Contrasting patterns of presence-absence variation of NLRS within *S. chilense* are mainly shaped by past demographic history

Gustavo A. Silva-Arias[1,2,*] (ORCID: 0000-0002-7114-9916)

Edeline Gagnon[3,4,5*] (ORCID: 0000-0003-3212-9688)

Surya Hembrom[1] (no orcid)

Alexander Fastner[5] (no orcid)

Muhammad Ramzan Khan[6,7] (ORCID: 0000-0001-9167-6556)

Remco Stam[5,#] (ORCID: 0000-0002-3444-6954)

Aurélien Tellier[1,#] (ORCID: 0000-0002-8895-0785)

[1] Professorship for Population Genetics, TUM School of Life Sciences, Technical University of Munich, Liesel-Beckmann Strasse 2, 85354 Freising, Germany

[2] Instituto de Ciencias Naturales, Facultad de Ciencias, Universidad Nacional de Colombia - sede Bogotá, Ciudad Universitaria, 111321, Bogotá, Colombia

[3] Department of Integrative Biology, College of Biological Science, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada, N1G 2W1

[4] Chair of Phytopathology, TUM School of Life Sciences, Technical University of Munich, Emil-Ramman-St. 2, 85354, Freising, Germany

[5] Department of Phytopathology and Crop Protection, Institute of Phytopathology, Faculty of Agricultural and Nutritional Sciences, Christian Albrechts University, Hermann Rodewald Str 9, 24118, Kiel, Germany

[6] National Institute for Genomics and Advanced Biotechnology, National Agricultural Research Centre, Islamabad, Pakistan

[7] PARC Institute for Advanced Studies in Agriculture, NARC, Islamabad, Pakistan

*,# = Equal contribution

Author for correspondence: Aurélien Tellier (aurelien.tellier@tum.de), Edeline Gagnon (edeline.gagnon@uoguelph.ca), Gustavo A. Silva-Arias (gasilvaa@unal.edu.co)

Abstract

Understanding the evolution of pathogen resistance genes (also known as NLRs) within a species requires a comprehensive examination of factors that affect gene loss and gain. We present a new reference genome of *Solanum chilense*, that leads to an increased number and more accurate annotation of NLRs. Next, using a target-capture approach, we quantify the presence-absence variation (PAV) of NLR *loci* across 20 populations from different habitats. We build a rigorous pipeline to validate the identification of PAV of NLRs, then show that PAV is larger within populations than between populations, suggesting that maintenance of NLR diversity is linked to population dynamics. Furthermore, the amount of PAV is not correlated with the NLR presence in gene clusters in the genome, but rather with the past demographic history of the species, with loss of NLRs in diverging populations at the distribution edges and smaller population sizes. Finally, using a redundancy analysis, we find limited evidence of PAV being linked to environmental gradients. Our results contradict the classic assumptions of the important selective role of PAV for NLRs, and suggest that NLRs PAV is driven by random processes (and weak selection) in an outcrossing plant with high nucleotide diversity.

Introduction

NLR (also known as NB-LRR) genes are vital components of disease resistance in plants in all plant families[1,2]. Their diversity has been subjected to intense study because of the desire to understand and improve resistance of plants to devastating pathogens. More than 500 loci have been verified as pathogen resistance genes in various plant species[3]. From a molecular and evolutionary point of view, studying their diversity and evolution remains a challenge due to their repetitive nature and high variation in copy number (from single to hundreds of copies) across the genome in different plant species[1,4]. NLR are functionally diverse and part of complex gene signalling networks. NLR genes, called “helpers” are highly connected (hubs), while so-called “sensors” are more peripheral[5]. Sometimes NLR interactions can be detrimental to plant fitness as intra- and inter-specific crosses can lead to autoimmunity reactions known as hybrid necrosis[6,7]. The evolvability of NLR loci depends thus on their function and interactions within the network[5,8].

Interspecific comparison of NLR sequences suggest that there is rapid evolutionary change of NLR families both as presence-absence of paralogs and at the nucleotide level[2,9], explained as a birth-and-death (turn-over) process of gene duplication, natural selection and deletion of unnecessary paralogs[10], yielding the definition of pan-NLRome at the inter-specific level[11,12]. However, in order to quantify the speed of NLR turn-over defined by random versus selective processes, it is necessary to study NLR diversity at the intra-specific level.

Studies in the model system *Arabidopsis thaliana* and various crops report high rates of intra-specific presence-absence variation (PAV) as well as clustering of NLRs in the genome[1,4]. For example, the NLR locus (RPP8) is constituted of a unique triad of orthologs exhibiting presence-absence variation (PAV)[13]. Using target capture of 64 accessions, van de Weyer et al.[14] confirmed the high level of PAV for many NLRs in this species, with the number of identified NLR ranging from 167 to 251 per accession, only half of the NLR being present in most accessions. Such extensive PAV at NLR loci led to the notion of intra-specific pan-NLRome defining the core set of loci shared between individuals, which is constituted of hub genes less prone to duplication/deletion events[14]. Thus, PAV in NLR clades is assumed to result from a dynamic process of gene duplication/deletion generating different homeologs on which local selection can act (birth and death process[10]), and would underpin adaptation of populations to different pathogen pressures[1,14]. However, there rarely has been proof that selection (positive and/or balancing [16–19]) is acting upon these to promote neo-subfunctionalization[15], nor that it is the main force driving the presence-absence diversity (in contrast to neutral processes such as demography and spatial structuring).

Early studies of *A. thaliana*[17,20–23], find modest evidence for pervasive positive or balancing selection on polymorphisms at NLR loci, as confirmed by later genome scans for footprints of selection[24]. When studying pairs of tandem NLRs, van de Weyer et al.[14] show a correlation in population genetics statistics between a sensor and a helper gene. However, such an effect may also be due to linkage disequilibrium and variable rates of recombination along the genome. Lee & Chae[25] took these analyses further and showed that there are differences in PAV across NLR clades and that radiations (clade expansion) of NLRs were not common. It remains unclear whether PAV between populations of *A. thaliana* across different environments can be attributed solely to neutral effects of 1) demographic events of population bottleneck and post-glacial age expansion, and/or 2) the relatively small local effective population sizes due to high selfing rate and absence of seed banks[26,27]. Studies investigating NLR diversity and evolution in non-model wild species are rare. Sequencing of few resistance genes of the *Pto* gene complex revealed the action of positive or balancing selection in several wild tomato species[28,29]. Seong et al.[11] studied the NLR composition of 18 plants representing four self-compatible wild tomato species, and found less PAV than in *A. thaliana*: NLR numbers range from 264 to 333, yet most accessions were within 10 NLRs of the median. Nonetheless, several of their defined orthogroups showed variation within single or multiple species, whereas others only showed inter-specific variation.

We suggest here two hypotheses which summarise the arguments above. In one scenario, local adaptation at NLRs relies on PAV, the various gene copies being subsequently subjected to positive or balancing selection (e.g. varying pathogen pressure in different habitats). Here, we expect PAV diversity to be more important than nucleotide diversity at NLR genes to generate novel recognition

features, so that PAV at different NLRs would be strongly associated with given pathogens in different habitats (defined by environmental variables), irrespective of (i.e. controlling for) the demographic history of the species. Conversely, if PAV at NLRs is rather shaped by neutral processes (genetic drift, spatial structure and gene flow), we would not expect a correlation between the PAV and specific habitats (environmental variables), but rather with the demographic history of the species. There may be selection at the polymorphism level for peculiar variants, as nucleotide diversity within orthologs would be the main process generating novel recognition features. A main determinant of the NLR evolutionary trajectory, is therefore the effective population size of populations and species, as it determines the rate at which duplication/deletion and gene conversion[18] occur and the strength of selection (coevolution with pathogens) versus genetic drift[19].

S. chilense is an obligate outcrossing species which presents strong patterns of local adaptation with known past demographic events of colonisation of various habitats around the Atacama desert[8,30,31]. It exhibits high local effective population sizes, due to the seed banks[32] and mild bottleneck of colonisation, and thus high genetic diversity at the SNP level. Selection occurring at genes for abiotic adaptation has been documented at SNP[30,31,33] and PAV levels[34,35]. We previously found few NLRs under positive selection for local adaptation based on SNPs[8]. However, our study had low resolution due to the used PoolSeq method and the gene annotation on a fragmented genome assembly of *S. chilense*[36].

Here we quantify the extent of PAV at NLRs in *S. chilense* as to 1) compare with other self-compatible tomato species, and 2) assess the occurrence of selection pressure for PAV. We first complete a new genome reference of *S. chilense* with chromosome-size scaffolds, annotate genes in a reliable way, finding a set of 278 putative NLRs of which 170 are assigned a full length. We then sequence by a RenSeq method the set of NLRs across 200 plants of *S. chilense* from 20 populations to study PAV at NLR loci, focusing on conserved NB-ARC domains. Our careful annotation, resequencing, and *de novo* reconstruction of the NLR set reveals moderate PAV at these genes. We finally demonstrate weak correlation between PAV of NLRs and main climatic and environmental variables across the range of the species, contrasting with the consistency of PAV with the past demographic history of the species. Overall, despite previous evidence of adaptation of NLR loci to different habitats, PAV of NLR loci does not seem to be linked to climatic variables or their occurrence as clusters on the genome, and shows stronger links to their putative molecular functions in the genome.

Results

Dovetail scaffolding resolves the **Solanum chilense** *genome to near chromosome level*

We provide a new reference genome sequence for *S. chilense* through scaffolding a previous version[36] using high-throughput chromosome conformation capture (Hi-C) technology (Table 1). The assembly has 12 chromosome-level scaffolds ranging from 110.22 to 53.75 Mb and 12,626 additional unplaced scaffolds resulting in an L50/N50 of 6/71.61Mb. The BUSCO scores (based on *Solanales_odb10*) are 95% complete (93% single copy and 2% duplicated), 1.1% fragmented and 3.9% missing BUSCO genes, supporting the improved quality and completeness compared to other assemblies in *S. chilense* and other species belonging to the tomato clade (Table 1). We annotate 40,113 protein-coding features, of which 74% show InterPro functional annotations (Dataset 1).

The improved reference genome allows the identification of additional NLRs

We identified 278 NLR sequences across the 12 main chromosomes (Fig. 1a): 170 sequences are annotated as complete, thus containing a C terminal domain, the NB-ARC domain and LRR repeats (Fig. 1b). To assign the genes to NLR-clades, we built a phylogeny including NLRs from both assemblies (Table S1). From this point on, we focus exclusively on NB-ARC containing loci, as *de novo* assembly of the LRR domains of NLRs can be complicated without long-read sequencing and manual curation[1,37,38]. The NB-ARC domain has a high conservation across the plant kingdom[39] and can be confidently resolved[1]. This approach has been successfully used to study the PAV of NLRs in *A. thaliana* accessions[25] and *de novo* transcriptome analyses of NLR[40]. This comparison allowed us to identify additional NLRs, but in similar proportions to annotations carried out on the previous genome assembly (Table S1), with a total of 123 (72%) NB-ARCs belonging to the CNL clade, whereas 27 and 20 genes belong to the TNL and RNL clades.

Physical clustering of NLR genes and composition of clusters

We assessed the number of physical clusters for our NB-ARC-loci using a previous cluster definition (genes < 200 kb apart[41]) and our full set of 285 NLRs (278 full loci + partial loci): 104 loci (61%) are found in 57 clusters, and the top 20 clusters contained a third of NLRs (97/284, 34%). Also, 48% of complete NLR genes are on chromosome 4, 5 and 11, similarly to 45% in domesticated tomato[42] and other Solanaceae species[43]. Most clustered NB-ARCs are CNLs (78%, or 82/104 loci), 15 are TNLs, and 7 are RNLs. Cluster size ranges from 2 to 10 loci (median =2). Ten of these clusters are composed of a single NLR type (CNL or TNL), presumably originating from simple tandem duplication. We hypothesise that the remaining clusters originated from more complex genomic recombination events,

13 other clusters have mixed compositions of different NB-ARC loci (including partial sequences; Dataset S2).

NB-ARC loci show considerable PAV in S. chilense

Like previous studies, we find PAV at NLRs between individuals of the same species[11,14,25]. Across the 186 samples (after filtering low quality samples and loci [HC dataset; see methods]), we document an average of 137 NB-ARC loci (median=137, SD=3.96; max: 145, min: 123) per individual (Fig. 2a), 30 loci (20%) were present across all individuals. The remaining loci were absent in at least one sample, with 12 loci present in less than half of all the individuals sampled (Fig. 2b). Unlike in *A. thaliana*, we do not find a typical separation of core vs. cloud NLRs[14], but rather a geometric abundance distribution with very few NLRs occurring less frequently (Fig. 2b). CNLs show higher PAVs than TNLs, with CNL loci being on average present in 160 individuals (median=180, SD=39.3; max:186, min:18) and TNLs in 182, with a smaller variance (median=181.6, SD=10.5; max: 186, min:143). Differences in PAV frequencies are detected amongst different CNL clades: the CNL1/9 clade shows the greatest amount of variation, contrasting strongly for example with CNL-RPW8 (also known as RNLs; Fig. 2c). An overview of PAV for each NLR in each individual can be found in Fig. S3.

We found significant differences in the frequencies of loci with PAV based on NLR attributes for CNL vs. TNL loci, helper vs. sensor loci, and for sensor vs non-sensor loci (Table 2), but not when comparing loci that are clustered vs. singleton (Table S3).

High within-population variation, yet low variation between populations

When recalculating the PAV of NB-ARC loci across the 186 individuals from our 20 populations (Fig. 3a), we find that variation amongst populations is lower than compared to variation across all individuals. We identify an average of 153 NB-ARC loci (median=154, SD=2.68; max: 156, min: 145) per population (Fig. 3b), a total of 133 loci (85%) being found across all populations. From the remaining loci, 22 are found in most populations (12 to 19) and a single locus is found in just eight populations. Looking at PAV within these populations, we observe variation in the mean number of recovered NB-ARC loci per population of 130.9 to 140.6 loci, with small SD values of 2.1 up to 5, with the lowest average and median found in the SC population (mean=130.9, median=131, SD: 3.89) (Fig. 3c).

NLR loci are generally maintained within geographical groups

Next, we examined PAV for the six geographically defined regions included in this study, four valleys in the central region (CV1, CV2, CV3 and CV4) and two regions from the southern range (SC and SH; Fig. 3a). All 156 loci are at least present in one or more individuals in the four CV regions, with

the exception of one locus lost in CV4 and SH (Fig. 3d). Ten loci are not found in the SH and SC regions (Fig. 3d), and four loci are lost only in the SC region. One locus is lost only in the SH region, and four loci are lost shared by the SC and SH regions (Fig. 3e). All of these loci are CNLs, the majority from the CNL1/9 and CNL22 categories. The majority of these genes are not sensor or helper NLRs (8 out of 10) and were found physically clustered in the genome (6 out of 10).

Lack of correlation between NLR PAV and environmental variables

We assessed whether PAV events are associated with environmental/climatic variables, both between and within geographical regions. The PCA of environmental variables (Fig. 4a) shows that populations in CV1, CV2 and CV3 tend to cluster and are ordered along an altitudinal gradient. The most southern Central group (CV4) formed its own cluster, as did SC and SH, confirming that these locations differ in their environment. In the PCA of NB-ARC frequencies across 20 populations (Fig. 4b), we found three main clusters: one including the populations of CV1, one with populations of CV4, and a third containing nearly all populations from CV2 and CV3. Coastal and Highland populations, along with a single population from CV3 occupied a distinct position in the PCA scatterplot. These results suggest that there is strong population structure of NB-ARC frequencies, which needs to be taken into account when considering whether PAV of NB-ARC might be correlated with environmental variables.

The redundancy analysis was not significant in the ANOVA permutation test that we carried out, for both the dataset containing all the 156 NB-ARC loci (p=0.184, Fig. 4c), and the second analysis containing only the more variable 112 CNL loci (p=0.218, figure not shown). This result remained unchanged when accounting for population structure in the Central Valley groups in the partial RDA analysis (ANOVA permutation test: p=0.09, Fig. 4d). Consistent with these results, we did not observe a clear distinction between populations in the RDA and partial RDA biplots (Figs. 4c-d).

Discussion

The new reference genome of *S. chilense* and refined annotation of coding regions allowed us to identify 278 NLRs, of which 170 are full length with NB-ARC domain, in line with previous reports of NLR numbers in wild tomatoes (204 to 265 across six self-compatible wild tomato species and the self-incompatible *S. habrochaites* [11,44]). By sequencing this set of NLRs across 186 plants from 20 populations to study PAV at these *loci*, we reveal that only few NLRs clades exhibit PAV across geographic locations. We did not find any significant correlation between PAV of NB-ARC and main environmental variables across the range of the species. The pattern of PAV is rather consistent with the past demographic history of the species, namely the loss of genes occurs chiefly in Southern

Coastal and Highland populations, that are derived from two colonisation events to novel habitats, characterised by mild bottlenecks[8,30,31] of which especially SC contains isolated populations with relatively small effective population sizes[30,31,45].

We find variation in the conservation of NLR loci based on functional identity and position in the NLR network when looking at PAV across different populations of *S. chilense*. Contrary to previous hypotheses , we do not find the clustered NB-ARC *loci* in the genome to have significant differences in PAV compared to singleton NB-ARC *loci* (Table 2). This result suggests that the rates of duplication and deletion events in *S. chilense* are irrespective of their arrangement in clusters along the genome, and the absence of NLR "hotspots" of diversification. A possible explanation may lie in the difference in evolutionary potential based on recombination or mutation rates between *A. thaliana* and *S. chilense* (see below).

Our study shows a lack of correlation of PAV of NLR loci with environmental variables based on the RDA analysis (Fig 4). Evidence from early studies finding cluster size differences between different ecotypes provide limited evidence for selection on NLR loci with regard to environment variables [46,47]. To our knowledge, PAV in NLR genes according to an altitudinal gradient has only been reported once at the *CHS3* and *CSAI* NLR genes in *A. thaliana* fixed in high-altitude populations [48]. As both genes also play a role in the response to different environmental cues, such as chilling stress, such environmental correlations may be very rare events. More recent studies by Lee and Chae[25] investigating the cluster diversity and size of NLR based on the *A. thaliana* panNLRome have not found any correlation between cluster size or cluster size expansion with altitude, latitude or longitude. We conclude that PAV at NLRs in *S. chilense* (and *A. thaliana*) are likely not driven by presence-absence of different pathogens across the range of habitats of the species.

Based on our first hypothesis from the introduction, we generalise that for species with small effective population size such as *A. thaliana* or crops (due to high selfing, fragmented populations and history of bottleneck and colonisation of habitats), PAV may be an important source of variability for NLRs. We note that it is yet unknown to which extent PAV is driven by neutral or selective processes. In *A. thaliana*, the population recombination rate is five times higher than the population mutation rates[27,49,50]. Recombination would thus be more efficient at creating novel NLRs variants than DNA mutation. Subsequently, PAV may generate additional non-functional (or partial) NLR paralogs, as large numbers would hinder the integrity of the genome. Purifying selection against these extra-numerous, possibly deleterious, NLR variants is relatively weak and genetic drift can maintain or even lead to fixation of such partial gene copies. We speculate that the large PAV and hot-spot effect of clustering in *A. thaliana*, may be in large part due to neutral processes, until (weak) selective and demographic processes can be disentangled.

Our second hypothesis states that a highly heterozygous, outcrossing plant species with high effective population sizes, such as *S. chilense*, exhibits a lot more diversity of SNPs and much lower levels of PAVs. In *S. chilense* where the effective population size is on the same order of magnitude as the population recombination rate [31,32], novel DNA mutations are as frequent as recombination events to generate novel variants. Partial (even mildly deleterious) gene copies may then be strongly counter-selected. The genetic diversity on which selection can act is more based on the diversity of SNPs present in these loci, rather than via PAV. Variation between individuals and maintenance of diversity at NLR loci across the range of the species could be due to balancing selection by pathogens or to seedbanking. These are hypothesised to be particularly important due to El Niño climatic fluctuations which leads to long-term cyclical rain and population dynamics, and to intermittent gene flow and dispersal of seeds within the different valleys[32]. However, from previous studies[8] and our results, we conclude that most PAV and polymorphism variation may likely be due to neutral demographic processes in *S. chilense.* This stems from the weak selection pressure for the plant to adapt to pathogens in such very arid habitats as indicated by the large and variable resistance response to infection by various pathogens[51–53]. In *S. chilense*, for example, (putative) natural pathogens have been identified, but more work would be needed to quantify the natural pathogen pressure and pathogen diversity across these different populations in experimental and natural set-ups[51,53,54].

Our hypotheses may help explain the observation of a low number of NLRs and little variation in PAV in other species[55], as the evolutionary dynamics are likely to vary according to different evolutionary and ecological contexts, and the complexity and redundancy of the NLR/pathogen interaction network within each species[56]. For example, in *Amborella*, weak PAV of genes related to biotic stress was observed, in sharp contrast with abiotic stress genes[57], presumably through lack of pathogen pressure and limited distribution in remote oceanic islands, not unlike *S. chilense*. Expanding these types of studies to more species is important for questions of how fragmentation and habitat loss can potentially affect plant health and resistance in the wild.

Material and methods

Genome scaffolding, annotation and visualisation

High-molecular weight DNA was sent to Dovetail Genomics (Santa Cruz, CA, USA) to construct Chicago libraries. The Chicago libraries were sequenced on an Illumina HiSeqX with 150 bp paired-end reads. Using the draft assembly as input [36], the HiRise scaffolding pipeline, used to build super scaffolds, was done at Dovetail Genomics using their proprietary protocol. Short-read sequences generated from Chicago and HiRise libraries are available at ENA (PRJNA508893). Using Pilon v1.23 [58] we ran one round of polishing onto the 12 chromosome-size contigs. Using BWA v0.7.17 we

first mapped the paired short-read sequences of the same individual (LA3111_t13) from Stam et al.[36]. The sorted alignments were then used as input in Pilon in full-correction mode. The quality of the genome assembly was evaluated in Quast v5.2.0 [59] which calculated basic statistics such as total length, GC content, N50/90, L50/90 and the number and size of the contigs. We also evaluated the completeness of the assembly using BUSCO v5.4.5 [60] based on the database Solanales v10 [61] (creation date: 2020-08-05, number of genomes: 11, number of BUSCOs: 5950). The assignment of chromosome-size scaffolds to each respective chromosome was done by aligning the 12 biggest scaffolds of the new reference to the genome sequence of *Solanum pennellii*[62], using the web application D-GENIES[63] and minimap v2[64].

Before annotating the assembly, we soft-masked the repetitive sequences using the benchmarking pipeline EDTA[65]. The repeat analysis shows that 60.06% of the *S. chilense* genome comprise transposable elements, the majority being long terminal repeat (LTR) retroelements (36.1%), mostly Gypsy-LTRs (26.09%). We performed structural gene annotation using BRAKER v2.1.5[66] which relies on GeneMark-ET[67] and AUGUSTUS v3.2.3[68]. We first performed *ab initio* gene prediction. Subsequently, we performed 10 runs of evidence-based gene prediction in BRAKER using transcriptomic paired-end libraries from leaf tissues of six individuals of *S. chilense* LA3111 (PRJNA474106[36]). For that, we mapped the RNA reads to the reference for genome annotation using BBmap[69], after filtering the adapters and low-quality reads with the Captus pipeline[70]. Additionally, we used Liftoff v1.3.0[71] with options -a 0.9 -s 0.8 -copies, to transfer the existing gene model annotations from *S. chilense*[36] and *S. pennellii* (GCF_001406875.1) into the new assembly. The three sets of predicted/transferred genes were then merged to generate a nonredundant reference gene set using EvidenceModeler v1.1.1[72]. For functional annotations, Blastp v2.12.0+[73] (with a threshold E-value of <1e−5) was used to align the protein sequences to the UniRef90 database[74], and only the best-matched targets were retained. InterProScan v5.52-86.0[75] was used to annotate motifs and domains by searching against ProDom[76], CDD, Gene3D, PRINTS, PFAM, SMART, PANTHER, SUPERFAMILY, TIGRFAM, and PROSITE databases. Gene Ontology[77] annotations were obtained based on the InterPro entries.

Annotation of NLR loci in the reference genome

The NLR Annotator pipeline [78] was used to identify NLR loci in the new reference genome, using default commands. Focusing solely on the protein alignments of the nucleotide-binding site domain-encoded genes (NB-ARC), we recovered a total of 177 loci, of which seven were eliminated because they represented duplicated or overlapping NLR loci (Dataset S2).

The output from NLR-annotator [78] was first used to reconstruct a phylogenetic tree based on the protein alignments of the NB-ARC domains of the identified 170 NLRs; the outgroup sequence used

in NLR-annotator (NP_001021202.1, Cell Death Protein 4 CDP4), was also included in the phylogenetic analyses. Protein sequences were aligned using default settings in MAFFT v.7, through the online portal service [79], with manual curation subsequently carried out in Jalview [80]. A maximum likelihood tree of the NB-ARC domain alignment was done with IQ-Tree2 [81], using ModelFinder to find the best substitution model for the alignment (in all cases the model JTT+F+R6 is selected according to the Bayesian Information Criterion (BIC)). In addition, a total of 1000 UltraFastBootstrap were carried out to calculate branch support. Visualisation of the phylogeny was carried out in R and Inkscape, using the packages ggtree [82], ggtreeExtra [83] and ggplot2 [84].

Previous NLR sequences from the Stam et al. [36] genome assembly identified a total of 134 NB-ARC domains that can be grouped into 15 different clades. To understand whether the new reference genome was able to recover these same sequences, the 170 NB-ARC loci from the new reference were aligned with the NB-ARC loci from the previous assembly using MAFFT v.7. A maximum likelihood phylogeny was calculated and visualised, using the same methods as described above. The NB-ARC loci from the new reference were assigned to different NLR clades based on whether they formed solid sister pairs with annotated NB-ARC loci from the previous assembly (bootstrap support above 95%). For all other sequences, the assignment to NLR clades was based on a visual inspection of the phylogenetic relationships (Fig 1a and S1).

In order to visualise and compare the assemblies of the previous and new references, as well as the location of the NB-ARC domains along the genome, we used the package "circulize" in R [85]. We also visualise the density of mRNA genes annotated along each of the scaffolds. For the new reference genome, we recovered 12 major scaffolds (corresponding to 12 chromosomes), in which the 170 detected NB-ARC loci were all found (Fig. 1b). In comparison, the previous assembly had 134 NB-ARC loci across a subset of the genome composed of 113 scaffolds.

Evaluation of NB-ARC presence-absence variation (PAV)

Sampling – We sampled 200 plants from 38 accessions and grew each plant in individual pots under standard glasshouse conditions at the GHL Dürnast plant research facilities of the TUM School of Life Sciences. The sampling includes 18 accessions (hereafter populations) from the central region (10 samples each) arranged in four central valleys (CV) representing replicates of elevation gradients (Fig. 3a). The remaining 20 plants were sampled from scattered populations (one plant per locality) from the two divergent lineages (south-highland SH, and south-coast SC[30]). Plants were grown from individual seeds obtained from the Tomato Genetics Resource Center (TGRC, University of California, Davis, USA; tgrc.ucdavis.edu). Genomic DNA was extracted using the DNAeasy extraction kit from Qiagen following the instructions of the supplier.

Target enrichment and sequencing – We designed 1959 probes targeting coding sequences of 170 NB-ARC regions annotated in this study. Probe synthesis, library preparation, and sequencing were carried out at RAPiD Genomics LLC (Gainesville, CA) using SureSelectXT (Agilent Technologies, CA) enrichment system followed by Illumina Hiseq 2000 sequencing to generate paired-end 150-bp reads.

Assembly of NB-ARC loci, PAV detection and validation – We used the Captus pipeline [70] that automatizes the processes of raw read quality check and trimming, *de novo* assembly, and extraction of the NB-ARC target loci from the full set assembled contigs for each sample. Reference loci for the extraction process corresponded to the 170 NB-ARC domains previously identified with NLR-Annotator. Briefly, the Captus pipeline uses Scipio to identify gene structures given a protein sequence. Captus also identifies intron-exon borders and splice sites, and can cope with loci assembled over multiple contigs (fragmented assemblies). We performed the extraction with a relaxed set of parameters (minimum Scipio score = 0.13, minimum identity percentage to reference proteins 65%, and minimum coverage percentage of reference protein to consider a hit by a contig = 20%).

The identification of PAV was further refined with the identification of cut-off values of identity and coverage statistics. To establish the cut-off values, the distribution of identity and coverage was evaluated for all extracted loci and visualised in 2D density plots with ggplot in R (Fig. S4). The distribution of the statistics for the complete data set showed that most of the assembled contigs have high coverage (> 95%). Within them, we found two clearly defined peaks of identity values. In order to avoid erroneous captures of paralogs we delimited a strict identity cutoff value of 94% which points to the group of assembled sequences of high confidence.

To compare the patterns of presence/absence variation (PAV) of NB-ARC loci identified with the Captus pipeline, we applied the above-described method to 1) the original ~130x short-read data used to assemble the first version of the genome of *S. chilense* [36] (LA3111_REFERENCE in Fig. S3), and 2) 30 whole genome sequencing samples [31]; marked with WGS suffixes in Fig. S3). The loci were annotated for the individual LA3111_REFERENCE. We found nine loci missing in the reference sample which are likely due to technical error during sequencing or assembly. These nine loci were thus removed from further analysis. Preliminary data exploration led to the exclusion of 14 samples from the target capture data set, which consistently showed a lower number of total NLR loci and lower read depth of coverage after short-read alignment (mapping) to the reference genome. Furthermore, a total of 14 NB-ARC loci were also excluded, as they showed discordant PAV patterns between the target capture and WGS data or very low or no read coverage after short-read alignment (Fig. S6), likely because target capture failed to capture them due to the lack of ad hoc probes for these loci. Then we remove them to avoid over-interpretation of the data. The resulting high-

confidence (HC) dataset consisted of a total of 156 NB-ARC loci sequenced for 186 sequenced individuals across 20 populations (Dataset S3).

Finally, we scored PAV using two different methods: (1) we assigned a binary character for each locus, namely a locus that either did not show any presence-absence variation across the entire dataset, or showed absence for at least one individual); (2) we assigned a quantitative variable per locus indicating the locus frequency (presence) across a set of sequences. We calculated both quantities for several sets of sequences: across all 186 individuals, within each of the 20 populations sampled, as well as within the six geographic regions. We also tested our approach on the RENseq long-read data from [11] using our set of 156 NB-ARC loci annotated in *S. chilense* revealing consistent patterns of PAV as our interspecific data set, but showing more absences in more variable NLR clusters according to the divergence of the included species (Fig. S5).

Attributes of NB-ARC loci

For each of the loci with complete NB-ARC domain, we scored them for attributes which possibly influence their PAV across populations:

(1) The location of NLRs as physical clusters. We used the definition by Andolfo et al. [42], identifying a cluster if genes are at a maximum distance of 200,000 bp from each other. We calculated the number of clusters based on two different approaches: 1) based on the position of the NB-ARC complete domains (170 loci), and 2) based on all recovered NLRs, including loci with incomplete NB-ARC domains (284 loci).

(2) The putative function of the NLR as being either helper, sensor, or non-sensor genes (*sensu* Wu et al [5]). We defined helper genes as loci that were sister to the core four core NRC sequences previously identified in *A. thaliana* and tomato species [5,8,36] leading to an expanded clade of 10 loci (Fig 1b, Dataset S2; hereafter referred to as "Helper+").

(3) The annotation of NLRs as TNL or CNL, based on the annotation of NLR loci using the phylogenetic methods described above.

Visualisation and chi-square tests

To visualise PAV across individuals and the 20 populations sampled, barplots, boxplots and UpSet plots were produced, using the following R packages: ggplot2, dplyr, stringr, tidyverse, ggtext, and UpSetR [84,86–90]. In addition, we carried out $\chi 2$-tests (chisq.test, as implemented in the "stats" package in R) to determine whether there were any significant differences in the frequency of various attributes.

Environmental correlation analysis with PAV

Environmental data and occurrence records - We tested for correlation between PAV and frequency of NLR loci and particular environments by selecting five environmental variables reflecting strong constraints for growth and survival of *S. chilense*. A proxy for cold stress was represented by the minimum temperatures of the coldest month (bio6), and temperature seasonality is summarised by the temperature annual range (bio7). Moisture availability due to high seasonality of rainfall and fog (in Lomas) in these ecosystems was summarised by precipitation amount of the wettest month (bio13); intra-annual variability of cloud frequency (MODCF_intraannualSD), and mean cloud frequency of September (MODCF_monthlymean_09). September is the month with the strongest coastal fogs in the Lomas ecosystems[91]. Temperature and precipitation variables were derived from CHELSA climate data, whereas cloud data was taken from the Global 1-km Cloud Cover dataset[92] downloaded from the EarthEnv portal [93]. All maps were at a 30 arcsec spatial resolution (c. 1 km).

Occurrence records of the 38 localities present in our samples were downloaded from the Tomato Genetic Resource Center and double-checked. Altitude values were re-extracted based on these (from the EarthEnv portal[93]) . Correlations among variables were verified and a standard PCA of the environmental data was carried out, using custom R scripts. Finally, because our analyses focused on 20 populations, we took the average value of all populations of the Southern Highland Group and the Southern Coastal Group. This was justified given the environmental distinctiveness of these localities based on our preliminary PCA results.

Environmental analyses - First, we ran an initial PCA analysis of the frequency of the 149 NLR loci across the 20 populations, and then tested for association between frequency of specific alleles to environmental conditions using a redundancy analysis (*rda* function from the VEGAN package in R[94]). Prior to running the analysis, the NLR frequency data was transformed using the Hellinger distance, and the frequency of these NLRs is modelled across the 20 localities as a function of the environmental variables. The significance of RDA-constrained axes was assessed using the *anova.cca* function, and the significant axes were then used to identify candidate loci ($P < 0.001$). Finally, we ran a partial RDA, with the four central valleys (CV1 to CV4), the SH and SC groups as covariable factors using the entire 15,649 NB-ARC dataset or a sub-set of genes with the 113 NB-ARC loci from the CNL clade.

Acknowledgements

We thank Edgardo Ortiz for advice in the processes with Captus. EG acknowledges support by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 899987. AT acknowledges funding from DFG

(Deutscheforschungsgemeinschaft) grant no.: 317616126 (TE809/7-1). AT and MRK were supported by a DAAD Germany - Pakistan exchange grant; GAS-A was funded by the Technical University of Munich. RS acknowledges DFG grant no.: 170483403 (SFB924). We thank the Tomato Genetics Resource Center (TGRC) of the University of California, Davis for generously providing us with the seeds of the accession included in this study.

Competing interests

The authors do not declare any competing interests.

Author contributions

GS, AT, EG and RS planned and conceptualized the research; EG, GS, AF, SH carried out the analyses in the paper; AT, MRK and RS obtained funding; EG, GS, AT, RS all contributed significantly to the writing of the manuscript, with all co-authors reviewing and approving the final version.

Data availability

The data that support the findings of this study are openly available in the European Nucleotide Archive at https://www.ebi.ac.uk/ena with the following accession numbers. SRA ERR11268525 (Chicago sequencing), SRA ERR11268526 (Hi-C sequencing) and BioProject PRJEB61272 (Target sequencing data). Reference genome sequence and annotation files are available on the Solgenomics website (https://solgenomics.net/ftp/genomes/Solanum_chilense/Gustavo/). The code for implementing the analyses used in this paper can be found on our GitLab repository: https://gitlab.lrz.de/population_genetics/Schilense_newref

References

1. Barragan, A. C. & Weigel, D. Plant NLR diversity: the known unknowns of pan-NLRomes. *Plant Cell* **33**, 814–831 (2021).
2. Gao, Y. *et al.* Out of Water: The Origin and Early Diversification of Plant R-Genes. *Plant Physiol.* **177**, 82–89 (2018).
3. Kourelis, J., Sakai, T., Adachi, H. & Kamoun, S. RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. *PLOS Biol.* **19**, e3001124 (2021).
4. Wersch, S. van & Li, X. Stronger When Together: Clustering of Plant NLR Disease resistance Genes. *Trends Plant Sci.* **24**, 688–699 (2019).
5. Wu, C.-H. *et al.* NLR network mediates immunity to diverse plant pathogens. *Proc. Natl. Acad. Sci.* **114**, 8113–8118 (2017).
6. Yamamoto, E. *et al.* Gain of deleterious function causes an autoimmune response and Bateson–Dobzhansky–Muller incompatibility in rice. *Mol. Genet. Genomics* **283**, 305–315 (2010).
7. Bomblies, K. *et al.* Autoimmune Response as a Mechanism for a Dobzhansky-Muller-Type Incompatibility Syndrome in Plants. *PLOS Biol.* **5**, e236 (2007).

8. Stam, R., Silva-Arias, G. A. & Tellier, A. Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. *New Phytol.* **224**, 367–379 (2019).
9. Liu, Y. *et al.* An angiosperm NLR Atlas reveals that NLR gene reduction is associated with ecological specialization and signal transduction component deletion. *Mol. Plant* **14**, 2015–2031 (2021).
10. Michelmore, R. W. & Meyers, B. C. Clusters of Resistance Genes in Plants Evolve by Divergent Selection and a Birth-and-Death Process. *Genome Res.* **8**, 1113–1130 (1998).
11. Seong, K., Seo, E., Witek, K., Li, M. & Staskawicz, B. Evolution of NLR resistance genes with noncanonical N-terminal domains in wild tomato species. *New Phytol.* **227**, 1530–1543 (2020).
12. Li, X. *et al.* PlantNLRatlas: a comprehensive dataset of full- and partial-length NLR resistance genes across 100 chromosome-level plant genomes. *Front. Plant Sci.* **14**, (2023).
13. MacQueen, A. *et al.* Population Genetics of the Highly Polymorphic RPP8 Gene Family. *Genes* **10**, 691 (2019).
14. Van de Weyer, A.-L. *et al.* A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell* **178**, 1260-1272.e14 (2019).
15. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
16. Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature* **400**, 667–671 (1999).
17. Bergelson, J., Kreitman, M., Stahl, E. A. & Tian, D. Evolutionary Dynamics of Plant R-Genes. *Science* **292**, 2281–2285 (2001).
18. Hörger, A. C. *et al.* Balancing Selection at the Tomato RCR3 Guardee Gene Family Maintains Variation in Strength of Pathogen Defense. *PLOS Genet.* **8**, e1002813 (2012).
19. Tellier, A., Moreno-Gámez, S. & Stephan, W. Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. *Evolution* **68**, 2211–2224 (2014).
20. Tian, D., Araki, H., Stahl, E., Bergelson, J. & Kreitman, M. Signature of balancing selection in Arabidopsis. *Proc. Natl. Acad. Sci.* **99**, 11525–11530 (2002).
21. Allen, R. L. *et al.* Host-Parasite Coevolutionary Conflict Between Arabidopsis and Downy Mildew. *Science* **306**, 1957–1960 (2004).
22. Bakker, E. G., Toomajian, C., Kreitman, M. & Bergelson, J. A Genome-Wide Survey of R Gene Polymorphisms in Arabidopsis. *Plant Cell* **18**, 1803–1818 (2006).
23. Huard-Chauveau, C. *et al.* An Atypical Kinase under Balancing Selection Confers Broad-Spectrum Disease Resistance in Arabidopsis. *PLOS Genet.* **9**, e1003766 (2013).
24. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18**, 86 (2017).
25. Lee, R. R. Q. & Chae, E. Variation Patterns of NLR Clusters in Arabidopsis thaliana Genomes. *Plant Commun.* **1**, 100089 (2020).
26. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
27. Sellinger, T. P. P., Awad, D. A., Moest, M. & Tellier, A. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genet.* **16**, e1008698 (2020).
28. Rose, L. E., Michelmore, R. W. & Langley, C. H. Natural Variation in the Pto Disease Resistance Gene Within Species of Wild Tomato (Lycopersicon). II. Population Genetics of Pto. *Genetics* **175**, 1307–1319 (2007).
29. Rose, L. E., Grzeskowiak, L., Hörger, A. C., Groth, M. & Stephan, W. Targets of selection in a disease resistance network in wild tomatoes. *Mol. Plant Pathol.* **12**, 921–927 (2011).
30. Böndel, K. B. *et al.* North–south colonization associated with local adaptation of the wild tomato species *Solanum chilense*. *Mol. Biol. Evol.* **32**, 2932–2943 (2015).
31. Wei, K., Silva-Arias, G. A. & Tellier, A. Selective sweeps linked to the colonization of novel habitats and climatic changes in a wild tomato species. *New Phytol.* **237**, 1908–1921 (2023).
32. Tellier, A., Laurent, S. J. Y., Lainer, H., Pavlidis, P. & Stephan, W. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 17052–17057 (2011).

33. Mboup, M., Fischer, I., Lainer, H. & Stephan, W. Trans-species polymorphism and allele-specific expression in the CBF gene family of wild tomatoes. *Mol. Biol. Evol.* **29**, 3641–3652 (2012).
34. Fischer, I., Camus-Kulandaivelu, L., Allal, F. & Stephan, W. Adaptation to drought in two wild tomato species: the evolution of the Asr gene family. *New Phytol.* **190**, 1032–1044 (2011).
35. Wei, K., Stam, R., Tellier, A. & Silva-Arias, G. A. Copy number variations shape genomic structural diversity underpinning ecological adaptation in the wild tomato Solanum chilense. 2023.07.21.549819 Preprint at https://doi.org/10.1101/2023.07.21.549819 (2023).
36. Stam, R. *et al.* 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 Genet.* **9**, 3933–3941 (2019).
37. Witek, K. *et al.* A complex resistance locus in Solanum americanum recognizes a conserved Phytophthora effector. *Nat. Plants* **7**, 198–208 (2021).
38. Jupe, F. *et al.* 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 (2013).
39. Shao, Z.-Q. *et al.* Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. *Plant Physiol.* **170**, 2095–2109 (2016).
40. Safdari, P., Höckerstedt, L., Brosche, M., Salojärvi, J. & Laine, A.-L. Genotype-Specific Expression and NLR Repertoire Contribute to Phenotypic Resistance Diversity in Plantago lanceolata. *Front. Plant Sci.* **12**, 675760 (2021).
41. Andolfo, G., Dohm, J. C. & Himmelbauer, H. Prediction of NB-LRR resistance genes based on full-length sequence homology. *Plant J.* **110**, 1592–1602 (2022).
42. Andolfo, G., D'Agostino, N., Frusciante, L. & Ercolano, M. R. The Tomato Interspecific NB-LRR Gene Arsenal and Its Impact on Breeding Strategies. *Genes* **12**, 184 (2021).
43. Di Donato, A., Andolfo, G., Ferrarini, A., Delledonne, M. & Ercolano, M. R. Investigation of orthologous pathogen recognition gene-rich regions in solanaceous species. *Genome* **60**, 850–859 (2017).
44. Seong, K. *et al.* A draft genome assembly for the heterozygous wild tomato Solanum habrochaites highlights haplotypic structural variations of intracellular immune receptors. 2022.01.21.477156 Preprint at https://doi.org/10.1101/2022.01.21.477156 (2022).
45. Raduski, A. R. & Igić, B. Biosystematic studies on the status of Solanum chilense. *Am. J. Bot.* **108**, 520–537 (2021).
46. Botella, M. A. *et al.* Three Genes of the Arabidopsis RPP1 Complex Resistance Locus Recognize Distinct Peronospora parasitica Avirulence Determinants. *Plant Cell* **10**, 1847–1860 (1998).
47. Noël, L. *et al.* Pronounced Intraspecific Haplotype Divergence at the RPP5 Complex Disease Resistance Locus of Arabidopsis. *Plant Cell* **11**, 2099–2111 (1999).
48. Günther, T., Lampei, C., Barilar, I. & Schmid, K. J. Genomic and phenotypic differentiation of Arabidopsis thaliana along altitudinal gradients in the North Italian Alps. *Mol. Ecol.* **25**, 3574–3592 (2016).
49. Strütt, S., Sellinger, T., Glémin, S., Tellier, A. & Laurent, S. Joint inference of evolutionary transitions to self-fertilization and demographic history using whole-genome sequences. *eLife* **12**, e82384 (2023).
50. Nordborg, M. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).
51. Stam, R., Scheikl, D. & Tellier, A. The wild tomato species Solanum chilense shows variation in pathogen resistance between geographically distinct populations. *PeerJ* **5**, e2910 (2017).
52. Kahlon, P. S. *et al.* Laminarin-triggered defence responses are geographically dependent for natural populations of Solanum chilense. *J. Exp. Bot.* erad087 (2023) doi:10.1093/jxb/erad087.
53. Kahlon, P. S. *et al.* Population studies of the wild tomato species Solanum chilense reveal geographically structured major gene-mediated pathogen resistance. *Proc. R. Soc. B Biol. Sci.* **287**, 20202723 (2020).
54. Schmey, T. *et al.* Small-spored Alternaria spp. (section Alternaria) are common pathogens on wild tomato species. *Environ. Microbiol.* **n/a**, (2023).
55. Baggs, E. L. *et al.* Convergent Loss of an EDS1/PAD4 Signaling Pathway in Several Plant

Lineages Reveals Coevolved Components of Plant Immunity and Drought Response[OPEN]. *Plant Cell* **32**, 2158–2177 (2020).
56. Adachi, H. & Kamoun, S. NLR receptor networks in plants. *Essays Biochem.* **66**, 541–549 (2022).
57. Hu, H. *et al.* Amborella gene presence/absence variation is associated with abiotic stress responses that may contribute to environmental adaptation. *New Phytol.* **233**, 1548–1555 (2022).
58. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
59. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
60. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 227–245 (Springer, 2019). doi:10.1007/978-1-4939-9173-0_14.
61. Kriventseva, E. V. *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
62. Bolger, A. *et al.* The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **46**, 1034–1038 (2014).
63. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
64. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
65. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
66. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3**, lqaa108 (2021).
67. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119 (2014).
68. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
69. Bushnell, B. BBTools. A suite of bioinformatics tools used for DNA and RNA sequence data analysis. *DOE Joint Genome Institute* https://sourceforge.net/projects/bbmap/ (2014).
70. Raza, M., Ortiz, E. M., Schwung, L., Shigita, G. & Schaefer, H. Resolving the phylogeny of *Thladiantha* (Cucurbitaceae) with three different targeted-capture pipelines. Preprint at https://doi.org/10.21203/rs.3.rs-2760642/v1 (2023).
71. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
72. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
73. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
74. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
75. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
76. Bru, C. *et al.* The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* **33**, D212–D215 (2005).
77. Harris, M. A., Lomax, J., Ireland, A. & Clark, J. I. The Gene Ontology project. in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (John Wiley & Sons, Ltd, 2005). doi:10.1002/047001153X.g408202.
78. Steuernagel, B. *et al.* The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire1 [OPEN]. *Plant Physiol.* **183**, 468–482 (2020).
79. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
80. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version

2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
81. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
82. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
83. Xu, S. *et al.* ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Mol. Biol. Evol.* **38**, 4039–4042 (2021).
84. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2016).
85. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
86. Wickham, H., François, R., Henry, L. & Müller, Kirikk. dplyr: A Grammar of Data Manipulation. (2022).
87. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. (2022).
88. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
89. Wilke, C. O. & Wiernik, B. M. ggtext: Improved Text Rendering Support for 'ggplot2'. (2022).
90. Gehlenborg, N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. (2019).
91. Ruhm, J. *et al.* Two sides of the same desert: floristic connectivity and isolation along the hyperarid coast and precordillera in peru and Chile. *Front. Ecol. Evol.* **10**, (2022).
92. Wilson, A. M. & Jetz, W. Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLOS Biol.* **14**, e1002415 (2016).
93. Amatulli, G. *et al.* A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Sci. Data* **5**, 180040 (2018).
94. Oksanen, J. *et al.* vegan: Community ecology package. R package version 2.3-0. (2015).
95. Li, N. *et al.* Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet.* 1–9 (2023) doi:10.1038/s41588-023-01340-y.
96. Hosmani, P. S. *et al.* An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. 767764 Preprint at https://doi.org/10.1101/767764 (2019).

Figures

Figure 1. (a) Genome assembly of *S. chilense.* (1) Density of mRNA genes; (2) NLRs found in the *S. chilense genome* are shown, loci in black were found in the previous version of the genome [36] , whereas loci in red are new to this version; (3) annotation labels indicating the clades to which the Nb-Arc loci correspond (coloured as in Fig. 1B). **(b) Phylogeny of the NLR domains**. Maximum Likelihood phylogeny of the NB-ARC domains recovered from the Dovetail genome assembly in *S. chilense*, using NLR-Annotator. Clades are coloured using the same colour scheme as used in Stam et al. (2019a). Nodes with UltraFastBootstrap support of 95% and above are indicated with black circles, whereas bootstrap support above between 80% and 94% are indicated with white circles. Nodes without circles have branch support values below 80%. For each of the 15 NLR clades, we indicated the UltraFastBootstrap value next to the crown node of each respective clade. NB-ARC tip names in bold represent sequences that formed a strongly supported sister pair (95% Bootstrap support and above) with NB-ARC sequences identified in Stam et al. [36].

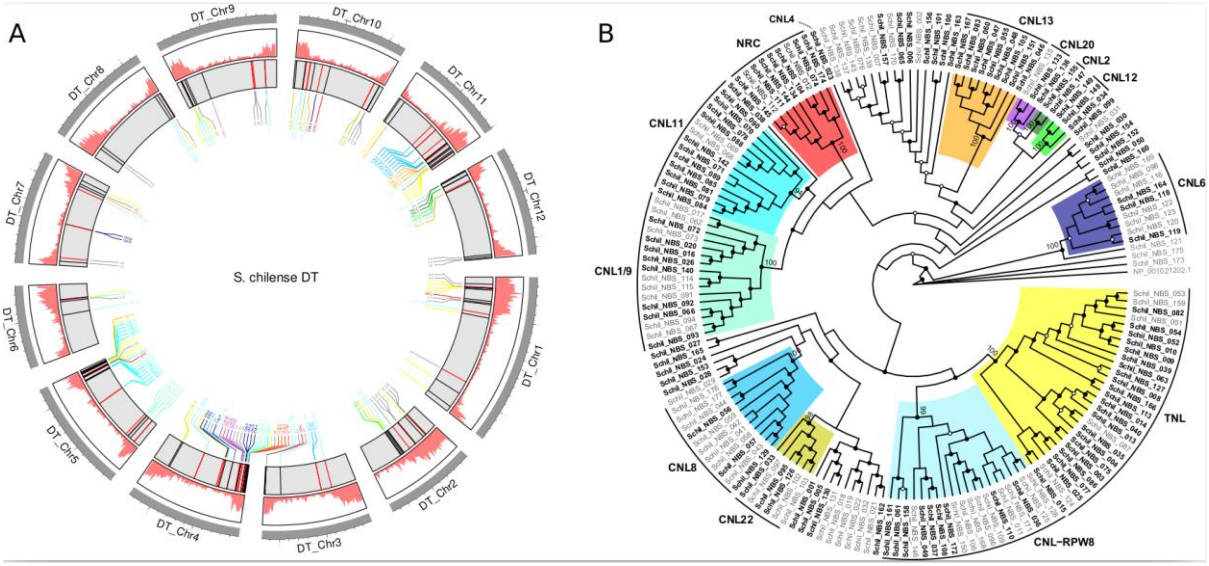


Figure 2. PAV in NB-ARC loci within *S. chilense.* (a) Total number of detected NB-ARC containing loci (y axis) for each of the sequenced individuals, each bar on the x axis represents a single individual; (b) Histogram showing the frequency distribution (y axis) of NB-ARC loci through our data set, shown as the number of individuals in which a locus occurs (x axis); (c) Barplot of NLR genes per population, with each barplot showing the number of loci belonging to different NB-ARC clusters.

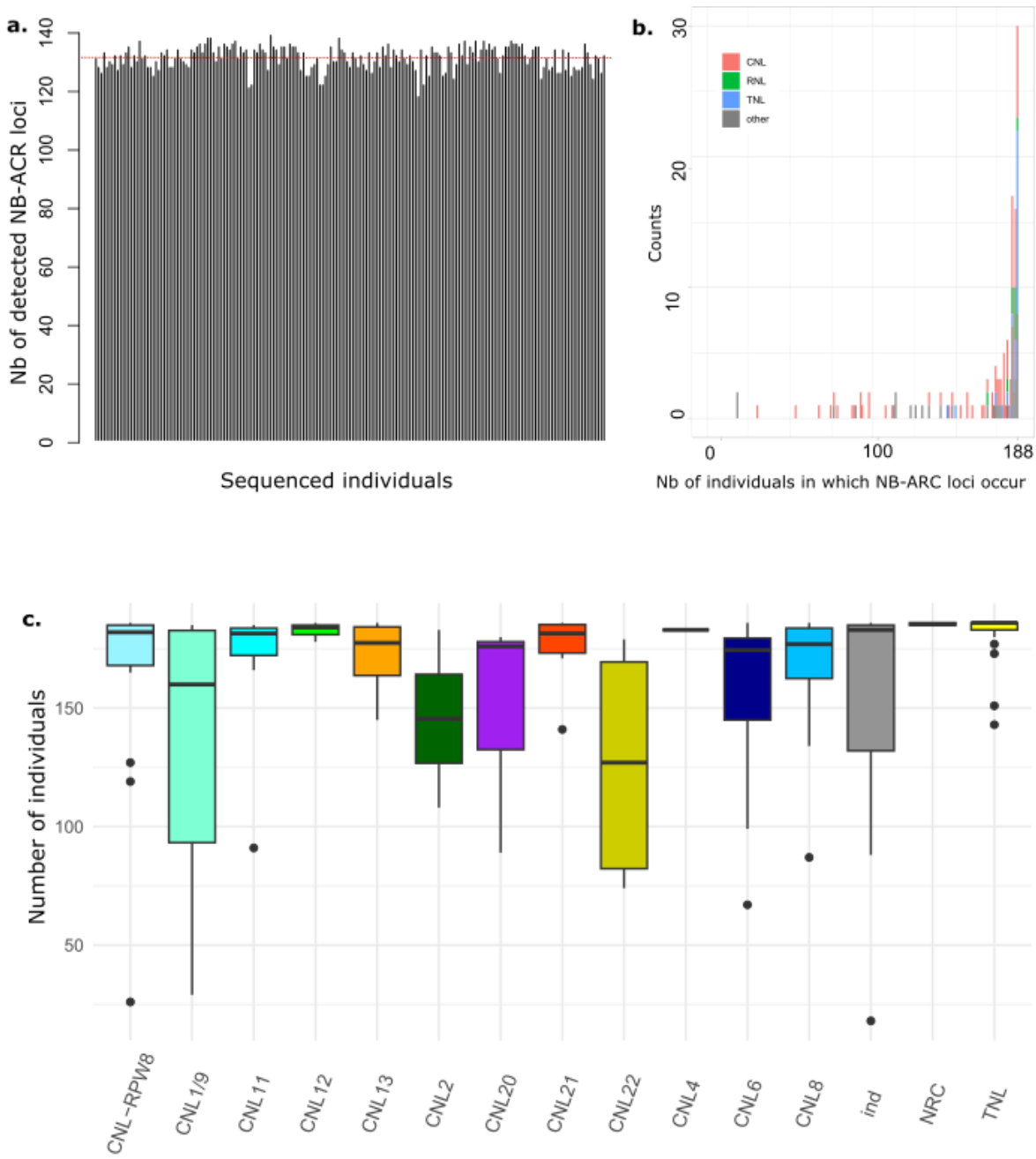


Figure 3. **PAV of NB-ARC loci on a geographical scale.** (a) Map of the localities of the TGRC samples that were sampled in this study, from the Central Valley, Southern Highland and Southern Coastal regions; colored dots represent known, active collection, the one with black dots are included in our study (b) Barplot of NLR genes per population, with each barplot showing the number of loci belonging to different NB-ARC clusters; (c) Boxplots of the total number of NB-ARC loci per population; colours correspond to CV, SH and SC regions as shown in Fig 3A. (d) Upset plot of NB-ARC loci present across the six metapopulations of *S. chilense*, showing how many loci are shared between the six metapopulations, from the Central Valleys 1 to 4 (CV1 to CV4), the Southern Highland region (SHG), and Southern Coastal Group (SCG).

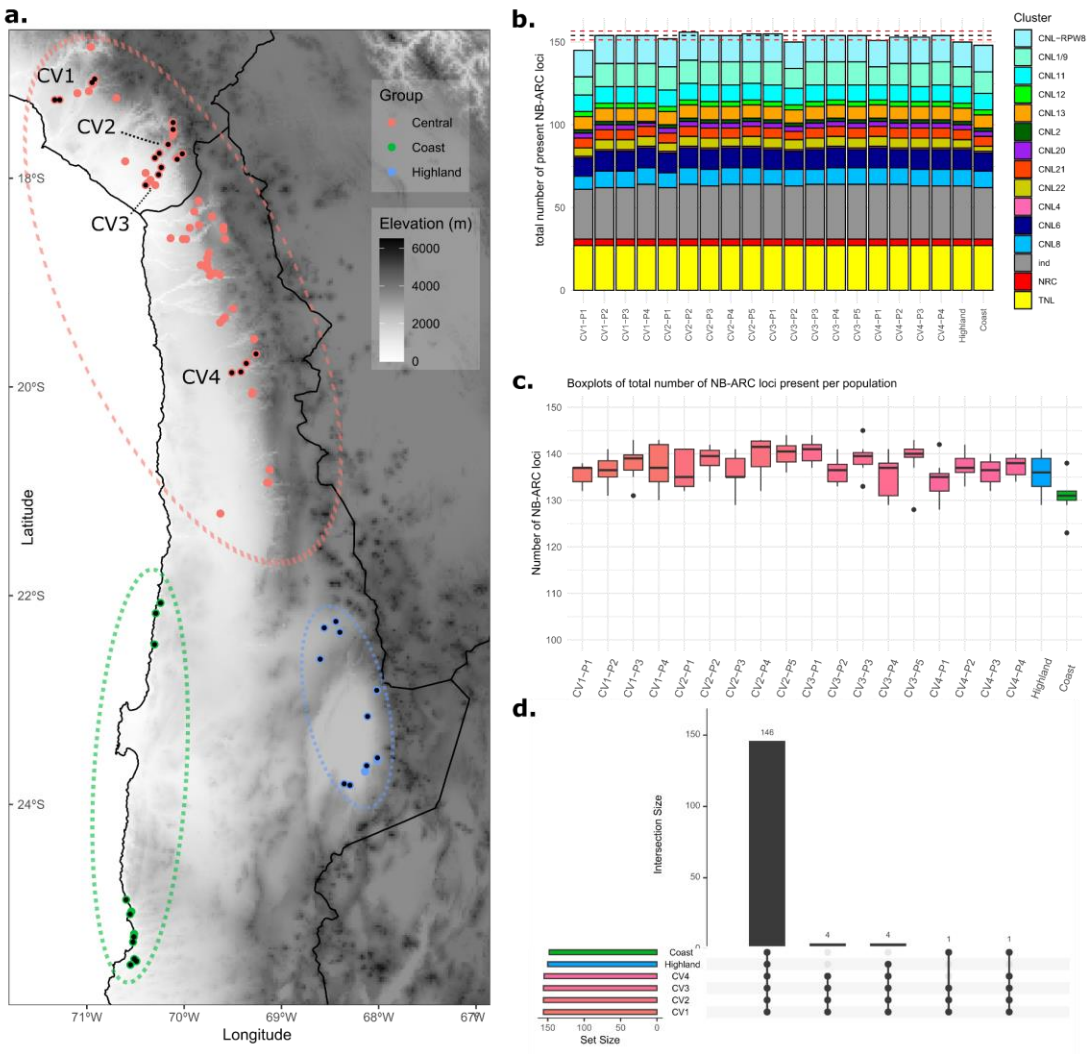
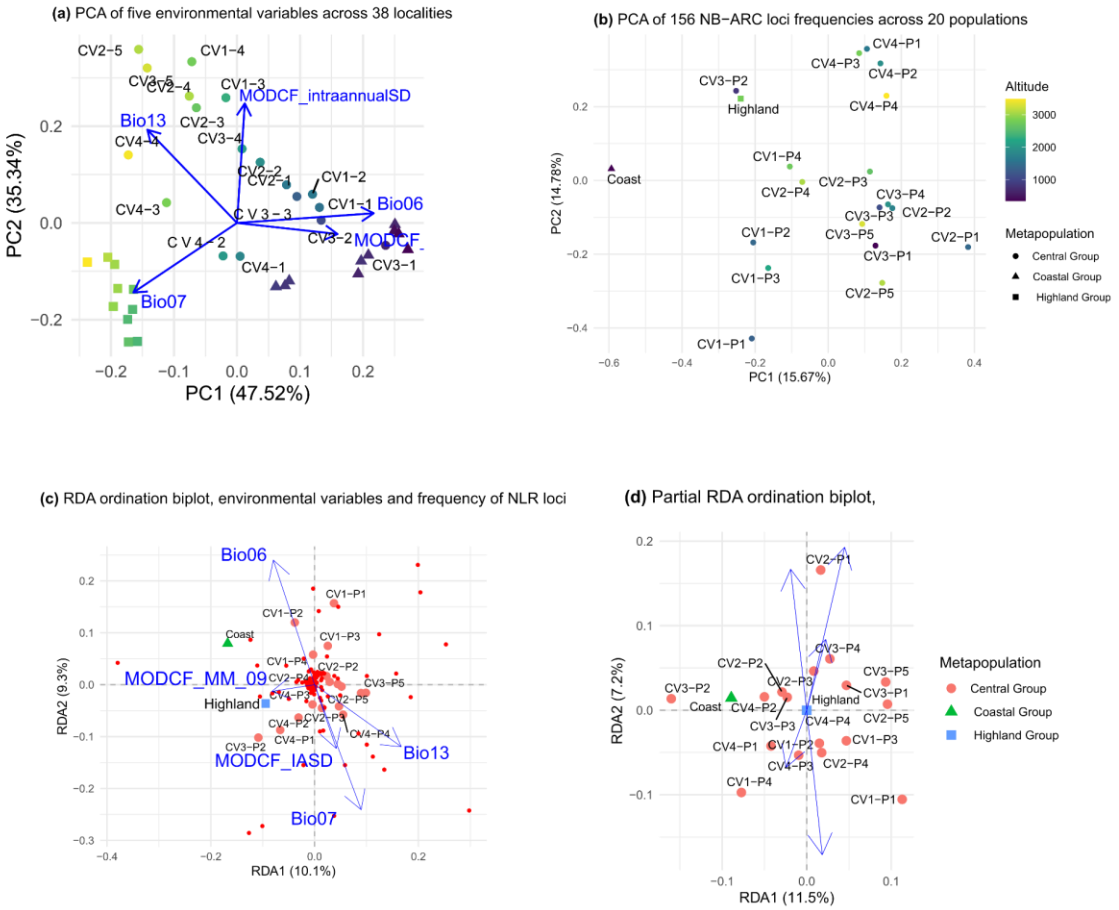


Figure 4. (a) PCA based on the five environmental variables, across 38 localities of *S. chilense* present in this study. The three main groups are represented by different shapes (circle: Coastal Group; triangle: Southern Coastal Group; square: Southern Highland group), and the dots are coloured according to elevation; (b) PCA of the frequency of the 156 NB-ARC loci across 20 populations; (c) RDA analysis of all 156 NB-ARC loci, scaling 1; (d) Partial RDA analysis of all 156 NB-ARC loci.



Tables

Table 1. Assembly statistics of the new version of *Solanum chilense* reference and comparison with the previous assembly[36], *S. chilense* (LA1969; Li et al., 2023a), *S. pennellii* [62] and *S. lycopersicum*[96].

Assembly statistics	*S. chilense* (this study)	*S. chilense* v0.2	*S. chilense* (LA1969)	*S. pennellii*	*S. lycopersicum*
# contigs	12,638	81,307	2055	12	12
Total length (Mb)	920	914	917	926	782
Largest contig (Mb)	110.2	1.1	92.4	109.3	98.9
GC (%)	33.8	33.7	34.6	34.5	34.3
N50 (Mb)	71.61	0.07	67.67	77.99	65.27
N90 (Mb)	53.75	0.004	0.13	60.73	53.47
L50	6	3144	7	6	6
L90	12	22,768	92	11	11
N's per 100 Kb	21,529.82	21,576.39	27.33	7,671.92	5.72
Complete BUSCO* [single copy, duplicated] (%)	95.0 [93.0, 2.0]	92.1 [89.9, 2.2]	96.2 [86.7, 9.5]	97.9 [95.8, 2.1]	97.9 [96.1, 1.8]
Partial BUSCO* (%)	1.1	2.2	0.5	0.3	0.4
Missing BUSCO* (%)	3.9	5.7	3.3	1.8	1.7

*The BUSCO statistics are based on the lineage dataset *solanales_odb10* (Creation date: 2020-08-05, number of genomes: 11, number of BUSCOs: 5950).

Table 2. Chi-square test results, comparing various attributes of NB-ARC loci. P-values below the threshold of 0.5 are indicated with an asterisk (*).

Comparison	**X-square test**	**P-value**
Cnl/Tnl vs. PAV	X-squared = 24.981, df = 1	p-value = 5.789e-07*
Cnl/Tnl vs. Clustered/Singleton	X-squared = 0.410, df = 1	p-value = 0.5222
PAV vs. Clustered/Singleton	X-squared = 0.320 df=1	p-value = 0.5715
Helper+/Sensor vs. PAV	X-squared = 7.511, df = 1	p-value = 0.006131*
Helper+/Sensor vs. Clustered/Singleton	X-squared = 2.205, df = 1	p-value = 0.1375
Sensor/Non-Sensor vs. PAV	X-squared = 7.045, df = 1	p-value = 0.007951*
Sensor/Non-Sensor vs. Clustered/Singleton	X-squared = 0.057, df = 1	p-value = 0.8109

Supporting Information

Figure S1. Dotplot representation of sequence alignment of the new assembly of *Solanum chilense* and *S. pennellii*.

Figure S2. Maximum Likelihood phylogeny of the NB-ARC domains recovered from the Dovetail and the short-read sequence assembly from *S. chilense*, using NLR-Annotator. Clades are coloured according to the same scheme as used in [36]. Nodes with UltraFastBootstrap support of 95% and above are indicated with white circles, whereas branch support of 80% to 94% are indicated with black circles. Numbers next to the left of the nodes indicate ultrafast bootstrap support, although these are not represented for nodes with support of 95% and above. Tip names in red represent NB-ARC sequences identified in [36].

Figure S3. NBS loci presence-absence variation matrix in 186 samples of *Solanum chilense* sequenced using the targeted sequencing approach, short-read data from the *S. chilense* sample LA3111_t13 used to assemble the reference genome in [36], and 30 whole-genome sequencing samples from Wei et al. [31].

Figure S4. 2D density plot showing the distribution of Identity percentage and Coverage percentage of all assembled contigs to the 170 NBS reference protein sequences annotated in the new reference. Dashed lines indicate the threshold values chosen to filter the extraction.

Figure S5. Distribution of depth of coverage for each NBS locus annotated in the new reference genome of *Solanum chilense* measured after read alignment of the 200 samples from the targeted sequencing (left panels) and 30 samples from whole-genome sequencing (right panels).

Figure S6. NBS loci presence-absence variation matrix obtained with long-read RENseq data from Seong et al. 2020 [11] including *S. lycopersicum*, five wild tomatoes, and two additional (non-*Solanum*) Solanaceae species.

Table S1. Comparison of the 170 NB-ARC-containing NLRs recovered in the new versus the previous version *S. chilense* genomes, for each NLR clade. Numbers in parenthesis indicate the number of new sequences that were identified in the new genome reference sequence.

Dataset S1. Predicted genes that were annotated and considered functional based on searches against the InterPro and UniRef90 protein signature databases

Dataset S2. Protein-coding features from the new reference genome

Dataset S3. Presence Absence table for 187 individuals and 149 loci (indicate excluded pops here).

Dataset S4. Alignments + Phylogenies