A draft genome assembly for the heterozygous wild tomato *Solanum habrochaites* highlights haplotypic structural variations of intracellular immune receptors

Kyungyong Seong[1,*], China Lunde Shaw[1], Eunyoung Seo[1,2], Meng Li[1,2], Ksenia V Krasileva[1,*] and Brian Staskawicz[1,2,*]

[1] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA.

[2] Innovative Genomics Institute, University of California, Berkeley, CA 94704, USA.

[*]Corresponding author: s.kyungyong@berkeley.edu, kseniak@berkeley.edu and stask@berkeley.edu

Abstract

Solanum habrochaites LA1353 is a self-incompatible, highly heterozygous wild tomato that is a useful germplasm resource for the study of metabolism, reproduction and disease resistance. We generated a draft genome assembly with PacBio HiFi reads and genome annotations, which underscored the expansion of gene families associated with metabolite-production, self-incompatibility, DNA regulation and immunity. After manually curating intracellular nucleotide-binding leucine-rich repeat immune receptors (NLRs), we found that *S. habrochaites* LA1353 has a larger NLR inventory than other wild tomato species. A great number of heterozygous local copy number variations (CNVs) driven by haplotypic structural variations further expands the inventory, both enhancing NLR diversity and providing more opportunities for sequence evolution. The NLRs associated with local CNVs predominantly appear in the helper NLR (NRC)-related phylogenetic clades and are concentrated in a few physical NLR gene clusters. Synteny analysis points out that these genomic regions correspond to the known NLR clusters from which experimentally validated, functional NLRs, such as *Hero*, *Mi-1.2* and *Rpi-amr1*, have been identified. Producing and incorporating Resistance Gene Enrichment Sequencing (RenSeq) data across wild tomato species, we reveal that the regions with local CNVs might have been shaped nearly equally by recent NLR gains and losses, along with enhanced sequence diversification that diminishes one-to-one orthology between heterozygous alleles. Our analysis suggests that these genomic regions may have accelerated evolutionary dynamics for NLR diversity generation in *S. habrochaites* LA1353.

Key words: Genome assembly, Intracellular immune receptors, NLRs, RenSeq, resistance genes, Solanaceae, Tomato, wild tomato

Introduction

Solanum habrochaites LA1353 is a self-incompatible, heterozygous wild tomato that grows in a suboptimal environment at 2,650 meters above sea level in the Andes (Fig. 1; Fig. S1) (Bauchet and Causse, 2012; Broz, Randle, *et al.*, 2017; Broz *et al.*, 2021). Besides its unique morphological, metabolic and reproductive features (Bedinger *et al.*, 2011; Fan *et al.*, 2017; Kilambi *et al.*, 2017; Tohge *et al.*, 2020), genetic diversity of *S. habrochaites* associated with disease resistance has drawn researchers' interest (Rick and Chetelat, 1995). Many resistance (R) genes identified from this species have been introgressed into the cultivated tomatoes (*Solanum lycopersicum*) (Chaudhary and Atamian, 2017). Novel disease resistance is continuously searched against diverse pathogens, including oomycete *Phytophthora infestans* (Li *et al.*, 2011; Nowakowska *et al.*, 2014; Copati *et al.*, 2019), fungi *Botrytis cinerea* (Finkers *et al.*, 2007) and bacteria *Pseudomonas syringae* (Bao *et al.*, 2015; Thapa *et al.*, 2015).

Intracellular nucleotide-binding leucine-rich repeat immune receptors (NLRs) are essential components of the R gene-mediated immunity. Tomato and its wild relatives encode about 300 NLRs that have originated from diverse modes of duplications (Seo *et al.*, 2016; Kim *et al.*, 2017), and their multi-domain architecture (MDA) typically includes the central NB-ARC (NB) domain and C-terminal leucine rich-repeats (LRRs) in common, as well as N-terminal coiled-coil (CC), RPW8 domain or TIR domain. NLRs can be functionally classified as sensors that directly or indirectly perceive pathogen molecules (effectors) or helpers that execute a hypersensitive response in coordination with some sensors (Baggs *et al.*, 2017). Sensors and helpers appear in distinct phylogenetic clades. In the CNL (CC-NB-LRRs) clades, the NRC clade members function as helpers with sensors in the NRC-dependent CNL superclade (Wu *et al.*, 2017).

R gene enrichment sequencing (RenSeq) enables efficient study of NLRs through their selective amplification and sequencing from large genomes (Jupe *et al.*, 2013; Witek *et al.*, 2016; Stam *et al.*, 2016; Seong *et al.*, 2020). However, RenSeq outputs lack sufficient genomic contexts around NLRs. NLR clusters, as a result, tend to be fragmented into multiple contigs unless NLRs appear with close intergenic regions. Because of the missing context, it is infeasible to distinguish with RenSeq data of self-incompatible species heterozygous NLR alleles from recently duplicated NLRs in a haplotype. Thus, RenSeq data from highly heterozygous species does not align well with those from homozygous species that technically focus on haplotypic NLR variations. For this reason, our previous study relied only on the self-compatible, homozygous wild tomatoes and necessitated whole genome sequencing and assembly data for heterozygous wild tomatoes to investigate NLR evolution (Seong *et al.*, 2020).

Although a genome assembly produced with Illumina short reads is available for *S. habrochaites*, it is highly fragmented and lacks annotations, making it difficult to utilize the assembly for understanding NLR evolution (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014). In this study, we generated a draft genome of *S. habrochaites* LA1353 with PacBio High-Fidelity (HiFi) reads and annotations, which can be useful for diverse studies on this wild tomato. We report that copy number variations (CNVs) between haplotypes are concentrated in a few NLR gene clusters, which may be hotspots of NLR evolution in *S. habrochaites*.

Materials and Methods

RenSeq with single-molecule real-time sequencing on wild tomatoes

We obtained the seeds of *Solanum arcanum* LA2152, *Solanum huaylasense* LA1365, *Solanum peruvianum* LA0446, *Solanum corneliomulleri* LA1677, *Solanum chilense* LA1932, *Solanum pennellii* LA1272 and *S. habrochaites* LA1353 from the Tomato Genetics Resource Center (TGRC) at the University of California, Davis (Table S1). We used gDNA from young leaf tissues to perform RenSeq in combination with single-molecule real-time sequencing (SMRT), following the previously published protocol (Seong *et al.*, 2020). The library was sequenced using the PacBio Sequel platform at the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley. We used ccs v6.0.0 to generate HiFi reads from the sequencing output (three full-passes and > 99% accuracy) (Table S1). 70 nucleotide sequences were trimmed from both ends of the HiFi reads with cutadapt v1.16 to remove barcodes and adapters, and the trimmed reads of each accession were assembled with HiCanu v2.2 (genomeSize=7M -pacbio-hifi) (Martin, 2011; Nurk *et al.*, 2020).

Whole genome sequencing, assembly and scaffolding

We conducted whole genome sequencing (WGS) for *S. habrochaites* LA1353 with 10X Genomics and PacBio sequencing. The 10X Genomics sequencing was performed with the gDNA collected from the same plant used for the RenSeq. After the DNA repair and BluePippin DNA size selection, we sequenced the gDNA as 150 bp linked-reads using the HiSeq X Ten platform at the University of California, Davis. The linked-reads were converted to interleaved reads with longranger v2.2.2 and assembled with Supernova v2.1.1 (https://www.10xgenomics.com).

We extracted gDNA from young leaf tissues of another plant and sequenced the gDNA using the Sequel II platform at the Vincent J. Coates Genomics Sequencing Laboratory. HiFi reads were generated with ccs v6.0.0 and assembled with hifiasm v0.12 (-l2) (Cheng *et al.*, 2021). The purge_dups package was used to separate out remaining haplotigs from the primary assembly, which were then concatenated with the alternative contigs (Guan *et al.*, 2020). We scaffolded the primary contigs with the 10X Genomics linked-reads, relying on ARCS v1.1.1 and LINKS v1.8.7 (Warren *et al.*, 2015; Yeo *et al.*, 2018). We removed contigs and scaffolds that represent chloroplast and mitochondrial DNAs based on the similarity to the known plasmid DNAs from *S. lycopersicum*.

Structural and functional annotation and annotation quality assessment

To run BRAKER v2.1.5 and MAKER v3.01.03 for structural annotations (Cantarel *et al.*, 2008; Brůna *et al.*, 2021), we collected protein sequences of annotated Solanaceae species from Sol Genomics and paired-end transcriptome data for *S. habrochaites* LA0407, LA1223, LA1777, LA2098 and LA2119 from the Sequence Read Archive (SRA) (Table S2) (Mueller *et al.*, 2005; Fernandez-Pozo *et al.*, 2015; Pease *et al.*, 2016; Broz, Guerrero, *et al.*, 2017; Arnoux *et al.*, 2021). We filtered the adapters and low-quality reads from the paired-end libraries with Trim Galore v0.6.4 (-q 20 --illumina --paired) and Cutadapt v2.4. The remaining reads were mapped to the genome with BWA-MEM v2.0pre1 (Vasimuddin *et al.*, 2019). Additionally, we performed transcriptome assembly with Trinity v2.11.0 using the filtered libraries pooled for each accession (Grabherr *et al.*, 2011). TransDecoder v5.5.0 was then used to select coding sequences (CDS) with complete open reading frames from super-transcripts (https://github.com/TransDecoder).

The *S. habrochaites* repeat library was generated with RepeatModeler v2.0.1 and used to soft-mask the assemblies with RepeatMasker v4.0.7 (Smit *et al.*, 2013; Flynn *et al.*, 2020). The mapped transcriptomic data and the collected protein sequences were used as evidence in BRAKER (--etpmode --softmasking), which relied on GeneMark v4.65_lic and AUGUSTUS v3.3.3 (Stanke *et al.*, 2006; Brůna *et al.*, 2020). We used MAKER to capture gene models supported by the assembled transcripts and protein evidence mainly relying on exonerate without ab initio predictors (est2genome=1; prot2genome=1) (Slater and Birney, 2005). The gene models with ≥ 98% bidirectional coverage against the top homologous hits in the protein annotation of *S. lycopersicum* (ITAG 4.1) were retained. A subset of these MAKER gene models with AED scores < 0.07 were used to train SNAP (Korf, 2004); randomly selected 2,000 gene models from the subset were used to train AUGUSTUS. We ran MAKER again with the external gene models from BRAKER (pred_model), as well as two ab initio predictors, SNAP and AUGUSTUS. The BRAKER and MAKER gene models with ≥ 98% bidirectional coverage against the top ITAG models were passed to the model_pred.

We sequentially selected final gene models from the exonerate-driven MAKER gene models and BRAKER gene models with ≥ 98% bidirectional coverage against the top ITAG annotations, MAKER gene models with AED scores ≤ 0.35 or InterPro matches, and BRAKER gene models supported by any hints. For this purpose, InterProscan v5.52-86.0 was used to search the sequences against Pfam v33.1, SUPERFAMILY v1.75 and Gene3D v4.3.0 (Wilson *et al.*, 2009; Jones *et al.*, 2014; Dawson *et al.*, 2017; Mistry *et al.*, 2021). Gene models composed only of transposable element (TE)-related domains were removed. The quality of the final annotation was assessed with BUSCO v5.2.2 and the Solanales datasets v10 (Kriventseva *et al.*, 2019; Seppey *et al.*, 2019).

NLR manual curation

We identified putative NLR loci by translating the *S. habrochaites* genome assemblies in the six reading frames and searching for the NB-ARC domain with hmmsearch v3.3 (-E 1e-4 --domE 1e-4) (Eddy, 2011), as well as mapping RenSeq HiFi reads to the assemblies with BWA-MEM v2.0pre. We extracted the candidate loci with 10,000 flanking regions and manually annotated them as described previously (Seong *et al.*, 2020). For other wild tomato species, we curated only a subset of contigs that contain NLRs belonging to the three phylogenetic clades of interest, G1, G8 and G14. These contigs were identified by searching the manually curated intact G1, G8 and G14 NLRs of *S. habrochaites* against the RenSeq contigs of wild tomatoes with exonerate v2.2.0 and selecting the best seven matches for each query (Slater and Birney, 2005).

NLR classification

The NLRs were classified based on their MDA by the previously constructed pipeline (Seo *et al.*, 2016). Gene models containing the NB-ARC domain were initially identified. Based on the N-terminal sequence homology to CC, RPW8 domain and TIR domain and the presence of C-terminal LRRs, the NLRs were classified to CNLs, RNLs, TNLs, NLs and Ns. We relied on NLR-parser to detect the major motifs on the NB-ARC domain and assigned an NLR as intact if its NB-ARC domain is equal to or larger than 160 amino acids and has three out of four major motifs present in order (Steuernagel *et al.*, 2015). An NLR was classified as incomplete otherwise, and we did not analyze incomplete NLRs.

For phylogenetic classification, we collected experimentally validated NLRs in Solanaceae from RefPlantNLR (Kourelis *et al.*, 2021). The NB-ARC domains of the intact NLRs from *S. habrochaites* LA1353, the reference NLRs and an outgroup CED-4 were aligned by MAFFT v7.313 (--globalpair --maxiterate 1000) (Katoh and Standley, 2013). The multiple sequence alignment was trimmed with TrimAl v1.4.rev22 (-gt 0.2) and used to infer a phylogenetic tree with RAxML v8.2.12 with 500 rapid bootstrapping (-p 12345 -x 12345 -m PROTGAMMAJTTF) (Capella-Gutierrez *et al.*, 2009; Stamatakis, 2014). We then followed the previous study for phylogenetic grouping (Seong *et al.*, 2020).

Gene family analysis

We identified with OrthoFinder v2.5.4 (default) orthologous protein groups between *S. lycopersicum* Heinz (ITAG 4.1), *S. chilense* LA3111, *S. pimpinellifolium* LA2093, *S. pennellii* LA0716 and *S. habrochaites* LA1353 (Bolger *et al.*, 2014; Hosmani *et al.*, 2019; Stam *et al.*, 2019; Emms and Kelly, 2019; X., Wang *et al.,* 2020). We then used CAFE v5 to select orthologous groups that rapidly expanded in *S. habrochaites* in comparison to another tomato species (Mendes *et al.*, 2021). The enrichment test was performed with the enricher function in clusterProfiler on PFAM domains of the expanded gene family members (Yu *et al.*, 2012). Significant hits required p-value < 0.05 and q-value < 0.05. Any TE-related results were removed. For comparison of gene counts with PFAM domains of interest, we counted the genes only if the detected region of domain covers 60% or more of the full domain to prevent incompletely annotated gene fragments from inflating the counts.

Results

HiFi reads lead to good quality heterozygous genome assembly and annotation for *S. habrochaites* LA1353

We generated 175.1X of PacBio subreads and 56.6X of 10X Genomics sequencing data, given 2Gbp of the wild tomato genome size for both haplotypes (Table 1). From the subreads, 21.6 Gbp of HiFi reads were produced, providing 10.8X coverage for each haplotype in genome assembly. The final draft genome was 981.2 Mbp composed of 794 scaffolds with a N50 of 6.7 Mbp (Table 1). RepeatMasker detected 559.1 Mbp (57.0%) and 63.7

(6.5%) Mbp as retroelements and DNA transposons, respectively, indicating a high level of repeat content. *S. habrochaites* LA1353 is self-incompatible and heterozygous; consistently, GenomeScope estimated 1.31% of heterozygosity based on the 21-mer distribution of our Hifi reads (Fig. S2) (Marçais and Kingsford, 2011; Ranallo-Benavidez *et al.*, 2020). The alternative assembly was 928.6 Mbp in size and consists of 6,306 contigs with a N50 of 524 kbp. Genome annotation captured 40,207 gene models from 268 scaffolds for primary assembly, and 35,412 gene models from 3,211 contigs for alternative assembly (Table 1). Each assembly showed 96.9% and 83.4% BUSCO completeness, collectively supporting high heterozygosity.

The expanded gene families highlight the biological features of *S. habrochaites* LA1353

To examine whether our genome annotation supports the known biological features of *S. habrochaites*, we identified from the primary annotation set expanded orthologous groups and analyzed enriched PFAM domains (Fig. 2; Table S3). S. habrochaites is known as a rich source of disease resistance genes (Rick and Chetelat, 1995). Consistently, genes that contain domains found in NLRs, such as NB-ARC (PF00931) and Rx N-terminal domain (PF18052), and cell surface immune receptors, such as D-mannose binding lectin (PF01453) and S-locus glycoprotein domain (PF00954) were expanded. The number of self-incompatibility factors (PF05938) was the greatest in *S. habrochaites* LA1353, and DNA regulation and replication-associated domains were also enriched. *S. habrochaites* LA1353 develops long, dense trichomes and produces volatile metabolites that emanate a unique scent distinct from domesticated tomato, which together may function as protective barriers and repellents against pests (Bergau *et al.*, 2015; F., Wang *et al.*, 2020). Enzymes possibly involved in cell wall regulation and metabolite production were consistently expanded. Reflecting the adaptation of *S. habrochaites* LA1353 to a suboptimal environment at a high altitude of the Andes, Chlorophyll binding protein (PF00504) and genes associated with ATP production and usage appeared as significant hits.

The genome of *S. habrochaites* LA1353 displays high completeness for NLR loci assembly and annotation

Some NLRs proliferate through tandem duplication and with repetitive elements, forming complex NLR gene clusters (Kim *et al.*, 2017; Barragan *et al.*, 2019; Krasileva, 2019). Assembling such clusters and annotating these NLRs can be challenging for short reads and automatic annotation pipeline. For instance, we assembled 218 kbp of NLR gene clusters with HiFi reads in which we manually curated 18 NLRs and one NB-ARC fragment (Fig. 3A). In this region, automatic annotations fragmented, fused or only partially captured the NLR gene models. RepeatMasker classified some of the NLR loci as repeats (Bayer *et al.*, 2018). Genome assembly we performed with 150 bp paired-end reads and Supernova assembler for the same accession failed to recover most NLR loci despite sufficient sequencing coverage.

We accessed the quality of NLR loci assembly for the selected wild tomato genomes by mapping non-redundant SMRT RenSeq data of wild tomatoes, which contained translated sequences homologous to the NB-ARC domain, to the respective genomes (Fig. 3B). Over 90% of the read mapping rate was reported for *S. habrochaites* LA1353 and *S. pimpinellifolium* LA2093 genomes produced with PacBio long reads. This *S. habrochaites* genome clearly displayed higher completeness than the one we generated with 10X Genomics sequencing reads and the one previously assembled with Illumina sequencing reads (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014). About 80% of the HiFi reads were mapped to the *S. pennellii* genomes. The heterozygous *S. chilense* genome assembled with Illumina reads only had 45%. Although accession-level NLR variations would affect the mapping statistics, this result suggested that our draft genome of *S. habrochaites* LA1353 displays high completeness for NLR loci assembly compared to other wild tomato genomes.

The structural annotation pipeline predicted 198 proteins with the NB-ARC domain from the primary assembly. However, we manually curated 349 gene models that contain the NB-ARC domain by correcting chimeric gene models and capturing unpredicted NLRs (Fig. 3A and 3C). Such improvement was significant given that *S. pimpinellifolium* LA2093 ended up with only 141 putative NLRs annotated from automatic gene prediction while the actual NLR count should likely be comparable to *S. pimpinellifolium* LA1269 from our previous RenSeq data (Fig. 3B) (Seong *et al.*, 2020). As our gene models include putatively pseudogenized NLRs and NB-ARC fragments without proper MDAs, we selected 265 intact NLRs that have three or more major motifs over the NB-ARC domain equal to or longer than 160 amino acids. In the six tomato species studied with SMRT RenSeq, the intact NLR number varied from 204 to 241; the 265 intact NLR in *S. habrochaites* LA1353 is greater in number than those of other tomato species.

Most heterozygous local copy number variations appear in NRC-clade-related phylogenetic groups

In heterozygous diploid genome assembly, one of the haplotypes builds a contiguous primary assembly together with collapsed, nearly homozygous regions; the other haplotype is separated into an alternative assembly (Fig. 4A) (Cheng *et al.*, 2021). To investigate heterozygous NLR variations, we manually curated 206 intact NLRs from the alternative contigs (Fig. 3C). We then classified the entire intact NLRs into five categories based on the sequence similarity and phylogenetic relationship of the NLRs, as well as synteny between the primary and alternative assemblies (Fig. 4B; Table S4): (i) an NLR annotated only from a primary scaffold without any assembled alternative contigs is ‘common’ for both haplotypes; the two NLRs annotated from both primary and alternative assemblies with one-to-one orthology are (ii) ‘common’ if their CDS are identical or (iii) ‘genetically variable’ otherwise; NLRs annotated from either a primary scaffold or an alternative contig because the other has lost the NLR loci are characterized by (iv) ‘PAVs’ (presence/absence variations) or (v) ‘local CNVs’ if these NLR do not or do have a paralog within the 50 kb regions in 3’ or 5’ end, respectively.

During the classification, we removed any short or highly fragmented NLR gene models, because such gene structures are abnormal in comparison to closely related NLRs’ gene structures and may be indicative of pseudogenes. The majority of fragmented NLR gene models appeared in putative NRC-independent CNL clades, such as G4, G5, G7 and G10 (Fig. 4C and Table S4). Of the remaining gene models, 78 and 118 NLR pairs were common and genetically variable, respectively, and they appeared across all phylogenetic groups (Fig. 4D). To compare sequence evolution of the genetically variable NLR pairs, we randomly selected 118 pairs of heterozygous alleles from single-copy BUSCO genes or non-BUSCO genes that have at least one paralog in the genome. These heterozygous pairs had one-to-one orthology and followed a similar size distribution of the genetically variable NLRs. We then compared synonymous (dS) and non-synonymous (dN) substitution rates of these groups, respectively. All groups did not display noticeable differences in the dS distribution (Fig. 4E); however, increased dN significantly altered the dN/dS ratio of the NLRs, suggesting their unique haplotypic divergence (Fig. 4F). Besides, among heterozygous alleles were examples of frameshift mutations or transposon insertion only in one of the two alleles. Such events appeared across the phylogeny, which altogether possibly suggested distinct evolutionary dynamics of some heterozygous alleles.

Four and 54 NLRs were classified to have PAVs and local CNVs, respectively (Fig. 4D). This indicated that haplotype-specific gain or loss of NLRs not only is common but also preferentially appears in NLR gene clusters rather than isolated NLRs (Fig. 4B). Additionally, the local CNVs were phylogenetically biased (Fig. 4G). For instance, NRC-dependent CNL clades G1 and G14 had noticeably greater numbers of local CNVs than the other clades. For instance, despite the large clade size, no local CNVs were observed in the TNL group (GT) (Fig. 4C). Also, G3, G6 and G9 that are evolutionarily closely related to G1 and G14 only displayed single local CNV. Therefore, we conjectured that the frequent local CNVs in G1 and G14 are not simply attributed to the clade size but may be related to clade-specific NLR properties.

A few known NLR gene clusters display most of the heterozygous local CNVs with complex evolution

The heterozygous local CNVs observed in clades G1, G5 and G14 are concentrated on a few NLR gene clusters, and these clusters correspond to the loci from which experimentally validated, functional NLRs were identified (Fig. 5A). For instance, a cluster in scaffold109 and two closely located clusters in scaffold47 together had all 14 CNVs found in G1. Based on synteny to the tomato reference genome, the NLR loci in these scaffolds corresponded to *Hero* and *Mi-1.2* clusters (Fig. 5A) (Vos *et al.*, 1998; Ernst *et al.*, 2002). Similarly, single clusters in scaffold102 and scaffold109 contained all 13 and 6 local CNVs of G14 and G5, respectively, and these loci corresponded to *Rpi-amr1* (G14) and *Rpi-blb3* (G5) gene clusters from wild potatoes (Lokossou *et al.*, 2009; Witek *et al.*, 2021).

The synteny between homologous primary and alternative assemblies clearly indicated frequent homology breaks on the NLR loci (Fig. S3). Consistently, both the genetic organization and phylogenetic relationships of the NLRs with local CNVs hinted at complex evolution that has shaped these loci (Fig. 5B-D; Fig. S4; Table S5). Loss of one-to-one orthology appeared in all clusters. Some pairs, such as A12/B13, could be explained by accumulated mutation in a heterozygous allele and a recent duplication to B20 (Fig. 5B; Fig. S4A). In the *Mi-1.2* and *Rpi-amr1* clusters, only two heterozygous pairs have retained one-to-one orthology. C05/D06 and E01/F02 pairs, for instance, were phylogenetically distant although they occupy genomically proximal regions (Fig. 5C and 5D). The scenarios which could explain such findings are that the original orthologous allele was entirely swapped out by a paralogous NLR, or if attributed to sequence diversification, this would require more drastic processes like gene conversion through recombination. Phylogenetic distance is not necessarily congruent with physical distance (e.g. A12/A14 and C03/C04), suggesting that some paralogs did not result from duplication of the closest NLR (Fig. S4). Although differences exist in the TE profiles of primary and alternative contigs (Fig. 5B-D), we could not pinpoint particular TE-associated mechanisms that might have led to the cluster evolution.

NLR gains and losses may nearly equally contribute to the cluster diversification

To examine how heterozygous NLR variations arose in the *Hero*, *Mi-1.2* and *Rpi-amr1* clusters, we incorporated interspecies NLR data to provide evolutionary contexts. The data include previously produced NLR annotations for six self-compatible tomato species (Fig. 2C) (Seong *et al.*, 2020), as well as newly generated ones for heterozygous species with SMRT RenSeq (Table S1). As it is not possible to distinguish heterozygous NLR alleles from recently duplicated haplotypic paralogs, we focused on PAVs of orthologous NLRs in other tomato species. We then examined scenarios that can explain the architecture of the NLR clusters and phylogenetic relationships of NLRs with a minimum number of gene gains and losses (Fig. 6; Fig. S5).

The presence of orthologous genes in other tomato species suggests origin of the NLR prior to speciation. The absence of a heterozygous allele in this case can mean a haplotypic loss event (e.g., allele losses associated with D08 and F02 groups) (Fig. 6A and 6E; Fig. S5A). The presence of a paralogous copy with a monophyletic relationship with other *S. habrochaites* NLRs can point to recent haplotypic duplications (Fig. 6B, 6C, 6F and 6G). On the other hand, the emergence of phylogenetic singletons, such as C09 and E07, can be more plausibly attributed to duplication, followed by drastic sequence diversification, than independent multiple losses of orthologous alleles in other wild tomato species (Fig. 6D and 6G). In these scenarios, we explained the local CNVs with a minimum number of NLR loss and gain events. In the *Mi-1.2* cluster, 3 haplotypic NLR gains and losses were required to explain the heterozygous local CNVs. Similarly, 5 gains and 6 losses were mapped to the *Rpi-amr1* cluster, collectively suggesting that gains and losses of NLRs nearly equally contribute to diversifying the NLR clusters and shaping the local CNVs.

Discussion

S. habrochaites is a useful germplasm source of disease resistance genes that could enhance resistance of domesticated tomatoes against diverse pathogens. RenSeq is a rapid and cost-effective approach to study disease resistance genes; however, its application to heterozygous species is not trivial and requires more careful consideration as RenSeq may not fully resolve complex NLR loci during assembly and annotation. In comparison to RenSeq data from homozygous species, the data from heterozygous ones generated two to three times the numbers of contigs and gene models, which contained redundant duplicates of one another. This was likely because RenSeq only captures a small subset of genomic regions, while resolving both heterozygous and haplotypic NLR loci requires more extensive genomic contexts. Therefore, elucidating NLR evolution in heterozygous species requires whole genome sequencing data.

Whole genome sequencing data alone, however, is not sufficient to study NLRs. Resolving complex NLR gene clusters depends on genome assembly with long, accurate reads and manual curation (Fig. 3A). In fact, the whole genome assembly we performed with short reads from 10X Genomics sequencing for the same accession failed to recover the majority of the *Hero* gene cluster. It is possible that masking NLR loci as repetitive elements and automatic annotation pipeline resulted in a low number of annotated NLRs in our genome and other long read assemblies (Fig. 3A and 3C). The quality of genome assembly and annotation would, therefore, lead to different conclusions about NLR evolution. For instance, previous studies have suggested that some wild tomatoes, such as *S. chilense* and *S. pennellii*, have lost a subset of NLRs (Stam *et al.*, 2016; Stam *et al.*, 2019). However, the relatively low mapping rates of our RenSeq data on the genomes assembled with short reads suggested that many NLR loci might have not been properly assembled (Fig. 3B). Therefore, we conclude that higher-quality genome assembly and annotation would be necessary to confirm the significant NLR loss events in certain lineages of wild tomatoes and convey a complete story of NLR evolution.

The *S. habrochaites* genome assembly presented in our study also captures drastic heterozygous NLR variations. Our study found that local CNVs appear more commonly than PAVs, indicating that structural variations are largely concentrated on NLR gene clusters, not isolated NLRs. Local CNVs predominantly appear in the NRC-related phylogenetic families. Two CNL sensor clades, G1 and G14, contain half of the NLRs associated with local CNVs (Fig. 4G). Although expected to have limited expansion (Wu *et al.*, 2017), the conserved NRC helper clade (G8) had the third largest local CNVs. This could be due to co-evolution between sensors and helpers as hinted by the *Hero* gene cluster (Fig. 5B). All the local CNVs of G1 and G14 existed in *Hero*, *Mi-1.2* and *Rpi-amr1* (Fig. 5), which may have altered evolutionary dynamics with frequent duplications and rearrangement as suggested for *Arabidopsis* (Jiao and Schneeberger, 2020). Examining the PAVs of orthologous genes across 13 tomato species, we found that NLR gain and loss events may almost equally contribute to diversifying the gene clusters with possible accelerated sequence diversification for some loci (Fig. 6). Such processes diminish one-to-one orthology; the *Mi-1.2* and *Rpi-amr1* gene clusters, in particular, only have two pairs of NLRs that have maintained one-to-one orthology, suggesting each haplotype is possibly divergently evolving. The complexity of evolution suggested that events like unequal crossovers, gene conversions and strand slippage might be also actively involved in diversifying the NLR gene clusters (Barragan and Weigel, 2021). Overall, our data collectively suggests that highly heterozygous species often harbor two distinct NLR haplotypes; divergence of these haplotypes may provide more opportunities for heterozygous wild tomatoes to diversify resistance against pathogens.

The limitation of our data is that the haplotypes are not fully resolved. Future studies may sequence parental DNAs and generate haplotype-resolved genome assemblies with the trio-binning assembly. Controlled propagation of progeny and targeted sequencing of the *Hero*, *Mi-1.2* and *Rpi-amr1* clusters may also provide further insights into the evolution of complex NLR gene clusters. Additionally, not all *S. habrochaites* accessions are highly heterozygous. Unlike self-incompatible accessions like *S. habrochaites* LA1353 that populate southern Peru, the transition to self-compatibility is observed for the accessions found in central Peru and Ecuador (Broz, Randle, *et al.*, 2017). Comparative genomics between self-compatible and self-incompatible *S. habrochaites* accessions could provide more contexts on how heterozygosity contributes to the NLR diversity generation.

Data Availability

The *S. habrochaites* genome sequencing data, including PacBio raw reads, HiFi reads and linked-reads are available in the SRA in the National Center for Biotechnology Information (NCBI) under the BioProject accession PRJNA753882. Primary and alternative assemblies are deposited under PRJNA795504 and PRJNA795505. The HiFi reads from RenSeq can be accessed via PRJNA795506. All data, including genome assemblies, structural annotations and manually curated NLRs, can be accessed via Zenodo (10.5281/zenodo.5080564) or Sol Genomics Network (https://solgenomics.net).

Acknowledgements

We thank the members at the genome center at the University of California Davis and the Vincent J. Coates Genomics Sequencing Lab at the University of California Berkeley for DNA extraction and sequencing. Seed was kindly provided by the Tomato Genetics Resource Center (TGRC) at the University of California, Davis. We appreciate the advice from Roger T. Chetelat at the TGRC on growing *S. habrochaites*. We thank Sarah Song for some photography and Erin Baggs and Pierre Joubert for reviewing the manuscript. This research relied on the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley. This work was supported by a grant from the Gordon and Betty Moore Foundation to the 2Blades Foundation (GBMF4725) and the Innovative Genomics Institute. Kyungyong Seong is supported by the Berkeley fellowship. Ksenia V Krasileva and China Lunde Shaw are supported by the Innovative Genomics Institute and NIH Director's Award.

Contribution

K.S. conceived and conducted the research and wrote the manuscript. M.L. performed DNA extraction and library preparation for RenSeq on the wild tomatoes. C.L.S. conducted pollination and field work. E.S. performed NLR classification. B.J.S. and K.V.K. supervised the research. All authors contributed to editing and reviewing this manuscript.

References

Arnoux, S., Fraïsse, C. and Sauvage, C. (2021) Genomic inference of complex domestication histories in three Solanaceae species. *J Evol Biol*, **34**, 270–283.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C. and Segata, N. (2015) Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, **3**, e1029.

Baggs, E., Dagdas, G. and Krasileva, K. (2017) NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Current Opinion in Plant Biology*, **38**, 59–67.

Bao, Z., Meng, F., Strickler, S.R., Dunham, D.M., Munkvold, K.R. and Martin, G.B. (2015) Identification of a Candidate Gene in *Solanum habrochaites* for Resistance to a Race 1 Strain of *Pseudomonas syringae* pv. *tomato*. *Plant Genome*, **8**.

Barragan, A.C. and Weigel, D. (2021) Plant NLR diversity: the known unknowns of pan-NLRomes. *The Plant Cell*, **33**, 814–831.

Barragan, C.A., Wu, R., Kim, S.-T., *et al.* (2019) RPW8/HR repeats control NLR activation in *Arabidopsis thaliana* G. Coaker, ed. *PLoS Genet*, **15**, e1008313.

Bauchet, G. and Causse, M. (2012) Genetic Diversity in Tomato (*Solanum lycopersicum*) and Its Wild Relatives. In M. Caliskan, ed. *Genetic Diversity in Plants*. InTech.

Bayer, P.E., Edwards, D. and Batley, J. (2018) Bias in resistance gene prediction due to repeat masking. *Nature Plants*, **4**, 762–765.

Bedinger, P.A., Chetelat, R.T., McClure, B., *et al.* (2011) Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. *Sex Plant Reprod*, **24**, 171–187.

Bergau, N., Bennewitz, S., Syrowatka, F., Hause, G. and Tissier, A. (2015) The development of type VI glandular trichomes in the cultivated tomato *Solanum lycopersicum* and a related wild species *S. habrochaites*. *BMC Plant Biol*, **15**, 289.

Bolger, A., Scossa, F., Bolger, M.E., *et al.* (2014) The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat Genet*, **46**, 1034–1038.

Broz, A.K., Guerrero, R.F., Randle, A.M., Baek, Y.S., Hahn, M.W. and Bedinger, P.A. (2017) Transcriptomic analysis links gene expression to unilateral pollen-pistil reproductive barriers. *BMC Plant Biol*, **17**, 81.

Broz, A.K., Randle, A.M., Sianta, S.A., Tovar□Méndez, A., McClure, B. and Bedinger, P.A. (2017) Mating system transitions in *Solanum habrochaites* impact interactions between populations and species. *New Phytol*, **213**, 440–454.

Broz, A.K., Simpson□Van Dam, A., Tovar□Méndez, A., Hahn, M.W., McClure, B. and Bedinger, P.A. (2021) Spread of self□compatibility constrained by an intrapopulation crossing barrier. *New Phytol*, **231**, 878–891.

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

Brůna, T., Lomsadze, A. and Borodovsky, M. (2020) GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, **2**, lqaa026.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A. and Yandell, M. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.

Capella-Gutierrez, S., Silla-Martinez, J.M. and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Chaudhary, R. and Atamian, H. (2017) Resistance-Gene-Mediated Defense Responses against Biotic Stresses in the Crop Model Plant Tomato. *J Plant Pathol Microbiol*, **08**.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, **18**, 170–175.

Copati, M.G.F., Alves, F.M., Dariva, F.D., Pessoa, H.P., Dias, F.O., Carneiro, P.C.S., Carneiro, D.J.H. and Nick, C. (2019) Resistance of the wild tomato *Solanum habrochaites* to *Phytophthora infestans* is governed by a major gene and polygenes. *An. Acad. Bras. Ciênc.*, **91**, e20190149.

Darling, A.E., Mau, B. and Perna, N.T. (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement J. E. Stajich, ed. *PLoS ONE*, **5**, e11147.

Dawson, N.L., Sillitoe, I., Lees, J.G., Lam, S.D. and Orengo, C.A. (2017) CATH-Gene3D: Generation of the Resource and Its Use in Obtaining Structural and Functional Annotations for Protein Sequences. In C. H. Wu, C. N. Arighi, and K. E. Ross, eds. *Protein Bioinformatics*. Methods in Molecular Biology. New York, NY: Springer New York, pp. 79–110.

Eddy, S.R. (2011) Accelerated Profile HMM Searches W. R. Pearson, ed. *PLoS Comput Biol*, **7**, e1002195.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*, **20**, 238.

Ernst, K., Kumar, A., Kriseleit, D., Kloos, D.-U., Phillips, M.S. and Ganal, M.W. (2002) The broad-spectrum potato cyst nematode resistance gene (Hero) from tomato is the only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. *Plant J*, **31**, 127–136.

Fan, P., Miller, A.M., Liu, X., Jones, A.D. and Last, R.L. (2017) Evolution of a flipped pathway creates metabolic innovation in tomato trichomes through BAHD enzyme promiscuity. *Nat Commun*, **8**, 2080

Fernandez-Pozo, N., Menda, N., Edwards, J.D., *et al.* (2015) The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Research*, **43**, D1036–D1041.

Finkers, R., Heusden, A.W. van, Meijer-Dekens, F., Kan, J.A.L. van, Maris, P. and Lindhout, P. (2007) The construction of a *Solanum habrochaites* LYC4 introgression line population and the identification of QTLs for resistance to Botrytis cinerea. *Theor Appl Genet*, **114**, 1071–1080.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*, **117**, 9451–9457.

Grabherr, M.G., Haas, B.J., Yassour, M., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, **29**, 644–652.

Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y. and Durbin, R. (2020) Identifying and removing haplotypic duplication in primary genome assemblies A. Valencia, ed. *Bioinformatics*, **36**, 2896–2898.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hosmani, P.S., Flores-Gonzalez, M., Geest, H. van de, *et al.* (2019) An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv*. doi: https://doi.org/10.1101/767764

Jiao, W.-B. and Schneeberger, K. (2020) Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun*, **11**, 989.

Jones, P., Binns, D., Chang, H.-Y., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Jupe, F., Witek, K., Verweij, W., *et al.* (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J*, **76**, 530–544.

Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kilambi, H.V., Manda, K., Rai, A., Charakana, C., Bagri, J., Sharma, R. and Sreelakshmi, Y. (2017) Green-fruited *Solanum habrochaites* lacks fruit-specific carotenogenesis due to metabolic and structural blocks. *Journal of Experimental Botany*, **68**, 4803–4819.

Kim, S., Park, J., Yeom, S.-I., *et al.* (2017) New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol*, **18**, 210.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Kourelis, J., Sakai, T., Adachi, H. and Kamoun, S. (2021) RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family X. Dong, ed. *PLoS Biol*, **19**, e3001124.

Krasileva, K.V. (2019) The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Current Opinion in Plant Biology*, **48**, 18–25.

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A. and Zdobnov, E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, **47**, D807–D811.

Li, J., Liu, L., Bai, Y., *et al.* (2011) Identification and mapping of quantitative resistance to late blight (*Phytophthora infestans*) in *Solanum habrochaites* LA1777. *Euphytica*, **179**, 427–438.

Lokossou, A.A., Park, T., Arkel, G. van, *et al.* (2009) Exploiting Knowledge of *R/Avr* Genes to Rapidly Clone a New LZ-NBS-LRR Family of Late Blight Resistance Genes from Potato Linkage Group IV. *MPMI*, **22**, 630–641.

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.*, **17**, 10.

Mendes, F.K., Vanderpool, D., Fulton, B. and Hahn, M.W. (2021) CAFE 5 models variation in evolutionary rates among gene families P. Robinson, ed. *Bioinformatics*, **36**, 5516–5518.

Mistry, J., Chuguransky, S., Williams, L., *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Research*, **49**, D412–D419.

Mueller, L.A., Solow, T.H., Taylor, N., *et al.* (2005) The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond. *Plant Physiology*, **138**, 1310–1317.

Nowakowska, M., Nowicki, M., Kłosińska, U., Maciorowski, R. and Kozik, E.U. (2014) Appraisal of Artificial Screening Techniques of Tomato to Accurately Reflect Field Performance of the Late Blight Resistance M. Gijzen, ed. *PLoS ONE*, **9**, e109328.

Nurk, S., Walenz, B.P., Rhie, A., *et al.* (2020) HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.*, **30**, 1291–1305.

Pease, J.B., Haak, D.C., Hahn, M.W. and Moyle, L.C. (2016) Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation D. Penny, ed. *PLoS Biol*, **14**, e1002379.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*, **11**, 1432.

Rick, C.M. and Chetelat, R.T. (1995) UTILIZATION OF RELATED WILD SPECIES FOR TOMATO IMPROVEMENT. *Acta Hortic.*, 21–38.

Seo, E., Kim, S., Yeom, S.-I. and Choi, D. (2016) Genome-Wide Comparative Analyses Reveal the Dynamic

Evolution of Nucleotide-Binding Leucine-Rich Repeat Gene Family among Solanaceae Plants. *Front. Plant Sci.*, **7**.

Seong, K., Seo, E., Witek, K., Li, M. and Staskawicz, B. (2020) Evolution of NLR resistance genes with noncanonical N□terminal domains in wild tomato species. *New Phytol*, **227**, 1530–1543.

Seppey, M., Manni, M. and Zdobnov, E.M. (2019) BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar, ed. *Gene Prediction*. Methods in Molecular Biology. New York, NY: Springer New York, pp. 227–245.

Slater, G. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Smit, A.F., Hubley, R. and Green, P. (2013) *RepeatMasker Open-4.0*, Available at: http://www.repeatmasker.org.

Stam, R., Nosenko, T., Hörger, A.C., Stephan, W., Seidel, M., Kuhn, J.M.M., Haberer, G. and Tellier, A. (2019) The *de Novo* Reference Genome and Transcriptome Assemblies of the Wild Tomato Species *Solanum chilense* Highlights Birth and Death of NLR Genes Between Tomato Species. *G3 Genes|Genomes|Genetics*, **9**, 3933–3941.

Stam, R., Scheikl, D. and Tellier, A. (2016) Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population. *Genome Biol Evol*, **8**, 1501–1515.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, **34**, W435–W439.

Steuernagel, B., Jupe, F., Witek, K., Jones, J.D.G. and Wulff, B.B.H. (2015) NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics*, **31**, 1665–1667.

Thapa, S.P., Miyao, E.M., Michael Davis, R. and Coaker, G. (2015) Identification of QTLs controlling resistance to Pseudomonas syringae pv. tomato race 1 strains from the wild tomato, *Solanum habrochaites* LA1777. *Theor Appl Genet*, **128**, 681–692.

The 100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., *et al.* (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J*, **80**, 136–148.

Tohge, T., Scossa, F., Wendenburg, R., *et al.* (2020) Exploiting Natural Variation in Tomato to Define Pathway Structure and Metabolic Regulation of Fruit Polyphenolics in the Lycopersicum Complex. *Molecular Plant*, **13**, 1027–1046.

Vasimuddin, Md., Misra, S., Li, H. and Aluru, S. (2019) Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. Rio de Janeiro, Brazil: IEEE, pp. 314–324.

Vos, P., Simons, G., Jesse, T., *et al.* (1998) The tomato Mi-1 gene confers resistance to both root-knot nematodes and potato aphids. *Nat Biotechnol*, **16**, 1365–1369.

Wang, F., Park, Y.-L. and Gutensohn, M. (2020) Glandular trichome-derived sesquiterpenes of wild tomato accessions (*Solanum habrochaites*) affect aphid performance and feeding behavior. *Phytochemistry*, **180**, 112532.

Wang, X., Gao, L., Jiao, C., *et al.* (2020) Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat Commun*, **11**, 5817.

Warren, R.L., Yang, C., Vandervalk, B.P., Behsaz, B., Lagman, A., Jones, S.J.M. and Birol, I. (2015) LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaSci*, **4**, 35.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. and Gough, J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, **37**, D380–D386.

Witek, K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D. and Jones, J.D.G. (2016) Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. *Nat Biotechnol*, **34**, 656–660.

Witek, K., Lin, X., Karki, H.S., *et al.* (2021) A complex resistance locus in *Solanum americanum* recognizes a conserved Phytophthora effector. *Nat. Plants*, **7**, 198–208.

Wu, C.-H., Abd-El-Haliem, A., Bozkurt, T.O., Belhaj, K., Terauchi, R., Vossen, J.H. and Kamoun, S. (2017) NLR network mediates immunity to diverse plant pathogens. *Proc Natl Acad Sci USA*, **114**, 8113–8118.

Yeo, S., Coombe, L., Warren, R.L., Chu, J. and Birol, I. (2018) ARCS: scaffolding genome drafts with linked reads C. Sahinalp, ed. *Bioinformatics*, **34**, 725–731.

Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, **16**, 284–287.

Figures and Tables

Table 1. Genome sequencing, assembly and annotation statistics

Sequencing coverage was estimated based on the expected genome size of 2 Gbp for both haplotypes. Genome assembly statistics was calculated with QUAST (Gurevich *et al.*, 2013). BUSCO was run on the protein annotation sets with the Solanales_odb10 database.

Figure 1. The morphological feature of *Solanum habrochaites* LA1353

A. S. *habrochaites* growing in the field at Berkeley, California, USA in November 2021. The plants were about three-month-old. **B**. Harvested S. *habrochaites* leaves, flowers and fruits. **C**. Flowers, fruits and leaves of *Solanum lycopersicum* Heinz and *S. habrochaites* LA1353. The white bars next to the organs indicate 1 cm. *S. habrochaites* fruits were about one month old.

Figure 2. Expanded gene families in *S. habrochaites* LA1353

A. Enriched PFAM domains in rapidly expanded gene families identified in *S. habrochaites* LA1353 by CAFE. After orthologous groups were identified for all tomato species used for the analysis by OrthoFinder, the gene counts were compared in a pairwise manner between *S. habrochaites* LA1353 and another tomato species by CAFE to identify expanded orthologous groups in *S. habrochaites*. PFAM enrichment tests were performed for all rapidly expanded gene families using clusterProfiler. The cut-offs for p-value and q-value were 0.05, and the value of -$\log_{10}$(q-value) is capped at 15 for visualization. The empty boxes indicate no enrichment. **B**. The counts of the genes with the given PFAM entry. To prevent incompletely annotated gene fragments from inflating the gene counts, the genes were considered significant only if the predicted domain region is equal or larger than 60% of the full domain. Some of the PFAM domain names were simplified for visualization.

Figure 3. The completeness of NLR loci assembly and annotation of the *S. habrochaites* LA1353 draft genome

A. The 218 kb NLR gene cluster in scaffold 109 (346,000-564,000). 1st row: manually curated NLR gene models. Alternating colors are to indicate separate gene models. Exon and intron structures are distinguished by the height of the box, with larger boxes indicating exons. 2nd and 3rd rows: automatically annotated NLR gene models by BRAKER and in the final annotation set. 4th row: repetitive elements annotated by RepeatMasker. Only the known classes of transposons ≥ 300 nucleotides are indicated. 5th row: the contigs assembled with Supernova and 150 bp linked-reads from 10X Genomics sequencing mapped to the region. Supplementary alignments are not shown. 6th row: the sequencing coverage from 10X Genomics sequencing reads used for Supernova assembly. **B**. The estimated completeness of NLR loci assembly. The RenSeq data were obtained for *S. pimpinellifolium* LA0411, *S. chilense* LA1932, *S. pennellii* LA1272 and *S. habrochaites* LA1353. The RenSeq data typically display inconsistent coverage across NLR loci and also include non-NLR regions. Therefore, the HiFi reads were filtered to reduce redundancy, and only those that contain homologous sequences to the NB-ARC domain were selected and mapped. Of the total number of the reads, those mapped as primary alignments with the coverage 80% were counted. **C**. The hypothetical species tree reconstructed based on the previous study (Bedinger et al., 2011) and the number of putative NLRs and intact NLRs. The branch length is not on an evolutionary scale. The NLR annotations were obtained from either SMRT RenSeq or whole genome sequencing (WGS) and assembly with long reads, such as PacBio or Nanopore, and short reads, such as Illumina. The total gene model was identified primarily by searching for the gene models that contain the NB-ARC domain. Those that contain NB-ARC domain with three major motifs in order and ≥ 160 amino acids were counted as intact. Dashed lines indicate missing genome annotations or RenSeq data for the species.

Figure 4. **Heterozygous NLR variations in *S. habrochaites* LA1353**

A. The schematics of heterozygous diploid genome assembly and NLR annotation. As haplotypes could not be fully resolved in the absence of parental DNAs, nearly or completely homozygous regions build contiguous primary assembly with one haplotype, while the other haplotype is separated to alternative assembly. **B**. The five categories of NLR variations. The provided phylogeny and synteny are representative but may not be the sole cases to explain each class. The phylogeny was created with GraPhlAn (Asnicar *et al.*, 2015). **C.** A phylogenetic tree inferred for intact NLRs annotated from primary and alternative assemblies. The tree includes experimentally validated NLRs as references, and they are listed on the right. At the outer ring, the five classes given for each NLR are indicated with colors. **D**. The number of NLRs that belong to each class given in B. The colors correspond to the outer ring legend given in C. **E** and **F.** The distribution of synonymous substitution rates (dS) and nonsynonymous substitution rates (dN)/dS of three groups. 118 pairs of heterozygous alleles with one-to-one orthology were randomly selected from complete single-copy BUSCO genes or non-BUSCO genes that have at least one paralog in the genome (Others), following the size distribution of the genetically variable NLR pairs. According to the two-sided Kolmogorov-Smirnov test, the distribution of dS did not differ significantly, whereas the distribution of dN/dS was ($P = 1E^{-09}$ for Others vs. NLRs and $P = 2E^{-10}$ for Busco vs. NLRs). **G**. The number of PAVs and local CNVs found in the clades G1, G5, G8 and G14 that have the largest number of local CNVs.

Figure 5. NLR gene clusters that display large numbers of heterozygous CNVs

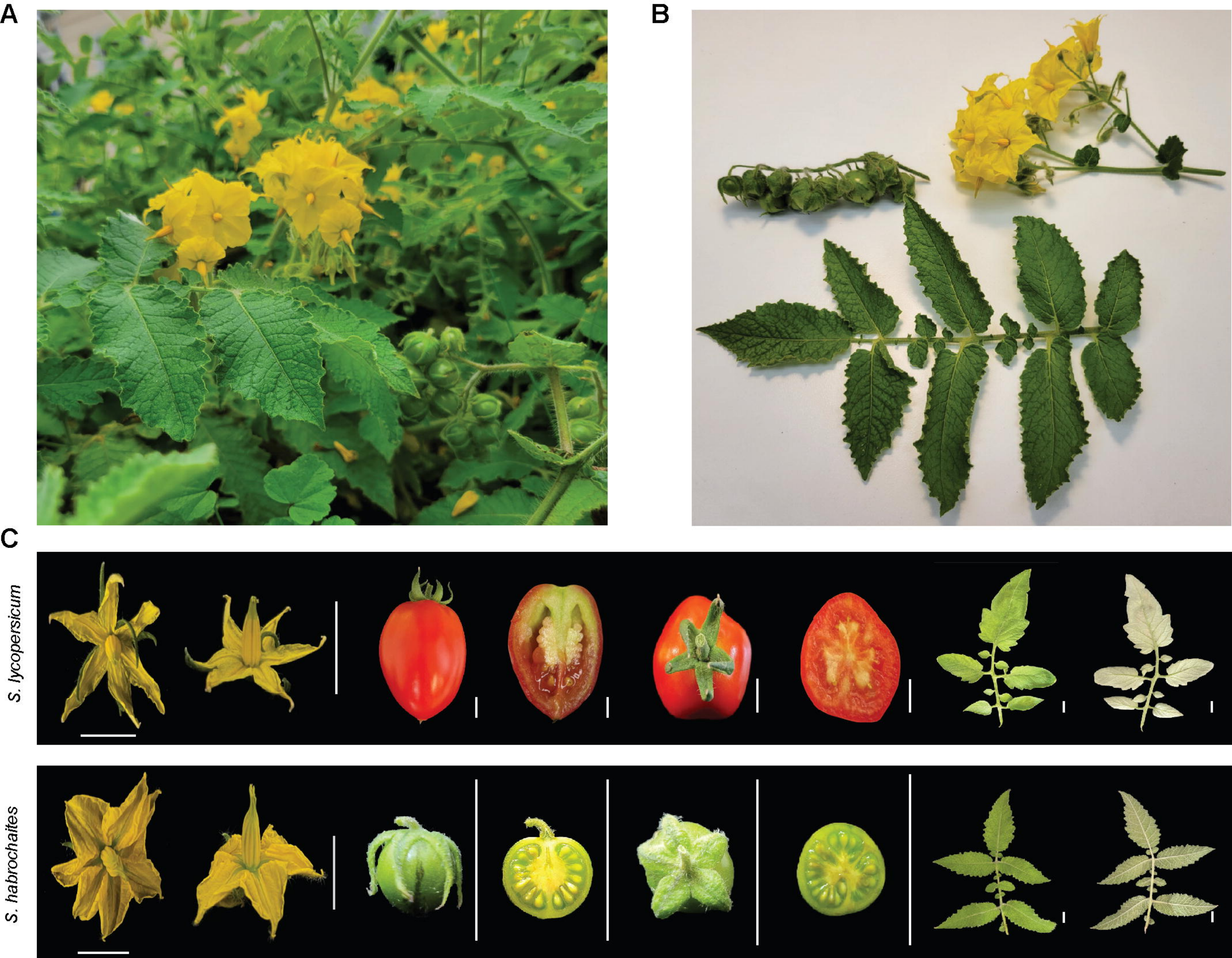
A. Synteny between reference chromosomes of *Solanum lycopersicum* or *Solanum tuberosum* and primary scaffolds of *S. habrochaites* LA1353. Although the *Rpi-blb3* cluster is indicated in chromosome 4 of *S. lycopersicum*, this cluster was initially identified in potato. **B-D**. Schematics for NLR organization in the *Hero*, *Mi-1.2* and *Rpi-amr1* NLR clusters in *S. habrochaites* LA1353. The blocks indicate manually curated NLRs and transposable elements (TEs) annotated by RepeatMasker. The NLR blocks reflect gene lengths but not exon-intron structures; the TE blocks do not reflect the orientation. Genomic synteny was detected with progressiveMauve (Darling *et al.*, 2010) and is highlighted in yellow. NLRs with one-to-one orthology were identified from phylogenetic trees (Fig. S4) and are highlighted in blue. The figures were generated with gggenomes (https://github.com/thackl/gggenomes). **B**. A17, A18 and A19 belong to the common class, as no alternative assembly that covers this region was found. Helper NLRs (G8) are indicated with red labels.

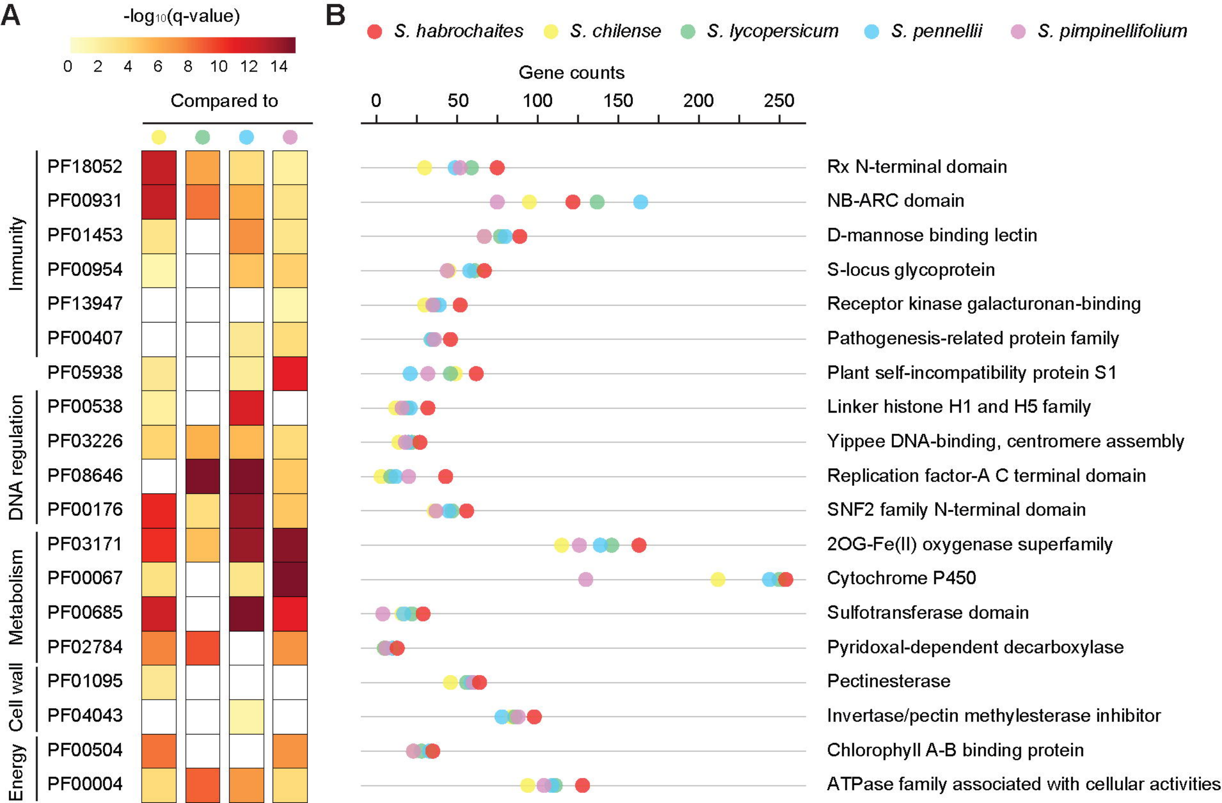
Figure 6. Interspecies NLR variations and suggested mechanism of heterozygous variations

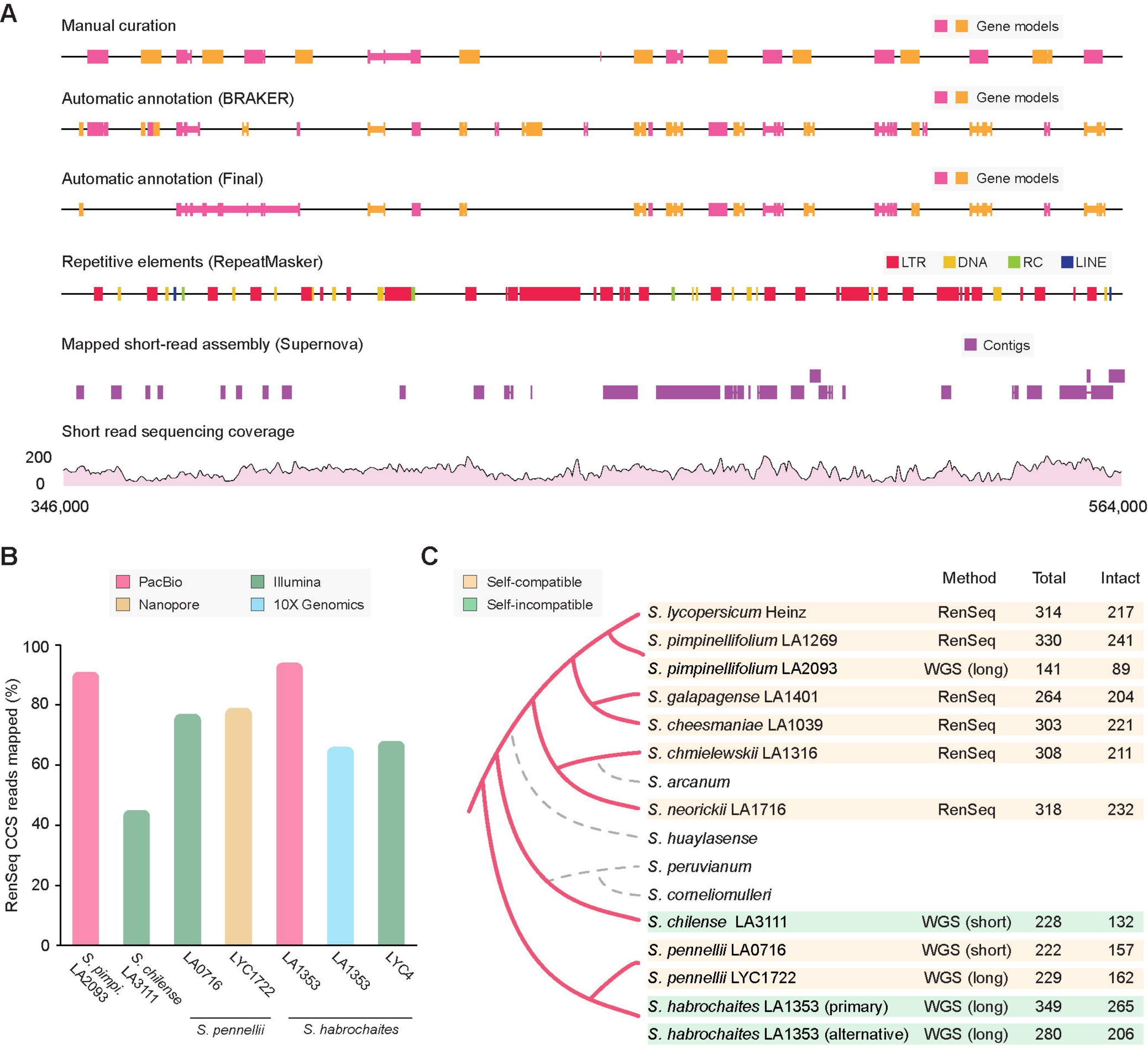
A and **E**. Interspecies PAVs and evolution of *Mi-1.2* and *Rpi-amr1* clusters in *S. habrochaites* LA1353. NLRs in *Mi-1.2* and *Rpi-amr1* clusters are depicted in the hypothetical coordinates. The labels of NLRs are consistent with those in Figure 5. Orthologous genes were phylogenetically identified from other 12 tomato species, and their presence is indicated with colored boxes, regardless of the copy numbers. Putative evolutionary events—gain, loss and diversification—are mapped for NLRs from *S. habrochaites* LA1353 to explain haplotypic NLR variations and organization of NLRs. **B, C, D, F** and **G**. Simplified subsets of NLR gene trees. As RenSeq data for heterozygous species cannot be fully resolved, we focused on presence/absence variations. Therefore, one copy of NLR was chosen from the RenSeq data of each of the 12 tomato species, and the trees were pruned accordingly. The names are indicated only for NLRs from *S. habrochaites* LA1353. Red dots on the nodes indicate bootstrap ≥ 70.

Table 1. Genome sequencing, assembly and annotation statistics

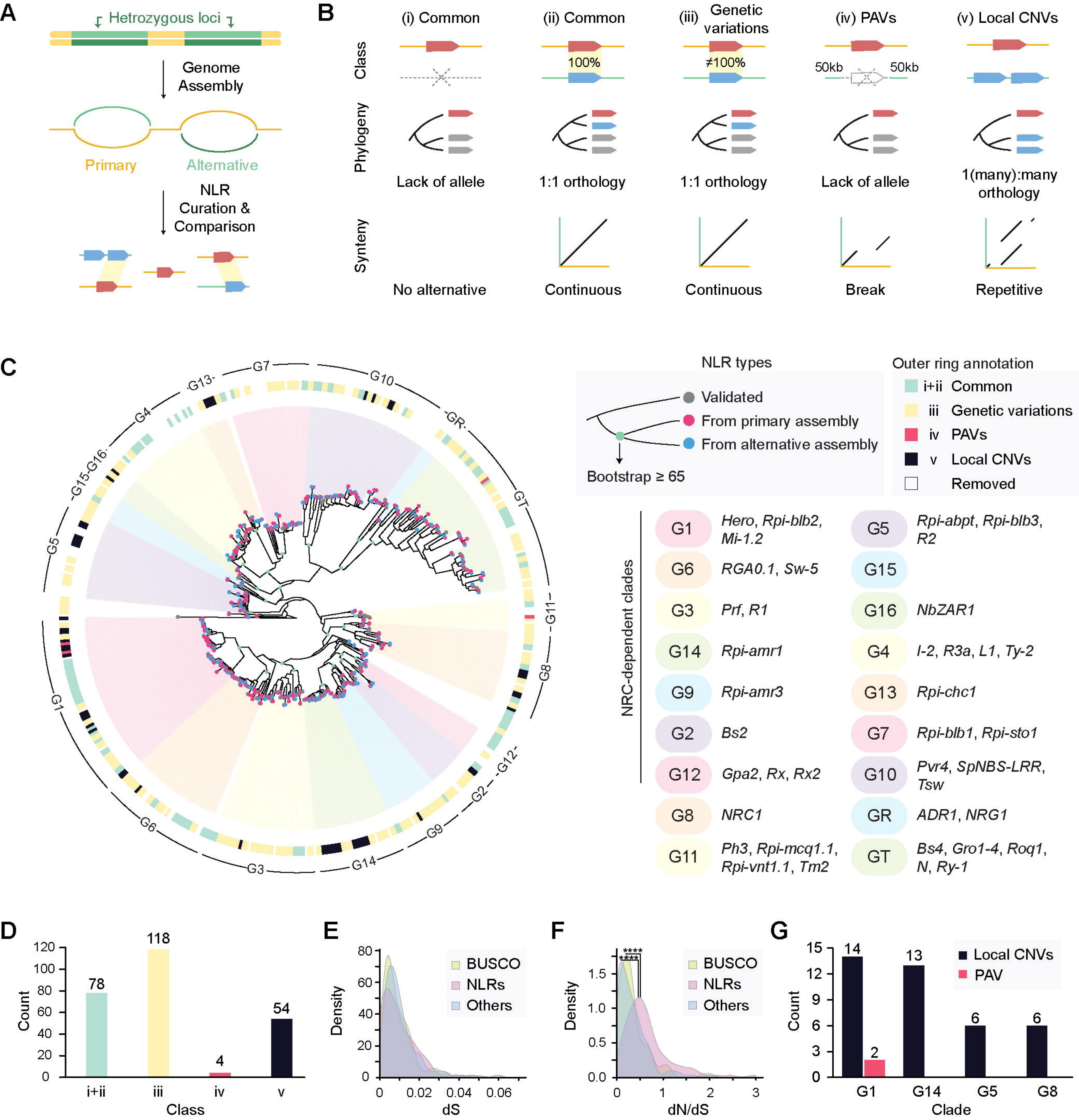
Sequencing coverage	For each haplotype	
PacBio (subreads)	175.1 X	
PacBio (HiFi)	10.8 X	
10X Genomics	56.6 X	
Genome assembly	**Primary assembly**	**Alternative assembly**
Size	981.2 Mb	928.6 Mbp
Contigs/scaffolds	794	6,306
N50	6.7 Mbp	525 kbp
L50	42	469
Repetitive elements	**Primary assembly**	**Alternative assembly**
Retroelements	559.1 Mb	521.0 Mb
DNA transposons	63.7 Mb	58.1 Mb
Total masked	719.4 Mb	689.2 Mb
Genome annotation	**Primary assembly**	**Alternative assembly**
Gene model	40,207	35,412
Contigs/scaffolds with gene models	268	3,211
BUSCO		
Complete, single-copy	5,604 (94.2%)	4,739 (79.6%)
Complete, duplicated	162 (2.7%)	223 (3.7%)
Fragmented	65 (1.1%)	86 (1.4%)
Missing	119 (2.0%)	902 (15.3%)

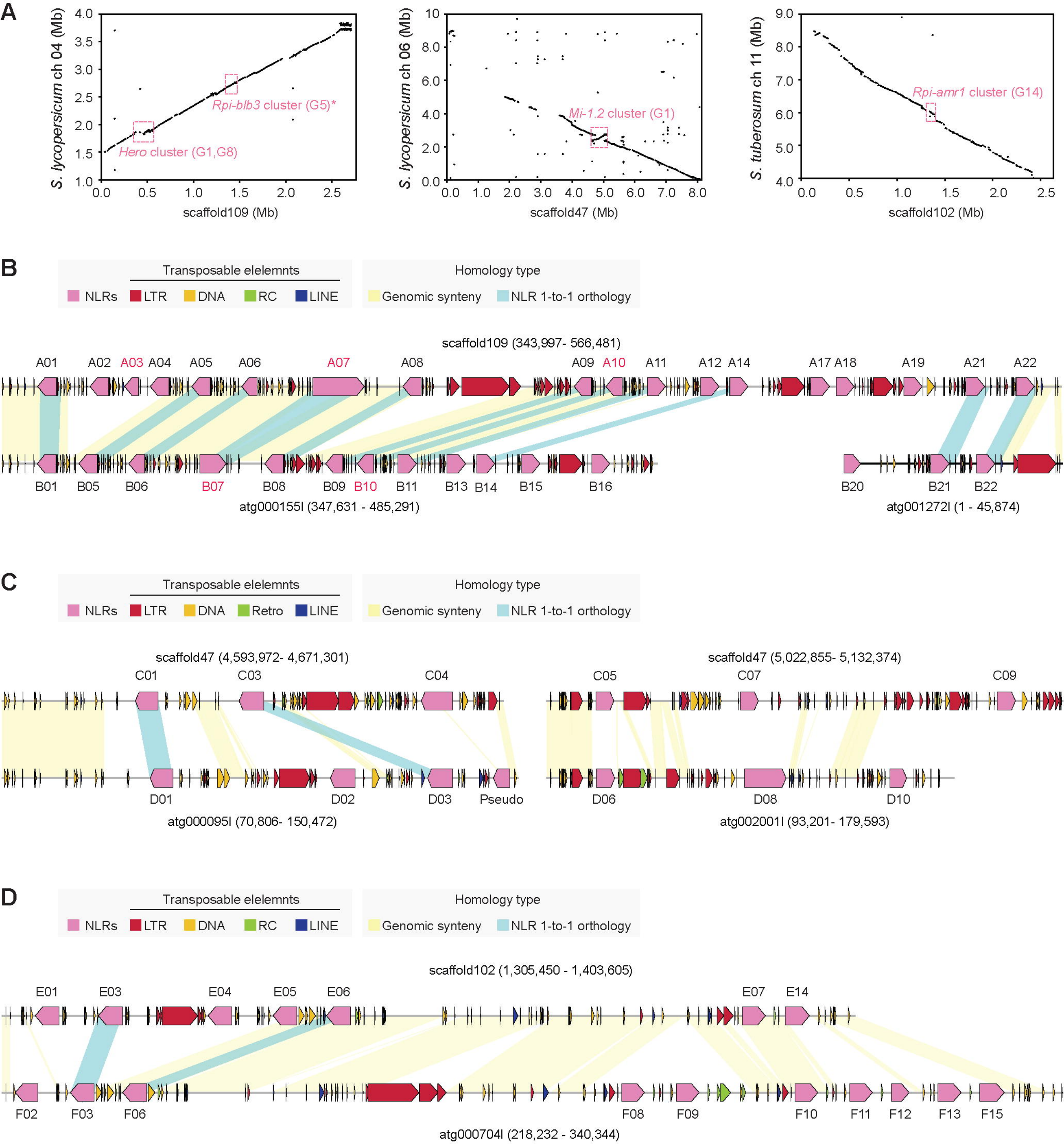

A
B
C
S. lycopersicum
S. habrochaites

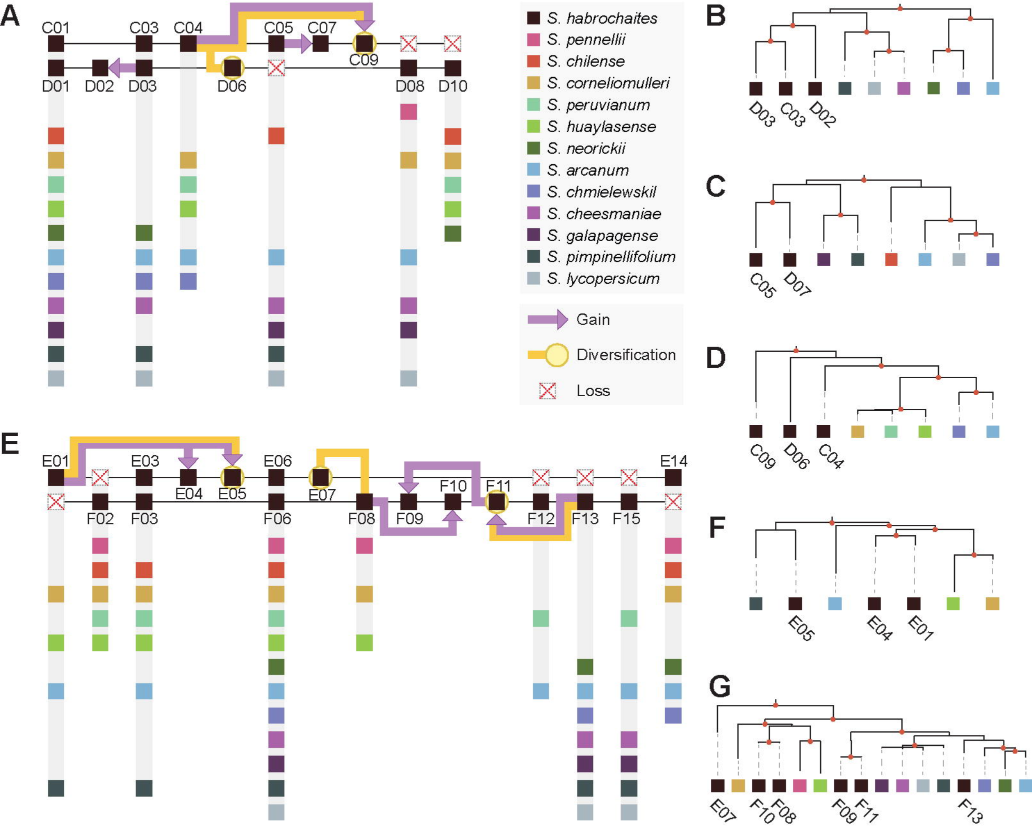

A
-log10(q-value)
0 2 4 6 8 10 12 14
Compared to
Immunity
DNA regulation
Metabolism
Cell wall
Energy
PF18052
PF00931
PF01453
PF00954
PF13947
PF00407
PF05938
PF00538
PF03226
PF08646
PF00176
PF03171
PF00067
PF00685
PF02784
PF01095
PF04043
PF00504
PF00004
B
S. habrochaites
S. chilense
S. lycopersicum
S. pennellii
S. pimpinellifolium
Gene counts
0 50 100 150 200 250
Rx N-terminal domain
NB-ARC domain
D-mannose binding lectin
S-locus glycoprotein
Receptor kinase galacturonan-binding
Pathogenesis-related protein family
Plant self-incompatibility protein S1
Linker histone H1 and H5 family
Yippee DNA-binding, centromere assembly
Replication factor-A C terminal domain
SNF2 family N-terminal domain
2OG-Fe(II) oxygenase superfamily
Cytochrome P450
Sulfotransferase domain
Pyridoxal-dependent decarboxylase
Pectinesterase
Invertase/pectin methylesterase inhibitor
Chlorophyll A-B binding protein
ATPase family associated with cellular activities



	Method	Total	Intact
S. lycopersicum Heinz	RenSeq	314	217
S. pimpinellifolium LA1269	RenSeq	330	241
S. pimpinellifolium LA2093	WGS (long)	141	89
S. galapagense LA1401	RenSeq	264	204
S. cheesmaniae LA1039	RenSeq	303	221
S. chmielewskii LA1316	RenSeq	308	211
S. arcanum			
S. neorickii LA1716	RenSeq	318	232
S. huaylasense			
S. peruvianum			
S. corneliomulleri			
S. chilense LA3111	WGS (short)	228	132
S. pennellii LA0716	WGS (short)	222	157
S. pennellii LYC1722	WGS (long)	229	162
S. habrochaites LA1353 (primary)	WGS (long)	349	265
S. habrochaites LA1353 (alternative)	WGS (long)	280	206


A
Hetrozygous loci
Genome Assembly
Primary
Alternative
NLR Curation & Comparison
B
(i) Common
(ii) Common
(iii) Genetic variations
(iv) PAVs
(v) Local CNVs
Class
100%
≠100%
50kb
50kb
Phylogeny
Lack of allele
1:1 orthology
1:1 orthology
Lack of allele
1(many):many orthology
Synteny
No alternative
Continuous
Continuous
Break
Repetitive
C
G7
G10
GR
GT
G11
G8
G12
G2
G9
G14
G3
G6
G1
G5
G15-G16
G4
G13
NLR types
Validated
From primary assembly
From alternative assembly
Bootstrap ≥ 65
Outer ring annotation
i+ii Common
iii Genetic variations
iv PAVs
v Local CNVs
Removed
NRC-dependent clades
G1 Hero, Rpi-blb2, Mi-1.2
G6 RGA0.1, Sw-5
G3 Prf, R1
G14 Rpi-amr1
G9 Rpi-amr3
G2 Bs2
G12 Gpa2, Rx, Rx2
G8 NRC1
G11 Ph3, Rpi-mcq1.1, Rpi-vnt1.1, Tm2
G5 Rpi-abpt, Rpi-blb3, R2
G15
G16 NbZAR1
G4 I-2, R3a, L1, Ty-2
G13 Rpi-chc1
G7 Rpi-blb1, Rpi-sto1
G10 Pvr4, SpNBS-LRR, Tsw
GR ADR1, NRG1
GT Bs4, Gro1-4, Roq1, N, Ry-1
D
Count
120
100
80
60
40
20
0
78
118
4
54
i+ii
iii
iv
v
Class
E
Density
80
60
40
20
0
0
0.02
0.04
0.06
dS
BUSCO
NLRs
Others
F
Density
1.5
1.0
0.5
0.0
****
****
0
1
2
3
dN/dS
BUSCO
NLRs
Others
G
Count
15
12
9
6
3
0
14
2
13
6
6
G1
G14
G5
G8
Clade
Local CNVs
PAV


A
S. lycopersicum ch 04 (Mb)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.0
0.5
1.0
1.5
2.0
2.5
scaffold109 (Mb)
Rpi-blb3 cluster (G5)*
Hero cluster (G1,G8)
S. lycopersicum ch 06 (Mb)
10
8.0
6.0
4.0
2.0
0.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
scaffold47 (Mb)
Mi-1.2 cluster (G1)
S. tuberosum ch 11 (Mb)
9.0
8.0
7.0
6.0
5.0
4.0
0.0
0.5
1.0
1.5
2.0
2.5
scaffold102 (Mb)
Rpi-amr1 cluster (G14)
B
Transposable elelemnts
NLRs
LTR
DNA
RC
LINE
Homology type
Genomic synteny
NLR 1-to-1 orthology
scaffold109 (343,997- 566,481)
A01
A02
A03
A04
A05
A06
A07
A08
A09
A10
A11
A12
A14
A17
A18
A19
A21
A22
B01
B05
B06
B07
B08
B09
B10
B11
B13
B14
B15
B16
B20
B21
B22
atg000155l (347,631 - 485,291)
atg001272l (1 - 45,874)
C
Transposable elelemnts
NLRs
LTR
DNA
Retro
LINE
Homology type
Genomic synteny
NLR 1-to-1 orthology
scaffold47 (4,593,972- 4,671,301)
C01
C03
C04
scaffold47 (5,022,855- 5,132,374)
C05
C07
C09
D01
D02
D03
Pseudo
atg000095l (70,806- 150,472)
D06
D08
D10
atg002001l (93,201- 179,593)
D
Transposable elelemnts
NLRs
LTR
DNA
RC
LINE
Homology type
Genomic synteny
NLR 1-to-1 orthology
scaffold102 (1,305,450 - 1,403,605)
E01
E03
E04
E05
E06
E07
E14
F02
F03
F06
F08
F09
F10
F11
F12
F13
F15
atg000704l (218,232 - 340,344)


A
C01
C03
C04
C05
C07
C09
D01
D02
D03
D06
D08
D10
S. habrochaites
S. pennellii
S. chilense
S. corneliomulleri
S. peruvianum
S. huaylasense
S. neorickii
S. arcanum
S. chmielewskii
S. cheesmaniae
S. galapagense
S. pimpinellifolium
S. lycopersicum
Gain
Diversification
Loss
B
D03
C03
D02
C
C05
D07
D
C09
D06
C04
E
E01
E03
E04
E05
E06
E07
E14
F02
F03
F06
F08
F09
F10
F11
F12
F13
F15
F
E05
E04
E01
G
E07
F10
F08
F09
F11
F13