes (Flynn et al. 2020). This custom library, combined with the Repbase library (Version 26.10.2018) and the Dfam library of transposable element (TE) sequence alignments and hidden Markov models (HMMs) v3.7, was utilized in RepeatMasker v4.1.5 (totoools v1.87) to identify and soft mask these repetitive elements within the genome assemblies. The genome assemblies with soft-masked repetitive regions were subsequently used for structural annotation.

Braker3 v3.0.3 was used to structurally annotate the *C. annuum*, *C. chacoense*, and *C. galapagoense* de novo assembled genomes (Brůna et al. 2021; Buchfink et al. 2015; Gabriel et al. 2023, 2021; Gotoh 2008; Hoff et al. 2016; Iwata and Gotoh 2012; Kim et al. 2019; Kovaka et al. 2019; Pertea and Pertea 2020; Stanke et al. 2008; Stanke and Waack 2003). Publicly available RNA-Seq data from various tissues and treatments were selected as input for Braker3 (SRR10011918, SRR10011923, SRR10011924, SRR11874047, SRR13488410, SRR17837298, SRR17837305, SRR771924, SRR771925, SRR771927, SRR771928, SRR771929, SRR771932, SRR771938, SRR771939, SRR771952, SRR8692596). Additionally, protein sequences from the Viridiplantae odb11 and the proteins of *C. annuum* UCD10Xv1.1 (NCBI RefSeq GCF 002878395.1) were used. Gene models were also predicted with Helixer v0.3.2 (Stiehler et al. 2020) using the land plant lineage

and default parameters. The final gene models were produced by combining and refining the Braker3 and Helixer gene models with EvidenceModeler (EVM) v2.1.0 (Haas et al. 2008). To enhance accuracy, we integrated assembled transcriptomic data from public RNA Seq libraries, along with assembled Iso-Seq data and the protein mappings from Viridiplantae odb11, UCD10Xv1.1 (NCBI RefSeq GCF 002878395.1) and Zunla (GCF 000710875.1). Finally, UTR regions were predicted, and we modeled alternative splicing using PASA v2.5.3 (Haas et al. 2008). Single exon genes without a homolog in another *Capsicum* genome and incomplete gene models were discarded. During the homology grouping analysis, we identified several inconsistencies in the gene models. Based on expression data validation, we replaced PASA genes that had been incorrectly merged with their corresponding EVM models. We assigned functions to the final set of predicted genes with functional annotation of the predicted gene models using InterProScan v5.66-98.0 (Jones et al. 2014).

4.9 | Capsicum Graph Pangenome Construction

Capsicum genome assemblies and annotations for *C. annuum* Early California Wonder (ECW), *C. annuum* UCD-10X-F1, *C. annuum* CM334, *C. annuum* Dempsey, *C. annuum* Zhangshugang, *C. annuum* Zunla, *C. annuum* Takanotsume, *C. annuum* 59 (using the updated annotation from Shi et al. 2024), *C. annuum* G1-36576, *Capsicum annuum* var. *glabriusculum*, *C. chinense* PI159236, *C. chacoense* CGN21477, *C. galapagoense* CGN22208, *C. baccatum* var. *pendulum* PI632928, *C. baccatum* var. *baccatum* PBC81, and *C. pubescens* Grif 1614 were collected to build a *Capsicum* clade graph pangenome. From the collected genomes, contigs smaller than 5 Kb were removed, and from the annotations, only the longest isoforms, encoding proteins larger than 30 amino acids were selected using the PanTools QC pipeline (<https://github.com/PanUtils/pantools-qc-pipeline>). The chromosomes and/or larger scaffolds of the input genomes were oriented based on their longest alignment with the respective chromosomes of *C. annuum* Zhangshugang. PanTools v4.3.1 functions build_pangenome, add_annotations, add_functions, and add_phenotypes were used to construct a genus-level pangenome from the selected 16 *Capsicum* genome assemblies, including structural and functional annotations with accompanying phenotype data (Jonkheer et al. 2022; Sheikhezadeh Anari et al. 2018; Sheikhezadeh et al. 2016). Homology grouping parameters were optimized with the optimal_grouping function using the solanales_odb10 lineage to monitor false and true positives and negatives, while the grouping of the proteins in HGs was performed using --relaxation 4. Homology groups were classified using the gene_classification function into core, unique and accessory categories, with further subdivision based on wild and domesticated phenotypes (Table S4).

4.10 | Phylogeny

Full genome core SNP phylogeny was constructed by classifying genes as core, accessory or unique using PanTools. For the core SNP phylogeny, a multiple sequence alignment of the single-copy core genes in the pangenome was used to generate a maximum likelihood phylogenetic tree, representing the

core genome phylogeny using the PanTools core_phylogeny function. Genome sequence distances between genomes were calculated using Mash v2.3 and the triangle function (Ondov et al. 2016).

4.11 | GO Term Enrichment Analysis

Gene ontology analysis for molecular function (MF), cellular component (CC), and biological process (BP) was performed with GOATOOLS with a p value <0.05 and FDR correction method Benjamini/Hochberg (Klopfenstein et al. 2018).

4.12 | Structural Variation Analysis

Genome assembly alignments were aligned using minimap2 (Li 2021; Li 2018), and alignments were visualized using the R package pafr (<https://github.com/dwinter/pafr>).

4.13 | Region of Interest Analysis

PanTools v4 commit ID a8d2b4e and extract_homology_groups was used to extract the regions of interest based on gene homology. Homology groups that contain potential gene copies of the genes of interest were identified by performing a BLASTP search of the query protein, against the pangenome protein database (threshold $\geq 50\%$ identity and $\geq 25\%$ alignment length of input sequence). The region of interest was visualized using the R package gggenomes (<https://github.com/thackl/gggenomes>).

Author Contributions

Christina Papastolopoulou: data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, writing – review and editing. **Ronald Nieuwenhuis:** data curation, formal analysis, investigation, methodology, software. **Sven Warris:** data curation, formal analysis, investigation, methodology, software. **Linda V. Bakker:** data curation, formal analysis, investigation, methodology, software. **Jan van Haarst:** data curation, formal analysis, investigation, methodology, software. **Jan Cordewener:** investigation. **Thamara Hesselink:** investigation. **Hetty van den Broeck:** investigation. **Willem van Dooijeweert:** resources. **Hans de Jong:** conceptualization, writing – original draft. **Julapark Chunwongse:** conceptualization, writing – original draft. **Sara Diaz Trivino:** writing – original draft. **Elio Schijlen:** investigation. **Dick de Ridder:** supervision, writing – review and editing. **Sandra Smit:** supervision, writing – review and editing. **Sander A. Peters:** conceptualization, funding acquisition, project administration, writing – original draft, writing – review and editing.

Acknowledgments

We would like to thank Zijiang Yang and Dirk-Jan van Workum for their annotation and pangenomics support. This work was funded by the Top consortium for Knowledge and Innovation (TKI) program with Project Number LWV2004.

Funding

This project was supported by the Netherlands Top Consortium for Knowledge and Innovation (TKI-U LWV2004).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The raw sequencing data (PacBio CLR, Illumina paired end), genome assemblies and annotations, are publicly accessible and can be retrieved from the National Center for Biotechnology Information (NCBI) under the Project ID PRJNA944626.

References

- Astashyn, A., E. S. Tvedte, D. Sweeney, et al. 2023. “Rapid and Sensitive Detection of Genome Contamination at Scale With FCS-GX.” *bioRxiv* 2023.06.02.543519. <https://doi.org/10.1101/2023.06.02.543519>.
- Back, S., J. M. Kim, H. Choi, et al. 2024. “Genetic Characterization of a Locus Responsible for Low Pungency Using EMS-Induced Mutants in *Capsicum Annuum* L.” *Theoretical and Applied Genetics* 137: 101. <https://doi.org/10.1007/S00122-024-04602-3>.
- Brown, M. R., P. M. Gonzalez de La Rosa, and M. Blaxter. 2025. “Tidk: A Toolkit to Rapidly Identify Telomeric Repeats From Genomic Datasets.” *Bioinformatics* 41, no. 2: btaf049. <https://doi.org/10.1093/bioinformatics/btaf049>.
- Brûna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky. 2021. “BRAKER2: Automatic Eukaryotic Genome Annotation With GeneMark-EP+ and AUGUSTUS Supported by a Protein Database.” *NAR Genomics and Bioinformatics* 3: lqaa108. <https://doi.org/10.1093/nargab/lqaa108>.
- Buchfink, B., C. Xie, and D. H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12: 59–60. <https://doi.org/10.1038/NMETH.3176>.
- Carrizo García, C., M. H. J. Barfuss, E. M. Sehr, et al. 2016. “Phylogenetic Relationships, Diversification and Expansion of Chili Peppers (*Capsicum*, *Solanaceae*).” *Annals of Botany* 118: 35–51. <https://doi.org/10.1093/AOB/MCW079>.
- Chen, W., W. Xiangfeng, J. Sun, et al. 2024. “Two Telomere-to-Telomere Gapless Genomes Reveal Insights Into Capsicum Evolution and Capsaicinoid Biosynthesis.” *Nature Communications* 15: 1–14. <https://doi.org/10.1038/s41467-024-48643-0>.
- Chetelat, R. T., P. Cisneros, L. Stamova, and C. M. Rick. 1997. “A Male-Fertile *Lycopersicon Esculentum* x *Solanum Lycopersicoides* Hybrid Enables Direct Backcrossing to Tomato at the Diploid Level.” *Euphytica* 95: 99–108. <https://doi.org/10.1023/A:1002958030799> / METRICS.
- D’Agostino, N., R. Tamburino, C. Cantarella, et al. 2018. “The Complete Plastome Sequences of Eleven *Capsicum* Genotypes: Insights into DNA Variation and Molecular Evolution.” *Genes* 9: 503. <https://doi.org/10.3390/GENES9100503>.
- English, A. C., S. Richards, Y. Han, et al. 2012. “Mind the Gap: Upgrading Genomes With Pacific Biosciences RS Long-Read Sequencing Technology.” *PLoS ONE* 7: e47768. <https://doi.org/10.1371/JOURNAL.PONE.0047768>.
- Flynn, J. M., R. Hubley, C. Goubert, et al. 2020. “Repeatmodeler2 for Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences of the United States of America* 117: 9451–9457. <https://doi.org/10.1073/pnas.1921046117>.
- Gabriel, L., T. Brûna, K. J. Hoff, et al. 2023. “BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence With GeneMark-ETP, AUGUSTUS and TSEBRA.” *BioRxiv*. <https://doi.org/10.1101/2023.06.10.544449>.
- Gabriel, L., K. J. Hoff, T. Brûna, M. Borodovsky, and M. Stanke. 2021. “TSEBRA: Transcript Selector for BRAKER.” *BMC Bioinformatics* 22: 1–12. <https://doi.org/10.1186/s12859-021-04482-0>.

- Gotoh, O. 2008. "A Space-Efficient and Accurate Method for Mapping and Aligning cDNA Sequences Onto Genomic Sequence." *Nucleic Acids Research* 36: 2630–2638. <https://doi.org/10.1093/NAR/GKN105>.
- Haas, B. J., S. L. Salzberg, W. Zhu, et al. 2008. "Automated Eukaryotic Gene Structure Annotation Using EVIDENCEModeler and the Program to Assemble Spliced Alignments." *Genome Biology* 9: 9. <https://doi.org/10.1186/GB-2008-9-1-R7>.
- Hazarika, G., R. Phukan, D. Sarma, et al. 2023. "Genetic Relatedness Among Interspecific Hybrids in the Genus *Capsicum* and Their Implications in Breeding King Chilli." *South African Journal of Botany* 163: 744–755. <https://doi.org/10.1016/J.SAJB.2023.11.006>.
- Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke. 2016. "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation With GeneMark-ET and AUGUSTUS." *Bioinformatics* 32: 767–769. <https://doi.org/10.1093/bioinformatics/btv661>.
- Hulse-Kemp, A. M., H. Ashrafi, J. Plieske, et al. 2016. "A HapMap Leads to a *Capsicum Annuum* SNP Infinium Array: A New Tool for Pepper Breeding." *Horticulture Research* 3: 3. <https://doi.org/10.1038/HORTRES.2016.36>.
- Hulse-Kemp, A. M., S. Maheshwari, K. Stoffel, et al. 2018. "Reference Quality Assembly of the 3.5-Gb Genome of *Capsicum annuum* From a Single Linked-Read Library." *Horticulture Research* 5: 1–13. <https://doi.org/10.1038/s41438-017-0011-0>.
- Iwata, H., and O. Gotoh. 2012. "Benchmarking Spliced Alignment Programs Including Spaln2, an Extended Version of Spaln That Incorporates Additional Species-Specific Features." *Nucleic Acids Research* 40: e161. <https://doi.org/10.1093/NAR/GKS708>.
- Jarret, R. L., G. E. Barboza, F. R. d. Costa Batista, et al. 2019. "*Capsicum*—An Abbreviated Compendium." *Journal of the American Society for Horticultural Science* 144, no. 1: 3–22. <https://doi.org/10.21273/jashs04446-18>.
- Jones, P., D. Binns, H. Y. Chang, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30: 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>.
- Jonkheer, E. M., D.-J. M. van Workum, S. Sheikhezadeh Anari, et al. 2022. "PanTools v3: Functional Annotation, Classification and Phylogenomics." *Bioinformatics* 38: 4403–4405. <https://doi.org/10.1093/BIOINFORMATICS/BTAC506>.
- Khan, A. W., V. Garg, M. Roorkiwal, A. A. Golicz, D. Edwards, and R. K. Varshney. 2020. "Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement." *Trends in Plant Science* 25: 148–158. <https://doi.org/10.1016/J.TPLANTS.2019.10.012>.
- Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping With HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37: 907–915. <https://doi.org/10.1038/S41587-019-0201-4>.
- Kim, M. S., G. Y. Chae, S. Oh, et al. 2021. "Comparative Analysis of *De Novo* Genomes Reveals Dynamic Intra-Species Divergence of NLRs in Pepper." *BMC Plant Biology* 21: 1–12. <https://doi.org/10.1186/s12870-021-03057-8>.
- Kim, S., J. Park, S. I. Yeom, et al. 2017. "New Reference Genome Sequences of Hot Pepper Reveal the Massive Evolution of Plant Disease-Resistance Genes by Retroduplication." *Genome Biology* 18: 1–11. <https://doi.org/10.1186/S13059-017-1341-9>.
- Kim, S., M. Park, S. I. Yeom, et al. 2014. "Genome Sequence of the Hot Pepper Provides Insights Into the Evolution of Pungency in *Capsicum* Species." *Nature Genetics* 46: 270–278. <https://doi.org/10.1038/ng.2877>.
- Klopfenstein, D. V., L. Zhang, B. S. Pedersen, et al. 2018. "GOATOOLS: A Python Library for Gene Ontology Analyses." *Scientific Reports* 8: 10872. <https://doi.org/10.1038/s41598-018-28948-z>.
- Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37, no. 5: 540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
- Kovaka, S., A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea. 2019. "Transcriptome Assembly From Long-Read RNA-Seq Alignments With StringTie2." *Genome Biology* 20: 20. <https://doi.org/10.1186/S13059-019-1910-1>.
- Laetsch, D. R., and M. L. Blaxter. 2017. "Blobtools: Interrogation of Genome Assemblies." *F1000Research* 6: 1287. <https://doi.org/10.12688/f1000research.12232.1>.
- Lee, J. H., J. Venkatesh, J. Jo, et al. 2022. "High-Quality Chromosome-Scale Genomes Facilitate Effective Identification of Large Structural Variations in Hot and Sweet Peppers." *Horticulture Research* 9: uhac210. <https://doi.org/10.1093/HR/UHAC210>.
- Lefebvre, V., S. Pflieger, A. Thabuis, et al. 2002. "Towards the Saturation of the Pepper Linkage Map by Alignment of Three Intraspecific Maps Including Known-Function Genes." *Genome* 45: 839–854. <https://doi.org/10.1139/G02-053>.
- Li, H. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34: 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- Li, H. 2021. "New Strategies to Improve minimap2 Alignment Accuracy." *Bioinformatics* 37: 4572–4574. <https://doi.org/10.1093/bioinformatics/btab705>.
- Liao, Y., J. Wang, Z. Zhu, et al. 2022. "The 3D Architecture of the Pepper Genome and Its Relationship to Function and Evolution." *Nature Communications* 13: 1–18. <https://doi.org/10.1038/s41467-022-31112-x>.
- Liu, F., J. Zhao, H. Sun, et al. 2023. "Genomes of Cultivated and Wild *Capsicum* Species Provide Insights Into Pepper Domestication and Population Differentiation." *Nature Communications* 14: 1–14. <https://doi.org/10.1038/s41467-023-41251-4>.
- Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows Along With Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." *Molecular Biology and Evolution* 38: 4647–4654. <https://doi.org/10.1093/MOLBEV/MSAB199>.
- Manni, M., M. R. Berkeley, M. Seppey, and E. M. Zdobnov. 2021. "BUSCO: Assessing Genomic Data Quality and Beyond." *Current Protocols* 1: e323. <https://doi.org/10.1002/CPZ1.323>.
- Marks, P., S. Garcia, A. M. Barrio, et al. 2019. "Resolving the Full Spectrum of Human Genome Variation Using Linked-Reads." *Genome Research* 29: 635–645. <https://doi.org/10.1101/GR.234443.118>.
- Mongkolporn, O., and P. W. J. Taylor. 2011. *Wild Crop Relatives: Genomic and Breeding Resources*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20450-0_4.
- Moscone, E. A., M. A. Scaldaferrro, M. Grabile, et al. 2007. "The Evolution of Chili Peppers (*Capsicum* -Solanaceae): A Cytogenetic Perspective." *Acta Horticulturae* 745: 137–170. <https://doi.org/10.17660/ActaHortic.2007.745.5>.
- Ondov, B. D., T. J. Treangen, P. Melsted, et al. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17: 1–14. <https://doi.org/10.1186/s13059-016-0997-x>.
- Ou, L., D. Li, J. Lv, et al. 2018. "Pan-Genome of Cultivated Pepper (*Capsicum*) and Its Use in Gene Presence-Absence Variation Analyses." *New Phytologist* 220: 360–363. <https://doi.org/10.1111/NPH.15413>.
- Pertea, G., and M. Pertea. 2020. "GFF Utilities: GffRead and GffCompare." *F1000Research* 9: 304. <https://doi.org/10.12688/f1000research.23297.2>.
- Pertuzé, R. A., Y. Ji, and R. T. Chetelat. 2002. "Comparative Linkage Map of the *Solanum Lycopersicoides* and *S. Siliens* Genomes and Their

- Differentiation From Tomato." *Genome* 45: 1003–1012. <https://doi.org/10.1139/G02-066>.
- Peters, S. A., J. W. Bargsten, D. Szinay, et al. 2012. "Structural Homology in the Solanaceae: Analysis of Genomic Regions in Support of Synteny Studies in Tomato, Potato and Pepper." *Plant Journal* 71: 602–614. <https://doi.org/10.1111/J.1365-313X.2012.05012.X>.
- Pickersgill, B. 1991. *Cytogenetics and Evolution of Capsicum L. Developments in Plant Genetics and Breeding*. Vol. 2, 139–160. Elsevier. <https://doi.org/10.1016/B978-0-444-88260-8.50013-6>.
- Qin, C., C. Yu, Y. Shen, et al. 2014. "Whole-Genome Sequencing of Cultivated and Wild Peppers Provides Insights Into Capsicum Domestication and Specialization." *Proceedings of the National Academy of Sciences* 111: 5135–5140. <https://doi.org/10.1073/pnas.1400975111>.
- Sanatombi, K. 2024. "Genetics and Evolutionary Insights From Solanaceae Genome Sequences." *Plant Systematics and Evolution* 310: 1–16. <https://doi.org/10.1007/s00606-023-01883-x>.
- Saxena, A., R. Raghuwanshi, V. K. Gupta, and H. B. Singh. 2016. "Chilli Anthracnose: The Epidemiology and Management." *Frontiers in Microbiology* 7: 215999. <https://doi.org/10.3389/fmicb.2016.01527>.
- Schreiber, M., M. Jayakodi, N. Stein, and M. Mascher. 2024. "Plant Pangenomes for Crop Improvement, Biodiversity and Evolution." *Nature Reviews Genetics* 25, no. 8: 563–577. <https://doi.org/10.1038/s41576-024-00691-4>.
- Sheikhzadeh Anari, S., D. de Ridder, M. E. Schranz, and S. Smit. 2018. "Efficient Inference of Homologs in Large Eukaryotic Pan-Proteomes." *BMC Bioinformatics* 19: 340. <https://doi.org/10.1186/s12859-018-2362-4>.
- Sheikhzadeh, S., M. E. Schranz, M. Akdel, D. De Ridder, and S. Smit. 2016. "PanTools: Representation, Storage and Exploration of Pan-Genomic Data." *Bioinformatics* 32: i487–i493. <https://doi.org/10.1093/bioinformatics/btw455>.
- Shi, C., H. Yu, L. Song, et al. 2024. "Comprehensive Re-Annotation and Transcriptome Analysis Provide Insights Into Pepper Development." *Scientia Horticulturae* 336: 113406. <https://doi.org/10.1016/J.SCIENTA.2024.113406>.
- Shi, L., H. Chen, M. Jiang, et al. 2019. "CPGAVAS2, an Integrated Plastome Sequence Annotator and Analyzer." *Nucleic Acids Research* 47: W65–W73. <https://doi.org/10.1093/NAR/GKZ345>.
- Shirasawa, K., M. Hosokawa, Y. Yasui, A. Toyoda, and S. Isobe. 2023. "Chromosome-Scale Genome Assembly of a Japanese Chili Pepper Landrace, *Capsicum annuum* 'Takanotsume'." *DNA Research* 30: dsac052. <https://doi.org/10.1093/dnares/dsac052>.
- Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler. 2008. "Using Native and Syntenically Mapped cDNA Alignments to Improve De Novo Gene Finding." *Bioinformatics* 24: 637–644. <https://doi.org/10.1093/BIOINFORMATICS/BTN013>.
- Stanke, M., and S. Waack. 2003. "Gene Prediction With a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19, no. Suppl 2: ii215. <https://doi.org/10.1093/BIOINFORMATICS/BTG1080>.
- Stewart, C., B. C. Kang, K. Liu, et al. 2005. "The Pun1 Gene for Pungency in Pepper Encodes a Putative Acyltransferase." *Plant Journal* 42: 675–688. <https://doi.org/10.1111/J.1365-313X.2005.02410.X>.
- Stewart, C., M. Mazourek, G. M. Stellari, M. O'Connell, and M. Jahn. 2007. "Genetic Control of Pungency in *C. Chinense* via the Pun1 Locus." *Journal of Experimental Botany* 58: 979–991. <https://doi.org/10.1093/jxb/erl243>.
- Stiehler, F., M. Steinborn, S. Scholz, D. Dey, A. P. M. Weber, and A. K. Denton. 2020. "Helixer: Cross-Species Gene Annotation of Large Eukaryotic Genomes Using Deep Learning." *Bioinformatics* 36, no. 22–23: 5291–5298. <https://doi.org/10.1093/bioinformatics/btaa1044>.
- Su, X., B. Wang, X. Geng, et al. 2021. "A High-Continuity and Annotated Tomato Reference Genome." *BMC Genomics* 22: 1–12. <https://doi.org/10.1186/s12864-021-08212-x>.
- Uliano-Silva, M., J. G. R. N. Ferreira, K. Krashenninnikova, et al. 2023. "MitoHiFi: A Python Pipeline for Mitochondrial Genome Assembly From PacBio High Fidelity Reads." *BMC Bioinformatics* 24: 1–13. <https://doi.org/10.1186/S12859-023-05385-Y>.
- Warren, R. L., C. Yang, B. P. Vandervalk, et al. 2015. "LINKS: Scalable, Alignment-Free Scaffolding of Draft Genomes With Long Reads." *GigaScience* 4: 1–11. <https://doi.org/10.1186/S13742-015-0076-3>.
- Yeo, S., L. Coombe, J. Chu, R. L. Warren, and I. Birol. 2017. "ARCS: Assembly Roundup by Chromium Scaffolding." *BioRxiv*: 100750. <https://doi.org/10.1101/100750>.
- Zhou, W., C. E. Armijos, C. Lee, et al. 2023. "Plastid Genome Assembly Using Long-Read Data." *Molecular Ecology Resources* 23: 1442–1457. <https://doi.org/10.1111/1755-0998.13787>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Dot plot analyses showing pairwise genomic comparisons between (from top to bottom): (A) *C. annuum* cv. Zhangshugang versus *C. annuum* ECW, (B) *C. annuum* WUR versus *C. chacoense*, (C) *C. galapagoense* versus *C. chacoense*. **Figure S2:** Chromosome marker alignment comparison of physical (Phys) and genetic (Gen) positions of the genetic markers from Hulse-Kemp et al. 2016 against the *C. annuum* WUR genome assembly. The chromosome orientation has been decided based on the *C. annuum* Zhangshugang genome assembly. **Figure S3:** Plant telomeric repeat sequence (TTAGGG) frequency for the 12 chromosomes of the *C. annuum* ECW, showing the presence of nine telomeres. Each red dot marks the presence of at least 100 repeats in a 10 kb window. **Figure S4:** *C. annuum* ECW chloroplast annotated by CP2GAVAS. **Figure S5:** *C. chacoense* chloroplast annotated by CP2GAVAS. **Figure S6:** *C. galapagoense* chloroplast annotated by CP2GAVAS. **Figure S7:** Genome similarity for the *Capsicum* genomes. Mash *k*-mer distance highlights the sequence similarity within the *C. annuum* genomes. The similarity gradually decreases for the other species in the *Annuum* clade and becomes larger for *Pubescens* and *Baccatum* clades. **Figure S8:** *Capsicum* panproteome growth curve. The number of core and dispensable (unique and accessory) homology groups for the different genome combinations within in the *Capsicum* panproteome. **Figure S9:** Dot plot analyses showing pairwise genomic comparisons of chromosome 2 (from top to bottom): (A) *C. pubescens*, (B) *C. baccatum* var. *pendulum*, (C) *C. chacoense*, (D) *C. annuum* ECW, and (E) *C. galapagoense* versus *S. lycopersicum* SL4.0. **Figure S10:** Dot plot analyses showing pairwise genomic comparisons of chromosome 6 (from top to bottom): (A) *C. pubescens*, (B) *C. baccatum* var. *pendulum*, (C) *C. chacoense*, (D) *C. annuum* ECW, and (E) *C. galapagoense* versus *S. lycopersicum* SL4.0. **Figure S11:** Dot plot analyses showing pairwise genomic comparisons of chromosome 10 (from top to bottom): (A) *C. pubescens*, (B) *C. baccatum* var. *pendulum*, (C) *C. chacoense*, (D) *C. annuum* ECW, and (E) *C. galapagoense* versus *S. lycopersicum* SL4.0. **Figure S12:** Dot plot analyses showing pairwise genomic comparisons of chromosome 11 (from top to bottom): (A) *C. pubescens*, (B) *C. baccatum* var. *pendulum*, (C) *C. chacoense*, (D) *C. annuum* ECW, and (E) *C. galapagoense* versus *S. lycopersicum* SL4.0. **Figure S13:** Synteny plot of chromosome 2 across the 16 *Capsicum* genome assemblies of the dataset. The genomes are order based on core SNP phylogeny. **Figure S14:** Dot plot analysis showing pairwise genomic comparisons of scaffolds 6 and 12 of *C. galapagoense* versus chromosomes 2,4 and 7 of *C. annuum* ECW. The red horizontal line indicates the breakpoint in *C. galapagoense* scaffold 12, switching alignment from *C. annuum* ECW chromosome 2 to chromosome 4, while the green line indicates the breakpoint on *C. galapagoense* scaffold 6. **Figure S15:** Alignment of the *C. annuum* ECW genome assembly and *C. galapagoense* PacBio CLR reads against the *C. galapagoense* genome. (Top) Visualization of the breakpoint in the alignment between *C. annuum* and *C. galapagoense* scaffold 12 (red line). Spanning PacBio

CLR reads support the scaffold assembly. (Bottom) Visualization of the breakpoint in the alignment between *C. annuum* and *C. galapagoense* scaffold 6 (green line). Spanning PacBio CLR reads support the scaffold assembly. **Figure S16:** *Pun-1* locus for the 16 *Capsicum* genome assemblies. The *Pun-1* gene copies are organized in nine homology groups, ranging from one to nine *Pun1*-like gene copies per genome. The genomes are order based on core SNP phylogeny. **Figure S17:** A Maximum likelihood tree for the members of the nine homology groups with a blast hit with the Pun1 protein. The clustering of the sequences majorly confirms the homology grouping in the nine respective groups (shown in nine different colors) with group highlighted in dark blue containing the members of capsaicin synthase 2 (*C. annuum* CM334 homolog in bold). **Table S1:** Summary of sequencing data statistics for the de novo *Capsicum* genomes. **Table S2:** Summary statistics for filtered genome assemblies and structural annotations. **Table S3:** Repeat annotation for the three de novo genomes. **Table S4:** Input *Capsicum* genomes for *Capsicum* pangenome. **Table S5:** Gene count per class and phenotype. **Table S6:** GO term enrichment results for core homology groups. **Table S7:** GO term enrichment results for unique homology groups. **Table S8:** GO term enrichment results for accessory homology groups. **Table S9:** GO term enrichment results for domesticated exclusive homology groups. **Table S10:** GO term enrichment results for wild exclusive homology groups. **Table S11:** GO term enrichment results for *C. chacoense* and *C. galapagoense* exclusive homology groups. **Table S12:** *Pun1*-like gene copies per genome.