# Subsets of NLR genes drive adaptation of tomato to pathogens during colonisation of new habitats

Remco Stam[1,2*], Gustavo A. Silva-Arias[2], Aurelien Tellier[2]

1 Phytopathology, Technical University Munich, Germany

2 Population Genetics, Technical University Munich, Germany

*corresponding author: stam@wzw.tum.de

**Keywords:**

Resistance genes, Evolutionary Genomics, Population Genetics, Tomato, Network Evolution

## ABSTRACT

- Nucleotide binding site, Leucine-rich repeat Receptors (NLRs), are canonical resistance (R) genes in plants, fungi and animals, functioning as central (helper) and peripheral (sensor) genes in a signalling network. We investigate NLR evolution during the colonisation of novel habitats in a model tomato species, *Solanum chilense.*
- We used R-gene enrichment sequencing (RENSeq) to obtain polymorphism data at NLRs of 140 plants sampled across 14 populations covering the whole species range. We inferred the past demographic history of habitat colonisation by resequencing whole genomes from three *S. chilense* plants from three key populations, and performing Approximate Bayesian Computation using data from the 14 populations.
- Using these parameters we simulated the genetic differentiation statistics distribution expected under neutral NLR evolution, and identified small subsets of outlier NLRs exhibiting signatures of selection across populations.
- NLRs under selection between habitats are more often helper genes, while those showing signatures of adaptation in single populations are more often sensor-NLRs. Thus, centrality in the NLR network does not constrain NLR evolvability, and new mutations in central genes in the network are key for R gene adaptation during colonisation of different habitats.

## INTRODUCTION

Antagonistic interactions can generate endless coevolution between hosts and their pathogens. The Red Queen hypothesis predicts that the genomes of both interacting partners evolve to match each other's changes (Van Valen, 1973): pathogens evolve infectivity to overcome defences, while hosts evolve pathogen recognition and resistance to avoid infection. Changes in allele frequencies of different infectivity/resistance specificities occur over tens to thousands of generations at the loci determining the outcome of interaction. Two extreme types of dynamics have been proposed, which differ in their signatures at the phenotypic (Gandon *et al.*, 2008) and genotypic polymorphism levels (Woolhouse *et al.*, 2002): the arms race model (Bergelson *et al.*, 2001) is characterised by recurrent selective sweeps in both partners, and the trench warfare model (Stahl *et al.*, 1999) shows long-lasting balancing selection. These premises form the basis for genome scans to detect the genes under coevolution/selection in hosts and pathogens (Bakker *et al.*, 2006).

Plant species exhibit a spatial distribution across habitats, which influences coevolutionary dynamics (Gandon *et al.*, 2008; Parratt *et al.*, 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 pathogens and resistance in hosts (Thrall *et al.*, 2001; Caicedo & Schaal, 2004). Species expansion and colonisation of new habitats could in addition cause the host to encounter new pathogens and subsequently promote coevolutionary dynamics at single copy genes or gene families 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 basis of host-pathogen coevolution in spatially heterogeneous populations and during the colonisation of new habitats. A crucial issue for such studies is to disentangle the signatures of selection at a few genes from the genome-wide effect of demography in shaping diversity. The problem is especially difficult in searches of genes under selection (selective sweeps or balancing selection) during adaptation to new habitats, because colonisation events generate bottlenecks resulting in an increase of the variance of the measured nucleotide diversity over the genome (e.g. in *Arabidopsis thaliana* (Lee *et al.*, 2017; Exposito-Alonso *et al.*, 2018) and *Arabis alpina* (Laenen *et al.*, 2018)).

Resistance genes are the key players in host - pathogen interactions, as they are sensing pathogen molecules to activate immune responses. Canonical R genes are members of the NLR family NLR (nucleotide binding site, leucine-rich repeat containing receptor) that occurs in both plants and animals (Jones *et al.*, 2016). NLRs have a modular structure. NLRs can have a N-terminal TIR-domain (TNLs) or CC-

domain (CNLs), followed by a Nucleotide Binding Site and leucine rich repeats). In *A. thaliana*, some NLRs appear to show signatures of positive or balancing selection (Bakker *et al.*, 2006) and overall NLRs seem to show more positive selection than other defence-related gene families (Mondragón-Palomino *et al.*, 2017). Yet, detailed studies of NLR evolution in wild pathosystems are lacking. In most cases only few candidate NLR genes have been studied. For example, in the common bean (*Phaseolus vulgaris*) the NLR locus PRLJI1 shows slightly higher overall $F_{ST}$ and markedly different patterns of spatial differentiation within and between populations compared to the genome-wide average (AFLP markers) (De Meaux *et al.*, 2003). In wild emmer wheat (*Triticum dicoccum)* 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 the number of NLRs can differ dramatically between species suggesting that the NLR family experiences a rapid birth-and-death process (Michelmore & Meyers, 1998) driven by large scale gene duplication and deletion, whereas within species variation is hypothesised to be mainly found at the nucleotide level at a few key genes (Wu *et al.*, 2017b) or at a few duplicated genes (Hörger *et al.*, 2012). The evolutionary mechanism explaining the latter is termed as the recycling of existing NLRs (Holub, 2001) by generation of new specificity at a given locus entering the host-pathogen coevolutionary process.

These theoretical expectations are based on the evolution of NLRs as single genes “sensing” the presence of pathogens (either directly or indirectly, Kourelis & Hoorn, 2018). It has now been found that NLRs form a complex multi-layer signalling network (Wu *et al.*, 2018) to recognise pathogens and transduct the signal into the appropriate defence response. A major recent finding is that members of the NRC (NLR required for cell death) clade are for example central in the network and are required as “helpers” for the functioning of other, “sensor” NLRs (Wu *et al.*, 2017a, 2018). The sensor NLRs are more peripheral in the network and have less connectivity to other genes. Expanding on the previous questions, we want to investigate if all NLRs in the network have the same evolutionary potential when colonising new habitats and encountering new pathogens.

We designed our study to address the following questions. How many NLR genes are involved in coevolution with pathogens across populations? What is the time scale of coevolution in newly colonised habitats and which genes are involved? We are particularly interested in finding how many genes exhibit different selection pressures between the original and the newly colonised habitat, namely genes evolving neutrally in the original habitat and being under (positive or balancing) selection in the derived one. Finally, we also want


3

to know whether there are differences in evolutionary changes for the various annotated NLR classes. For example, are genes central in the network showing more evolutionary constraints?

We answer these questions by studying the sequence evolution of NLR genes in a wild tomato species, *Solanum chilense*. This species is particularly amenable to this approach: it exhibits a high effective population size ($N_e$), high nucleotide diversity (heterozygosity), and high recombination rates (Arunyawat *et al.*, 2007). These features are due to outcrossing, spatial structuring of populations linked by gene flow and the presence of seed banks (Arunyawat *et al.*, 2007; Tellier *et al.*, 2011). *S. chilense* occurs in southern Peru and northern Chile. Local adaptation to abiotic and 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; Böndel *et al.*, 2018), 2) balancing selection in several genes of the *Pto* resistance pathway providing resistance to *Pseudomonas sp*. (Rose *et al.*, 2011), and 3) variable resistant phenotypes against filamentous pathogens across populations (Stam *et al.*, 2017). *S. chilense* is also an established source of fungal and viral R genes used in breeding programmes (Tabaeizadeh *et al.*, 1999; Verlaan *et al.*, 2013). *S. chilense* consists of four clearly defined geographical groups. The central group, considered the centre of origin of the species, is found in the mesic part of its range in southern Peru and northern Chile. Two southern groups likely result from two distinct southward colonisation events around the Atacama desert, one towards the coastal part of northern Chile (southern coast group), and other through high altitudes of the Chilean Andes (southern mountain group) (Böndel *et al.*, 2015). The northern group (southern Peru) was derived from the central one and is found in sympatry with its sister species *S. peruvianum*. The bottlenecks during these colonisation events have been relatively mild, so that this species still exhibits high genetic diversity (and adaptive potential) after the range expansions. The southward colonisation events provide two independent replicates of the process of adaptation to new abiotic and biotic stresses. In a recent study, we sequenced the ~915 Mb reference genome and *de novo* transcriptome of *S. chilense* (Stam *et al.*, 2019). We annotated 25,885 high confidence gene models, 71% of them are supported by transcriptome data. Our annotation yielded 236 NLRs in *S. chilense,* 201 can be considered high quality annotations, and all previously identified NLR functional clades (Jupe *et al.*, 2013; Andolfo *et al.*, 2014) can be found , albeit some with different numbers compared to other tomato species. Additionally, we identified two newly expanded clades. Overal, the *S. chilense* NLR complement looked similar to that of *S. pennellii*, a wild tomato species for which we have previously shown that NLR sequence diversity is maintained within a single population (Stam *et al.*, 2016).


4

We derive a three-pronged approach to examine the adaptation of *S. chilense* NLR genes between populations of different habitats. We re-sequence all NLRs in 14 populations for ten plants per population. Then, we infer the colonisation and demographic history based on three full genomes representative of the three major habitat groups [central (centre of origin), southern coastal and southern mountain (derived)] and use these data to infer expected NLR diversity in all fourteen populations. Lastly we combine these data to identify NLRs under different selection in the derived groups compared to the original one. We conclude by discussing the selective pressures acting on and the evolvability of the host defence network when colonising new habitats in the light of the functional classes to which the NLRs belong.

## Methods

### Plant material and accessions

We grew ten plants for each of the 14 populations of *S. chilense* in our glasshouse (20°C, 16h light). Accession numbers: LA3111, LA4330, LA2932, LA1958, LA1963, LA2747, LA2755, LA2931, LA3784, LA3786, LA2750, LA4107, LA4117(A), LA4118. (Supplementary Notes, S1)

### Pooled R gene enrichment sequencing and SNP analysis (RENSeq)

Genomic DNA was extracted from ten mature plants per population and pooled. Sequencing was done at NGS@TUM. We performed the library preparation, read mapping and SNP calling as described before (Stam *et al.*, 2016) and (Supplementary Notes S2). The NLR probes were based on known R-genes in solanaceae and *A. thaliana* and have successfully been used before (Stam *et al.*, 2016). Mapping was done using Stampy (Lunter & Goodson, 2011), SNP calling using two callers: GATK (McKenna *et al.*, 2010) and Popoolation (Kofler *et al.*, 2011). We previously found 236 NLRs in *S. chilense* and focus here on 201 high quality ones (Stam *et al.*, *2019)*. To verify the stringency of the filters and the cut-off values, we compared the merged SNP calls to Sanger sequence data for three genes for all ten plants for several populations (Supplementary data 3). After comparison, cut-offs were adjusted to obtain the best true SNP calls and both callers were run again. The combined results of the last round were used. Summary statistics $\pi$, $\theta_W$, $\pi_N$ and $\pi_S$ were calculated with SNPGenie (Nelson *et al.*, 2015) $F_{ST}$ values were calculated for pairs of populations using the Hudson et al. (Hudson *et al.*, 1992) estimator: $F_{ST} = (\pi_{between} - \pi_{within})/ \pi_{between}$. We assure robustness of the $F_{ST}$ calculations by using only 91 NLR genes with high and even coverage between all compared populations. Significant differences were tested using ANOVA, with the Tukey HSD test and recorded when $p < 10^{-5}$, unless stated otherwise.

In addition, we sequenced 14 reference loci (hereafter CT loci), used in previous studies in *S. chilense* and *S. peruvianum* (e.g. Arunyawat *et al.*, 2007; Böndel *et al.*, 2015). The CT loci summary statistics were


5

compared to the results by Böndel et al. (2015) who used an overlapping set of populations (but not the same plants). Due to known difficulty to reliably assess allele frequencies in pooled data (Futschik & Schlotterer, 2010), our analyses are based on the nucleotide diversity statistics. These seem well estimated by our SNP call procedure when we compare at the CT loci our results to a previous study (see results below).

**Full Genome Resequencing**

Accessions LA4330 and LA2932, representing southern mountain and southern coast, respectively, were sequenced at Eurofins Genomics on a Illumina HiSeq 2500 with standard library size of 300bp. We mapped the sequenced reads of the three sequenced plants (our reference genome (LA3111), representing the central region and resequence data from LA4300 and LA2932 representing the southern mountain and southern coast populations, respectively), against our *S. chilense* reference genome (Stam *et al.*, 2019) using BWA (mem, call -M with default parameters). SNPCalling was done using samtools (mpileup -q 20 -Q 20 -C 50).

**Demographic inferences with MSMC, ABC and simulation of summary statistics**

We inferred the demographic history of three *S. chilense* populations LA3111, LA2932 and LA4330 using whole genome sequence data and the MSMC method (Schiffels & Durbin, 2014). MSMC relies on long genomic fragments, thus we restrict the variant calling to the 200 largest scaffolds of the *S. chilense* reference genome: ~79.6Mb of sequence (mean length=398Kb, min=294Kb, max=1.12Mb). We estimate the past changes in effective population ($N_e$) size per population and cross-coalescence rates, assuming a per site mutation rate of $5x10^{-8}$ and generation time of 5 years. (full details on data preparation and settings are given in Supplementary Notes S3). The latter rates compare the frequency at which the most recent common ancestor is found either within individual (diploid) genomes or between two individuals of different populations, and thus indicate the time of population split. To check robustness of the inference we simulated independent scenarios of demography and divergence using ms (Hudson, 2002). We tested the demographic estimation with simulated sequences of the same length as the *S. chilense* reference genome and same estimated values of the population mutation rate (based on $\theta_W$ values) and the population recombination rate (based on $\rho$ values). We assessed the ability of MSMC to estimate the correct demographic parameters (population sizes, time of split) for simple demographic models, and a model of population splits mimicking the southward colonisation events. Using the two simulated scenarios that better resembled the MSMC estimations obtained with the observed data, we simulated sequences with the same


6

192 features as our empiric NLR dataset to obtain neutral distributions of the summary statistics (gene length
193 =2149).

194 Given that the estimates obtained with MSMC do not assume migration between populations, a feature of
195 many wild plant species which likely occurs between populations of *S. chilense* (Tellier *et al.*, 2011; Böndel *et*
196 *al.*, 2015) we additionally implemented a more comprehensive demographic inference via an Approximate
197 Bayesian Computation (ABC) approach (Beaumont *et al.*, 2002). This allows us to take into account post-
198 split gene flow between populations and test for the most likely divergence scenario (Supplementary Notes
199 S4). Three demographic models of geographic group divergence were tested to assess the order of the
200 splits. We then estimated $N_e$, divergence times and migration rates under the best supported model. The
201 data used for the ABC consist of synonimous sites of the 91 high quality NLRs and 14 CT reference loci at
202 the 14 populations. The ABC is conducted with ms (Hudson, 2002) and the R package abc (Csilléry et al.,
203 2012). From the ABC posterior parameter estimations we generated a set of neutral distributions of $F_{ST}$
204 values for all population pairwise comparisons which, based on 30,000 loci defined by the average length of
205 our NLRs and genomic population recombination rate estimated with MSMC ($4.5x10^{-9}$ – $1.1x10^{-8}$ per site per
206 generation).

207 Using forward simulations (Supplementary Notes, S5) we tested that genes under different selective
208 pressures in different populations can be revealed by outlier high $F_{ST}$ values compared to the neutral
209 expected distributions from our neutral demographic scenario. For that, we ran simulations using SLiM
210 (Haller & Messer, 2019) assuming genes evolving neutrally in all populations, and changing from neutral to
211 either positive or balancing selection in the southward colonisation processes.

212

213 **Definition of outlier NLR**

214 For each pairwise comparison between the populations, we conservatively selected the NLRs that fell
215 outside the maximum simulated value (out of 30,000 simulations). Main habitat adaptation genes were
216 defined by selecting the genes that occur as outlier in at least one third of the possible pairwise population
217 comparisons between two groups. To test whether the relative abundance of the NLR classes in the main
218 and local adaptation groupings could arise by chance, we randomised the $F_{ST}$ values within 1) the whole data
219 set, and 2) the total set of selected outlier NLRs, and subsequently reran our analyses. Using these
220 randomisation outputs we calculated the average number of major genes that can be identified (under 1,000
221 whole dataset randomisation) or the mean number of NRC genes that are classified as major genes (in
222 1,000 randomisations following procedure 2 within outliers). We estimated the confidence interval for the
223 number of major genes to be found from the random sampling (mean ±2σ). (Supplementary Notes, S6)

224

## Results

### Enrichment sequencing provides high coverage and reliable summary statistics

Polymorphism data at NLRs were obtained by targeted enrichment sequencing of pooled DNA of ten plants for each of the 14 populations (Figure 1A). For each population one to two million read pairs passed trimming and quality controls (Supplemental Data 01). For all pooled samples, the coverage exceeds 100x for 80% of the targeted NLRs. To evaluate the short-read data quality, we also enriched and sequenced the set of 14 CT genes, which showed a coverage of more than 100x in most pooled samples (S Figure 1A). We called SNPs per gene against our LA3111 reference genome (Stam *et al.*, 2019) for the 201 high quality NLRs (out of 236 identified genes) and all 14 CT genes in each population. We calculated the statistic π, summarising nucleotide diversity, and $\pi_N$ and $\pi_S$ as the nucleotide diversity for non-synonymous or synonymous sites only (Supplemental Data 2). No significant correlation was found between the number of mapped reads or bases and the number of SNPs per population ($R^2$ = 0.46 and p = 0.1) or π per population (for read pairs: $R^2$ = 0.30, p = 0.30, for bases: 0.35 and 0.2) (S Figure 1B). Thus, our data is not biased for coverage differences between the samples.

To confirm our calculations, we computed the correlations for π, $\pi_s$ and $F_{ST}$ at the CT loci between our data and a previous study, which used different plants from the same populations (Böndel *et al.*, 2015). There is a strong and significant correlation for π ($R^2$ = 0.95, p = $3.7*10^{-6}$), $\pi_S$ ($R^2$ = 0.95, p = $5.8*10^{-6}$) and pairwise $F_{ST}$ between populations ($R^2$ = 0.94, p = $2.2*10^{-16}$; S Figure 1C-D). We could finally confirm the majority of SNPs in a subset of genes using Sanger sequencing (Supplemental Data 1), demonstrating the robustness of our SNP call approach and computation of diversity statistics for our pooled data.

### NLR genes show a wide range of diversity statistics

We find between 2,748 and 7,653 SNPs within each of the 14 sequenced populations. Across the set of 201 NLRs, 63.8 (± 0.48)% of SNPs are found on average to be non-synonymous, contrary to only 34 (± 3.26)% of non-synonymous SNPs at the CT genes. PCA analyses of the NLR SNPs show that most variation can be explained using the first two principal components, which reflect the geographical locations of the populations (Figure 1B). For each group, the median π is significantly higher for NLR than for CT genes (Figure 3A, p< $10^{-5}$). The reduced π values observed for the CT and NLR genes in southern mountain and coastal populations are indicative of the demographic consequences of the colonisation events that occurred during the species expansion southwards.

The $\pi_N/\pi_S$ values for most genes remain below one, indicative of purifying selection. However, NLR genes have significantly higher $\pi_N/\pi_S$ than CT genes (Figure 2B). Such higher $\pi_N/\pi_S$ values could indicate the


8

occurrence of weak positive or balancing selection but also relaxed constraints at the NLRs. Mean and median π values are similar between CT loci and NLRs, but the variance is larger in NLRs. Six NLR genes show very large (median >0.02) π, and 15 genes show high $\pi_N/\pi_S$ (median > 1) (S Figure 2A).

To compare the signatures of selection at the short time scale (polymorphisms within a species) with those at the longer time scale of divergence (between species), we compare, respectively, the $\pi_N/\pi_S$ within LA3111 to dN/dS calculated for our reference genome (LA3111) against *S. pennellii* LA0716. The dN/dS distribution over the NLRs does not differ from that at the CT loci (Böndel *et al.*, 2015) (t-test p = 0.17), nor when comparing between the different functional NLR clades (ANOVA, p = 0.6) (S Figure 3A). In the CT genes, which all have orthologs in *S. pennellii*, $\pi_N/\pi_S$ values are correlated to dN/dS (corr 0.65, p-value 0.02). This correlation is weaker at the NLRs for which orthologs in *S. pennellii* can be found (corr 0.33, p-value 0.004). Moreover, $\pi_N/\pi_S$ is significantly higher in NLRs which do not have any ortholog in *S. pennellii* than for the other NLRs (p-value = 0.003, S Figure 3B).

**Spatially heterogeneous selection pressure acting on different NLR functional classes.**

When NLRs are grouped by functional clades, we see that the CNL6 and NRC show very low $\pi_N/\pi_S$ and the newly identified clades (CNL20 and CNL21) show high values (S Figure 4A). Interestingly, contrasting patterns appear between the geographical groups (S Figure 4B). CNL11 shows the highest $\pi_N/\pi_S$ values in the coastal populations, whereas these values are lowest for CNL2 at the coast. Genes with $\pi_N/\pi_S > 1$ differ between groups and populations, indicating that genes of the functional NLR clades are under different evolutionary pressures in the different geographical regions. NRCs appear quite conserved at both the phylogenetic time scale (between species) (low median dN/dS ratio for the LA3111 genome compared to *S. pennellii*, S Figure 3C) and at the polymorphism time scale (within species) (low median $\pi_N/\pi_S$, S Figure 4B). We calculated the fixation index ($F_{ST}$) based on π for each gene between each pair of populations (Supplemental data 3). $F_{ST}$ can be interpreted as a measure of genetic differentiation. We assure robustness of the calculations by using only 91 NLR genes with high and even coverage between all compared populations. Median $F_{ST}$ values per NLR gene range between 0.12 and 0.7, with 17 genes having a median $F_{ST}$ over 0.5 (S Figure 5). As expected, $F_{ST}$ is lowest within geographic groups and highest between the coastal and the southern mountain populations (Figure 3).

**Genome-wide inference of the species' past demographic history**

We inferred the demographic history of three *S. chilense* populations LA3111 (central), LA2932 (southern coast) and LA4330 (southern mountain) using whole genome sequence data. We find consistent population expansion events for the three populations between 50 to 500 thousand years ago before reaching current


9

290 $N_e$, with a stronger expansion for the central group than for the other two populations (Figure 4A).
291 Divergence estimations support that the central group is the area of origin of the species (Figure 4B). The
292 species' dispersal towards the new habitats occurred via two separate colonisation events around the
293 Atacama desert: an older split between the central and coastal populations 0.2 to 1 million years ago, and a
294 more recent divergence between the central and southern mountain populations, 30 to 150 thousand years
295 ago. Note that the bottlenecks towards the south are relatively mild as $N_e$ remains above $10^4$.

296 We tested the power to estimate known demographic histories with our genomic data. We simulated two
297 (single-population) demographic scenarios for each population: one with a constant population size and one
298 with a recent bottleneck event (S Figure 6B-C). Subsequently, we simulated two plausible scenarios derived
299 from the interpretation of the observed data with both $N_e$ changes and population splits including all three
300 populations: one scenario limited to a single $N_e$ change (i.e. bottleneck), and a more complex scenario with
301 several $N_e$ changes during the divergence processes (S Figure 6D-E). MSMC estimations from the simulated
302 data verified the ability to recover known demographic parameters. We also use those simulations to
303 compare the obtained demographic estimates from the empirical data. We find that the latter simulated
304 scenario showed the best fitting to observed data (S Figure 6E).

305 To confirm that under the inferred demography of *S. chilense*, $F_{ST}$ statistics can be used as indicators of
306 different selective pressures between populations we used forward simulations to generate polymorphism
307 signatures of genes either under neutral, positive or balancing selection between populations during the
308 population divergence with mild bottlenecks. Genes under positive or balancing selection in the southern
309 populations can be differentiated from the neutral genes showing high value outliers in population pairwise
310 $F_{ST}$, in spite of the mild bottleneck effect that increases variance in $F_{ST}$ distributions (Figure 4C). Studying low
311 $F_{ST}$ values for evidence of genes with similar selection pressures across populations is not powerful enough
312 given our demographic history. We thus concentrate on high $F_{ST}$ outliers between populations, which
313 indicates novel and heterogeneous selective pressures (positive, balancing or relaxed constrains) in the
314 derived populations (Charlesworth *et al.*, 1997).

315

316 **Defining $F_{ST}$ cut-off values in a species-wide population structure**

317 To study selection at NLRs over the whole species range (*e.g.* in our 14 populations which includes also a
318 northern group of two populations) we additionally infer the past demographic history taking into account
319 post-divergence migration by means of an ABC approach. We tested three models that include different
320 scenarios for the divergence, while accounting for migration (Figure 5A). As observed summary statistics we

use data at synonymous sites from all 91 NLRs and 14 CT loci to compute π per population and all pairwise $F_{ST}$. In concordance with the results obtained with the whole-genome approach, the inference from ABC confirmed that the divergence of coast and mountain from the central group were two independent processes (Model 1; Supplementary Notes S4 - Figure1). This model showed strong support in five out of six rejection analyses performed (two rejection methods x three threshold values for simulations retained; Model 1 (Supplementary Notes S4 - Table1). As expected, posterior paramenter estimations showed higher $N_e$ for populations from the central region. Lowest $N_e$ value was estimated for south coastal populations. We also estimated higher gene-flow within the central group as well as among populations from the south mountain with the central group (Supplementary Notes S4, Table2).

We used the posterior distributions of the parameters based on the best supported model to simulate the pairwise $F_{ST}$ between 14 populations for 30,000 loci (approximately the number of genes in the genome) with a mean gene size equal to that of our NLRs. Our inference yields a good fit to the observed values (S Figure 7). Especially when using the ABC, we were able to simulate median values that are very close to those of the observed data (Figure 5B). We can thus use the maximum of the simulated values as a conservative cut-off for $F_{ST}$ based outlier detection.

**Revealing genes under selection as outlier loci: specific subgroups of NLRs evolve in each habitat**

We identify the outlier NLRs as those whose $F_{ST}$ values are found outside the simulated ranges for each pairwise population comparison (shown in Figure 6B for three population comparisons). We find a median of 7 NLRs to be outliers in all pairwise comparisons. In total 52 NLRs are found as outliers in at least one of the 91 possible pairwise comparisons (S Figure 8A). How often a gene is found as outlier in a pairwise comparison differs greatly. For example, eight NLRs appear as outliers in more than 15 pairwise comparisons, whereas six are identified only once or twice (S Figure 8B). When we sum the results per geographic group, we find a similar number of genes showing signatures of selection (due to genetic differentiation) in the southern coast or the southern mountain group and slightly less between the northern and the central group (Figure 6A). We also find NLRs under selection within the central group as well as some in the southern mountain group, but not within the northern or southern coastal groups.

**Main habitat adaptation NLRs and local adaptation NLRs belong to different functional classes**

We now define NLRs that are under strong evolutionary pressure in multiple comparisons between the geographical groups as “main habitat-adaptation” genes. They are found to exhibit common selective pressure in several populations of the derived groups compared to the central group. We suggest that the


11

pressure at these genes is shaped by global changes in habitat and/or pathogens during the early phase of colonisation. We analysed all pairwise $F_{ST}$ comparisons between populations and find that 17 of the 52 outlier NLRs are main habitat-adaptation genes. These are outliers in more than 1/3 of the possible comparisons between two groups (Figure 6B). The remaining 35 genes that are under selection, only appear in few population comparisons between groups or only within a geographical group. We define these as "local adaptation" genes, presumably responsible for population level adaptation. These are clear outliers based on our neutral demographic model but do not exhibit habitat specific patterns, but rather exhibit an heterogeneous geographic mosaic of selection.

Looking at functional classes, main adaptation NLRs are more often TNLs or likely to belong to the NRC clade. The local adaptation genes are found to contain more individual NLRs that do not belong to a clade, belong to clades not part of the NRC-network, or are sensor NLRs (Figure 6C). By performing a randomisation procedure, we confirmed that the observed clade distribution of the NLRs under selection is unlikely to have arisen by chance. The observed number of major genes we find (17) is much larger than the expectation which has mean 3.5 (and C.I. [0.1-7.2]). Similarly, the observed fraction of NRC genes amongst the main habitat adaptation genes is five and larger than the expected one NRC (CI. [0.6-2.3]).

## Discussion

### NLR show sequence diversity within and between populations

NLR genes are important in plant defence responses and some have been shown to be under selection between different *Arabidopsis* species or populations (Mondragon-Palomino & Gaut, 2005; Bakker *et al.*, 2006). We used R-gene enrichment sequencing to investigate the extent of adaptation in the NLR family across wild populations of a non-model species, *Solanum chilense*.

We calculated synonymous and non-synonymous polyphormism statistics to asses possible selection on the NLRs. dN/dS ratios can be used to assess divergence of genes between species, and $\pi_N/\pi_S$ is the preferred statistic within species (Kryazhimskiy & Plotkin, 2008). High genetic diversity (observed as $\pi$ and $\pi_S$ values) is prevalent throughout the species. Between species diversity ratios (dN/dS) (*S. chilense* – *S. pennellii*) does not correlate with diversity ratios within *S. chilense* ($\pi_N/\pi_S$). This suggests recent positive selection is acting on the NLRs.

The $\pi_N/\pi_S$ ratio remains below 1 for the majority of the NLR genes in all populations, suggesting purifying selection and that the function of most NLRs is conserved within and between populations. Differences in diversity (and of the ratios) can be observed between previously defined genetic groups, with lower diversity in the derived groups. Yet, in all groups some NLRs exhibit high (non-synonymous) diversity, indicating that novel specitifties at NLRs appear and are picked up by natural selection as proposed in the NLR recylcing


12

386 scenario (Holub, 2001) . Note, *S. chilense* exhibits two previously undefined functional clades of NLRs in the
387 *Solanaceae* (CNL 20 and CNL21), indicating the importance of birth and death process generating new
388 NLRs with novel function at the phylogenic time scale (Michelmore & Meyers, 1998).

389
390 **Demographical inferences support two independent southward colinization processes in *S. chilense***

391 We implemented two demographic approaches that support a southward colonisation process already
392 proposed by Böndel et al. (2015). This ocurred via two independent events over the last 200.000 years, one
393 through the coast and the other across the highlands, resulting in two new sub-specific lineages in
394 contrasting habitats. We find some discrepancies between ABC and MSMC in the divergence time
395 estimations. These are expected given the differences in the approaches, model assumptions (*i.e.*
396 considering or not migration) and the data used (*i.e.* set of genes vs. genome-wide) (Beichman *et al.*, 2017).
397 In addition, gene exchange during divergence leads to an increment of variance of coalescence time among
398 genes (Wakeley & Hey, 1997) causing discrepancies between population divergence and gene coalescence
399 time estimations, especially for scenarios of small divergence times compared to $N_e$ (Slatkin *et al.*, 2002).
400 Even taking into account intrinsic bias to the methods used, we consider that the two demographic
401 approaches provide complementary evidences. We were able to generate a wide neutral distribution of
402 genetic differentiation which is conservative enough to avoid false positives in the identification genes under
403 selection. When using the ABC, we were able to simulate median values that are very close to those of the
404 observed data. Furthermore, we provide evidence of the good fit of the demographic estimations to our data.
405 For the MSMC method, we demonstrate using simulations that the high amount of nucleotide diversity and
406 recombination rate found in *S. chilense* (Roselius *et al.,* 2005).

407

408 **NLR change selection within and between habitats**

409 We used demographic inference to establish a neutral distribution of the genetic differentiation to define
410 outlier NLRs that change selective pressure between populations. We found that during the intra-specific
411 differentiation NLRs not only change selection between different geographical groups, but also regularly
412 between populations within the same region, especially in the Central group. Böndel et al. (2015) already
413 found the central group to be genetically more diverse and noted that it should maybe not be treated it as a
414 single panmictic unit because its relative high climatic heterogenity. We postulated that the coastal and
415 southern environments differ in their biotic factors from the central region (Stam *et al.*, 2017). In the Coastal
416 region, we expected to observe a lack of selection on NLRs as we assumed the arid environment would be
417 void of phytopathogens. Contrary to our hypothesis, our data show selection towards the coast and thus
418 indicate that pathogens are historically present. This could for example be due to seasonally running rivers

as well as a regularly occurring sea-fog phenomenon in the early morning (Cereceda & Schemenauer, 1991).

**Major and local adaptation NLR.**

Our results allowed us to separate major habitat adaptation NLRs from local adaptation NLRs. The 17 major habitat adaptation show changes of selection throughout the species' distribution, with different major genes between each geographical group. Habitat adaptation NLRs more often belong to the class of helper-NLRs, called NRC (Wu *et al.*, 2017a), as well as to the TNL. NRCs are hypothesised to be under strong purifying selection due to their central role (hub) in the NLR-signaling network (Wu *et al.*, 2018). Indeed, NRCs showed low dN/dS values and overall, $\pi_N/\pi_S$ is low in the NRCs. High fixation ($F_{ST}$) for between populations for some NRC, indicates that minor changes in individual hub proteins could also be under strong selection.

In *A. thaliana* RPW8-like NLRs, ARD1 and NRG1, function as helper NLRs for TNLs and are required for functioning of NLRs against several well studied pathogens (Brendolise *et al.*, 2018; Qi *et al.*, 2018; Castel *et al.*, 2019). In our study, RPW8-like genes are not detected as outliers. This could be explained by the fact that ARD1 and NRG1 have no clear homologues in the *Solanum* genus and thus that TNL signaling in this genus is likely to function differently, possibly with a subclade of TNLs taking over the function of hubs.

Local adaptation NLRs are more often not assigned to known functional clusters, or smaller clades, like the newly defined CNL20, suggesting that new clades are involved in local fine tuning of the defence responses, generating a geographic mosaic (Thompson, 2005) of NLR variants that have co-evolved with local pathogens. It is know that the NRC-dependent R-gene Pto (and other genes of the Pto signalling network), indeed shows such large allelic variation and is under balancing selection within different wild tomato species, including *S. chilense* (Rose *et al.*, 2007, 2011).

**Scenarios leading to two-tiered selection of NLR in new habitats**

Changes in the NRC-clade dependent defence response thus rely on co-evolution of both the sensor and the helper NLR, rather than the evolution of the sensor alone. We hypothesize that within *S. chilense* each NRC co-evolves as a helper NLR with a specific set of sensor NLRs. In experimental evolution in yeast, major evolutionary and functional novelty has been shown to occur by changes in the hubs of a gene network (Koubkova-Yu *et al.*, 2018). The main genes underlying habitat adaptation are often "helpers" and do not on their own provide a specific recognition of the newly encountered pathogens (new species or genera), but improve signalling processes. Several single non-synonymous mutations have been shown to result in gain of function of NRC1 for downstream signalling activity (Sueldo *et al.*, 2015). Moreover, NLR functioning is known to be dependent on temperature (Cheng *et al.*, 2013) and other abiotic stresses (Ariga *et al.*, 2017). In


14

*S. chilense* different NRCs could, for example be responding to different temperatures between the coast and the mountains.

In a fixed habitat, genes that are well connected in the defence gene network would be expected to be under strong functional constraints (purifying selection). Such selection has, for example been described for the NRC-independent I2 gene in *S. pimpinellifolium* (Couch *et al.*, 2006). In newly colonised habitats, selection at these genes could be resulting from two possible scenarios. 1) The new mutations at the main habitat-adaptation genes enable their binding with different and previously unbound sensors or new binding abilities under different abiotic conditions. This scenario would explain the occurrence of new positive or balancing selection at the helper genes in the derived habitats, and that different NRC genes are under selection in the three new habitats. 2) The helper genes are under relaxed constraint because the associated sensor genes are not necessary as their specific associated pathogens are absent in the new habitat. The sensor could become non-functional, so that helper genes are free to evolve neutrally or even develop novel beneficial functions (neo-functionalisation), that are selected for in subseqtent generations. In both scenarios sensors NLR can freely evolve to optimizing the detection of the newly encountered pathogens in specific localities and would coevolve rapidly with the pathogens. Together, this would lead to the observed two-tiered selection process.

## Conclusions

Our work represents a first step in studying the dynamics of NLR evolution across space and across the gene/plant defence network at the population level. Our results strengthen the view that NLRs do not evolve on their own to sense/recognise pathogen molecules, but their evolution is constrained by their interaction with other genes in the network. Future work on reliable identification of functional R genes, as well as the effectors of natural pathogens present in the different populations, will allow us to study the population genetics of direct effector-target interactions (Terauchi & Yoshida, 2010) and thus provide insight into the molecular factors shaping the different plant-pathogen coevolutionary dynamics in nature.

## Acknowledgements

RS was supported by the Alexander von Humboldt foundation. Genome resequencing was funded by DFG grant TE 809/7-1 to AT. GSA acknowledges the TUM University Foundation Fellowship. We thank Christine Wurmser (NGS@TUM) for help with the re-sequencing and Christopher Huptas and Mareike Wenning (NGS@TUM) for help with the NLR sequencing. We also thank the TGRC at UC Davis (USA) for plant material, and Anja Hörger, Tetyana Nosenko and Wolfgang Stephan for feedback on the manuscript.


15

485 **References**

**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.

**Ariga H, Katori T, Tsuchimatsu T, Hirase T, Tajima Y, Parker JE, Alcázar R, Koornneef M, Hoekenga O, Lipka AE, *et al.* 2017**. NLR locus-mediated trade-off between abiotic and biotic stress adaptation in Arabidopsis. *Nature Plants* **3**: 17072.

**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.

**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.

**Beaumont MA, Zhang WY, Balding DJ**. **2002**. Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

**Beichman AC, Phung TN, Lohmueller KE**. **2017**. Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. *G3: Genes, Genomes, Genetics* **7**: 3605–3620.

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

**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.

**Böndel KB, Nosenko T, Stephan W**. **2018**. Signatures of natural selection in abiotic stress-responsive genes of Solanum chilense. *Royal Society Open Science* **5**.

**Brendolise C, Martinez-Sanchez M, Morel A, Chen R, Dinis R, Deroles S, Peeters N, Rikkerink EHA, Montefiori M**. **2018**. NRG1-mediated recognition of HopQ1 reveals a link between PAMP- and Effector-triggered Immunity. *bioRxiv*: 293050.

**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.

**Castel B, Ngou P-M, Cevik V, Redkar A, Kim D-S, Yang Y, Ding P, Jones JDG**. **2019**. Diverse NLR immune receptors activate defence via the RPW8-NLR NRG1. *New Phytologist* **222**: 966–980.

**Cereceda P, Schemenauer RS**. **1991**. The Occurrence of Fog in Chile. *Journal of Applied Meteorology* **30**: 1097–1105.

**Charlesworth B, Nordborg M, Charlesworth D**. **1997**. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* **70**: 155–174.

**Cheng C, Gao X, Feng B, Sheen J, Shan L, He P**. **2013**. Plant immune response to pathogens differs with changing temperatures. *Nature Communications* **4**: 2530.

**Couch BC, Spangler R, Ramos C, May G**. **2006**. Pervasive purifying selection characterizes the evolution of I2 homologs. *Molecular plant-microbe interactions: MPMI* **19**: 288–303.

**Csilléry K, Francois O, Blum MGB**. **2012**. abc: an R package for Approximate Bayesian Computation (ABC). *Methods in Ecology and Evolution* **3**: 475–479.

**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.

**Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, *et al.* 2018**. The rate and potential relevance of new mutations in a colonizing plant lineage. *PLOS Genetics* **14**: e1007155.

**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.

**Futschik A, Schlotterer C**. **2010**. Massively parallel sequencing of pooled DNA samples - the next generation of molecular markers. *Genetics*.

**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.

**Haller BC, Messer PW**. **2019**. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution* **36**: 632–637.

**Holub EB**. **2001**. The arms race is ancient history in Arabidopsis, the wildflower. *Nature reviews. Genetics* **2**: 516–27.

**Hörger AC, Ilyas M, Stephan W, Tellier A, van der Hoorn RAL, Rose LE**. **2012**. Balancing selection at the tomato RCR3 Guardee gene family maintains variation in strength of pathogen defense. (R Mauricio, Ed.). *PLoS genetics* **8**: e1002813.

**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.

**Jupe F, Witek K, Verweij W, Śliwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ, *et al.* 2013**. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *The Plant Journal* **76**: 530–544.

**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.

**Koubkova-Yu TC-T, Chao J-C, Leu J-Y**. **2018**. Heterologous Hsp90 promotes phenotypic diversity through network evolution. *PLOS Biology* **16**: e2006450.

**Kourelis J, Hoorn RAL van der**. **2018**. Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. *The Plant Cell* **30**: 285–299.

**Kryazhimskiy S, Plotkin JB**. **2008**. The Population Genetics of dN/dS. *PLoS Genet* **4**: e1000304.

**Laenen B, Tedder A, Nowak MD, Toräng P, Wunder J, Wötzel S, Steige KA, Kourmpetis Y, Odong T, Drouzas AD, *et al.* 2018**. Demography and mating system shape the genome-wide impact of purifying selection in Arabis alpina. *Proceedings of the National Academy of Sciences*: 201707492.

**Lee C-R, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, Alonso-Blanco C, Weigel D, Nordborg M**. **2017**. On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nature Communications* **8**: 14458.

**Lunter G, Goodson M**. **2011**. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* **21**: 936–939.

**McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010**. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

**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.

**Mondragón-Palomino M, Stam R, John-Arputharaj A, Dresselhaus T**. **2017**. Diversification of defensins and NLRs in Arabidopsis species by different evolutionary mechanisms. *BMC Evolutionary Biology* **17**: 255.

**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.

**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.

**Parratt SR, Numminen E, Laine A-L**. **2016**. Infectious Disease Dynamics in Heterogeneous Landscapes. *Annual Review of Ecology, Evolution, and Systematics* **47**: 283–306.

**Qi T, Seong K, Thomazella DPT, Kim JR, Pham J, Seo E, Cho M-J, Schultink A, Staskawicz BJ**. **2018**. NRG1 functions downstream of EDS1 to regulate TIR-NLR-mediated plant immunity in Nicotiana benthamiana. *Proceedings of the National Academy of Sciences* **115**: E10979–E10987.

**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.

**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.

**Roselius K, Stephan W, Städler T**. **2005**. The Relationship of Nucleotide Polymorphism, Recombination Rate and Selection in Wild Tomato Species. *Genetics* **171**: 753–763.

**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.

**Slatkin M, Veuille M, Malécot G**. **2002**. Modern developments in theoretical population genetics : the legacy of Gustave Malécot.

**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, Nosenko T, Hörger AC, Stephan W, Seidel M, Kuhn JMM, Haberer G, Tellier A**. **2019**. The de novo reference genome and transcriptome assemblies of the wild tomato species Solanum chilense. *bioRxiv*: 612085.

**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.

**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.

**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.

**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.

**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.

**Tellier A, Laurent SJY, Lainer H, Pavlidis P, Stephan W**. **2011**. 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.

**Terauchi R, Yoshida K**. **2010**. Towards population genomics of effector–effector target interactions. *New Phytologist* **187**: 929–939.

**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.

**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.

**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.

**Van Valen L**. **1973**. A New Evolutionary Law. *Evolutionary Theory* **1**: 1–30.

**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.

**Wakeley J, Hey J**. **1997**. Estimating ancestral population parameters. *Genetics* **145**: 847–855.

**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.

**Wu C-H, Abd-El-Haliem A, Bozkurt TO, Belhaj K, Terauchi R, Vossen JH, Kamoun S**. **2017a**. NLR network mediates immunity to diverse plant pathogens. *Proceedings of the National Academy of Sciences*: 201702041.

**Wu C-H, Derevnina L, Kamoun S**. **2018**. Receptor networks underpin plant immunity. *Science* **360**: 1300–1301.

**Wu Q, Han T-S, Chen X, Chen J-F, Zou Y-P, Li Z-W, Xu Y-C, Guo Y-L**. **2017b**. Long-term balancing selection contributes to adaptation in Arabidopsis and its relatives. *Genome Biology* **18**: 217.

**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.


19

Figure 1

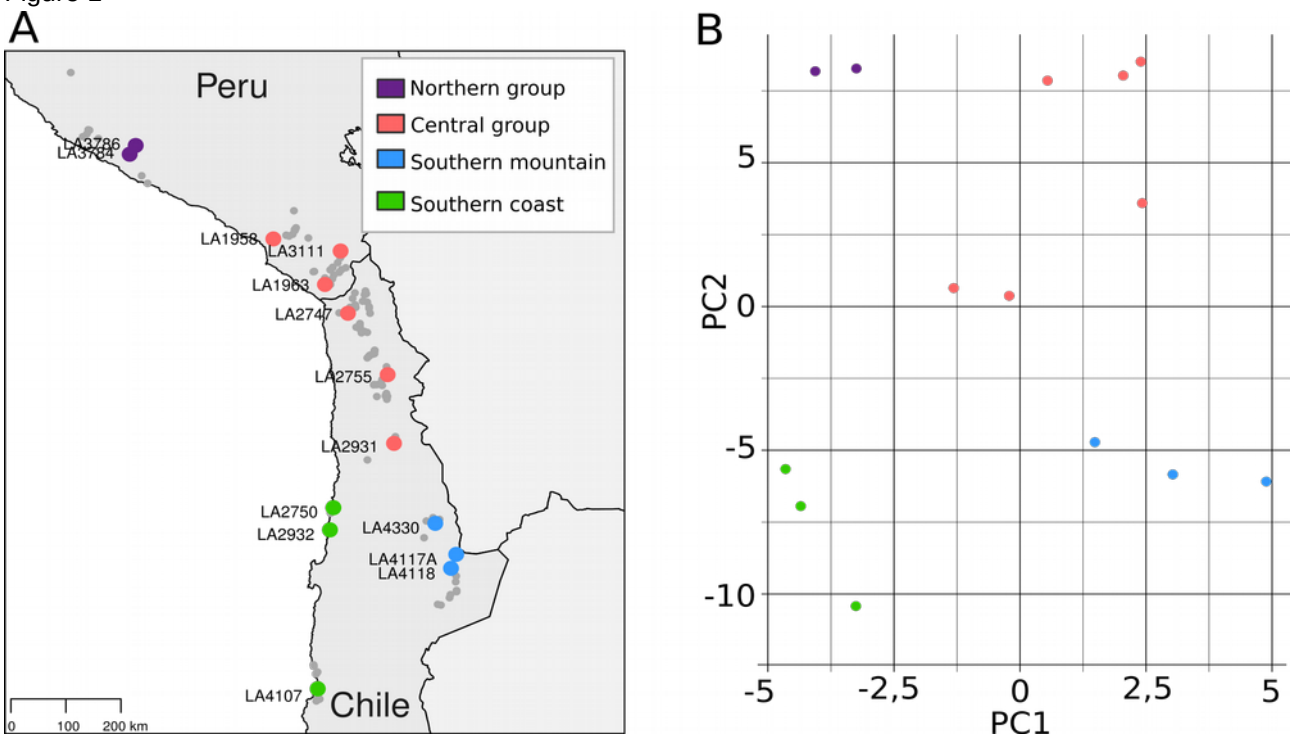


Figure 1

Overview of the studied populations and structuring of species-wide NLR diversity across the 14 populations. A) Map of the studied populations (colored by genotype group) compared to all reported *S. chilense* populations from the TGRC database, UC Davis, USA (grey dots). B) Principal component analysis of all SNPs in all sequenced NLR genes. First two components are shown and explain respectively 18 and 12% of the variance.

Figure 2

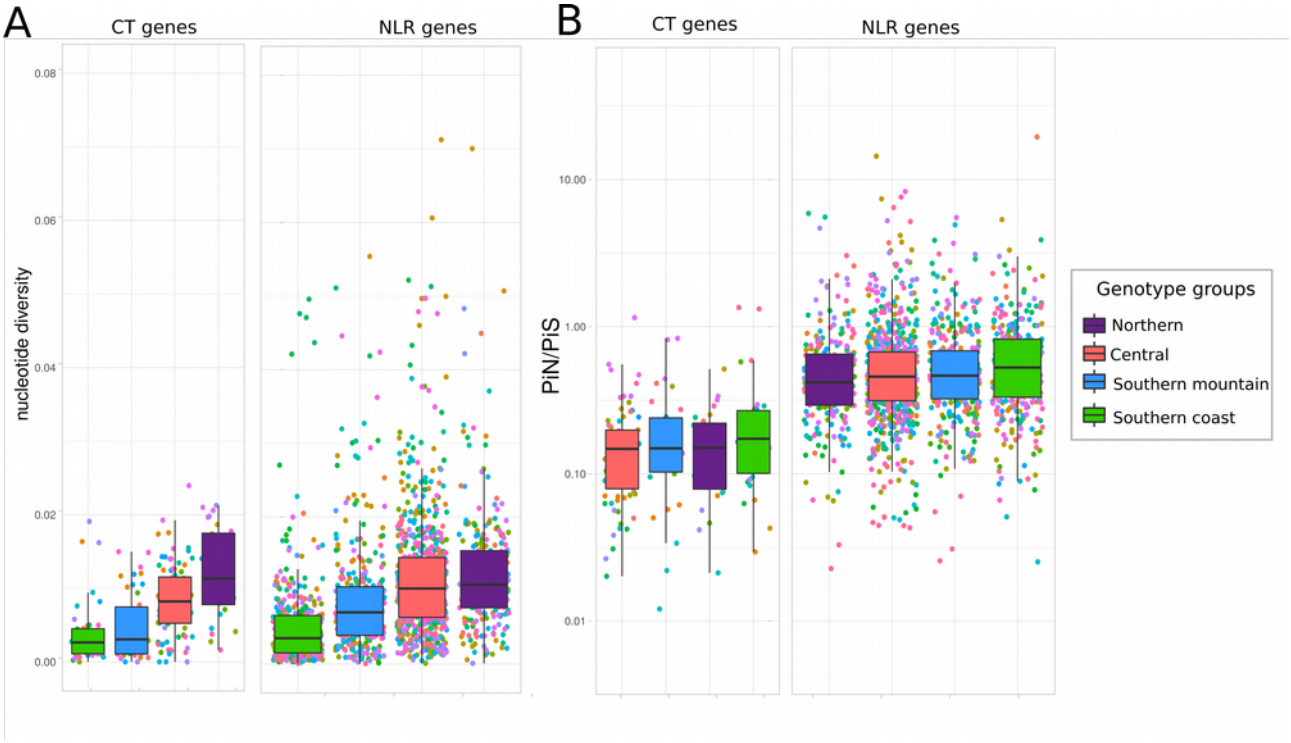


Figure 2

Population genetic statistics for NLR and CT loci

A) Nucleotide diversity ($\pi$) for each gene, plotted per geographic group. B) Non-synonymous over synonymous nucleotide diversity ($\pi_N/\pi_S$) for each gene, plotted per group. Box plot colours match those of the geographic groups on Figure 1. Each dot represents a single gene, colours are assigned randomly.

Figure 3

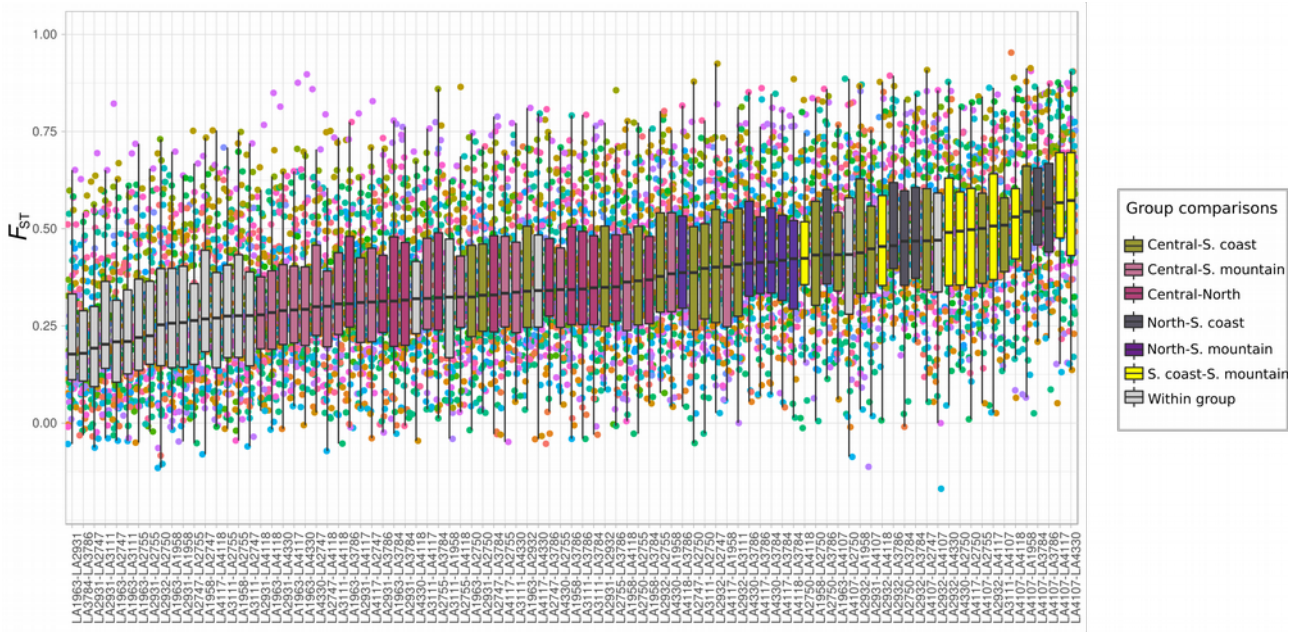


Figure 3
NLR Fixation index
Fixation index ($F_{ST,}$ y-axis), for each gene in each pairwise comparison between populations (x-axis). Colours of the boxes indicate the pairwise group comparisons. Each dot represents a single gene.

Figure 4

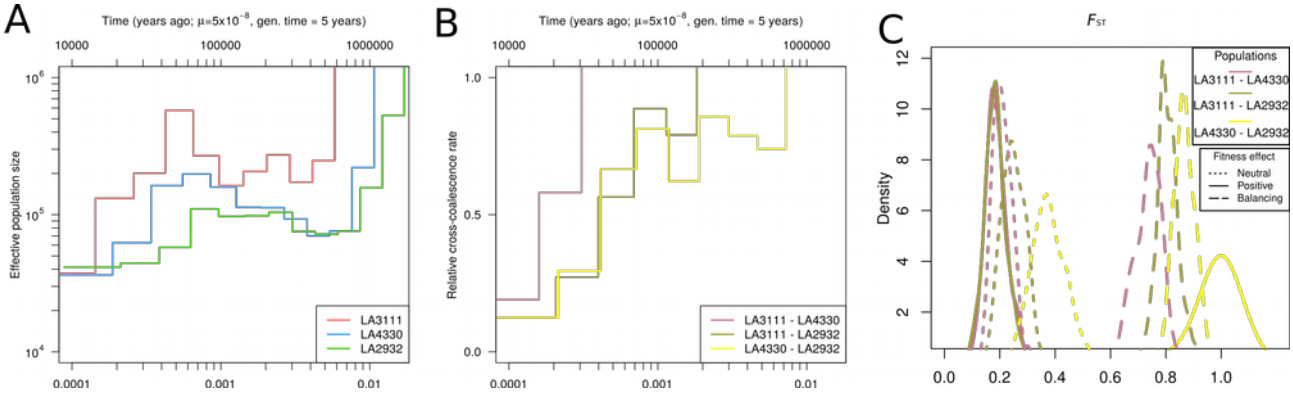


Figure 4

Historical demography reconstructions based on whole genome data of one individual from each of the central, southern coast and southern mountain populations.

A) Effective population size ($N_e$) through time estimations for central (LA3111; red line), southern coast (LA2932; green line) and southern mountain (LA4330; blue line) populations obtained with MSMC. Y-axis indicates the $N_e$, x-axis the time in years ago (top)

B) Estimation of the genetic divergence between pairs of populations through time: central-mountain (LA3111-LA4330; salmon line), central-coast (LA3111-LA2932; olive line) and mountain-coast (LA4330-LA2932; yellow line). The measures are based on the ratio between the cross-population and within-population coalescence rates (y-axis) as a function of time (x-axis). A rate of one indicates panmictic populations and rates of zero indicate fully separated populations.

C) Genetic differentiation distributions ($F_{ST}$) among the central, southern coast and southern mountain populations. Simulated genes evolve under neutrality in the central group and under either neutral, positive or balancing selection regimes in both southern populations following the colonization scenario and demography inferred with MSMC. $F_{ST}$ (x-axis)is plotted against the observed density (y-axis), The comparisons are coloured as in B.

Figure 5

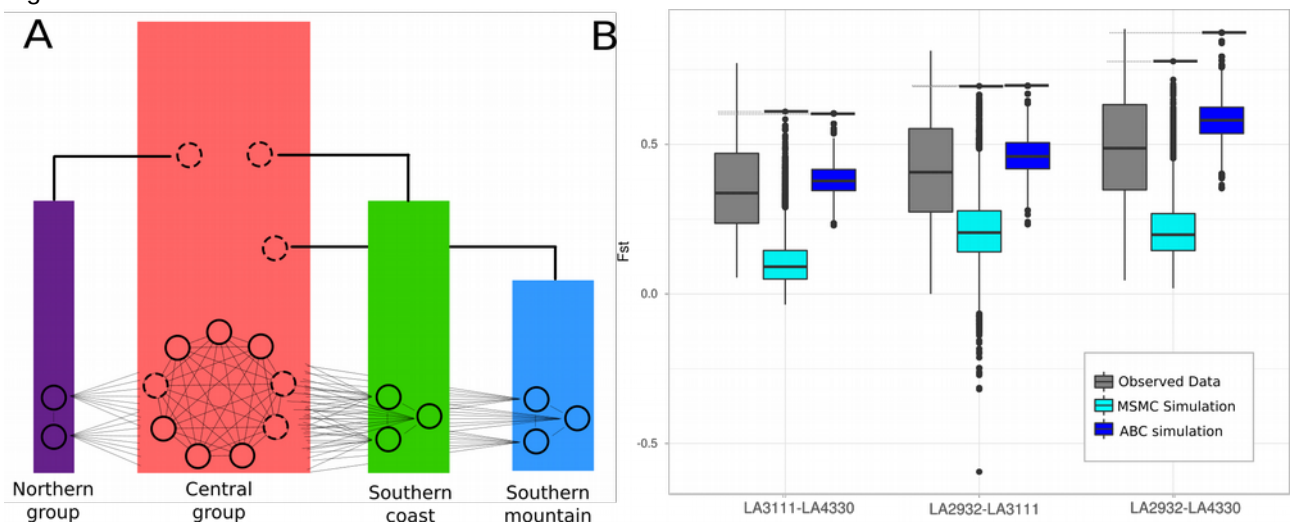


Figure 5

Coalescent simulations to identify $F_{ST}$ cut-off values

A) Coalescent model simulated for parameter estimation through the Approximate Bayesian Computation (ABC) approach. The model presents the same sampling for the empirical dataset with 14 populations from four regions (solid circles), as well as three unsampled “ghost” populations from the central region from which the populations of the other groups diverge. To illustrate that they also contribute to genetic diversity in the central population at present time they are presented twice in the figure (dashed circles). Populations evolve under the island model where migration among groups is smaller than migration within groups.

B) Boxplots indicating the similarity between the observed data (grey) and our simulations based on ABC inference (dark blue) or the MSMC inference (turquoise). The black horizontal bars (and dotted extension) indicate the maximum simulated values. The simulations are based on 30,000 genes under the model inferred by MSMC or ABC. The maximum values obtained under the ABC model (top lines), are used as $F_{ST}$ cut-offs for outlier selection.

Figure 6

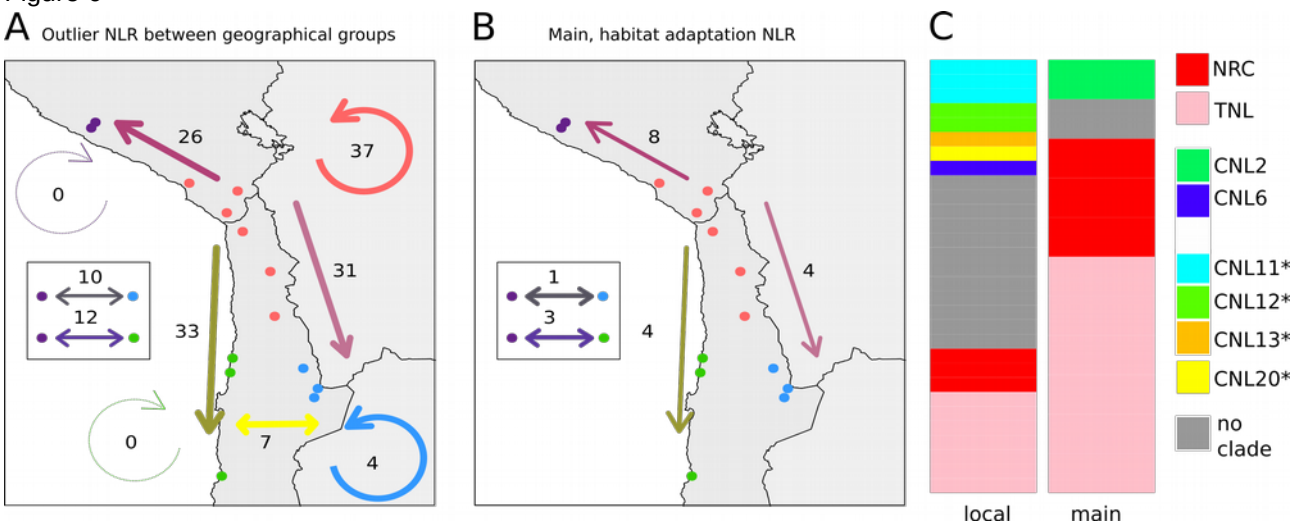


Figure 6

NLR genes under selection can be divided in main “habitat adaptation” and local “fine tuning “ adaptation NLRs.

A) Number of genes under selection ($F_{ST}$ outliers) found between the four main geographical groups (straight arrows) as well as within each of the groups (circular arrows), when summing all individual outliers. NLRs under selection between the north and mountain or north and coast are indicated in the box. In total 53 genes can be identified, many are common to several geographic groups.

B) Maps showing the number of main “habitat adaptation” NLRs between the different geographical groups. Habitat adaptation NLRs are defined as those that occur in more than one third of the possible population comparisons between the examined geographic groups.

C) Functional clade assignment (as fraction) of the main adaptation and local fine tuning NLRs. Clades marked with an * are expected to be NRC-dependent