A small subset of NLR genes drives local adaptation to pathogens in wild tomato

Remco Stam[1,2*], Gustavo A. Silva-Arias[2], Tetyana Nosenko[3], Daniela Scheikl[2], Anja C. Hörger[4], Wolfgang Stephan[5], Georg Haberer[3], Aurelien Tellier[2]

1 Phytopathology, Technical University Munich, Germany
2 Population Genetics, Technical University Munich, Germany
3 Plant Genome and Systems Biology, Helmholtz Zentrum München, Germany
4 Ecology and Evolution, University of Salzburg, Austria
5 Evolutionary Biology, LMU München, Germany

*corresponding author: remco.stam@tum.de

Abstract

In plants, defence-associated genes including the NLR gene family are under constant evolutionary pressure to adapt to pathogens. It is still unknown how many NLRs contribute to adaptation, and if the involved loci vary within a species across habitats. We use a three-pronged approach to reveal and quantify selection signatures at over 90 NLR genes over 14 populations of *Solanum chilense* a wild tomato species endemic to Peru and Chile found in different habitats. First, we generated a de novo genome of *S. chilense*. Second, by whole genome resequencing of three geographically distant individuals we infer the species past demographic history of habitat colonisation. Finally, using targeted resequencing we show that a small subset of NLRs, 7%, show signs of positive or balancing selection. We demonstrate that 13 NLRs change direction of selection during the colonisation of new habitats and form a mosaic pattern of adaptation to pathogens. We estimate that the turn over time of selection (Birth and Death rate) on NLRs is 18,000 years. Finally, our work identifies new NLRs under strong selective pressure between habitats, thus providing novel opportunities for R-gene identification.

Introduction

In nature and agriculture, plants face various abiotic and biotic challenges. Antagonistic host – pathogen interactions, are arguably amongst the most interesting phenomena because they can generate endless coevolution between plants and their pathogens. The Red Queen hypothesis indeed predicts that the genomes of both interacting partners evolve to match each others changes (Van Valen, 1973; Dawkins & Krebs, 1979): pathogens evolve infectivity to overcome plant defences, while plant evolve pathogen recognition and resistance to avoid infection. Such coevolutionary dynamics can be observed as changes in allele frequencies over time at the key loci determining the outcome of interaction between hosts and pathogens. Two extreme types of dynamics have beenproposed, which differ in their signatures at the phenotypic (Gandon et al., 2008) and polymorphism levels (Woolhouse et al., 2002): 1) the so-called arms race model (Bergelson et al., 2001) is characterized by recurrent selective sweeps in both partners, and 2) the trench warfare model (Stahl et al., 1999) shows long-lasting balancing selection. These simple predictions form the basis for current scans for selection in plant and parasite genomes or at candidate loci (e.g. Rose, Michelmore & Langley, 2007; Badouin et al., 2017), yet they suffer from two major drawbacks. First, theory suggests that a continuum of possible complex dynamics exists between these two extremes so these simple signatures of coevolution may not be typical (Tellier, Moreno-Gámez & Stephan, 2014). Second, these predictions are based on coevolution occurring within one population.

Most plant species exhibit a spatial distribution across homogeneous or heterogeneous habitats, which influences coevolutionary dynamics (theory in e.g. Gandon et al., 2008; Brown & Tellier, 2011; Moreno-Gamez, Stephan & Tellier, 2013; Parratt, Numminen & Laine, 2016). Diverse habitats generate differential pathogen pressure across space due to variation in 1) disease presence or absence and prevalence, 2) disease transmission between hosts, and 3) co-infection or competition between pathogen species. As a result, spatial heterogeneity is observed for infectivity in pathogen and resistance in hosts (*e.g.* Thrall, Burdon & Bock, 2001; Caicedo & Schaal, 2004). Species expansion and colonisation to new habitats could in addition cause the host to encounter new pathogens and subsequently promote coevolutionary dynamics at novel loci compared to the original habitat. Despite the wealth of studies at the phenotypic and ecological levels (Thrall & Burdon, 2003; Thrall et al., 2012; Tack & Laine, 2014), we know little about the genetic bases of host-pathogen coevolution in spatially heterogeneous populations. We want to know how many host genes are involved in coevolution across populations, and the time scale of coevolution in newly colonized habitats.

In order to address these questions in plants, the host species should be found in different habitats. Knowledge of past history of colonization, and availability of adequate polymorphism data for candidate loci for resistance are essential. We answer the above questions by studying the evolution of resistance related genes in a wild tomato species, *Solanum chilense*. This outcrossing species is ideal to study molecular evolution in heterogeneous habitats because it exhibits very high effective population size, reflected in very high nucleotide diversity (heterozygosity) and high recombination rates (Arunyawat, Stephan & Städler, 2007). These features are attributed to the spatial structuring of populations linked by gene flow and the presence of seed banks (Arunyawat, Stephan & Städler, 2007; Tellier et al., 2011b; Böndel et al., 2015). *S. chilense* occurs in southern Peru and northern Chile. The species has split from its nearest sister species *S. peruvianum* 1 to 2 mya (Städler, Arunyawat & Stephan, 2008). It has dispersed southward by colonising a range of diverse habitats around the Atacama desert (Böndel et al., 2015). Local adaptation to abiotic but also biotic stresses in *S. chilense* or its sister species is indicated by 1) signatures of positive selection in genes involved in cold and drought stress response (Xia et al., 2010; Fischer et al., 2013; Nosenko et al., 2016), 2) balancing selection in several genes of the *Pto* resistance pathway providing resistance to *Pseudmonas sp.* (Rose et al., 2011), and 3) variable resistant phenotypes against filamentous pathogens across populations (Stam, Scheikl & Tellier, 2017). *S. chilense* is also an extablished source of fungal and viral resistance genes (R genes) used in breeding programmes (Tabaeizadeh et al., 1999; Verlaan et al., 2013).

Canonical R genes are members of a large gene family called NLR (for Nod-like receptor or Nucleotide binding site, leucine rich repeat containing receptor) that occurs in both plants and animals (Jones, Vance & Dangl, 2016). NLRs have a modular structure and typically contain an N-terminal domain that either

resembles a Toll-Interleukin receptor (TIR) or a Coiled Coil (CC) domain, which give rise to the terms TNL and CNL. In addition, typical NLRs contain a Nucleotide Binding Site (NBS) domain and a number of C-terminal Leucine Rich Repeats (LRR). Most identified Resistance genes are “complete” NLRs (containing N-terminal TIR or CC-domain, NBS and LRRs), yet examples of functional partial NLRs (lacking one of those domains) also exist (reviewed in Baggs, Dagdas & Krasileva, 2017). A large number of NLRs also contain functional integrated domains (ID) that might act as the attractant or binding region for the pathogen effectors (Cesari et al., 2014; Wu et al., 2015b; Maqbool et al., 2015; Sarris et al., 2016) and the phylogenetic relationship of NLR appears to determine their functionality (Wu et al., 2017).

In *Arabidopsis thaliana*, some NLRs appear to evolve by positive or balancing selection (Mondragón-Palomino et al., 2002; Bakker et al., 2006). Overall, NLRs seem to show higher rates of evolution than other defence-related gene families (Mondragon et al., in revision). In other wild species, few candidate NLR genes have been studied. In the common bean (*Phaseolus vulgaris*) the NLR locus PRLJI1 show markedly different patterns of spatial differentiation, as measured by the fixation index F_{ST}, compared to AFLP markers (De Meaux et al., 2003). In the wild emmer wheat a marker-based analysis shows that NLRs exhibit higher differentiation (F_{ST} = 0.58) than other markers (F_{ST}= 0.38) (Sela et al., 2009). Within a single genus and especially among crops, the number of NLR can differ dramatically between species (Meyers et al., 1998; Pan, Wendel & Fluhr, 2000; Guo et al., 2011) suggesting that the NLR family experiences a rapid birth-and-death process (Michelmore & Meyers, 1998).

Yet, an exhaustive genome-wide NLR analysis in a non-model organism and in an ecological perspective is missing. This is partly because 1) computational identification of NLRs has been hampered by the complexity of the gene family, and 2) individual domains can also be found in other gene families. A HMM-based identification method was successfully applied to potato (Jupe et al., 2012) Due to the complexity of the NLR family and the highly repetitive nature of the LRR regions, studies rely on the availability of a high quality reference genome sequence and annotation. In cases where such reference genome is not available, a targeted resequencing method can be applied to enrich NLR-like regions and use depth of coverage on the genome as a guide to annotate NLRs. This method has been applied in potato (Jupe et al., 2013) and cultivated tomato (Andolfo et al., 2014). When this method was applied to (bulked) resistant and susceptible pools (Van Weymers et al., 2016; Witek et al., 2016), or to mutagenised samples (Steuernagel et al., 2016) it has led to the rapid identification of novel specific R-gene alleles. As a proof of principle, we previously used a pooled enrichment strategy on one population of the wild tomato species *S. pennellii* and showed that the recovery of Single Nucleotide Polymorphisms (SNPs) was accurate for NLRs. Despite the very little genetic diversity in this population, some R-genes show distinct patterns of past selection (Stam, Scheikl & Tellier, 2016).

In this study, we derive a three-pronged approach to quantify the adaptation of *S. chilense* at NLR genes. First we have generated the first reference genome sequence for this species. Second, with three full genomes for representatives of different habitat groups, we infer the colonisation and demographic history. This allows us to assess the divergence time between populations and thus the time necessary for adaptation to new biotic and abiotic environments. Third, we use pooled sequencing of 236 NLRs in 14 populations covering the whole species range to analyse variation in selective pressures due to difference in habitats and compafre these to our demographic expectations.

Results

First *S. chilense* reference genome

We used two complementary approaches to assess the phylogenetic relationship of three selected *S. chilense* accessions (LA3111, LA4330, LA2932), four *S. peruvianum* and three alleged *S. chilense* samples for which whole genome short read data have been published previously (The 100 Tomato Genome Sequencing Consortium et al., 2014; Lin et al., 2014): 1) ML phylogenetic analysis of chloroplast sequences using PhyML (Figure 1). 2) ML analysis of a nuclear genome dataset consisting of concatenated sequences for six nuclear control loci (CT), extracted from above-mentioned datasets supplemented with 53 sanger sequences from previous publications (Städler, Arunyawat & Stephan, 2008; Tellier et al., 2011b) (Figure 1 – Figure Supplement 1). It was previously suggested that *S. peruvianum* forms a polyphyletic clade, while *S. chilense* represents a monophyletic sister group (Pease et al., 2016). The chloroplast phylogeny shows that our three samples group together, whereas other alleged *S. chilense* samples are clustering in between known *S. peruvianum* samples. When looking at the CT loci, it is even more evident that previously sequenced specimens from Lin et al. and The 100 tomato genome sequencing consortium fall within the *S. peruvianum* clade. Only our three *S. chilense* samples group with other previously robustly assigned *S. chilense* samples. We conclude that our three accessions of *S. chilense* represent the first fully sequenced genomes of this species.

S. chilense genome properties

To obtain a comprehensive genome-wide list of NLR genes, we generated a total of ~134 Gb of short read genomic sequence from the accession LA3111 to build a draft genome for *S. chilense*. The final assembly consisted of 81,307 scaffolds which capture ~717 Mb of the genomic sequence and span ~914 Mb in total. The draft genome assembly has a N50 of 70.6 kb with more than 2,000 and 14,500 scaffolds exceeding a size of 100 and 10kb, respectively. We predicted 25,885 high confidence (hc) gene loci that show high homology and coverage to known proteins of tomato species. Next to their support by homology, approximately 71% (18290) of the hc genes are additionally supported by RNAseq data derived from leaf tissue samples. Complementary to the set of hc models, we also report the presence of 41,481 low confidence (lc) loci to maximize gene content information for researchers and to assist them in the identification of candidate genes in a region of interest. The lc gene models likely comprise in part functional genes for which insufficient sequence and/or evidence information impede the prediction of a complete model, but also truncated genes, pseudogenes or transposon related genes. Functionality for some of these models (6,569) is suggested by transcriptome evidence support from the leaf RNAseq data. Based on the presence or absence of 1,440 single-copy genes conserved among plants (BUSCO groups (Simão et al., 2015)) we estimated the gene content completeness at 91.8 and 93.3% for the genome assembly and gene annotation, respectively Fragments were found for 3.1 and 3.3% additional BUSCO orthologs, respectively.
According to the results of reciprocal similarity searches with e-value threshold 10^{-30}, 39683 and 40,769 *S. chilense* gene models find homologous genes in *S. lycopersicum* (24,651 genes) and *S. pennellii (*25,695 genes) genome sequence assemblies, respectively. Of them, 14,013 and 12,984 genes belong to syntenic gene blocks conserved between the *S. chilense* draft genome sequence and *S. lycopersicum* and *S. pennellii* genomes, respectively. To compare, 977 syntenic gene blocks detected between *S. lycopersicum* and *S. pennellii* genomes using the same parameters consist of 8,107 and 17,933 gene models respectively (Supplementary Data 1). Thus the *S. chilense* genome is largely organized as the cultivated tomato and S. pennellii genomes, though gene copy numbers vary slightly and small rearrangements did occur.

S. chilense demographic history

We inferred the demographic history of three populations from Central (LA3111, our reference), Southern (LA4330) and Coastal (LA2932) regions, based on whole-genome sequences of one diploid individual plant per population, by applying the Multiple Sequential Markovian Coalescent (MSMC) method (Schiffels & Durbin, 2014). We find a consistent population expansion event for the three populations between 50 to 500 thousand years ago. We estimated a stronger expansion for the Central group (10 times higher than current N_e) than for the other two populations (Figure 02). Cross-coalescence analysis confirmed that the Central group is likely the area of the species origin. The species dispersal occurred via two separate colonization events. An older split between the Central and Coastal populations 0.2 to 1 million years ago, and a more recent divergence between Central and Southern populations (30 to 150 thousand years ago). We also

demonstrate by simulations that considering the very high amount of heterozygosity and genetic diversity in *S. chilense*, the demographic inference is robust to the size of our scaffolds in the genome assembly (Figure 02 – Figure supplement 2-4).

NLR Identification

In addition to the automated genome annotation as described above, we used NLRParser (Steuernagel et al., 2015) to identify NLR genes within our new *S. chilense* genome. We identified 236 putative NLRs. We found 139 (42 complete) CNLs and only 35 (19 complete) TNLs. The majority of the 62 unclassified and partial NLR fragments represent loose LRR repeats, but we also identified NBS only fragments. (Supplementary data 02).NLR genes are highly repetitive which complicates their assembly and annotation especially if genes are located at the break-point of scaffolds. To avoid over-interpretation of our results, we assessed the quality of all identified NLRs and removed all genes that fall within 1,5 kb of scaffold edges or lacked RNASeq evidence. The remaining set consists of 201 high quality (HQ) NLR gene models. The syntenic blocks identified between the *S. chilense* and the *S. lycopersicum* and *S. pennellii* genomes include 69 and 50 HQ NLR genes, respectively, thus confirming the quality of NLR annotations. The NLR-containing syntenic blocks are distributed across all 12 chromosomes in both genomes, most of them being represented by a single NLR gene (Supplementary Data 02). To additionally check for relative completeness of the NLR set in *S. chilense*, we reconstructed a phylogeny for the gene family based on the NBS protein sequences, which defines functional clades of NLRs (see Jupe et al., 2013; Wu et al., 2017). The phylogenetic reconstruction of the gene family in *S. chilense* confirmed that all major NLR-clades are present in the genome (Figure 03) and revealed only some small, but interesting differences with other tomato spp. The CNL-4 and CNL-15 clusters contained four of five members in *S. lycopersicum,* yet in *S. chilense* each had only one member. In addition, two small new clades, CNL20 and CNL21, were identified. This confirms the high rate of molecular evolution of NLRs in tomato.

Enrichment sequencing provides high coverage of both NLR and control genes

To investigate the forces driving NLR gene evolution within and between our *S. chilense* populations, we sequenced 14 samples. Each sample consisted of pooled and enriched DNA from 10 diploid plants from a given population (Accessions are indicated on Figure 05) using Illumina MiSeq 250 bp paired-end sequencing was then used. For each population between 1 and 2 million read pairs passed trimming and quality controls (Supplementary Data 03); on average 1.15 million read pairs remained per sample with an insert size as expected of 417 +- 91. For all samples, the coverage exceeds 100x for 80% of the targeted NLRs. To evaluate the short read data quality, we also enriched and sequenced a set of 14 control (CT) genes which were extensively used as reference genes in previous studies on *S. chilense* (Arunyawat, Stephan & Städler, 2007; Böndel et al., 2015). These loci showed a coverage of more than 100x in most samples (Figure 04).

Pooled sequencing provides reliable summary statistics.

We called SNPs per gene per population and per geographic population group (previously mentioned Central, Coastal,and Southern groups and an additional Northern group), using a combination of SNP callers and filters (see Methods section and Stam, Scheikl & Tellier, 2016). We separated the analysis for NLR and CT genes. These CT genes, though not evolving entirely neutrally (Tellier et al., 2011a), have been used in previous studies to infer the spatial structure and demography of wild tomato species, and can thus be used to assess the robustness of our SNP calls and computations in our dataset (Böndel et al., 2015).

We calculated the statistics θ_W, π, π_N summarizing nucleotide diversity for the NLR genes in each population (Supplementary Data 04). No significant correlation was found between the number of mapped reads or bases and the number of SNPs per population (corr = 0.46 & p = 0.1). The correlations were even lower for π per population, (for read pairs: corr = 0.30, p = 0.30, for bases: 0.35 & 0.2) (Figure 04 – Figure Supplement 01). Thus our data is not biased for coverage differences between the samples.

Values of π, π_s and θ_W have previously been calculated for the for CT genes in several *S. chilense* populations, using high depth NGS and Sanger sequencing. To confirm the robustness of our calculations (Supplementary Data 05), we computed the correlations at the 14 CT loci between our data and previously published values (Böndel et al., 2015) for π, π_S, θ_W and the pairwise Fixation index (F_{ST}) between all populations. These comparisons reveal a very strong and significant correlation for π (corr = 0.95, p = $3.67*10^{-6}$), π_S (corr = 0.95, p = $5.80*10^{-6}$), θ (corr = 0.72, p = $9.0*10^{-3}$) as shown in (Figure 04 – Figure

Supplement 02) and for F_{ST} (corr = 0.94, p = $2.2*10^{-16}$; (Figure 04 – Figure Supplement 03)). Thus our approach and coverage are adequate to calculate population genetics summary statistics from pooled *S. chilense* data.

We calculated the folded site frequency spectra (SFS) for each geographic group (pooling of several populations) over all CT genes (Figure 04 – Figure Supplement 04). In the absolute SFS, populations from the Central region show much higher number of singletons compared to the rest of the geographical population groups, partly due to the larger number of populations in this group (the pooling effect, (Städler, Arunyawat & Stephan, 2008). The relative SFS shows some excess of intermediate frequency SNPs in the Coast and South groups of pooled populations and to a lesser extend in the North, which is in line with the demographic model of derived populations through bottleneck during colonization (Böndel et al., 2015).

For final verification of our methods, we cloned and Sanger sequenced fragments of two selected CT genes from all plants for one selected population (CT166 for LA3111 and CT192 for LA2755) as described in (Stam, Scheikl & Tellier, 2016). Figure 04 – Figure Supplement 05 shows that none of the cloned fragments of CT166 have any SNPs, indicating absence of false negatives. CT192 is a gene with high number of SNPs (18) at the 3' end in LA2755 of which we recover 100%, including two singletons. Besides the two CT genes, we also cloned parts of two predicted NLR genes, SOLCI000037500 from LA2931 and SOLCI002113000 from LA3111. HIn 10 Sanger sequenced alleles, we find 21 and 25 SNPs respectively, though 1 or 2 false negatives occurred. This set of controls confirms that our pooled sequencing and SNP calling correctly identified the majority of SNPs, and does not produce any noticeable bias in the estimates of nucleotide diversity (π) and fixation index between populations (F_{ST}).

NRC genes are the most conserved and few outlier NLRs are under selection

In our HQ NLR set, we find between 2,748 and 7,653 SNPs within each population. On average 63,8 (± 0.48) % of them are causing nonsynonymous mutations, contrary to only 34 (± 3.26) % of nonsynonymous SNPs in the CT genes. PCA analyses of NLR SNPs show that most variation can be explained using the first two principal components, showing a good fit to the geographical locations of populations (Figure 05A). This suggests that NLR genes' neutral evolution is driven by the species demography and history of colonization. Yet, to get a better insight into possible selection pressures, we calculated π and π_N/π_S per gene per population and per geographic group. For each population, median π is significantly higher for NLR than for CT genes (Figure 05, Figure 05 – Figure Supplement 1). The significantly reduced π values observed for the CT and NLR genes in Southern and Coastal populations is indicative of the previously reported bottleneck event that occurred during the species expansion southwards. A significant difference in π is observed for the NLR genes between the Southern and Coastal group, thus CT and NLR diversity are not fully correlated. There are no significant differences in median or average π_N/π_S in NLRs between any populations or groups as a whole.

The π_N/π_S values for most genes remain below 1 indicative of purifying selection, however, NLR genes have significantly higher π_N/π_S than CT genes (Figure 05B, Figure 05 – Figure Supplement 1) suggesting possible higher selective pressures or relaxed constrains on the NLRs. In fact, mean and median π values are very similar between CT loci and NLRs, but the variance is larger in NLRs. This explains why the number of high π_N/π_S outliers in NLRs is significantly larger (based on a 5% threshold) than one would expect based on a normal distribution. The highest π_N/π_S values are found in CNL and partial NLRs.

When split in previously assigned functional clades, we see that the CNL6 and NRC cluster shows overall very low π_N/π_S (Figure 5 – Figure Supplement 2A), and contrasting patterns appear between the geographical groups (Figure 5 – Figure Supplement 2B). CNL11 shows the highest π_N/π_S values in the Coastal populations, whereas these values are lowest for CNL2 in the coast. Moreover, genes with $\pi_N/\pi_S > 1$ appear to differ between groups and populations, thus functional NLR clades are under different evolutionary pressures in the different geographical regions.

At the individual gene level in the species as a whole, six NLR genes show very large (median >0.02) π, and 15 genes show high π_N/π_S (median > 1) (Figure 5 – Figure Supplement 3). Two genes, SOLCI002284200 and SOLCI002848100, are both highly diverse and show high π_N/π_S, but as we show below, these are possibly non-functional. The Coastal geographic group is characterized by low diversity but also by a high variance in π_N/π_S ratio (including outliers with very high π_N/π_S; Figure 5 – Figure Supplement 3). We found no correlation between π_N/π_S and π per gene (cor 0.09, p-value = 0.2). Variance of π_N/π_S increases slightly at low π (Figure 5 – Figure Supplement 4).

Seven percent of NLR genes drive adaptation between populations

We calculated the fixation index (F_{ST}) for each gene between each pair of populations (Figure 6 – Source data). Median F_{ST} values per NLR gene range between 0.12 and 0.7, with 17 genes having median F_{ST} over 0.5. (Figure 6). As expected from the spatial structure and demographic history, F_{ST} is generally lowest within geographic groups and highest between the Coastal and the Southern populations (Figure 6 – Figure supplement 1,2). We now derive a conservative test for local adaptation. We focus first on three populations, LA3111, LA4330 and LA2932 for which we have inferred above the demographic history. Based on these demographic parameters (time of expansion, time of split, and population sizes) we simulated a set of sequences of same length and sample size as our NLR dataset and obtain a 'null' neutral distribution of pairwise F_{ST} between these three populations (10.000 coalescent simulations) (Figure 7). Each pairwise comparison of populations reveals six or seven outlier NLR (13 in total), corresponding to about 7% of the testable NLR genes per comparison. These genes are conservatively assumed to be under selection as their pattern of differentiation is outside the expected neutral distribution. Two genes (SOLCI002229700 and SOLCI005342400) are found in all three comparisons (Table 1). To polarize the spatial change in selection, we computed $\Delta\pi_N/\pi_S$, the difference of π_N/π_S values between two populations. The values of $\Delta\pi_N/\pi_S$ are positive for all six genes in the LA2932-LA3111 and the LA2932-LA4330 comparison, indicating increasing π_N/π_S towards the coast (Figure 7, Figure Supplement 1). In the third pairwise comparison (LA4330 versus LA3111), $\Delta\pi_N/\pi_S$ is positive for some and negative for other outlier NLRs. LA3111 represents the ancestral region of the species, thus negative $\Delta\pi_N/\pi_S$ suggests relaxed selective pressure in the derived LA4330 population. To differentiate between types of selective pressures (in line with Bakker et al., 2006, we compare values for π, π_N and π_N/π_S within the populations (Figure 7 – Figure Supplement 2-4, Supplementary data 4). For example, SOLCI002229700 shows a decrease of π, but an increase of π_N and π_N/π_S towards LA2932 (Coast), suggestive of positive selection. At the other genes (SOLCI005342400, SOLCI001607700, SOLCI001766400, SOLCI002632400, SOLCI001524200) positive or balancing selection seems to occur in LA2932. Conversely, the decrease in π_N and π_N/π_S for SOLCI005342400 in LA4330, suggests stronger purifying selection in that population compared to LA3111.

Previously undescribed NLRs show local selection

We extend these analyses and conservatively select genes in the upper 7th percentile of the F_{ST} distribution for all pairwise comparison between all 14 populations. We conclude first that positive or balancing selection events have occurred chiefly in the Coastal populations, and to a lesser extend in the Southern ones (Figure 8 – Figure supplement 1). Second, we find nine NLRs that occur in more than 20 pairwise comparisons, including the genes described approach above. Out of 49 genes with outlier F_{ST}, 23 genes occur in less than five comparisons, thus showing different selective pressure only between a few populations or as the result of neutral stochastic processes (Figure 8 – Figure Supplement 2). This means that only few NLRs are indeed responsible for adaptation. Three genes from the TNL and NRC classes (SOLCI002229700 (TNL), SOLCI002632400 (NRC) and SOLCI001524200 (TNL)) show patterns of selection in the Coastal population as revealed by F_{ST} and π_N/π_S outlier values. (Figure 8A-C). Interestingly, the two members of the TNL pair SOLCI001524200 / SOLCI001524300 (genomically adjacent) show almost opposite selective patterns (Figure 8C,F). SOLCI001524300 (TNL) appears to have experienced selection in Northern and Southern populations (and not in the Coastal ones). In addition we highlight two partial genes, unassigned to known functional classes. SOLCI006700600, shows an atypical pattern of possible local selection in the Northern populations (Figure 8D). SOLCI001607700 (Figure 8E) is an example of a complex pattern, away from the Central group with possibly similar selection pressure between the Coastal and Southern populations.

Discussion

NLR genes are important players in plant defence responses and some of them have been shown to be under selection between different *Arabidopsis* species or populations (Mondragon-Palomino & Gaut, 2005; Bakker et al., 2006). Here we investigate for the first time the extend of a geographic mosaic of adaptation in the NLR family across wild populations of a non-model species, *Solanum chilense*. Ultimately we identify populations from different habitats in which coevolutionary dynamics (arms race or trench warfare) occurred along with the number of genes involved in the adaptation to new habitats/pathogens.

To establish the time scale of NLR evolution following habitat colonisation, and identify the genes under selection we used a combined approach of whole genome sequencing with *de novo* assembly, whole genome resequencing of three individuals and targeted resequencing of >200 NLR genes over 14 populations (corresponding to 140 plants).
Our de-novo genome assembly yielded ~717 Mb of genomic sequence and resulted in 25,885 high confidence gene models, 71% of which have good homology and RNA seq support. It is the second fully sequenced and annotated wild tomato species after *Solanum pennellii* (Bolger et al., 2014). The species is known to show both cold (Nosenko et al., 2016) and drought adaptation (Fischer et al., 2013). The availability of a full genome sequence will allow more extensive comparative genomics research related to such adaptations.

We identified a total of 236 NLR genes, including a number of “partial” NLRs near edges of our scaffolds. As noted in previous studies (Andolfo et al., 2014; Stam, Scheikl & Tellier, 2016) such annotation might indicate breakage points in the genome assembly. The actual number of full-length NLR genes is likely to increase in improved assemblies, yet the total number is unlikely to increase. A higher number of NLR genes in syntenic gene blocks between *S. chilense* and *S. lycopersicum* than between *S. chilense and S. pennellii* (71 and 51 NLRs, respectively), suggests that the difference in NLR gene numbers observed between tomato species should only partly be attributed to different quality of genome assembly and annotation. We conclude that the number of NLRs in *S. chilense* is significantly lower than in cultivated tomato (355)(Andolfo et al., 2014) and more similar to that of *Solanum pennellii* (216) (Bolger et al., 2014).These observations also agree with the fact *S. chilense* has been used as a source of introgression of resistance in cultivated tomato (*e.g.* Lin et al., 2014).

Our de-novo assembly of *S. chilense* allowed us to expand our targeted pooled sequencing approach (Stam, Scheikl & Tellier, 2016). π and π_N/π_S are significantly higher in NLR than in control genes. A π_N/π_S ratio of > 1 is an indication for positive or balancing selection. If recurrent selective sweeps would occur very frequently, π_N/π_S may become smaller than one. However, the latter is not probable in our data considering our short divergence times and previous theory of coevolution showing that this signature is unlikely to be observed in hosts (Tellier, Moreno-Gámez & Stephan, 2014). The π_N/π_S ratio remains below 1 for nearly 80% of the NLR genes, suggesting that the majority of NLR are under purifying selection and that the function of most NLRs is conserved within and between populations. This suggest that there are a common set of biotic stresses over the wide range of habitats of *S. chilense*. Looking at the different functional clusters of NLRs, we see that there are large differences in their rate of evolution. The TNLs have lower median π and π_N/π_S than CNL. We show that NLR clades thought to function as homodimers (e.g. CNL6, containing Rpi-blb1 and CNL-RWP8, containing RPS2/5; Qi, DeYoung & Innes, 2012), are very conserved, suggesting important functions for these R genes in each of the populations. Three smaller CNL classes (two of which unique to *S. chilense*) show the highest π_N/π_S. NRC are a class of helper NLRs, they dimerise with other NLR (Wu et al., 2015a) and are required to form functional signalling complexes for response to a multitude of pathogens. As such this NRC class is expected to be conserved at the phylogentic time scale (Wu et al., 2017). We show that this is also visible at the scale of few tens of thousands of years. Interestingly, in the Coastal populations an NRC1 homolog (SOLCI002632400) shows very high π_N/π_S. NRC1 is required for Rx, I-2, Mi-1 and Cf-4/Cf-9 resistance (Gabriels et al., 2007), but not for Pto associated resistance (Wu et al., 2015a). Several single nonsynonymous mutations resulted in gain of function of NRC1 for downstream signalling activity (Sueldo et al., 2015). We suggest therefore that the Coastal region is likely experiencing different pathogen pressures from the rest of the species that resulted in rapid fixation of such mutations at NRC1.

We used a robust combination of demographic inference and coalescent simulations to generate neutral expected patterns of the fixation index (F_{ST}). We confirm previous demography hypothesis (Böndel et al., 2015) and show that two likely independent colonization events occurred: one towards the coast (200,000 ya) and a second towards the Southern mountains (30,000ya). 13 NLR genes have changed selection pressures, 11 towards the coast. Thus a novel NLR gene enters the coevolutionary process, *i.e.* changes direction of selection, between plants and their pathogens every 18,000 years, which would correspond to the typical turn over rate of the suggested Birth and Death process (Michelmore & Meyers, 1998). However, that in *S. chilense*, this rate may be particularly high because of the high amount of standing genetic variation due to the presence of seed banks (Tellier et al., 2011b).

We hypothesised an heterogeneous spatial structure of coevolution across populations and habitats. By establishing $\Delta\pi_N/\pi_S$ per gene between the populations, we could determine the direction of the selection.
In the Coastal region, we expected to observe a lack of selection on NLRs as we assumed the arid environment would be void of phytopathogens. Contrary to our hypothesis, high F_{ST} and $\Delta\pi_N/\pi_S$ indicate that overall there is a strong selection on multiple NLR genes in the Coastal populations. This might be due more seasonal rivers as well as a regularly occurring sea-fog phenomenon (Cereceda & Schemenauer, 1991). These conditions are likely to make the micro-climate conductive for pathogens, so that disease prevalence is higher. The Coastal populations may be hot spots of coevolution (Thompson, 2005) with some pathogen species being possibly unique to these habitats and the host plant populations have adapted over 200.000 years by novel variants being selected across multiple NLRs (Figure 8, genes A-C, E).
Between the Central and the Southern mountain region, we find less evidence for selection at NLRs and outlier genes (for F_{ST} and $\Delta\pi_N/\pi_S$) were distributed across several populations. Both gain as well as loss of selection at few NLRs occurred towards the Southern mountain region. The reason can be that both regions show similar climates or that the divergence time is too recent.

An additional point on spatial heterogeneity is that variation in coevolutionary parameters between populations can generate balancing selection if migration is not too low. At low migration rate, different dynamics (trench warfare or arms race) can occur in the different populations which are uncoupled (Tellier & Brown, 2011; Moreno-Gamez, Stephan & Tellier, 2013). In the Central group where migration is high between populations, selection can be expected to be synchronized between populations. This is seen as several genes in several populations always show up in the F_{ST} comparisons (Figure 8, genes A, C-E). If migration is low, for example between groups, then local adaptation occurs with different dynamics in space (Figure 8, genes A & F). We demonstrate that few NLRs are responsible for a geographic mosaic of adaptation to pathogens, and that colonisation of new habitats promote a change in selection at different genes as expected under a Birth and Death process.

Future work and identification on the natural pathogens present will allow to link these data to pathogen polymorphisms and thus provide insight into the underpinning molecular factors shaping the different plant-pathogen coevolutionary dynamics in nature. Furthermore, we show that certain NLR families are under higher selective pressure and we find novel outlier candidates that seem to play a role in nature for resistance to pathogens, and thus could be valuable Resistance genes to be used in breeding.

Methods

Plant material and accessions

We ordered seeds from 14 accessions of *S. chilense* from the TGRC (UC Davis, California, USA, http://tgrc.ucdavis.edu) and grow 10 plants for each accession in our glasshouse (20°C, 16h light). As shown in previous work on this species, the seed set per accession represents a good sample of the original genetic diversity (Arunyawat, Stephan & Städler, 2007; Tellier et al., 2011a; Böndel et al., 2015). The chosen accessions were LA3111, LA4330, LA2932, LA1958, LA1963, LA2747, LA2755, LA2931, LA3784, LA3786, LA2750, LA4107, LA4117, LA4118. Note that these accessions were collected in different years and were multiplied different number of times at TGRC to produce seeds. However, it is shown that this experimental procedure does not affect the observed genetic diversity (Böndel et al., 2015). The accessions cluster by geographic and genetic proximity and we denote therefore four groups: Central (LA 1958, LA1963, LA2755, LA2931 and LA3111), Northern (LA3784 and LA3786), Coast (LA2750, LA2932 and LA4107), and Southern mountain (LA4117(A), LA4118 and LA4330). A geographic map with the locations of all sampled accessions is found in Figure 05.

Sequencing and available data

DNA was extracted from leaves of individuals of three lines using the DNAeasy Qiagen kit following the instructions of the supplier in order to obtain high quality DNA for sequencing.

Sequences of LA3111 (plant number 3111_t13) were employed to produce a draft genome. The sequencing was conducted at Eurofins Genomics with different library and sequencing techniques. Four libraries were produced for LA3111 with insert sizes of 300bp and 500-550bp for the paired-end, and 8kb and 20kb for the long jumping distance protocol. The 500bp fragment library was sequenced using the MiSeq protocol, and overlapping paired-end reads (~55%) were stitched to longer single reads using the software PEAR (Zhang et al., 2014). Remaining unstitched clusters (45%) were retained as paired-end reads. The other three libraries were analyzed on Illumina HiSeq2500 sequencers. Long jumping distance libraries were postprocessed and cleaned by Eurofins to separate paired-end contaminations and remove low-quality regions and sequence adapters.

In addition, a plant from each accession LA4330 and LA2932 was sequenced using Illumina HiSeq 2500 with standard library size of 300bp at the Sequencing service of Prof. Fries (Chair of Animal Breeding, TUM, Freising). This generated for LA4330 ~64,5M reads, and ~74,4M reads for LA2932. The reads are available on Short Read Archive (SRA) of NCBI. The data will be released upon acceptance of the manuscript.

We also downloaded publicly available data for 9 accessions (one individual per accession) from *S. peruvianum* and *S. chilense*. From Lin et al. we used *S. peruvianum* accessions TS402, TS403, TS404 and the presumably *S. chilense* TS408. From Aflitos et al. we used the *S. peruvianum* LA1278 and LA1954, and the presumed *S. chilense* CGN15530 and CGN15532.

Chloroplast and CT loci phylogenetic reconstructions

To reconstruct the phylogeny of the members of the *S. peruvianum* clade, we mapped our newly sequenced reads from 3 accessions, as well as from all 9 publicly available *S. peruvianum* and assumed *S. chilense* data against the *S. pennellii* reference genome (Bolger et al., 2014) using STAMPY (Lunter & Goodson, 2011), (substitution rate 0.01, may insert size 500), SNPCalling and filtering was done using samtools. (mpileup, call -m with default parameters) and the reconstructed alternative sequences were extracted from *S. pennellii* for the chloroplast region of each of the samples. These aligned sequences were used for phylogenetic tree construction using PhyML (Guindon et al., 2010) (GTR, 1000 bootstraps, Best of NNI&SPR, BioNJ). The resulting tree was visualised in and edited for publication using Figtree (Rambaut & Drummond, 2009).

We extracted for all 12 accessions the sequence at six CT loci (CT066, CT093, CT166, CT179, CT198, CT268). To account for heterozygosity, the two alleles were constructed randomly per individual. An alignment was performed and manually checked to which we added 53 sequences obtained by Sanger sequencing in previous work on *S. chilense* and *S. peruvianum* (Städler, Arunyawat & Stephan, 2008, Tellier et al., 2011b). The outgroup is *S. ochranthum* (accession LA2682). The phylogentic reconstruction (Figure 1 – Supplement 1) was obtained by Neighbour-Joining algorithm as implemented in the software Geneious (1000 Bootstrap). Similar results were obtained using UPMGA (in Geneious) and Maximum Likelihood method (implemented in RaxML).

***De novo* Assembly of *S. chilense* LA3111**

We used the Celera assembler (CAv8.3; https://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.3) employing stitched and unassembled MiSeq read data to generate contigs. The fragment correction module and the bogart unitigger of the celera assembler was applied with a graph and merge error rate of 5%. Minimal overlap length, overlap and merge error rates were set to 50bp and 6% each, respectively. The final contig assembly comprised 150,750 contigs ranging from 1 to 162kb totalling ~717.7 Mb of assembled genome sequence with a N50 of 9,755 bp. The resulting contigs were linked to scaffolds by SSPACE using all four available libraries of LA3111 (Boetzer et al., 2011). Scaffolds were further processed by 5 iterations of GapFiller and corrected by Pilon in full-correction mode (Boetzer & Pirovano, 2012; Walker et al., 2014). The 81,307 final scaffolds span a total size of 914 Mb with a N50 of 70.6 kb.

Gene prediction

We applied a previously described consensus approach (Wang et al., 2014) to derive gene structures from the Solanum chilense draft genome. Briefly, *de novo* genefinders Augustus, Snap, and GeneID were trained on a set of high confidence models that were derived from the LA3111 transcriptome assembly. Exisiting matrices for dicots and *Solanum lycopersicum* were used for predictions with Fgenesh (Salamov & Solovyev, 2000)and GlimmerHMM (Majoros, Pertea & Salzberg, 2004), respectively. Predictions were weighted by a decision tree using the JIGSAW software. Spliced alignments of known proteins and *S.chilense* transcripts of this study were generated by the GenomeThreader tool (Gremme et al., 2005). We used current proteome releases (status of August 2016) of *Arabidopsis thaliana, Medicago truncatula, Ricinus communis, Solanum lycopersicum, Glycine max, Nicotiana benthiamiana, Cucumis sativa* and *Vitis vinifera*. Spliced alignments required a minimal alignment coverage of 50% and a maximal intron size of 50kb under the Arabidopsis splice site model. Next, *de novo* and homology predictions were merged to top-scoring consensus models by their matches to a reference blastp database comprising Arabidopsis, Medicago and *S. lycopersicum* proteins. In a last step, we annotated the top-scoring models using AHRD ("A human readable description")-pipeline (Wang et al., 2014) and interproscan to identify and remove gene models containing transposon signatures. The resulting final models were then classified into high scoring models according to an alignment consistency of ≥90% for both the *S. chilense* query and a subject protein of a combined *S. lycopersicum* and *S. pennellii* database. Residual models were grouped into the low confidence class.

Gene validation

The completeness of the assembled genome was assessed using the BUSCO (Simão et al., 2015). Functional gene annotation and assignment to the GO term categories were performed using Blast2GO v. 4.1 (Conesa & Götz, 2008) based on the results of InterProScan v. 5.21 (Zdobnov & Apweiler, 2001) and BLAST (Altschul et al., 1997)similarity searches against the NCBI non-redundant sequence database. KEGG pathway orthology assignment of protein-coding genes was conducted using the KAAS (Moriya et al., 2007).

To assess synteny between genomes of three tomato species, *S. chilense*, *S. lycopersicum* (NCBI genome annotation release 102), and *S. pennellii* (NCBI genome annotation release 100), we identified orthologous pairs of protein-coding genes using reciprocal BLAST searches with an e-value threshold of 10^{-30} and maximum target sequence number 50. For *S. lycopersicum* and *S. pennellii*, the longest splice variant for each gene was used as a BLAST input. A spacial distribution of resulting ortologous gene pairs was analysed and gene blocks conserved between genomes (syntenic) were identified using iADHoRe (hybrid mode with minimum syntenic block size = 3; (Proost et al., 2012)). For tandem arrays of genes, a single representative was retained in syntenic blocks.

***De novo* assembly of *S. chilense* leaf transcriptome**

RNA-Seq data were generated for two *S. chilense* populations: LA3111 and LA2750. Replicates are obtained by propagating plants vegetatively. Total RNA was extracted from the leaf tissue samples from multiple mature plants under normal and stress (chilling) conditions using the RNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany) and purified from DNA using the TURBO DNA-free Kit (Ambion, Darmstadt, Germany). RNA concentration and integrity were assessed using the Bioanalyzer 2100 (Agilent Technologies, Waldbroon, Germany). The preparation of random primed paired-end Illumina HiSeq2500 libraries and sequencing were conducted by the GATC Biotech AG. Data for each population were

assembled *de novo* using Trinity (Grabherr et al., 2011) SOAPdenovo-Trans (Xie et al., 2014) and Oases-Velvet (Schulz et al., 2012); the resulting assemblies were combined (per-population) and processed using EvidentialGene pipeline (Gilbert, 2013).

Demographic Inference with MSMC and simulation of summary statistics

Sequence data and heterozygosity

We mapped the sequenced reads of the 3 accessions from Central (LA3111), South (LA4330), and Coastal (LA2932) populations against the *S. chilense* reference genome using BWA (mem, call -M with default parameters). SNPCalling was done using samtools (mpileup -q 20 -Q 20 -C 50). We restrict the variant calling to the 200 largest scaffolds of the *S. chilense* reference genome, comprising ~79.6Mb of sequence (mean length=398Kb, min=294Kb, max=1.12Mb). For LA3111 accession we found a total of 171,989 heterozygote sites (per scaffold mean: 859.94; sd: 921.88), 169,990 for LA4330 (mean: 849.95; sd: 894.62), and 176,630 for LA2932 (mean=883.15, sd=908.62).

Demographic inference

We used the multiple sequentially Markovian coalescent (MSMC v2) approach (Schiffels & Durbin, 2014) on unphased called genotypes to estimate both the effective population sizes and cross-coalescence rates over time using one individual form per population representing each of the three main regions (Central LA3111, South LA4330 and Coast LA2932). The selected genomic regions (200 largest scaffolds of the *S. chilense* reference genome) showed 26X mean coverage for LA3111, 20X for LA4330 and 23X for LA2932. As recommended in the MSMC documentation, in addition to the variant matrix, we generate for each scaffold one mask file in bed format which restrict the analysis to the adequately covered regions of a given individual genome. In addition, we generate a mappability mask using Heng Li's SNPable tool (http://lh3lh3.users.sourceforge.net/snpable.shtml) which masks out all regions on the scaffolds on which short sequences cannot be uniquely mapped. Briefly, we used the SNPable tool set to divide the reference genome (200 largest scaffolds in our case) into overlapping short sequences of 100bp and align it back to the genome using BWA (aln -R 1000000 -O 3 -E 3). Then, bed files for the masked regions were generated with a python script (https://gist.github.com/danielecook/cfaa5c359d99bcad3200) and used in MSMC analysis as negative-mask.

For the MSMC analysis we set the time segment patterning parameter (-p) to 15*1. For each run of the cross-coalescence analysis we used four haplotypes, two from each individual from the three populations using the -P 0,0,1,1. We plot the results assuming a mutation rate range of 5×10^{-8} per generation per base pair, and a generation time of 5 years (as assumed in Städler, Arunyawat & Stephan, 2008; Tellier et al., 2011b).

We generate a visual confidence interval of the MSMC curves through replicates of the analysis with subsets of the dataset. This is performed as 10 independent runs of MSMC (Figure 02 – Figure supplement 1), each one run with a random subset of 100 scaffolds (out of 200) using the same settings as for the complete data set.

Power analysis of the inference

The MSMC method results is sensitive to the size of the genomic regions used for the analysis (Schiffels & Durbin, 2014) (). Therefore, we test the ability of MSMC to recover the demographic history of simulated data sets with the same features as our empirical data for some specified demographic scenarios. We used the coalescent simulator ms (Hudson, 2002) to simulate independent scenarios of 1) the populations' demography for each individual, and 2) single scenarios for the divergence and demography of the three individuals. For each scenario we implement 200 independent simulations using the same length size of each scaffold (l) of our data set as well as the Theta (θ) and Rho (ρ) values estimated by MSMC. Population size and divergence time parameters for the simulations were set taking account the results obtained from the MSMC inference.

We simulated two sets of single individual scenarios:

1) Constant effective population size for each population (Central LA3111, South LA4330, and Coast LA2932; Figure 02 – Figure supplement 2). The N_e parameter was fixed to resemble the highest value detected by MSMC for each individual (i.e. 10* N_e for Central individual, 3* N_e for South individual, and 10*

N_e for Coast individual). The ms command line for each model model were:

Central: ms 2 1 -t 10*Ɵ -r ρ l

South: ms 2 1 -t 3*Ɵ -r ρ l

Coast: ms 2 1 -t 2*Ɵ -r ρ l

2) Single event of sudden population expansion. Effective population sizes were set to resemble the highest value detected by MSMC for each individual (i.e. 10* N_e for Central individual, 3* N_e for South individual, and 10* N_e for Coast individual) and then reduced (Figure 02 – Figure supplement 3). The ms command line for each model model were:

Central: ms 2 1 -t 10*Ɵ -r ρ l -eN 0.39 0.1

South: ms 2 1 -t 3*Ɵ -r ρ l -eN 0.36 0.33

Coast: ms 2 1 -t 2*Ɵ -r ρ l -eN 0.4 0.5

Finally, we simulated a scenario with three populations, one expansion process for each population, and independent split processes of Coastal and South populations from the Central population resembling the MSMC results for the empirical data (Figure02) and previous suggestions in Böndel et al. (2015). For each population, a single diploid individual was sampled. We consider this model represents accurately the demographic history of Central, South and Coastal groups (here represented by each one population; Figure 02 – Figure supplement 4). The ms command line was:

ms 6 1 -t Ɵ -r ρ l -I 3 2 2 2 -en 0.12 1 10 -en 0.11 2 4 -en 0.12 3 2.5 -en 0.57 1 1 -ej 0.25 2 1 -ej 0.58 3 1

All scenarios were simulated 10 times. Outputs of each simulation were converted to MSMC input using the ms2multihetsep.py script available in the MSMC-tools. Then, we run MSMC on these three simulated datasets with the same settings used for the empirical data and plotted the results compared with the empirical data (Figure 02 – Figures supplement 2-4).

NLR identification

Putative NLR genes were identified using NLRParser (Steuernagel et al., 2015) using cut-off thresholds as described before (Stam, Scheikl & Tellier, 2016). NLRParser was run on the whole genome of LA3111 and the data overlaid with CDS information and above mentioned RNASeq data. We manually inspected all regions with NLR motifs and updated the annotated open reading frames where this was required. The improved annotation was based on NLR motifs , sequence homology to known NLRs and expression evidence (from the RNAseq). Most CDS were supported by all three measures. Using theRNASeq data for LA3111, 15 NLR genes could be manually improved for annotation and CDS. In the remaining cases, frame shifts made it impossible to enhance the gene model. For these genes the computationally predicted CDS were retained. Members of the NRC clade of NLRs were identified by selecting the best hits of a BLASTP search using cultivated tomato (*S. lycopersicum*) NRC1 and NRC2 (Solyc01g090430, Solyc10g047320). The phylogeny for the NLR was generated based on protein sequence of the NBS, and to define NLR clusters BLASTP searches were used to link new clusters to previously identified ones (Jupe et al., 2013). In one instance, members of our new cluster matched two previously defined clusters equally well, this cluster thus has double naming (CNL1/CNL9). The NLRs in two identified clusters did not match any NLRs that had been clustered previously, in these cases new cluster numbers were assigned (CNL20, CNL21).

Pooled enrichment sequencing

Leaf material was collected from 10 mature plants per population and genomic DNA was extracted and quantified using a Qubit fluorometer (Life technologies). The samples were diluted and pooled per population to obtain 3 ng of high quality DNA with equal amounts per individual. The total DNA is prepared in130 µl for gene enrichment. A sequencing library of enriched DNA was prepared using Agilents SureSelect XT with Custom Probes as in (Stam, Scheikl & Tellier, 2016). In short: DNA was sheared on a Covaris S220 to 600-800 bp. Size selection and cleaning was done using AMPure XP beads (beckman Coultier) in two steps using 1.9:1 and 3.6:2 fragment DNA to beads ratio. The quality was assessed using a Bioanalyzer 2100 (Agilent). End repair, adenylation and adaptor ligation were preformed as described by Agilent. Pre-capture amplification was done using Q5 high fidelity PCR mixes. The amplified library was again quality checked on a Bioanalyzer 2100.

Hybridisation was performed as suggested for libraries <3 Mb. The libraries were indexed with 8bp index primers A01-B08 using Q5 PCR mix, quality was assessed using the Bioanalyzer and individual samples were quantified using Qubit. Eight samples were pooled in equal DNA amounts and the resulting pool was quantified by QPCR using the NGSLibrary quantification kit for Illumina (Quanta biosciences) and diluted down to a final concentration of 20 nM. Sequencing was done on an Illumina MiSeq sequencer to obtain 250 bp paired end reads. Sequencing was done with ZIEL - Institute for Food & Health, Core Facility Microbiome/NGS.

Mapping and Population analysis

Read mapping and SNP calling were done using the same methods as described in Stam et al (2016), with minor modifications. We used bedtools to extract read coverage data and calculate the depth per covered fraction of the targeted region in each sample for CT and NLR genes separately. SNPs were called using two callers GATK (McKenna et al., 2010) and Popoolation (Kofler et al., 2011). To verify the stringency of the filters and the cut-off values used in both programs, we compared the merged SNP calls to Sanger sequence data for three genes for all ten plants for several populations. After comparison, cut-off were adjusted to obtain the best true SNP calls and both callers were run again. To obtain high quality data we assessed mapped reads for all genes for the populations. We then identified a subset of genes with uneven coverage and some instances or for which individual reads introduced large numbers of low frequency SNPs. These reads often overlapped with reads with missing mate pairs, suggesting mis-mapping, likely due to insertion or deletion events or partial gene duplications. To avoid false positive genes (i.e. an over-interpretation of the NLR diversity) these genes were removed from our analysis. Hence we identified a subset of genes with high confidence annotations and mapping Data in the main text are presented for this high quality subset of NLRs. Summary statistics π, θ, π_N and π_S were calculated per site and gene in each population with correction for ploidy in the sample using a method based on the one of Nei and Gojobori (Nei & Gojobori, 1986) as implemented in SNPGenie (Nelson & Hughes, 2015; Nelson, Moncla & Hughes, 2015).

Population genetics and diversity assessment

Per site values obtained for π, θ by SNPGenie were summarised per population statistics were calculated. Per gene π, θ, π_N and π_S statistics, were calculated over the HQ mapped and base-called gene length only to avoid bias due to missing data. Population averages were obtained by averaging per gene statistics. Further analyses were performed using R (R Core Team, 2015). We extracted per gene summary statistics and derived individual statistics per NLR class and genetic groups. F_{ST} values were calculated for pairs of populations using the Hudson et al. (1992) estimator: $F_{ST} = (\pi_{between} - \pi_{within})/ \pi_{between}$. We define π_{within} as the average π for two populations, and $\pi_{between}$ as the nuceotide diversity called on the two populations with all HQ reads merged together. To obtain $\pi_{between}$ the whole pipeline was repeated for all of the 91 possible pairwise comparisons. $\Delta\pi_N/\pi_S$ was calculated as the fold change between two populations: $\log_2((\pi_N/\pi_S)_{pop1} / (\pi_N/\pi_S)_{pop2})$. Results were analysed and visualised in R using the reshape (Wickham & Hadley, 2007) and ggplot2 (Wickham, 2009) packages. Maps were drawn using maps (Becker, Minka & Deckmyn, 2016) .
Statistical comparisons of groups or classes were done using analysis of variance as implemented in the aov() function. Where applicable post-hoc pairwise differences were calculated using tukeyHSD() to account for multiple testing. All comparisons are reported as significantly different when $p < 0.00001$, unless stated otherwise.

Calculation of neutral interval of pairwise F_{ST} measures

The most likely scenario of two independent population split described above (and infered using MSMC, Figure 02 – Figure supplement 4) was used to create a neutral 'null' distribution of the inter-population F_{ST} values. We performed 10,000 simulations using the mean size length of the R-genes assessed in this study (mean length = 2149 bp). We used ms with the following parameter and now for 10 diploid samples per population (ms 60 10000 -t 7.7 -r 0.94 2149 -I 3 20 20 20 -en 0.12 1 10 -en 0.11 2 4 -en 0.12 3 2.5 -en 0.57 1 0.1 -ej 0.25 2 1 -ej 0.58 3 1).
F_{ST} values were calculated using the Hudson et al. (Hudson, Slatkin & Maddison, 1992) estimator $F_{ST} = (\pi_{between} - \pi_{within})/ \pi_{between}$. The proportion of polymorphic sites within each population (π_{within}), and between populations ($\pi_{between}$) were calculated using a custom PERL script written by N. Takebayashi (available at: http://raven.iab.alaska.edu/~ntakebay/teaching/programming/coalsim/scripts/msSS.pl).

Acknowledgements

RS was supported by the Alexander von Humboldt foundation. *S. chilense* genome sequencing was funded by DFG grant TE809/7-1 to AT. Generating and sequencing of the *S. chilense* RNA-Seq data was supported by the DFG grant STE 325/15 to WS. GSA acknowledges funding from a TUM post-doc fellowship. We thank Christine Wurmser (TUM, Chair of Animal Breeding) for help with the re-sequencing of two accessions and Christopher Huptas and Mareike Wenning (ZIEL - Institute for Food & Health Core Facility Microbiome/NGS) for help with the NLR MiSeq sequencing,

Tables

Table1

		LA4330-LA3111		LA2932-LA3111		LA2932-LA4330	
Gene	Clade	Fst	DeltaPiNS	Fst	DeltaPiNS	Fst	DeltaPiNS
SOLCI002229700	**TNL**	**0.6759010399**	**-0.600460045**	**0.7558815644**	**1.33857402**	**0.833602581**	**1.939034062**
SOLCI005342400	**CNL8**	**0.7718180265**	**-0.081483721**	**0.7650123739**	**0.8219733**	**0.8541980893**	**0.903457021**
SOLCI001607700	N/A	**0.695225166**	**1.2184140018**	**0.728020504**	**1.00778738**	0.6908687	-0.21062663
SOLCI001766400	TNL	**0.6670090879**	**0.700714321**	**0.7011355**	**1.2681441**	0.7270852	0.56742978
SOLCI004489700	NRC	**0.7199353894**	**0.493228774**	0.6466276	2.40406057	0.5551948	1.9108318
SOLCI006416700	TNL	**0.6504433481**	**0.398601902**	0.6218002	0.6110404362	0.6909128	0.21243853
SOLCI002632400	NRC	**0.7412130563**	**-0.844470694**	0.6697512	0.4549557061	**0.8779047476**	**1.2994264**
SOLCI001524200	TNL	0.4516363	0.01403061	**0.7351870048**	**0.41213382**	**0.8032428672**	**0.39810321**
SOLCI002096700	CNL21	0.5052818	2.06013238	**0.8133546759**	**2.40449572**	0.6776456	0.34436334
SOLCI006863400	N/A	0.5852913	0.43224908	**0.7195543974**	**1.3074901**	0.7239007	1.73973918
SOLCI000135100	CNL2	0.186865	-1.02632006	0.6350716	4.45791678	**0.8333831827**	**5.484236839**
SOLCI003665500	ind	0.4207573743	1.3715821792	0.464927	-3.08444716	**0.7553752301**	**-4.45602934**
SOLCI006592800	TNL	0.5796519	0.04321498	0.5650659	0.2523890618	**0.7926682468**	**0.20917408**

References

Altschul SF., Madden TL., Schaffer AA., Zhang J., Zhang Z., Miller W., Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res.* 25:3389 – 402.

Andolfo G., Jupe F., Witek K., Etherington GJ., Ercolano MR., Jones JDG. 2014. Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biology* 14:120. DOI: 10.1186/1471-2229-14-120.

Arunyawat U., Stephan W., Städler T. 2007. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution* 24:2310–2322. DOI: 10.1093/molbev/msm162|ISSN.

Badouin H., Gladieux P., Gouzy J., Siguenza S., Aguileta G., Snirc A., Le Prieur S., Jeziorski C., Branca A., Giraud T. 2017. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Molecular Ecology* 26:2041–2062. DOI: 10.1111/mec.13976.

Baggs E., Dagdas G., Krasileva K. 2017. NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Current Opinion in Plant Biology* 38:59–67. DOI: 10.1016/j.pbi.2017.04.012.

Bakker EG., Toomajian C., Kreitman M., Bergelson J. 2006. A Genome-Wide Survey of R Gene Polymorphisms in Arabidopsis. *The Plant Cell* 18:1803–1818. DOI: 10.1105/tpc.106.042614.

Becker OS code by RA., Minka ARWR version by RBE by TP., Deckmyn A. 2016. *maps: Draw Geographical Maps*.

Bergelson J., Kreitman M., Stahl EA., Tian D. 2001. Evolutionary Dynamics of Plant R-Genes. *Science* 292:2281–2285. DOI: 10.1126/science.1061337.

Boetzer M., Henkel CV., Jansen HJ., Butler D., Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578–579. DOI: 10.1093/bioinformatics/btq683.

Boetzer M., Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biology* 13:R56. DOI: 10.1186/gb-2012-13-6-r56.

Bolger A., Scossa F., Bolger ME., Lanz C., Maumus F., Tohge T., Quesneville H., Alseekh S., Sørensen I., Lichtenstein G., Fich EA., Conte M., Keller H., Schneeberger K., Schwacke R., Ofner I., Vrebalov J., Xu Y., Osorio S., Aflitos SA., Schijlen E., Jiménez-Goméz JM., Ryngajllo M., Kimura S., Kumar R., Koenig D., Headland LR., Maloof JN., Sinha N., van Ham RCHJ., Lankhorst RK., Mao L., Vogel A., Arsova B., Panstruga R., Fei Z., Rose JKC., Zamir D., Carrari F., Giovannoni JJ., Weigel D., Usadel B., Fernie AR. 2014. The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature Genetics* 46:1034–1038. DOI: 10.1038/ng.3046.

Böndel KB., Lainer H., Nosenko T., Mboup M., Tellier A., Stephan W. 2015. North–South Colonization Associated with Local Adaptation of the Wild Tomato Species Solanum chilense. *Molecular Biology and Evolution* 32:2932–2943. DOI: 10.1093/molbev/msv166.

Brown JKM., Tellier A. 2011. Plant-parasite coevolution: bridging the gap between genetics and ecology. *Annual review of phytopathology* 49:345–67. DOI: 10.1146/annurev-phyto-072910-095301.

Caicedo AL., Schaal BA. 2004. Heterogeneous evolutionary processes affect R gene diversity in natural populations of Solanum pimpinellifolium. *Proceedings of the National Academy of Sciences of the United States of America* 101:17444–17449. DOI: 10.1073/pnas.0407899101.

Cereceda P., Schemenauer RS. 1991. The Occurrence of Fog in Chile. *Journal of Applied Meteorology* 30:1097–1105. DOI: 10.1175/1520-0450(1991)030<1097:TOOFIC>2.0.CO;2.

Cesari S., Bernoux M., Moncuquet P., Kroj T., Dodds PN. 2014. A novel conserved mechanism for plant NLR protein pairs: the “integrated decoy” hypothesis. *Frontiers in Plant Science* 5. DOI: 10.3389/fpls.2014.00606.

Conesa A., Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008:619832. DOI: 10.1155/2008/619832.

Dawkins R., Krebs JR. 1979. Arms Races between and within Species. *Proc. Royal Soc. London. Series B* 205:489–511.

De Meaux J., Cattan-Toupance I., Lavigne C., Langin T., Neema C. 2003. Polymorphism of a complex resistance gene candidate family in wild populations of common bean (Phaseolus vulgaris) in Argentina: comparison with phenotypic resistance polymorphism. *Molecular Ecology* 12:263–273. DOI: 10.1046/j.1365-294X.2003.01718.x.

Fischer I., Steige KA., Stephan W., Mboup M. 2013. Sequence Evolution and Expression Regulation of Stress-Responsive Genes in Natural Populations of Wild Tomato. *PLOS ONE* 8:e78182. DOI: 10.1371/journal.pone.0078182.

Gabriels SH., Vossen JH., Ekengren SK., van Ooijen G., Abd-El-Haliem AM., van den Berg GC., Rainey DY., Martin GB., Takken FL., de Wit PJ., Joosten MH. 2007. An NB-LRR protein required for HR signalling mediated by both extra- and intracellular resistance proteins. *Plant J* 50:14–28. DOI: TPJ3027 [pii] 10.1111/j.1365-313X.2007.03027.x.

Gandon S., Buckling A., Decaestecker E., Day T. 2008. Host–parasite coevolution and patterns of adaptation across time and space. *Journal of Evolutionary Biology* 21:1861–1866. DOI: 10.1111/j.1420-

9101.2008.01598.x.
Gilbert D. 2013. Gene-omes built from mRNA seq not genome DNA. In: *7th annual arthropod genomics symposium. Notre Dame*.
Grabherr MG., Haas BJ., Yassour M., Levin JZ., Thompson DA., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren BW., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* 29:644–652. DOI: 10.1038/nbt.1883.
Gremme G., Brendel V., Sparks ME., Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* 47:965–978. DOI: 10.1016/j.infsof.2005.09.005.
Guindon S., Dufayard J-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59:307–321.
Guo YL., Fitz J., Schneeberger K., Ossowski S., Cao J., Weigel D. 2011. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. *Plant Physiol* 157:757–769.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Hudson RR., Slatkin M., Maddison WP. 1992. Estimation of levels of gene flow from DNA-sequence data. *Genetics* 132:583–589.
Jones JDG., Vance RE., Dangl JL. 2016. Intracellular innate immune surveillance devices in plants and animals. *Science* 354:aaf6395. DOI: 10.1126/science.aaf6395.
Jupe F., Pritchard L., Etherington GJ., MacKenzie K., Cock PJ., Wright F., Sharma SK., Bolser D., Bryan GJ., Jones JD., Hein I. 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics* 13:75. DOI: 10.1186/1471-2164-13-75.
Jupe F., Witek K., Verweij W., Śliwka J., Pritchard L., Etherington GJ., Maclean D., Cock PJ., Leggett RM., Bryan GJ., Cardle L., Hein I., Jones JD. 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. *The Plant Journal* 76:530–544. DOI: 10.1111/tpj.12307.
Kofler R., Orozco-terWengel P., De Maio N., Pandey RV., Nolte V., Futschik A., Kosiol C., Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One* 6:e15925. DOI: 10.1371/journal.pone.0015925.
Lin T., Zhu G., Zhang J., Xu X., Yu Q., Zheng Z., Zhang Z., Lun Y., Li S., Wang X., Huang Z., Li J., Zhang C., Wang T., Zhang Y., Wang A., Zhang Y., Lin K., Li C., Xiong G., Xue Y., Mazzucato A., Causse M., Fei Z., Giovannoni JJ., Chetelat RT., Zamir D., Städler T., Li J., Ye Z., Du Y., Huang S. 2014. Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics* 46:1220–1226. DOI: 10.1038/ng.3117.
Lunter G., Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21:936–939. DOI: 10.1101/gr.111120.110.
Majoros WH., Pertea M., Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20:2878–2879. DOI: 10.1093/bioinformatics/bth315.
Maqbool A., Saitoh H., Franceschetti M., Stevenson CEM., Uemura A., Kanzaki H., Kamoun S., Terauchi R., Banfield MJ. 2015. Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor. *eLife* 4:e08709. DOI: 10.7554/eLife.08709.
McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297–1303. DOI: 10.1101/gr.107524.110.
Meyers BC., Shen KA., Rohani P., Gaut BS., Michelmore RW. 1998. Receptor-like Genes in the Major Resistance Locus of Lettuce Are Subject to Divergent Selection. *The Plant Cell Online* 10:1833–1846. DOI: 10.1105/tpc.10.11.1833.
Michelmore RW., Meyers BC. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research* 8:1113–1130.
Mondragon-Palomino M., Gaut BS. 2005. Gene Conversion and the Evolution of Three Leucine-Rich Repeat Gene Families in Arabidopsis thaliana. *Molecular Biology and Evolution* 22:2444–2456. DOI: 10.1093/molbev/msi241.
Mondragón-Palomino M., Meyers BC., Michelmore RW., Gaut BS. 2002. Patterns of Positive Selection in the Complete NBS-LRR Gene Family of Arabidopsis thaliana. *Genome Research* 12:1305–1315. DOI: 10.1101/gr.159402.
Moreno-Gamez S., Stephan W., Tellier A. 2013. Effect of disease prevalence and spatial heterogeneity on polymorphism maintenance in host-parasite interactions. *Plant Pathology* 62:133–141.
Moriya Y., Itoh M., Okuda S., Yoshizawa AC., Kanehisa M. 2007. KAAS: an automatic genome annotation

and pathway reconstruction server. *Nucleic Acids Research* 35:W182–W185. DOI: 10.1093/nar/gkm321.
Nei M., Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:418–426.
Nelson CW., Hughes AL. 2015. Within-host nucleotide diversity of virus populations: Insights from next-generation sequencing. *Infection, Genetics and Evolution* 30:1–7. DOI: 10.1016/j.meegid.2014.11.026.
Nelson CW., Moncla LH., Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* 31:3709–3711. DOI: 10.1093/bioinformatics/btv449.
Nosenko T., Böndel KB., Kumpfmüller G., Stephan W. 2016. Adaptation to low temperatures in the wild tomato species Solanum chilense. *Molecular Ecology* 25:2853–2869. DOI: 10.1111/mec.13637.
Pan Q., Wendel J., Fluhr R. 2000. Divergent Evolution of Plant NBS-LRR Resistance Gene Homologues in Dicot and Cereal Genomes. *Journal of Molecular Evolution* 50:203–213. DOI: 10.1007/s002399910023.
Parratt SR., Numminen E., Laine A-L. 2016. Infectious Disease Dynamics in Heterogeneous Landscapes. *Annual Review of Ecology, Evolution, and Systematics* 47:283–306. DOI: 10.1146/annurev-ecolsys-121415-032321.
Pease JB., Haak DC., Hahn MW., Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biology* 14:e1002379. DOI: 10.1371/journal.pbio.1002379.
Proost S., Fostier J., De Witte D., Dhoedt B., Demeester P., Van de Peer Y., Vandepoele K. 2012. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* 40:e11. DOI: 10.1093/nar/gkr955.
Qi D., DeYoung BJ., Innes RW. 2012. Structure-Function Analysis of the Coiled-Coil and Leucine-Rich Repeat Domains of the RPS5 Disease Resistance Protein1[W][OA]. *Plant Physiology* 158:1819–1832. DOI: 10.1104/pp.112.194035.
R Core Team 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rambaut A., Drummond A. 2009. *FigTree v1. 3.1*.
Rose LE., Grzeskowiak L., Hörger AC., Groth M., Stephan W. 2011. Targets of selection in a disease resistance network in wild tomatoes. *Molecular Plant Pathology* 12:921–927. DOI: 10.1111/j.1364-3703.2011.00720.x.
Rose LE., Michelmore RW., Langley CH. 2007. Natural variation in the Pto disease resistance gene within species of wild tomato (Lycopersicon). II. Population genetics of Pto. *Genetics* 175:1307–1319.
Salamov AA., Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome research* 10:516–522.
Sarris PF., Cevik V., Dagdas G., Jones JDG., Krasileva KV. 2016. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biology* 14:8. DOI: 10.1186/s12915-016-0228-7.
Schiffels S., Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* 46:919–925. DOI: 10.1038/ng.3015.
Schulz MH., Zerbino DR., Vingron M., Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)* 28:1086–1092. DOI: 10.1093/bioinformatics/bts094.
Sela H., Cheng J., Jun Y., Nevo E., Fahima T. 2009. Divergent diversity patterns of NBS and LRR domains of resistance gene analogs in wild emmer wheat populations. *Genome* 52:557–565. DOI: 10.1139/G09-030.
Simão FA., Waterhouse RM., Ioannidis P., Kriventseva EV., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212. DOI: 10.1093/bioinformatics/btv351.
Städler T., Arunyawat U., Stephan W. 2008. Population genetics of speciation in two closely related wild tomatoes (Solanum section Lycopersicon). *Genetics* 178:339–50. DOI: 10.1534/genetics.107.081810.
Städler T., Roselius K., Stephan W. 2005. Genealogical footprints of speciation processes in wild tomatoes: Demography and evidence for historical gene flow. *Evolution* 59:1268–1279.
Stahl EA., Dwyer G., Mauricio R., Kreitman M., Bergelson J. 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature* 400:667–671.
Stam R., Scheikl D., Tellier A. 2016. Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population. *Genome Biology and Evolution* 8:1501–1515. DOI: 10.1093/gbe/evw094.
Stam R., Scheikl D., Tellier A. 2017. The wild tomato species Solanum chilense shows variation in pathogen resistance between geographically distinct populations. *PeerJ* 5:e2910. DOI: 10.7717/peerj.2910.
Steuernagel B., Jupe F., Witek K., Jones JDG., Wulff BBH. 2015. NLR-parser: rapid annotation of plant NLR

complements. *Bioinformatics (Oxford, England)* 31:1665–1667. DOI: 10.1093/bioinformatics/btv005.
Steuernagel B., Periyannan SK., Hernández-Pinzón I., Witek K., Rouse MN., Yu G., Hatta A., Ayliffe M., Bariana H., Jones JDG., Lagudah ES., Wulff BBH. 2016. Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nature Biotechnology* 34:652–655. DOI: 10.1038/nbt.3543.
Sueldo DJ., Shimels M., Spiridon LN., Caldararu O., Petrescu A-J., Joosten MHAJ., Tameling WIL. 2015. Random mutagenesis of the nucleotide-binding domain of NRC1 (NB-LRR Required for Hypersensitive Response-Associated Cell Death-1), a downstream signalling nucleotide-binding, leucine-rich repeat (NB-LRR) protein, identifies gain-of-function mutations in the nucleotide-binding pocket. *New Phytologist* 208:210–223. DOI: 10.1111/nph.13459.
Tabaeizadeh Z., Agharbaoui Z., Harrak H., Poysa V. 1999. Transgenic tomato plants expressing a Lycopersicon chilense gene demonstrate improved resistance to Verticillium dahliae race 2. *Plant Cell Reports* 19:197–202. DOI: 10.1007/s002990050733.
Tack AJM., Laine A-L. 2014. Ecological and evolutionary implications of spatial heterogeneity during the off-season for a wild plant pathogen. *New Phytologist* 202:297–308. DOI: 10.1111/nph.12646.
Tellier A., Brown JKM. 2011. Spatial heterogeneity, frequency-dependent selection and polymorphism in host-parasite interactions. *BMC Evolutionary Biology* 11:319.
Tellier A., Fischer I., Merino C., Xia H., Camus-Kulandaivelu L., Städler T., Stephan W. 2011a. Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity* 107:189–199.
Tellier A., Laurent SJY., Lainer H., Pavlidis P., Stephan W. 2011b. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences of the United States of America* 108:17052–7. DOI: 10.1073/pnas.1111266108.
Tellier A., Moreno-Gámez S., Stephan W. 2014. Speed of Adaptation and Genomic Footprints of Host–Parasite Coevolution Under Arms Race and Trench Warfare Dynamics. *Evolution* 68:2211–2224. DOI: 10.1111/evo.12427.
The 100 Tomato Genome Sequencing Consortium, Aflitos S., Schijlen E., de Jong H., de Ridder D., Smit S., Finkers R., Wang J., Zhang G., Li N., Mao L., Bakker F., Dirks R., Breit T., Gravendeel B., Huits H., Struss D., Swanson-Wagner R., van Leeuwen H., van Ham RCHJ., Fito L., Guignier L., Sevilla M., Ellul P., Ganko E., Kapur A., Reclus E., de Geus B., van de Geest H., te Lintel Hekkert B., van Haarst J., Smits L., Koops A., Sanchez-Perez G., van Heusden AW., Visser R., Quan Z., Min J., Liao L., Wang X., Wang G., Yue Z., Yang X., Xu N., Schranz E., Smets E., Vos R., Rauwerda J., Ursem R., Schuit C., Kerns M., van den Berg J., Vriezen W., Janssen A., Datema E., Jahrman T., Moquet F., Bonnet J., Peters S. 2014. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *The Plant Journal* 80:136–148. DOI: 10.1111/tpj.12616.
Thompson JN. 2005. *The Geographic Mosaic of Coevolution*. Chicago: University of Chicago Press.
Thrall PH., Burdon JJ. 2003. Evolution of virulence in a plant host-pathogen metapopulation. *Science (New York, N.Y.)* 299:1735–1737. DOI: 10.1126/science.1080070.
Thrall PH., Burdon JJ., Bock CH. 2001. Short-term epidemic dynamics in the Cakile maritima–Alternaria brassicicola host–pathogen association. *Journal of Ecology* 89:723–735. DOI: 10.1046/j.0022-0477.2001.00598.x.
Thrall PH., Laine A-L., Ravensdale M., Nemri A., Dodds PN., Barrett LG., Burdon JJ. 2012. Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. *Ecology letters* 15:425–35. DOI: 10.1111/j.1461-0248.2012.01749.x.
Van Valen L. 1973. A New Evolutionary Law. *Evolutionary Theory* 1:1–30.
Van Weymers PSM., Baker K., Chen X., Harrower B., Cooke DEL., Gilroy EM., Birch PRJ., Thilliez GJA., Lees AK., Lynott JS., Armstrong MR., McKenzie G., Bryan GJ., Hein I. 2016. Utilizing “Omic” Technologies to Identify and Prioritize Novel Sources of Resistance to the Oomycete Pathogen Phytophthora infestans in Potato Germplasm Collections. *Crop Science and Horticulture*:672. DOI: 10.3389/fpls.2016.00672.
Verlaan MG., Hutton SF., Ibrahem RM., Kormelink R., Visser RGF., Scott JW., Edwards JD., Bai Y. 2013. The Tomato Yellow Leaf Curl Virus Resistance Genes Ty-1 and Ty-3 Are Allelic and Code for DFDGD-Class RNA–Dependent RNA Polymerases. *PLOS Genet* 9:e1003399. DOI: 10.1371/journal.pgen.1003399.
Walker BJ., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo CA., Zeng Q., Wortman J., Young SK., Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* 9:e112963. DOI: 10.1371/journal.pone.0112963.
Wang W., Haberer G., Gundlach H., Gläßer C., Nussbaumer T., Luo MC., Lomsadze A., Borodovsky M., Kerstetter RA., Shanklin J., Byrant DW., Mockler TC., Appenroth KJ., Grimwood J., Jenkins J., Chow J., Choi C., Adam C., Cao X-H., Fuchs J., Schubert I., Rokhsar D., Schmutz J., Michael TP., Mayer KFX., Messing J. 2014. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature Communications* 5:ncomms4311. DOI: 10.1038/ncomms4311.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
Wickham, Hadley 2007. Reshaping data with the reshape package. *Journal of Statistical Software* 21.
Witek K., Jupe F., Witek AI., Baker D., Clark MD., Jones JDG. 2016. Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nature Biotechnology* advance online publication. DOI: 10.1038/nbt.3540.
Woolhouse MEJ., Webster JP., Domingo E., Charlesworth B., Levin BR. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature genetics* 32:569–577. DOI: 10.1038/ng1202-569.
Wu C-H., Abd-El-Haliem A., Bozkurt TO., Belhaj K., Terauchi R., Vossen JH., Kamoun S. 2017. NLR network mediates immunity to diverse plant pathogens. *Proceedings of the National Academy of Sciences*:201702041. DOI: 10.1073/pnas.1702041114.
Wu C-H., Belhaj K., Bozkurt TO., Birk MS., Kamoun S. 2015a. Helper NLR proteins NRC2a/b and NRC3 but not NRC1 are required for Pto-mediated cell death and resistance in Nicotiana benthamiana. *New Phytologist*:n/a-n/a. DOI: 10.1111/nph.13764.
Wu C-H., Krasileva KV., Banfield MJ., Terauchi R., Kamoun S. 2015b. The “sensor domains” of plant NLR proteins: more than decoys? *Frontiers in Plant Science* 6. DOI: 10.3389/fpls.2015.00134.
Xia H., Camus-Kulandaivelu L., Stephan W., Tellier A., Zhang Z. 2010. Nucleotide diversity patterns of local adaptation at drought-related candidate genes in wild tomatoes. *Molecular Ecology* 19:4144–4154. DOI: 10.1111/j.1365-294X.2010.04762.x.
Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Huang W., He G., Gu S., Li S., Zhou X., Lam T-W., Li Y., Xu X., Wong GK-S., Wang J. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics (Oxford, England)* 30:1660–1666. DOI: 10.1093/bioinformatics/btu077.
Zdobnov EM., Apweiler R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.
Zhang J., Kobert K., Flouri T., Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30:614–620. DOI: 10.1093/bioinformatics/btt593.

In file Figures:
Figure 01

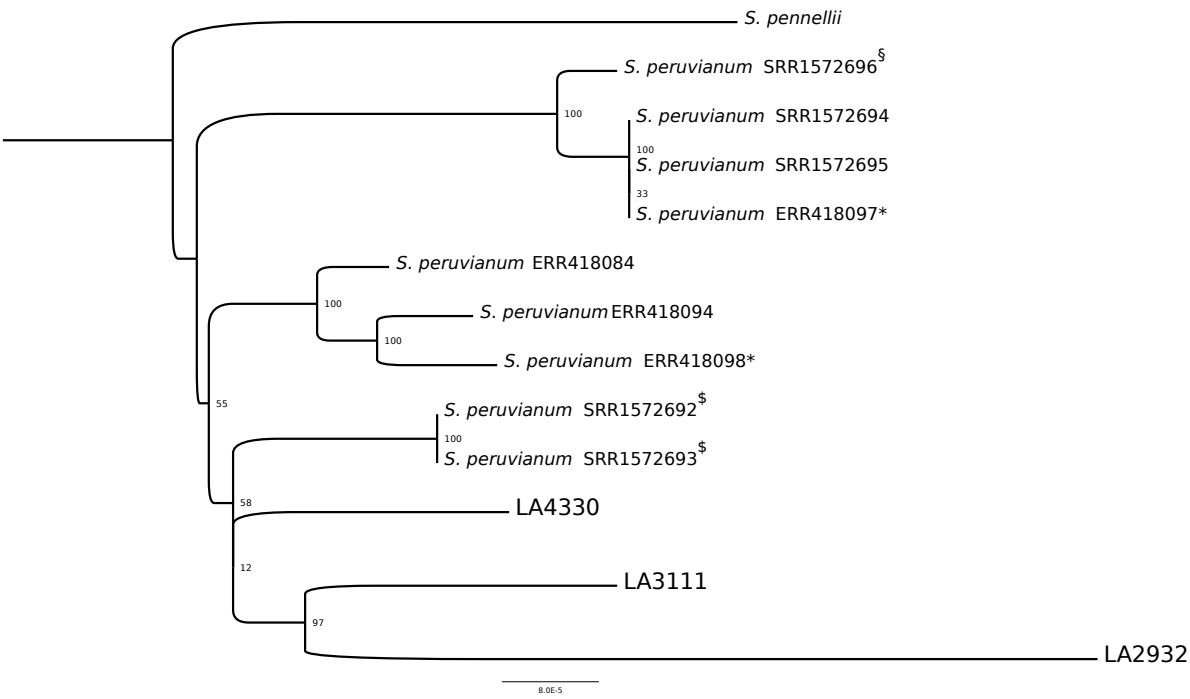
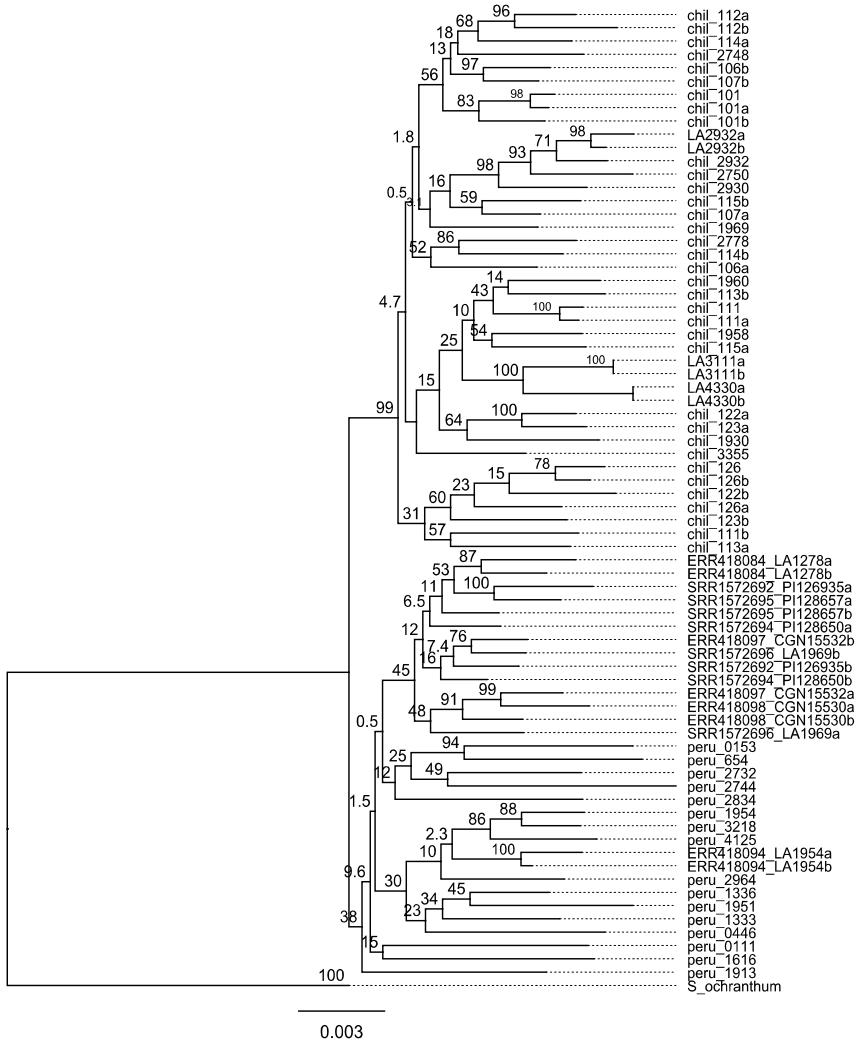


Figure 01

Phylogeny of Snps in chloroplasts extracted from our three *S. chilense* samples and previously sequenced *S. peruvianum* and alleged *S. chilense* samples.
Tree was constructed after extracting the data mapped to the *S. pennellii* reference genome. A tree was built for the aligned sequences using PhyML (JC, NNI, BioNJ, with 1000 bootstraps). Bootstrap values are reported on each of the branches.
§ Specimen was reported as *S. chilense* in the main text of the paper (Jin et al, 2014), but the authors confirm it is *S. peruvianum*, as is written in the supplementary data of their paper that contains all origin data.
$ Both specimens were deposited by Jin et al as coming from the same *S. peruvianum* individual (TS-402)
* Both specimens were labeled as *S. chilense* in the Aflitos et al, 2014, but this classification has since been withdrawn from the CGN database. The accompanying pictures on the CGN website are not showing *S. chilense* plants.

Figure 1 – Figure supplement 1

Figure 01 – Figure supplement 1



Phylogeny based on six CT loci (nuclear genes) extracted from our three *S. chilense* samples and previously sequenced *S. peruvianum* and alleged *S. chilense* samples.
The phylogeny was constructed after extracting the data mapped to the *S. pennellii* reference genome. A tree was built for the aligned sequences using the Neighbour-Joining algorithm (in the software Geneious, 1000 bootstrap replicates). Bootstrap values are reported on each of the branches.
Chil and *peru* indicates sanger sequence from *S. chilense* and *S. peruvianum,* respectively.

Figure 02 - MSMC

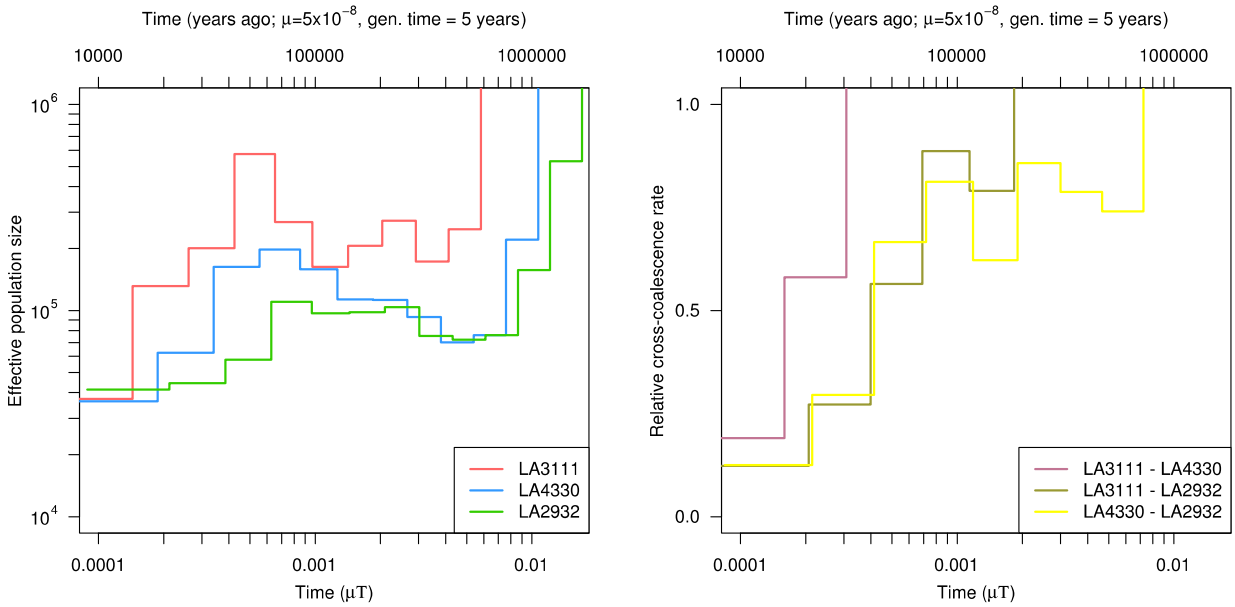


Figure 02. Historical demography reconstructions based on whole genome data of one individual from each of the Central, Coastal and South populations.
A. Effective population size through time estimations for Central (LA3111; red line), Coastal (LA2932; green line) and South (LA4330; blue line) populations obtained with MSMC.
B. Estimation of the genetic divergence between pairs of populations through time. Central-South (LA3111-LA4330; salmon line), Central-Coastal (LA3111-LA2932; green line) and South- Coastal (LA4330-LA2932; yellow line). The measure is based on the ratio between the cross-population and within-population coalescence rates (y axis) as a function of time (x axis). 1 indicate well mixed populations and rates = 0 indicate fully separated populations. The split of South from Central population is around 10 times later than the split of Coastal from Central population. The estimation of the separation of Coastal and South populations appear earlier suggesting that divergence of South and Coastal population should be independent processes and there is no gene flow between them. Times on the lower x-axes are given in units of divergence per base pair and times on the upper x-axes are given in years before present.

Figure 2 – Figure Supplement 1

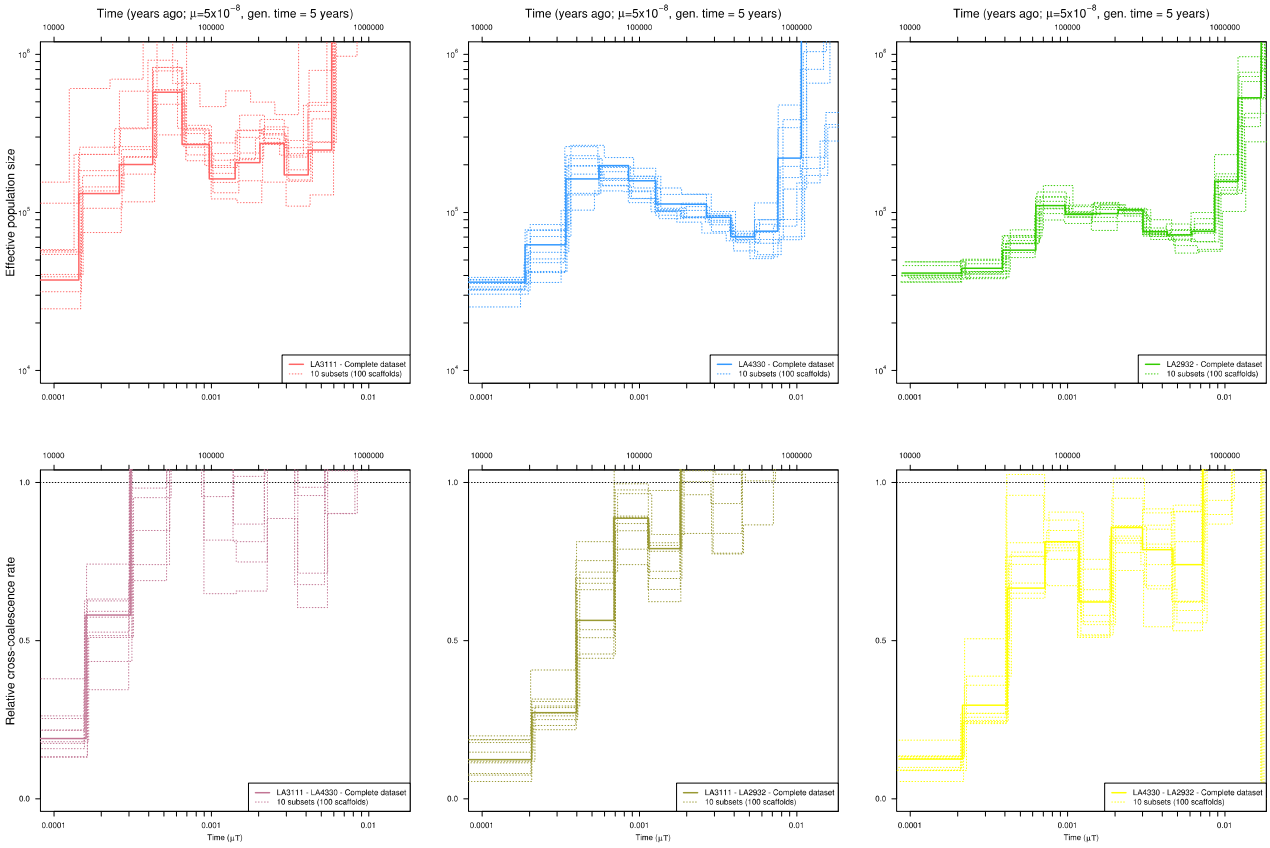


Figure 02 – Supplement 02. Confidence intervals of the demographic estimations obtained with MSMC. Upper row corresponds to the effective population size estimations for Central (LA3111; red line), Coastal (LA2932; green line) and South (LA4330; blue line) populations. Bottom row corresponds to the estimations of the genetic divergence between pairs of populations through time. Central-South (LA3111-LA4330; salmon line), Central-Coastal (LA3111-LA2932; green line) and South- Coastal (LA4330-LA2932; yellow line).
The solid lines represents the estimations obtained with complete dataset, whereas the dashed lines represent estimations based on a random sampling of 100 scaffolds of the *S. chilense* reference genome. Genetic divergence between populations is measured as relative cross-coalescence rates (y axis) as a function of time (x axis). 1 indicate well mixed populations and rates = 0 indicate fully separated populations.

Figure 02 - Supplement 02

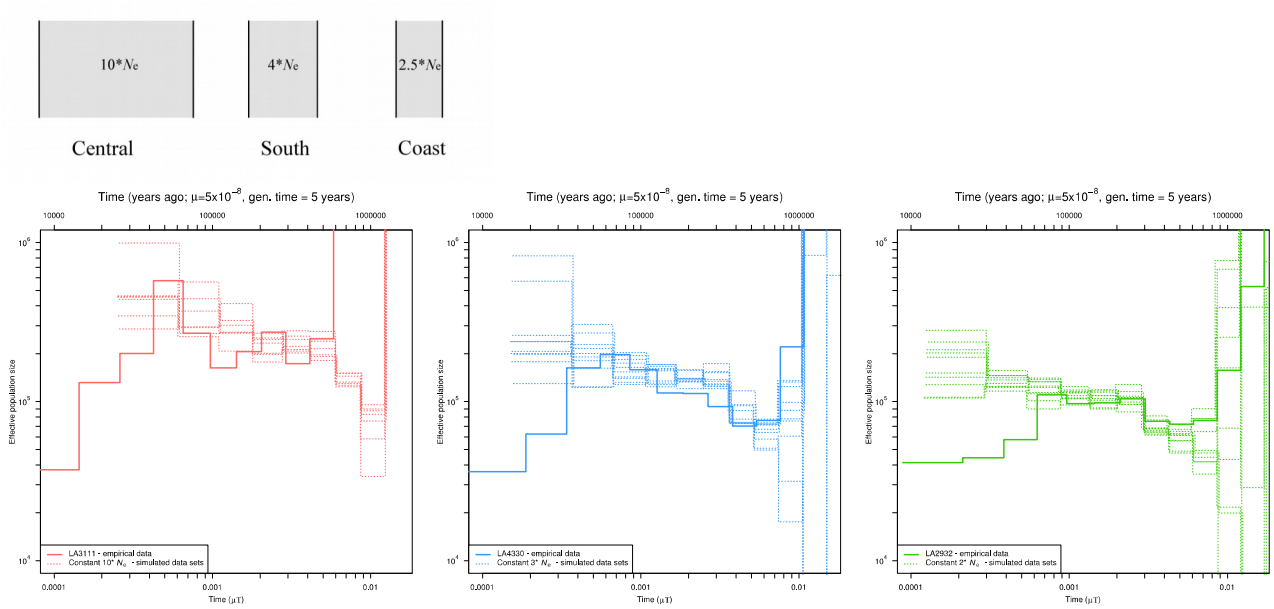


Figure 02 – Supplement 02. Effective population size (N_e) estimations for individual simulations of constant N_e scenarios.
The simulations were based on the current N_e value estimated with MSMC for the empirical genome data. The solid line represents the complete dataset, whereas the dashed lines represent the estimations for 10 independent simulations. Times on the lower x-axes are given in units of divergence per base pair and times on the upper x-axes are given in years before present.

Figure 2 – Supplement 03

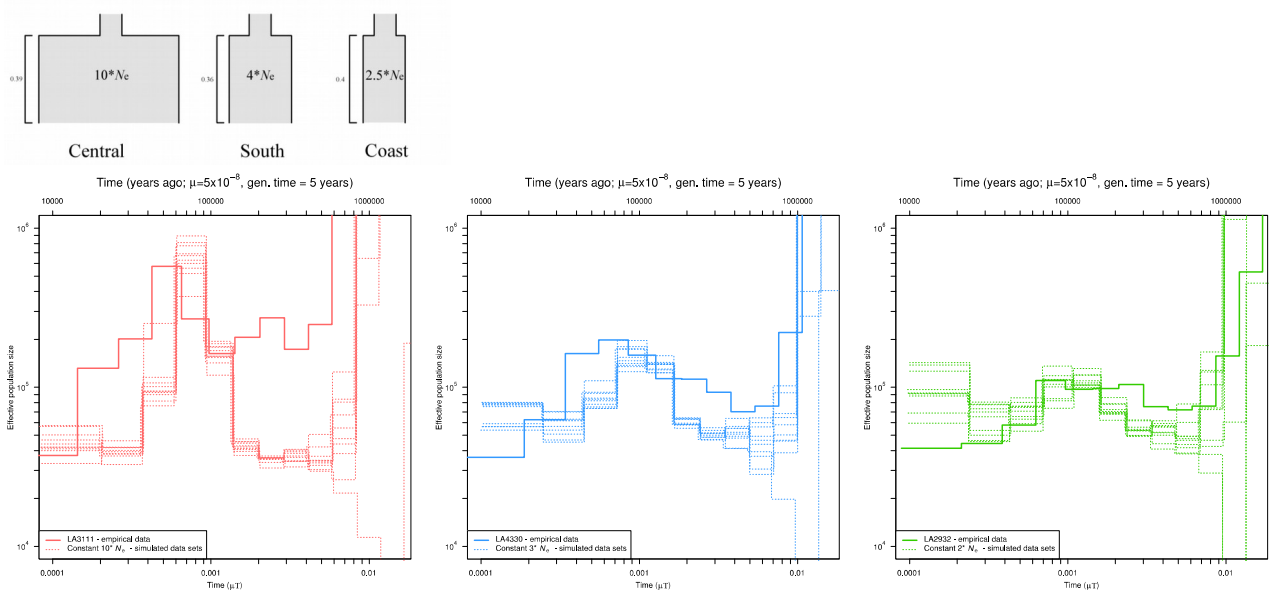


Figure 02 – Supplement 03. Effective population size (N_e) estimations for individual simulations of past expansion of N_e scenarios.
The simulations were based on the current N_e and the time of population expansion values estimated with MSMC for the empirical genome data. The solid line represents the complete dataset, whereas the dashed lines represent the estimations for 10 independent simulations. Times on the lower x-axes are given in units of divergence per base pair and times on the upper x-axes are given in years before present.

Figure 02 – Supplement 04

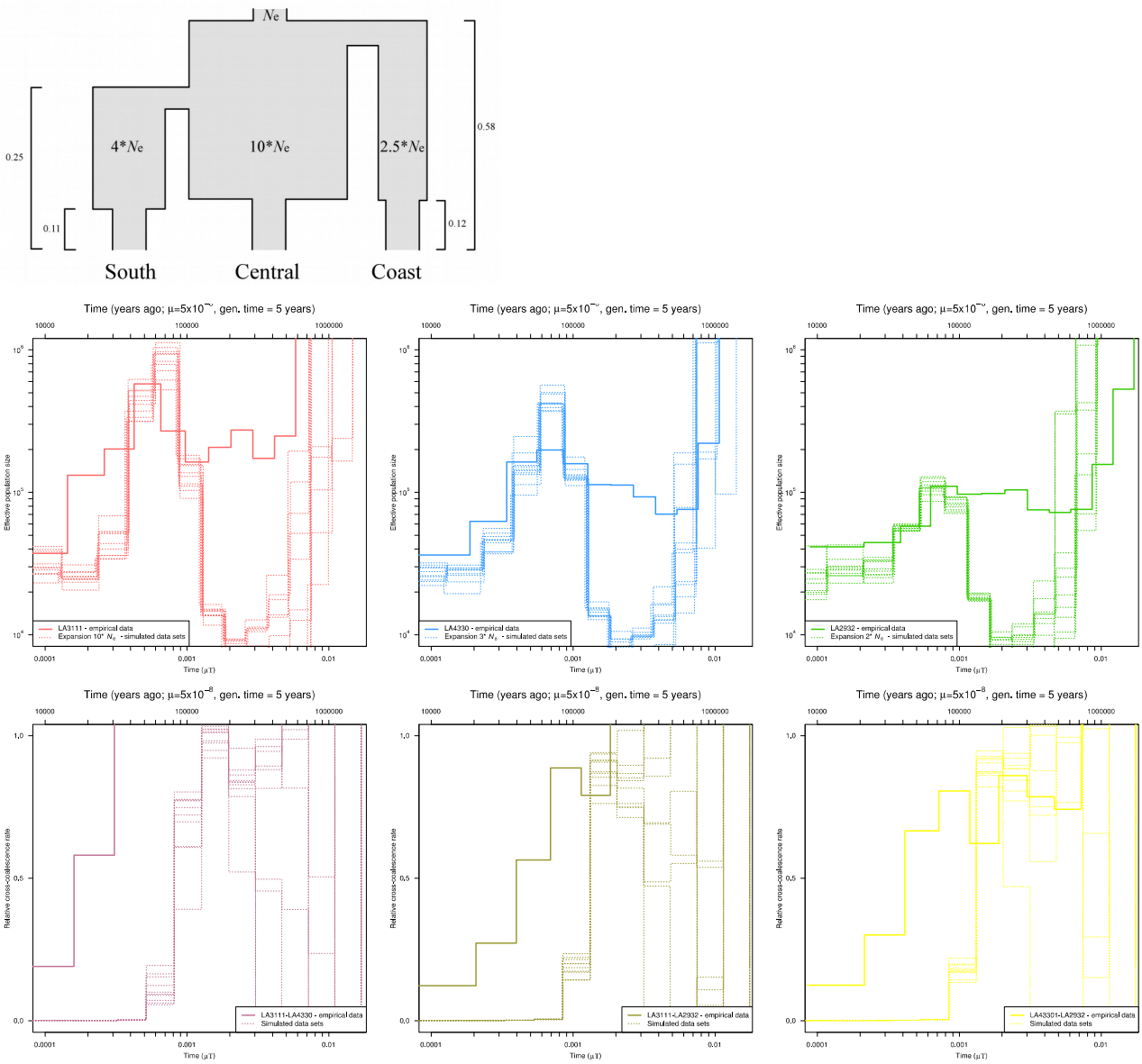


Figure 02 – Supplement 04. Effective population size (N_e) and genetic divergence estimations for individual simulations of a demographic scenario inferred from the MSMC results.
The simulations were based on the current N_e, the time of population expansion, and split times values estimated with MSMC for the empirical genome data. The solid line represents the complete dataset, whereas the dashed lines represent the estimations for 10 independent simulations. Upper row corresponds to the effective population size estimations for Central (LA3111; red line), Coastal (LA2932; green line) and South (LA4330; blue line) populations. Bottom row corresponds to the estimations of the genetic divergence between pairs of populations through time. Central-South (LA3111-LA4330; salmon line), Central-Coastal (LA3111-LA2932; green line) and South- Coastal (LA4330-LA2932; yellow line). Times on the lower x-axes are given in units of divergence per base pair and times on the upper x-axes are given in years before present.

Figure 03

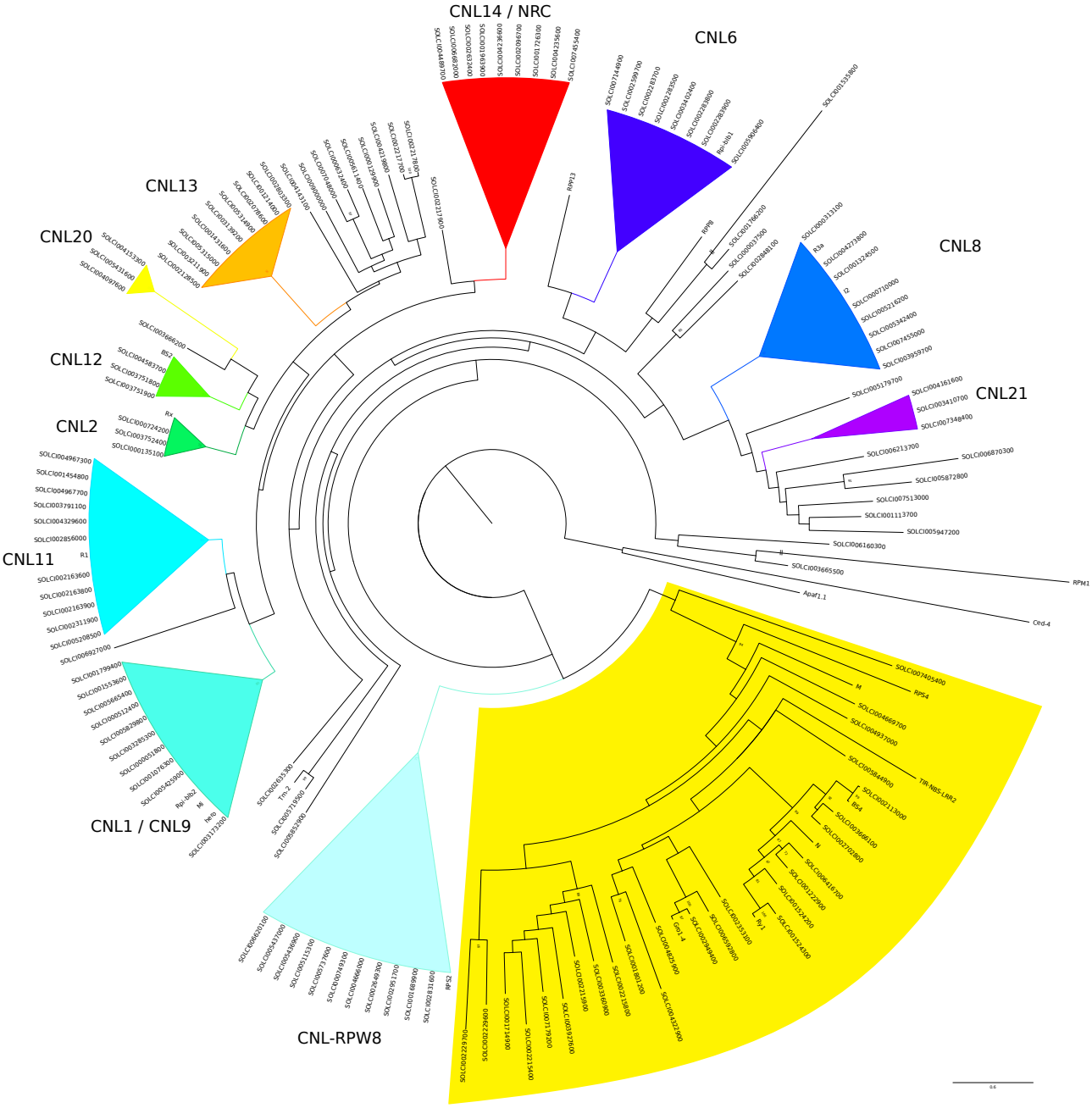


Figure 03

Phylogenetic tree for the NLR identified in *S. chilense*

The tree was made as described in (Stam, Scheikl, and Tellier 2016). Clades with high (>80%) bootstrap values are collapsed. Most previously described clades can be identified and are indicated as such. The TNL famimly is highlighted in yellow. Several previously identified NLR genes from different species are included for comparison and Apaf1.1 and Ced4 are used as an outgroup.

Figure 04

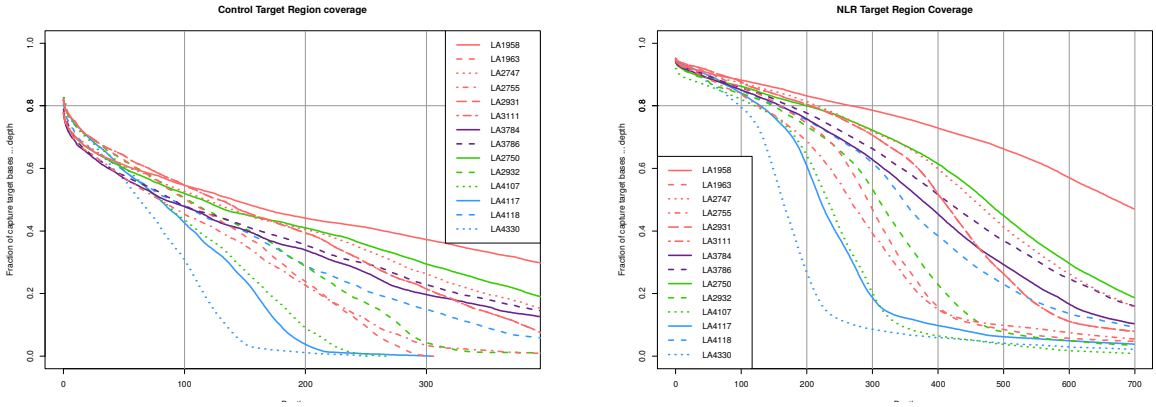


Figure 04
Coverage plots for enrichment sequencing. X-axis shows the depth of coverage, ranging from 0 to 400 or 700 for CT or NLR genes respectively. Y-axis shows the fraction of the targeted region (e.g. the part of the genome that was targeted by the probes) that has the indicated coverage or higher.

Figure 4 – Suppllement 01

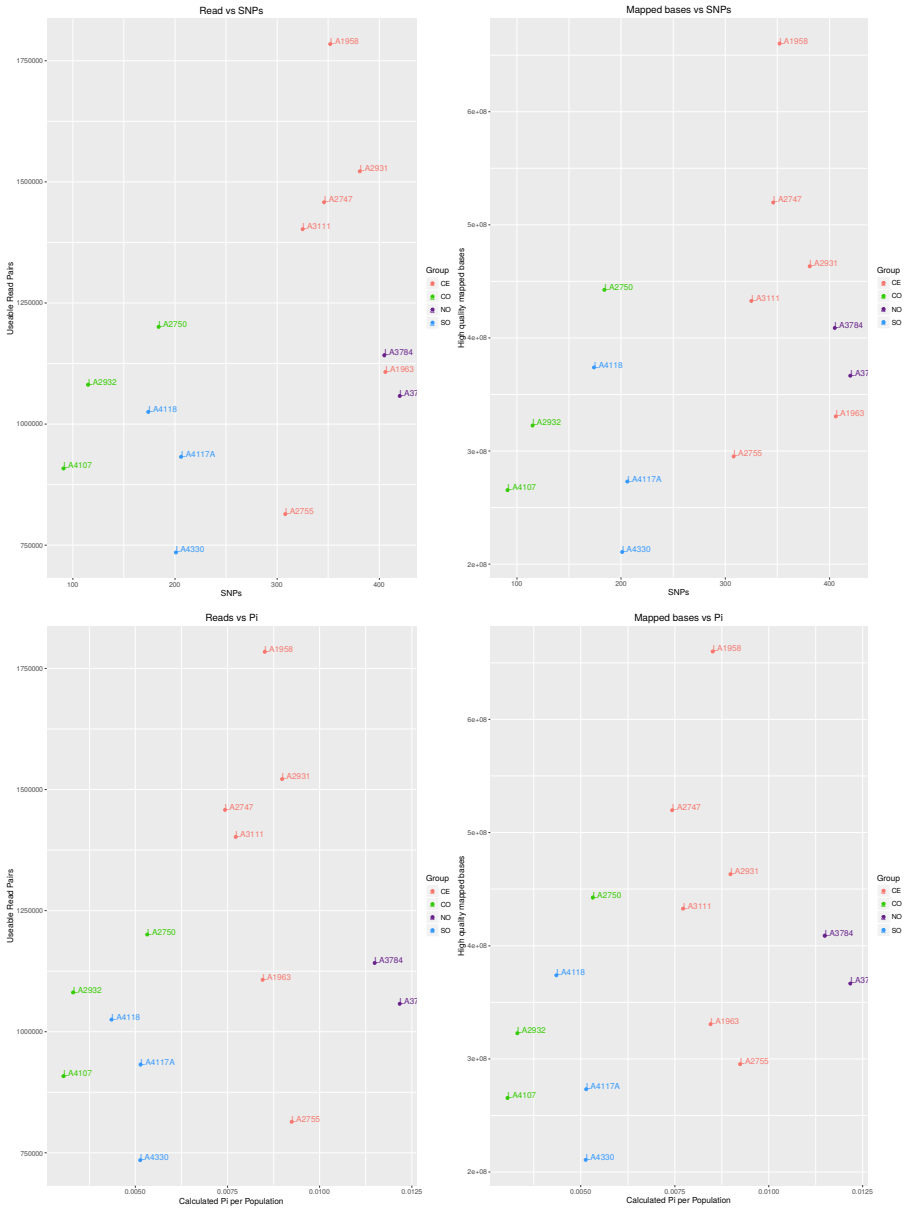


Figure 04 – Figure Supplement 01
Number of sequenced reads (left) or high quality mappable read pairs (right) plotted agains the total number of SNPs per sample (top) or π per population.

Figure 04 – Figure Supplement 02

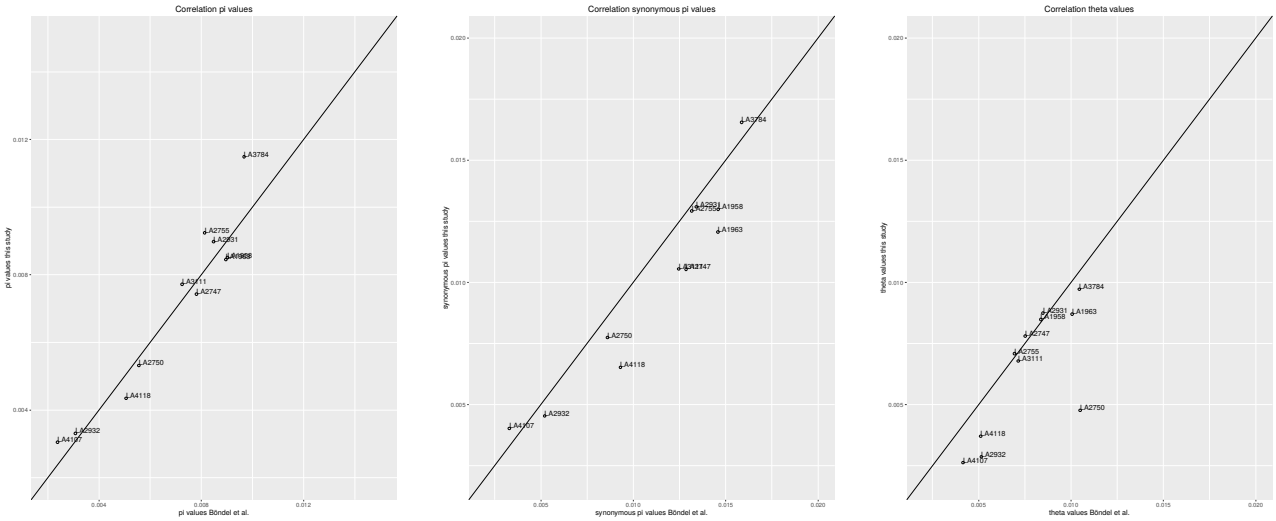


Figure 04 – Figure Supplement 02
Calculated values for π (left), π_s (centre) and θ (right) from this study (y-axis), plotted against the values obtained by Böndel et al (x-axis). The studies used the same geographical populations, but different plants.

Figure 4 – supplement 03

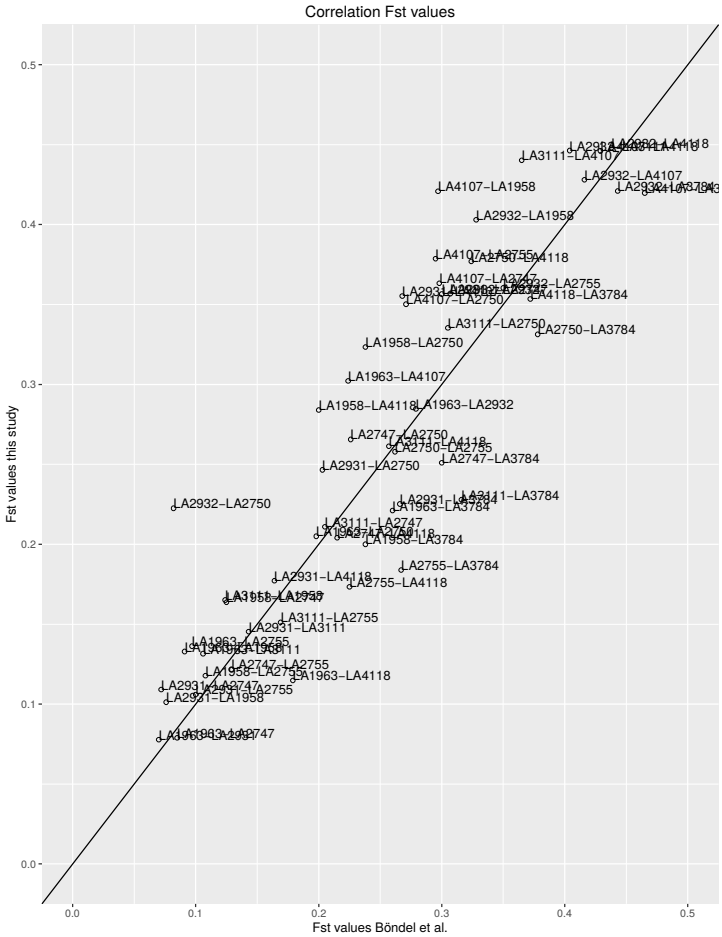


Figure 4 – supplement 03
Calculated values for Fst his study (y-axis), plotted against the values obtained by Böndel et al (x-axis). The studies used in part the same geographical populations, but different plants.

Figure 04 – Figure Supplement 04

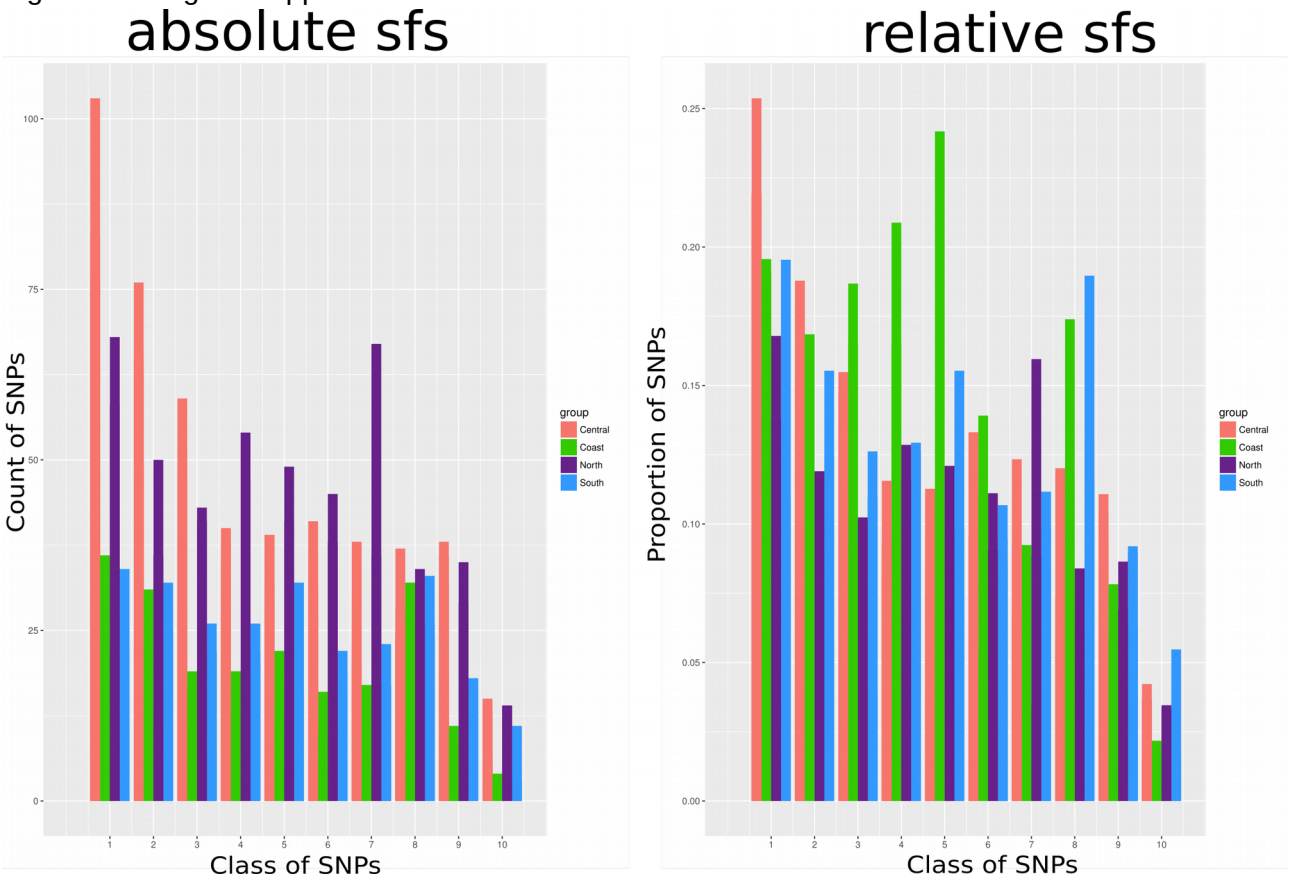
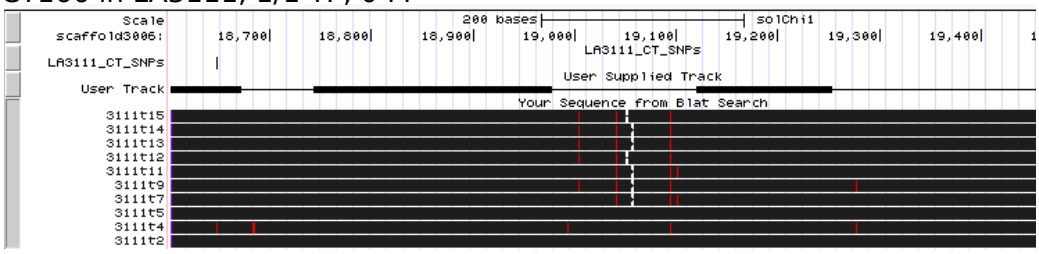


Figure 04 – Figure Supplement 04

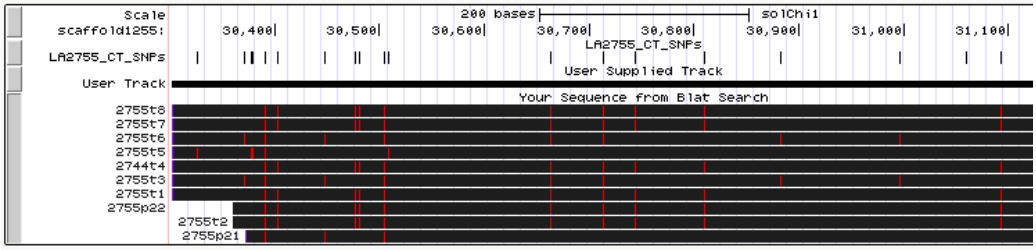
Site Frequency Sprectrum (SFS) calculated from all SNPS in all NLR samples. Left: absolute SFS, showing the total amount of SNPS observed in each group (y-axis) against the folded classes of SNPS (x-axis), Right, the relative SFS, corrected for the total amount of SNPs in each geographical group. Showing the proportion of SNPs (y-axis) against the folded classes of SNPs (x-axis). In the relative SFS, the northern and southern populations show an expected SFS, the Coastal and Southern group show a slight increase in interediate frequency SNPS, indicative of a bottleneck of stronger positive selection.

Figure 04 – Figure Supplement 05

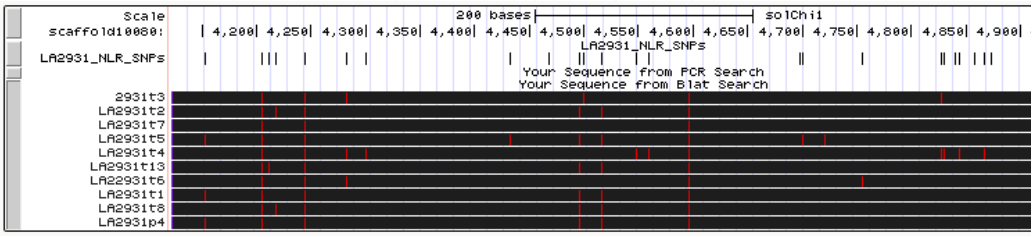
CT166 in LA3111, 1/1 TP, 0 FP



CT192 in LA2755, 18/18 TP, 2 FP



SOLCI000037500 in LA2931, 20/21 TP, 1 FN, 5 FP,



SOLCI002113000 in LA3111, 23/25 TP, 2 FN, 1 FP

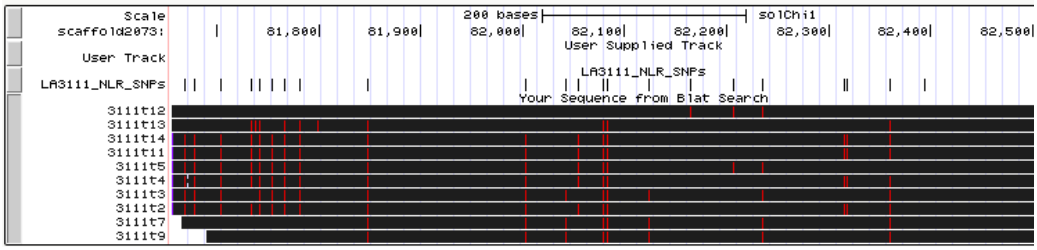


Figure 04 – Figure Supplement 05

Genome Browser Screenshots showing from top to bottom, two CT and two NLR gene fragments. Each panel shows in the top half the genomic position on the scaffolds the identified SNPs in the respective population (as vertical black dashes) and in the lower half black bars indicating the Sanger sequenced region, with the actual found SNPs relative to the reference as red vertical dashes. Each black bar represents an individual plant. TP: True Positive, FN; False Negative, FP: False Positive

Figure 05

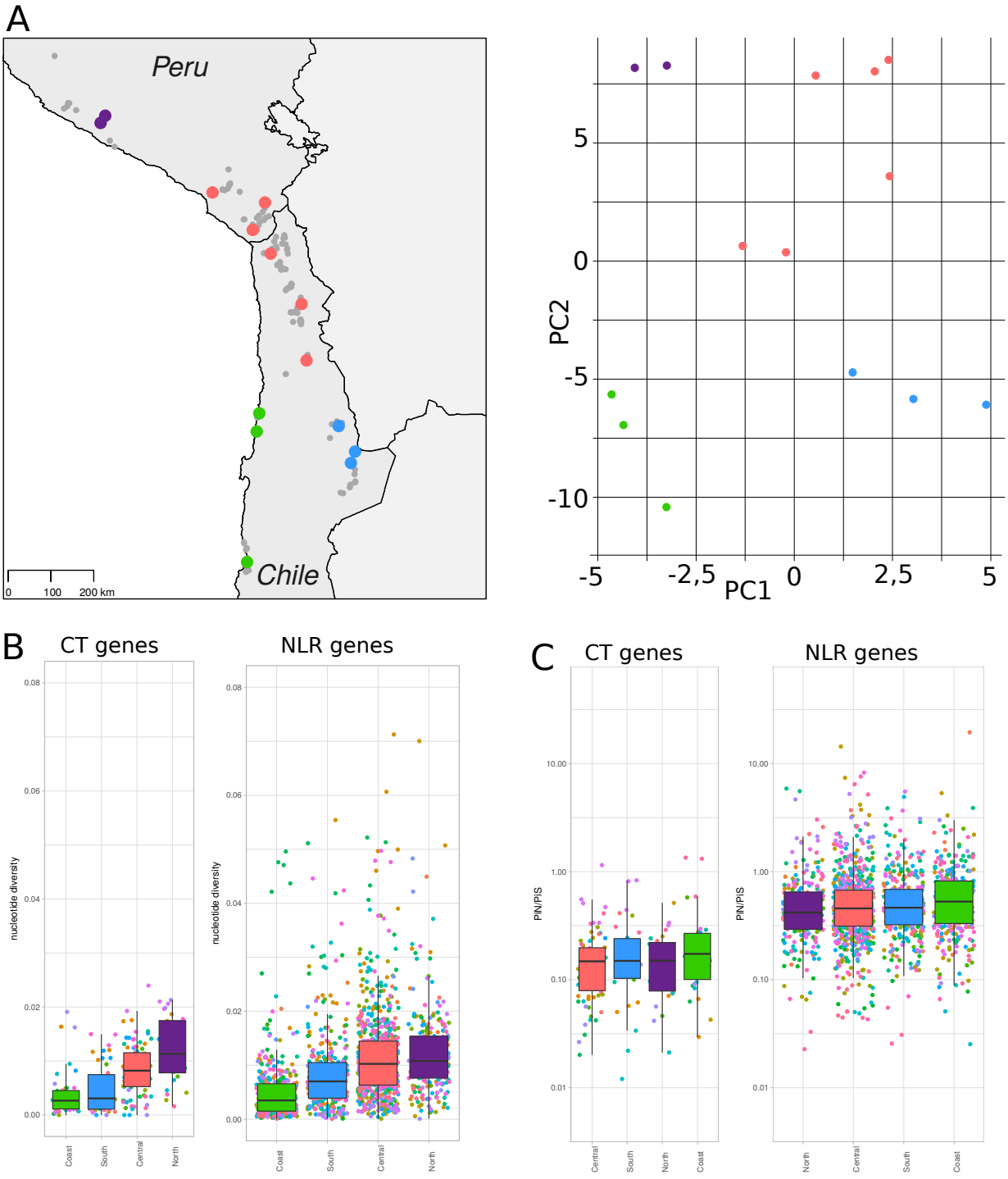


Figure 05
Species wide overview of diversity in NLR and control (CT) genes.
A) Principal Component Analysis of all SNPs in all NLR genes. First two components are shown and explain resp.
B) Nucleotide diversity and C) PiN/PiS for each gene plotted perregion. Each dot represents a single gene. Boxplots are colored as in panel A.

Figure 05 – Figure Supplement 1

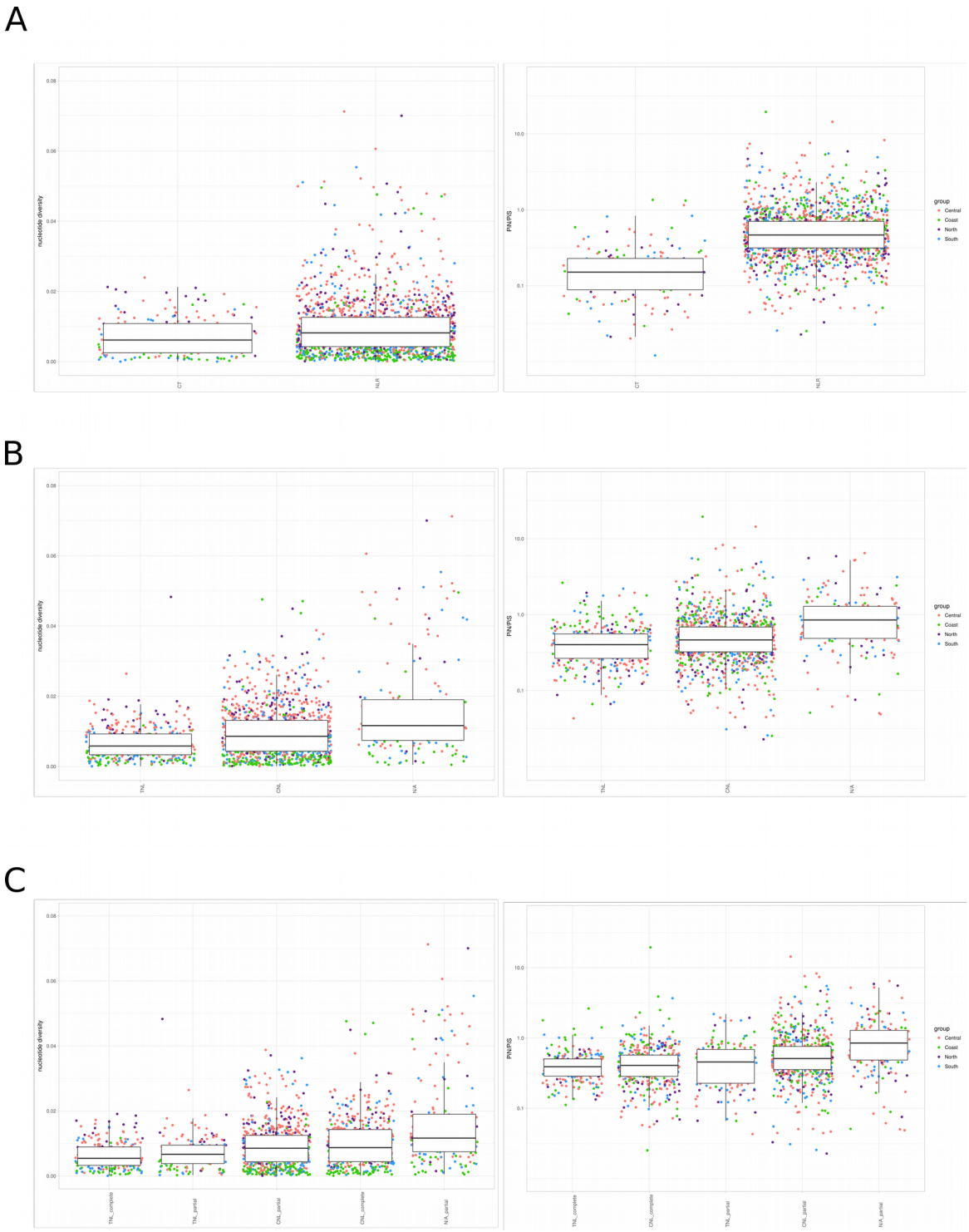


Figure 05 – Figure Supplement 1
Differences in nucleotide diversity (π) (left) and non-synonymous over synonymous diversity (π_n/ π_s) shown for different categories. A) overall values for CT genes and NLR genes, B) values for NLR genes split in CNL and TNL genes, and C) NLR genes split in CNL and TNL with an indication of their “completeness” as assessed by NLRParser. Y-axis shows the respecifive π and π_n/ π_s values, x-axis indicates the classes.

Figure 05 – Figure Supplement 2

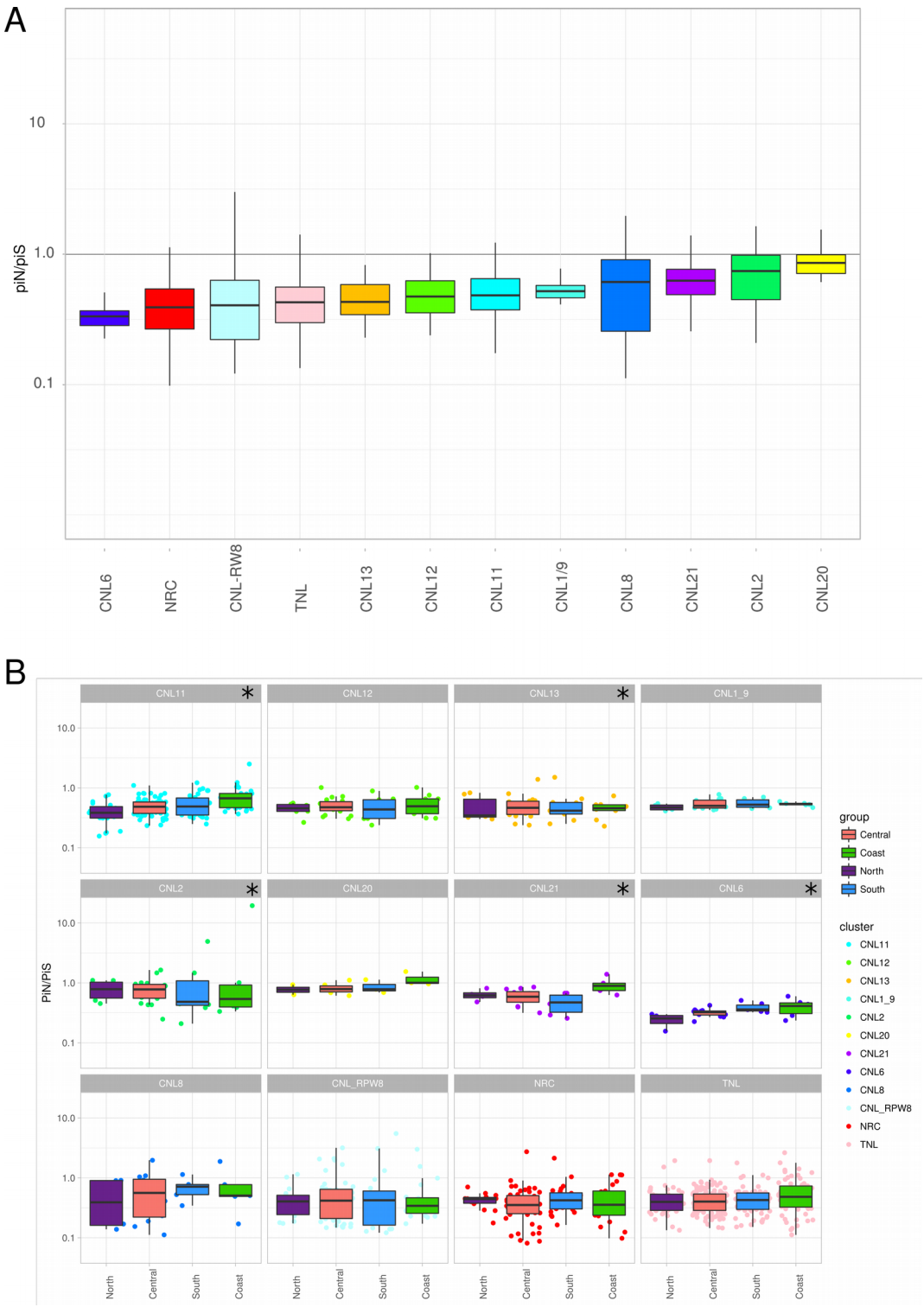


Figure 05 – Figure Supplement 2
Cluster specific evolution of NLR genes
A) Different NLR clusters, based on phylogenetic relationship, show large variation in piN/piS within the species as a whole, with three small classes having the highest variation ratio.
B) When split in different geographic regions, each NLR class shows a different pattern. Classes with significant differences between the regions are marked with *

Figure 05 – Figure Supplement 3

Figure 05 – Figure Supplement 3

Per gene overview for nucleotide diversity and the nonsynonymous over synonymous diversity ratio

Nucleotide diversity (π) (A) as well as π_N/π_S (B) ratios in individual NLR (right) does in the majority of cases not exceed the range observed in CT genes (left). However, several genes with higher values can be observed, as well as individual outliers of a gene in one or several populations. Each point represents a single gene in one population, coloured by geographic group. The boxplots show median and upper and lower quartiles and the NLRs are coloured per cluster.

Figure 05 – Figure Supplement 4

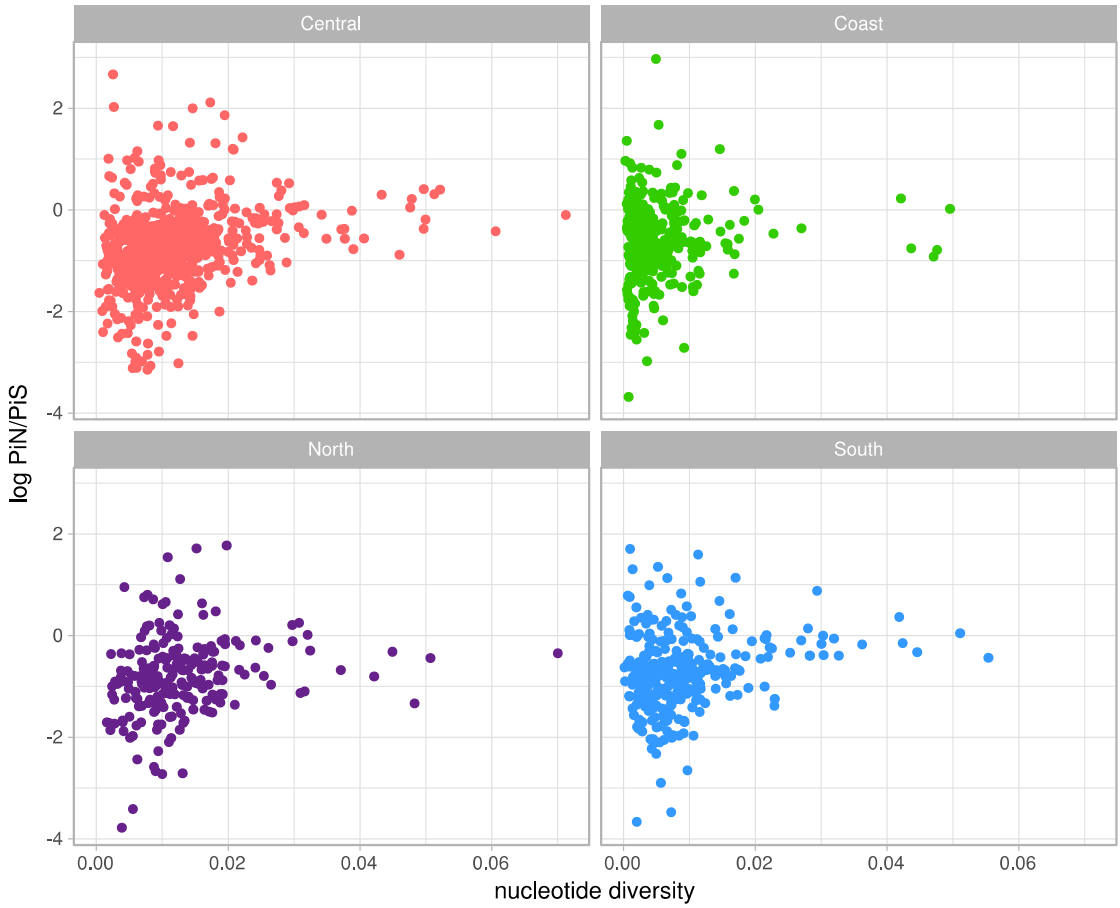


Figure 05 – Figure Supplement 4
Per gene values for π_N/π_S plotted against nucleotde diversity π. There is no clear correlation visible between these two parameters, indicating that no artefacts were introduced in the southern or coastal populations due to their bottlenecked demography.

Figure 6

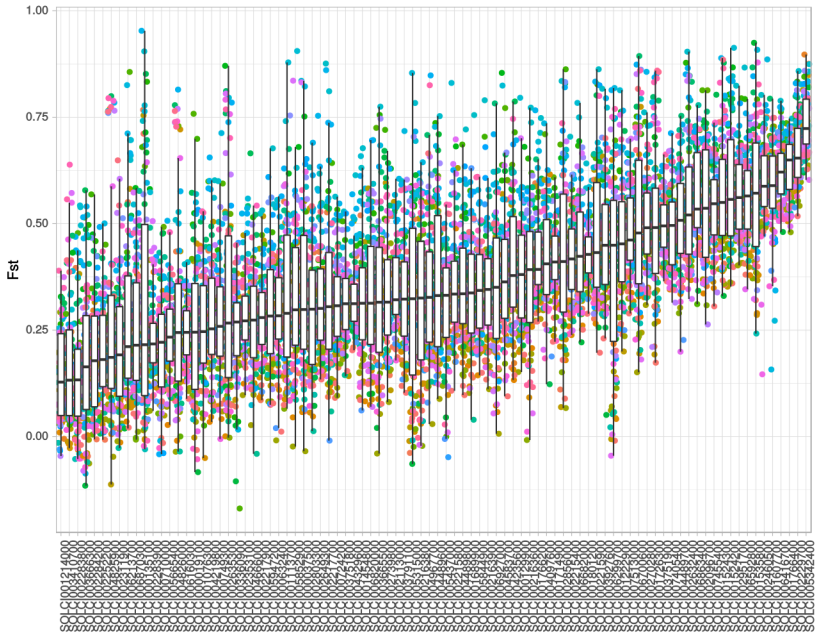


Figure 6 – Figure supplement 1

F_{ST} values (y-axis) are plotted for each NLR gene (x-axis) for each of 91 possible pairwise comparisons. Each comparison is individually coloured and overlain with boxplots to show generic trends.

Figure 6 – Figure supplement 1

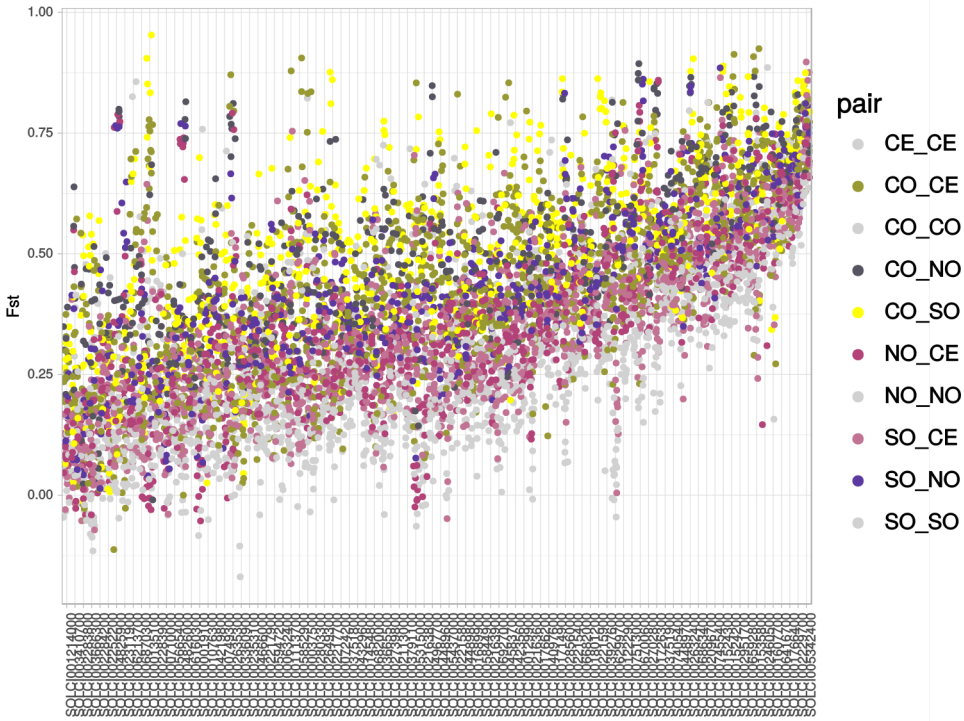


Figure 6 – Figure supplement 1

F_{ST} values (y-axis) are plotted for each NLR gene (x-axis) for each of 91 possible pairwise comparisons. Each comparison is coloured based on the genomic region of origin of each of the two populations in the pair (CE =Central, NO = North, So = South, Co = Coast).

Figure 6 – Figure supplement 2

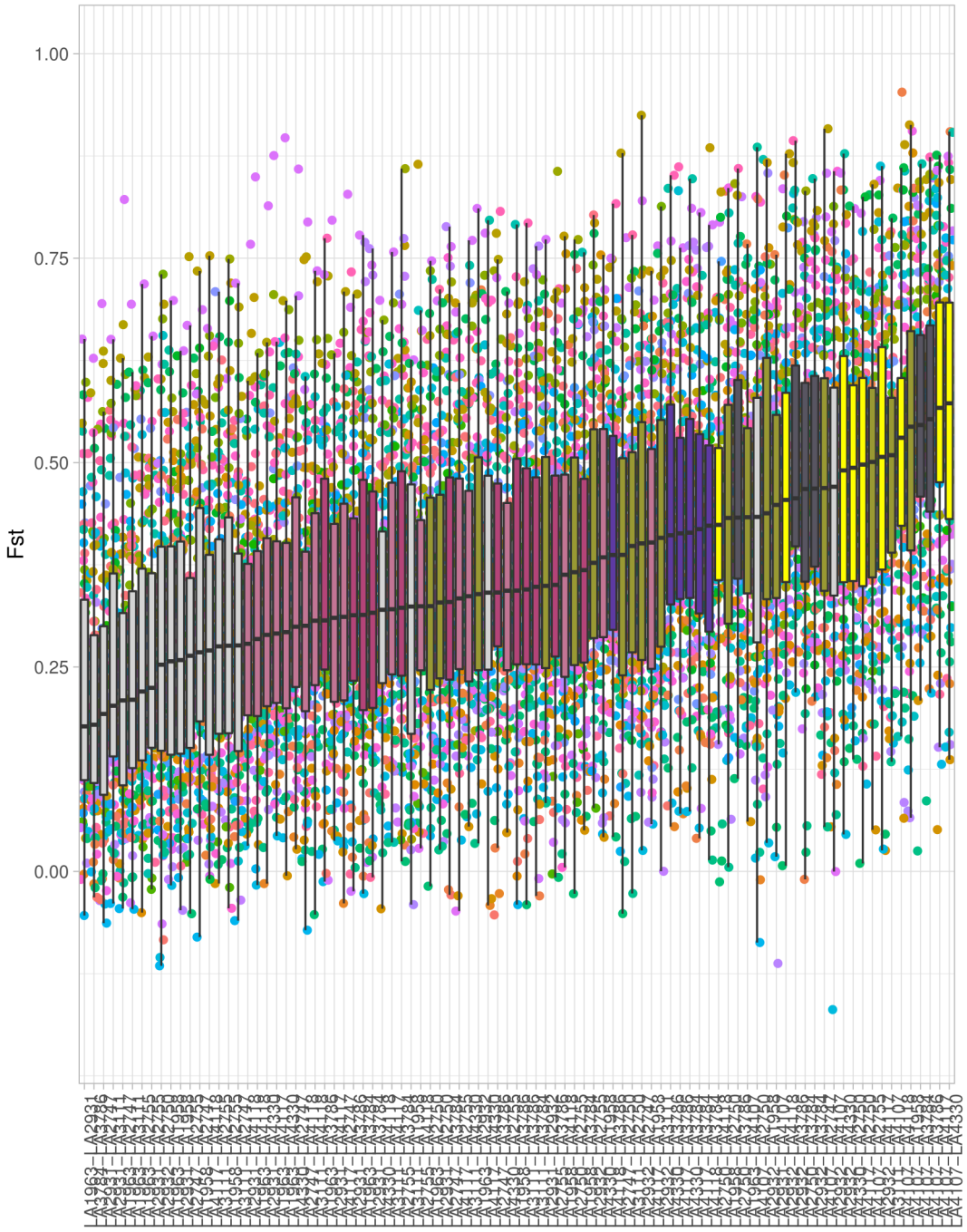


Figure 6 – Figure supplement 2

F_{ST} values (y-axis) are plotted for each pair-wise comparison of populations (x-axis) for each of the genes (individual dots). Genes are colored individually, boxplots are coloured as described in Figure 6 – supplement 1.

Figure 7

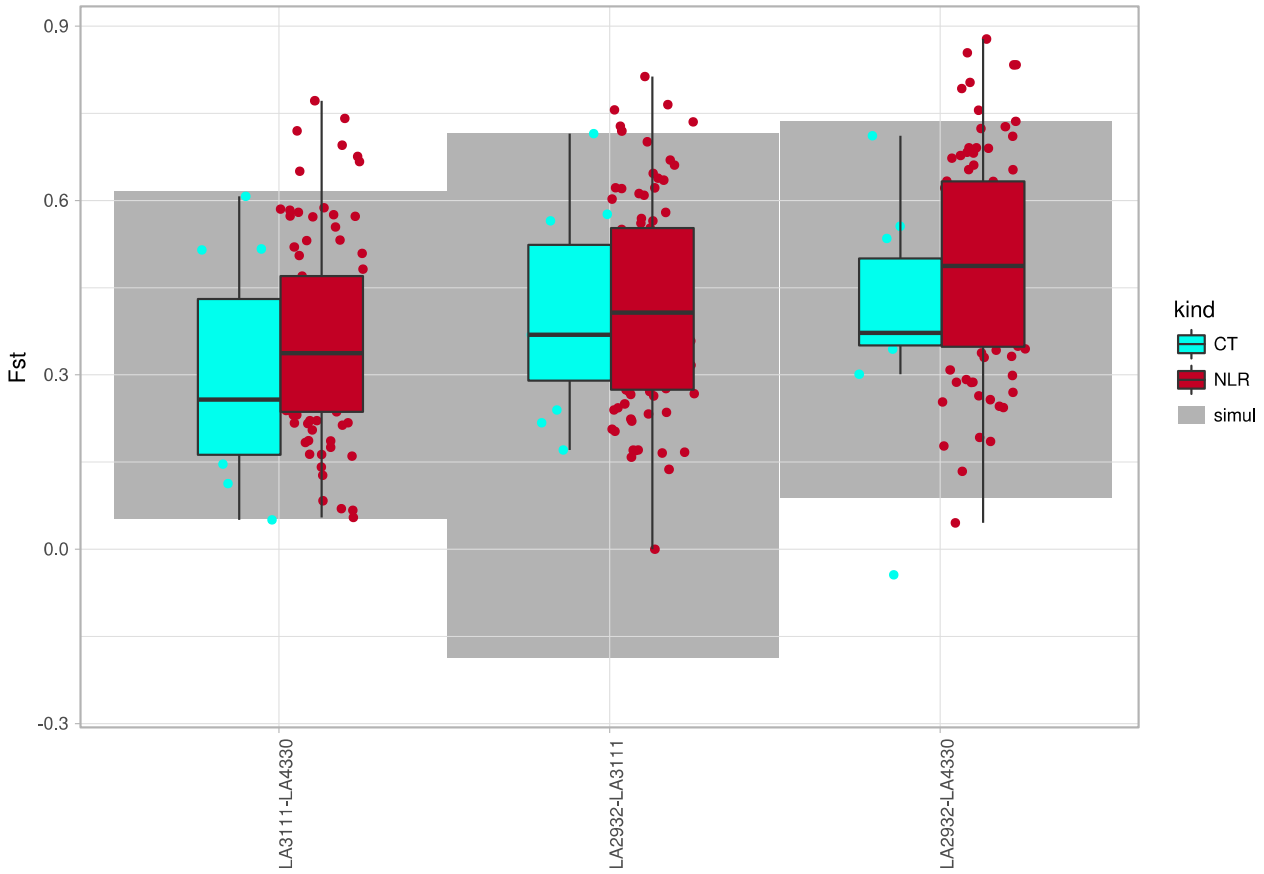


Figure 7
7% of NLR fall outside the expected F_{ST} distribution.
Grey boxes show the range of 100.000 simulations for Fixation index (F_{ST}), based on the species demography in each of the three pairwise comparisons. Blox plots show the quartiles and median for CT (blue) and NLR (red) genes. Values outside the 1st and 3rd quantile are plotted for each gene individually.

Figure 7 – Figure supplement 1

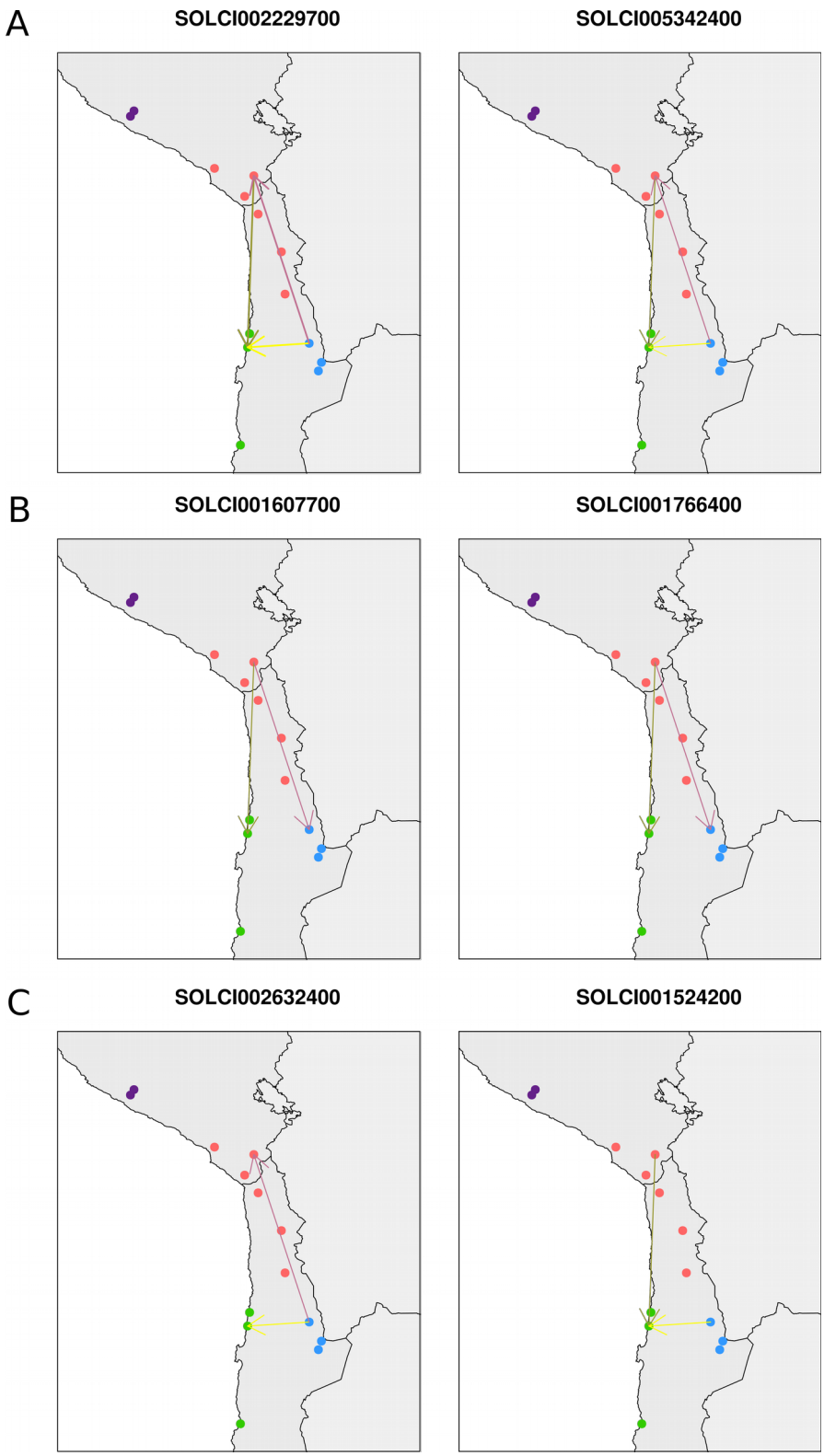


Figure 7 – Figure supplement 1
Direction of selective pressure drawn for the six NLR genes identified as outliers in more than one comparison. The direction of the arrow is based on the pairwise $\Delta\pi_N/\pi_S$ with the head being the population with the higher π_N/π_S., thus indicating the direction of positive or balancing selection, colour coding is the same as in Figure 6 – Figure supplement 1

Figure 7 – Figure supplement 2

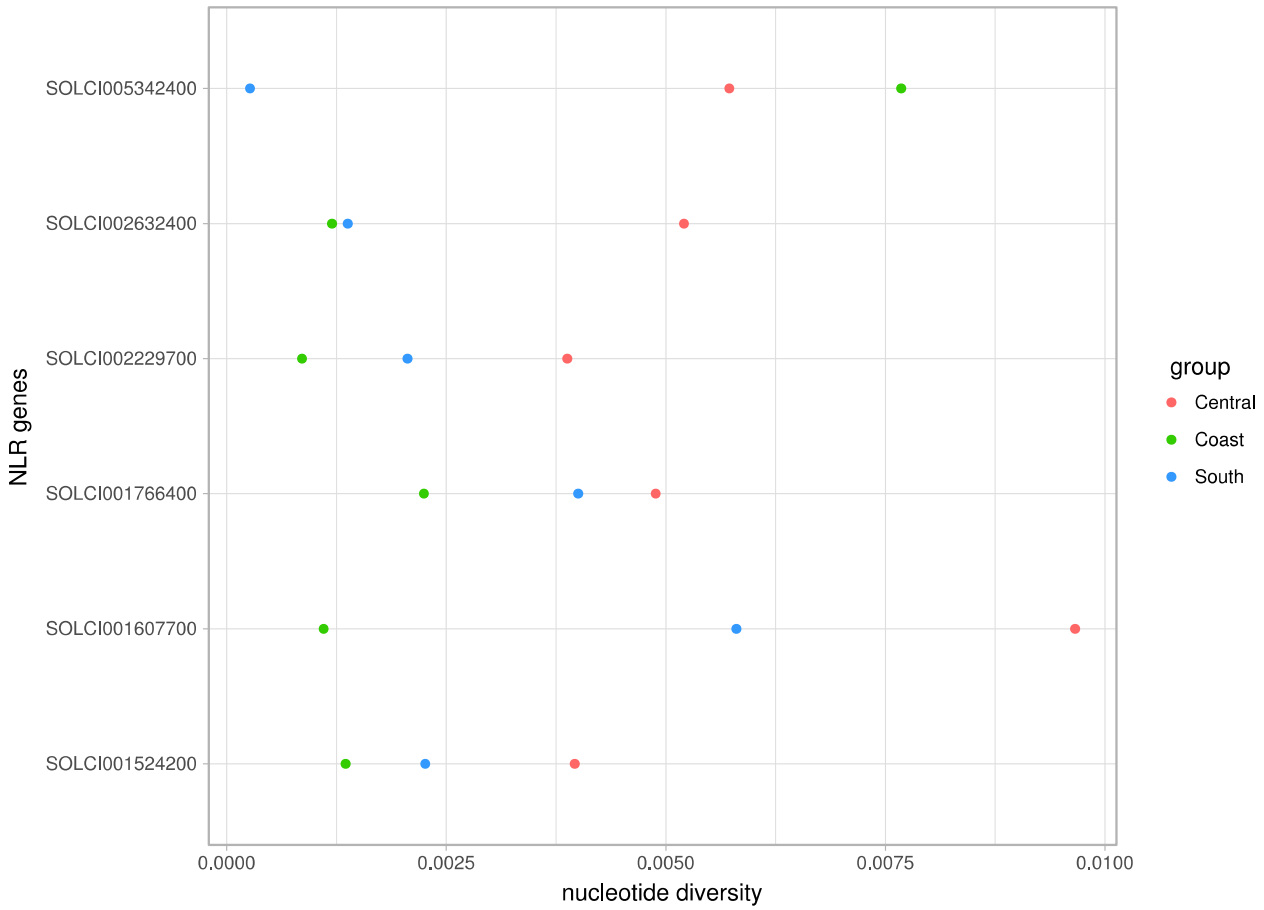


Figure 7 – Figure supplement 2

Changes in nucleotide diversity in the six highlighted gene from Figure 7. Gene names are indicated on the y-axis, Nucleotide diversity on the x-axis. The three individuals are coloured by the genetic group they represent.

Figure 7 – Figure Supplement 3

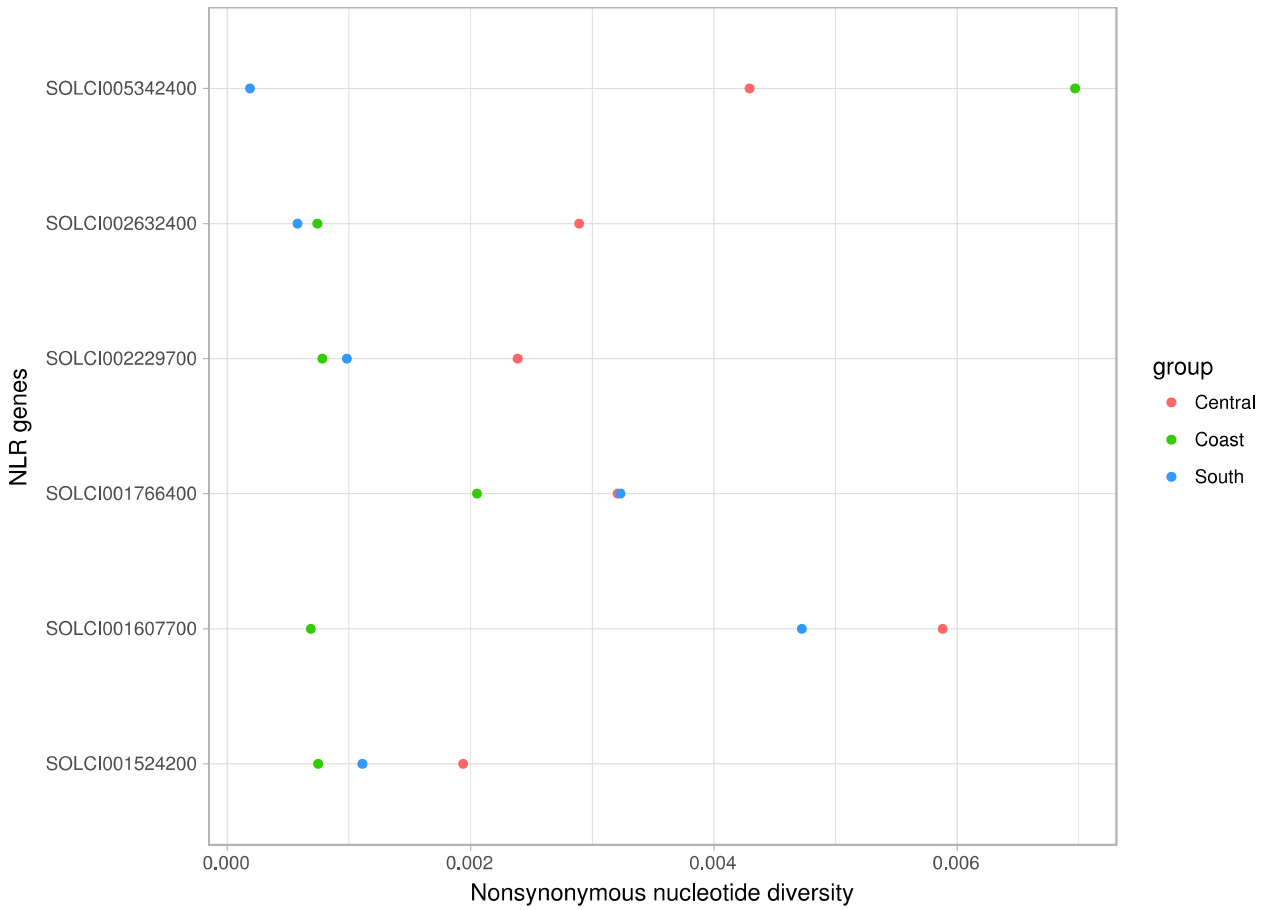


Figure 7 – Figure Supplement 3
Changes in Non-synonymous nucleotide diversity in the six highlighted genes from Figure 7. Gene names are indicated on the y-axis, Nucleotide diversity on the x-axis. The three individuals are coloured by the genetic group they represent.

Figure 8

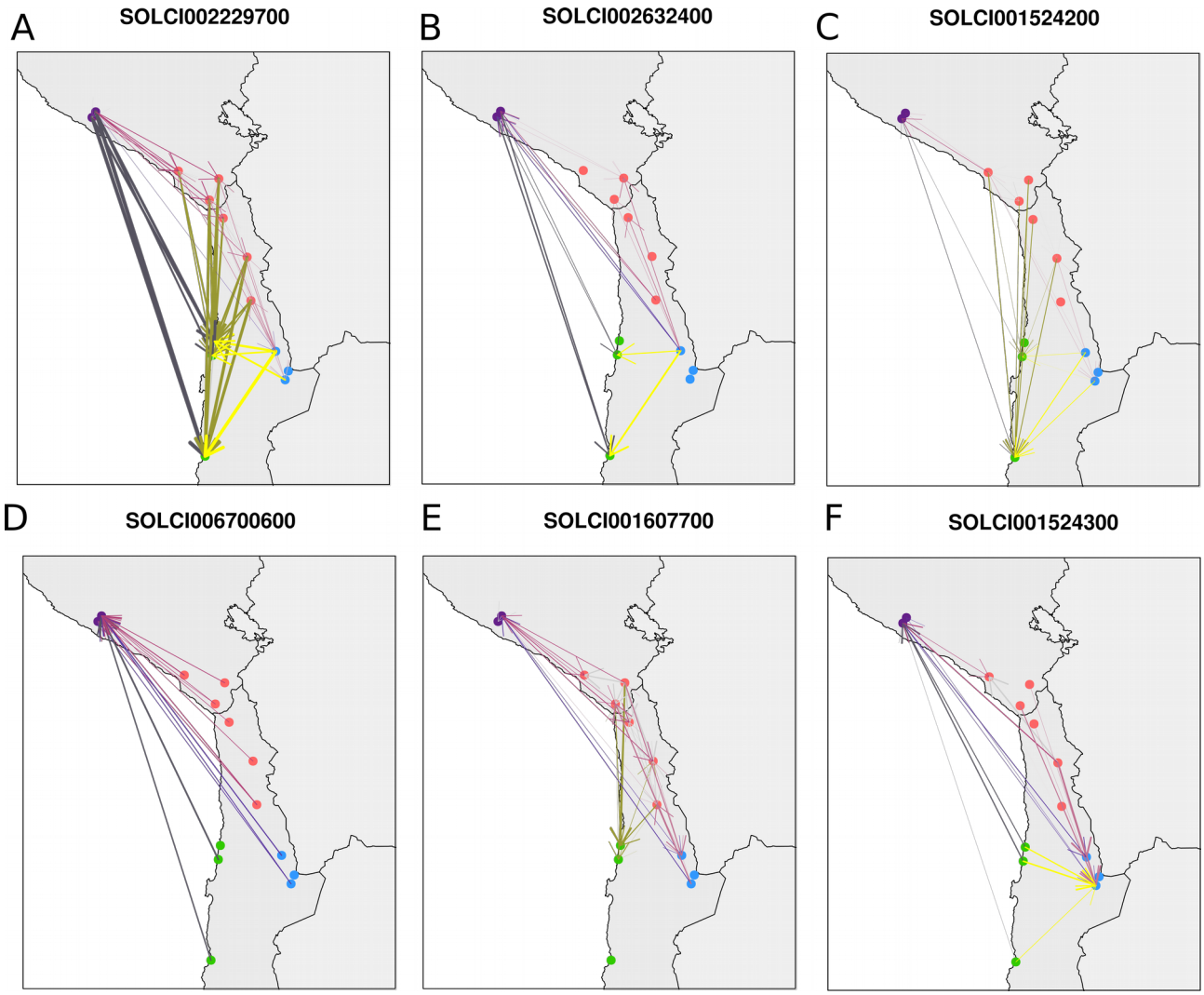


Figure 8
NLR genes show individual patterns of positive selection between populations
(A,B,C) Three examples of with selection pointing towards the coastal populations, indicating clear positive selection towards that region. D) An example of strong fixation and higher π_N/π_S S in the northern populations could indicate strong positive selection in these populations or loss of selection in the derived populations. E) More complicated mosaic selection patterns can be observed in SOLCI001607700. F) SOLCI001524300 shows selection away from the coast and central region and between the northern and southern populations.
Arrows represent those cases where F_{ST} values between two populations fall outside the expected distribution (cut-off based on Fig 7). The direction of the arrow is based on the pairwise $\Delta\pi_N/\pi_S$ with the head being the population with the higher π_N/π_S. Arrow width shows the actual value of $\Delta\pi_N/\pi_S$ in points, e.g. wider arrows have larger differences.

Figure 8 – Figure Supplement 1

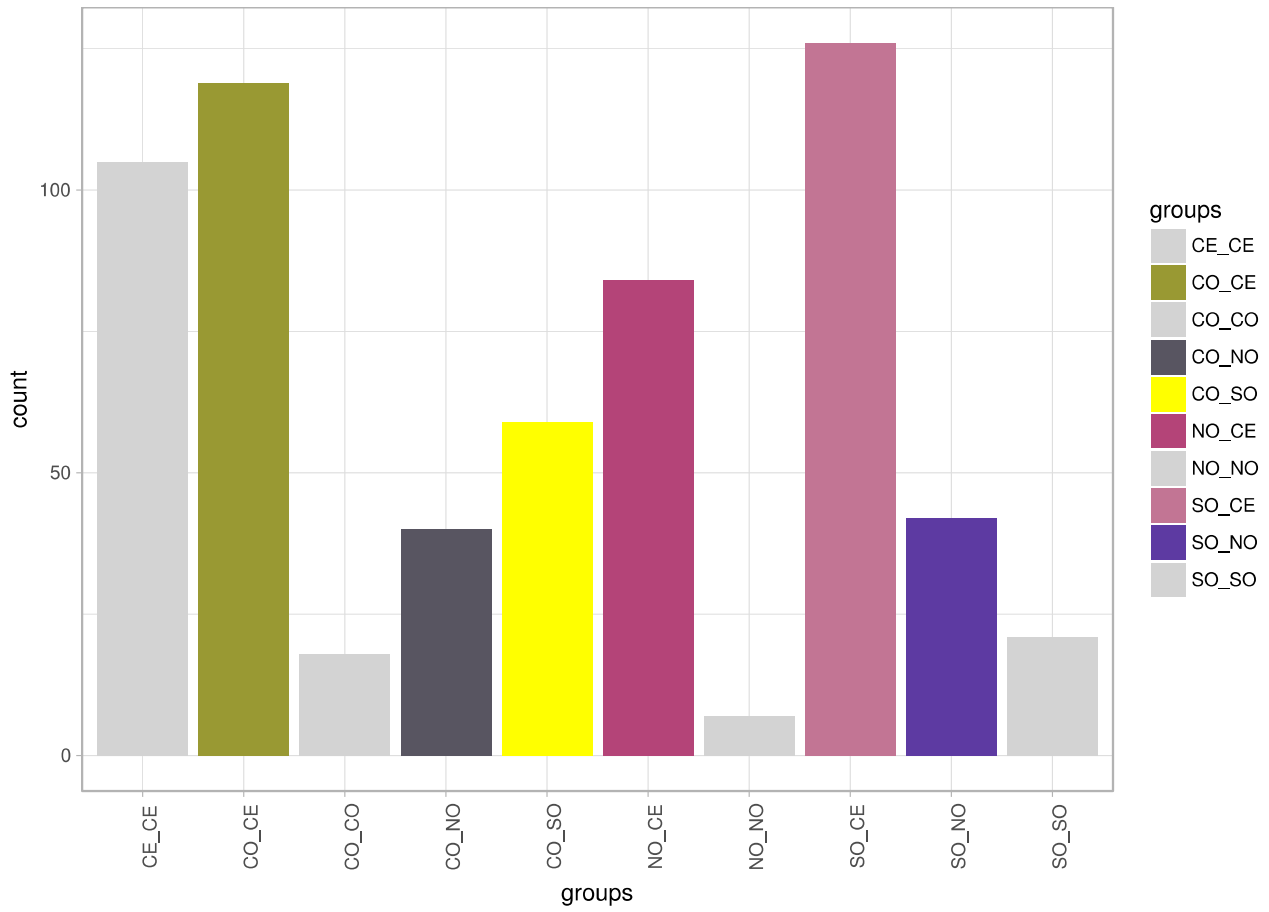


Figure 8 – Figure supplement 1
Occurrences of outlier Fst values between or within geographic regions. Pair-wise groups are indicatd on the x-axis. The y-axis shows the counts. Color codes are identical as those in Figure 6 and 7

Figure 8 – Figure supplement 2

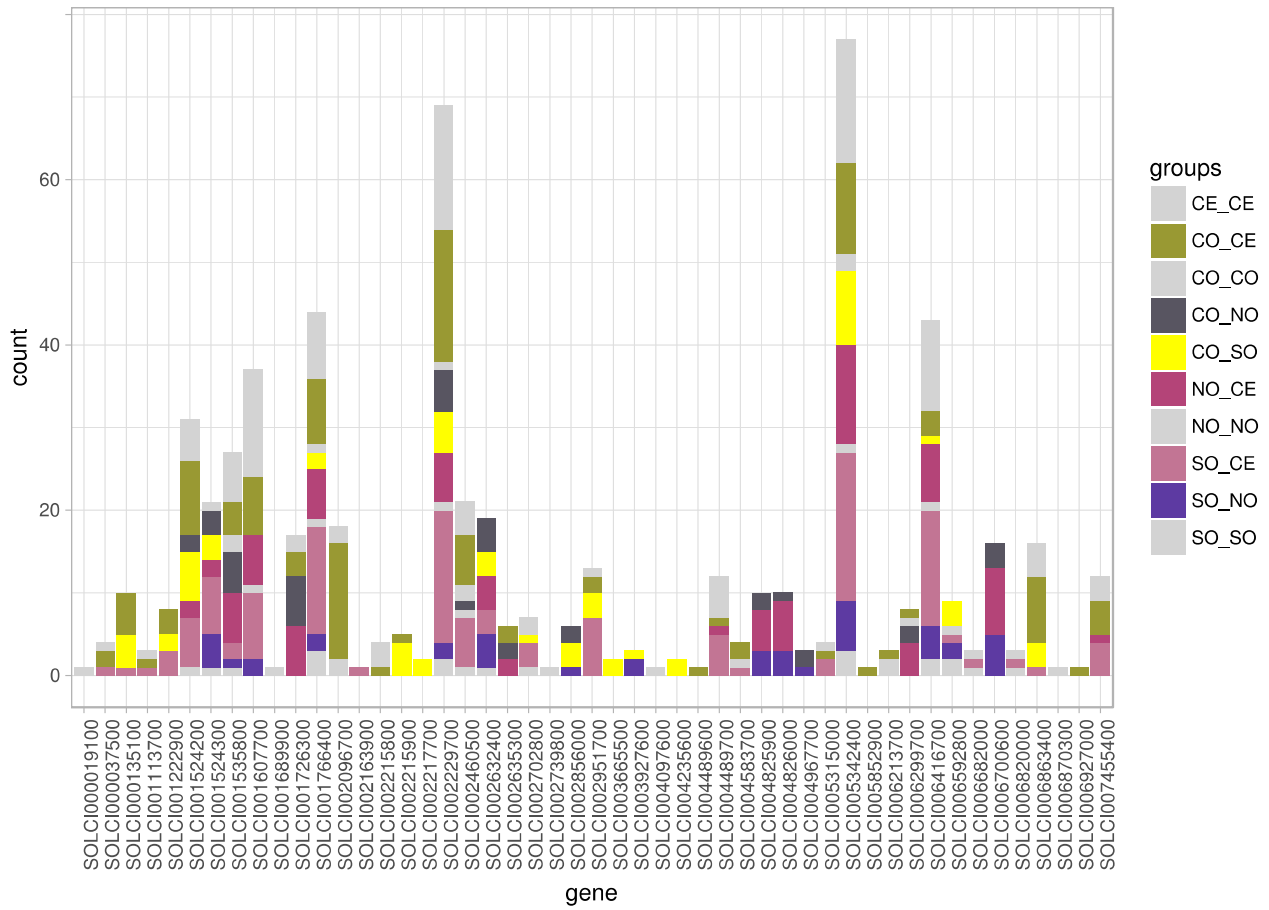


Figure 8 – Figure supplement 2
Count data (y axis) showing how often a certain NLR (x axis) is categorised as an outlier based on it's F_{ST} value. Counts are summed for each pair wise comparison between populations. Color codes are the same as in Figure 6 and 7.