Long-term balancing selection drives evolution of immunity genes in *Capsella*

Daniel Koenig[1,5]*, Jörg Hagmann[1,6], Rachel Li[1,7], Felix Bemm[1,8], Tanja Slotte[2], Barbara Neuffer[3], Stephen I. Wright[3], and Detlef Weigel[1]*

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany

[2]Department of Ecology, Environment, and Plant Sciences, Stockholm University, Stockholm, Sweden

[3]Department of Biology, University of Osnabrück, Osnabrück, Germany

[4]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

[5]Current address: Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

[6]Current address: Computomics GmbH, Tübingen, Germany

[7]Current address: Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

[8]Current address: KWS SE, 37574 Einbeck, Germany

*To whom correspondence should be sent: Detlef Weigel, weigel@tue.mpg.de; Daniel Koenig, dkoenig@ucr.edu

ABSTRACT

Genetic drift is expected to remove polymorphism from populations over long periods of time, with the rate of polymorphism loss being accelerated when species experience strong reductions in population size. Adaptive forces that maintain genetic variation in populations, or balancing selection, might counteract this process. To understand the extent to which natural selection can drive the retention of genetic diversity, we document genomic variability after two parallel species-wide bottlenecks in the genus *Capsella*. We find that ancestral variation preferentially persists at immunity related loci, and that the same collection of alleles has been maintained in different lineages that have been separated for several million years. Our data point to long term balancing selection as an important factor shaping the genetics of immune systems in plants and as the predominant driver of genomic variability after a population bottleneck.

INTRODUCTION

Balancing selection describes the suite of adaptive forces that maintain genetic variation for longer than expected by random chance. It can have many causes, including heterozygous advantage, negative frequency-dependent selection, and environmental heterogeneity in space and time. The unifying characteristic of these situations is that the turnover of alleles is slowed, resulting in increased diversity at linked sites (Charlesworth, 2006). In principle, it should be simple to detect the resulting footprints of increased coalescence times surrounding balanced sites, and many candidates have been identified using diverse methodology (Fijarczyk and Babik, 2015). However, theoretical models predict even strongly balanced alleles will be stochastically lost over long time spans, suggesting that most balanced polymorphism is short lived (Tellier et al., 2014).

The strongest evidence for balancing selection comes from systems in which alleles are maintained in lineages that are reproductively isolated and that have separated millions of years ago, resulting in trans-specific alleles with diagnostic trans-specific single nucleotide polymorphisms (tsSNPs). A few, well known genes fit this paradigm: the self-incompatibility loci of plants (Vekemans and Slatkin, 1994), mating loci of fungi (Wu et al., 1998), and the major histocompatibility complex (MHC) and ABO blood group loci in vertebrates (Lawlor et al., 1988; Mayer et al., 1988; McConnell et al., 1988; Ségurel et al., 2012; Watkins et al., 1990). Additional candidates have been proposed by comparing genome sequences from populations of humans and chimpanzees, and from populations of multiple *Arabidopsis* species. These efforts have revealed six loci in primates (Leffler et al., 2013; Teixeira et al., 2015) and up to 129 loci, that were identified by at least two shared SNPs each, in *Arabidopsis*

(Bechsgaard et al., 2017; Novikova et al., 2016), as potential targets of long-term balancing selection and/or introgression. In both systems, genes involved in host-pathogen interactions were enriched, which in *Arabidopsis* is consistent with previous findings that several disease resistance loci appear to be under balancing selection in this species, based on the analysis of individual genes (Bakker et al., 2006; Botella et al., 1998; Caicedo et al., 1999; Huard-Chauveau et al., 2013; Noël et al., 1999; Rose et al., 2004; Stahl et al., 1999; Tian et al., 2002; Todesco et al., 2010). However, whole-genome scans in *A. thaliana* have suggested a limited role for balancing selection in genome evolution (1001 Genomes Consortium, 2016; Cao et al., 2011), and the total number of examples with robust evidence across species remains small. Moreover, the importance of long-term balancing selection as a mode of evolution has been been questioned by theoretical models of host-pathogen interactions (Tellier et al., 2014).

One explanation for this paucity of evidence for pervasive and stable balancing selection is that cases of long-term maintenance of alleles are rare. However, there are good reasons to believe that many studies lacked the power to detect the expected effects (DeGiorgio et al., 2014; Fijarczyk and Babik, 2015). If one requires that alleles have been maintained in species separated by millions of years, then only targets of outstandingly strong selective pressures that remain the same over many millennia can be identified. Furthermore, recombination between deeply coalescing alleles will typically reduce the size of the genomic footprint to very short sequence stretches, thus limiting the opportunity for distinguishing old alleles from recurrent mutations.

We hypothesized that self-fertilizing species provide increased sensitivity to detect balancing selection based on two observations (Wiuf et al., 2004a; Wright et al., 2008). First, self fertilization greatly reduces the effective rate of recombination, thus potentially expanding the footprint of balancing selection. In addition, the transition to self fertilization is generally associated with dramatic genome-wide reductions in polymorphism, potentially making it easier to detect outlying loci that retain variation from the outcrossing, more polymorphic ancestor. In this study we sought to assess how strongly selection acts to maintain genetic diversity in the context of repeated transitions to self fertilization in the flowering plant genus *Capsella*. Like many plant lineages, the ancestral state of *Capsella* is outcrossing (found in the extant diploid species *C. grandiflora*), but selfing has evolved independently in two diploid species, *C. rubella* and *C. orientalis* (Figure 1A)(Bachmann et al., 2018; Foxe et al., 2009; Guo et al., 2009). The genomes of both species exhibit the drastic loss of genetic diversity typical for many selfers (Figure 1B-C) (Brandvain et al., 2013; Foxe et al., 2009; Guo et al., 2009; Slotte et al., 2013, 2012; St. Onge et al., 2011). In the younger species, *C. rubella*, loss of genetic diversity was initially thought to have occurred uniformly throughout the entire genome (Foxe

et al., 2009; Guo et al., 2009), but subsequent reports already hinted at some loci having increased diversity (Brandvain et al., 2013; Gos et al., 2012), motivating the present study.

RESULTS

Polymorphism discovery in *C. grandiflora* and *C. rubella*

The species *Capsella rubella* is young, only 30,000 to 200,000 years old, and was apparently founded when a small number of *C. grandiflora* individuals became self-compatible (Foxe et al., 2009; Guo et al., 2009). Previous studies had hinted at unequal representation of *C. grandiflora* alleles across the *C. rubella* genome (Brandvain et al., 2013; Gos et al., 2012), leading us to analyze this phenomenon systematically by comparing the genomes of 50 *C. rubella* and 13 *C. grandiflora* accessions from throughout each species' range (Figure 1 - figure supplement 1 and Figure 1 - figure supplement 2). Because the calling of trans-specific SNPs (tsSNPs) is particularly sensitive to mismapping errors in repetitive sequences, we applied a set of stringent filters, resulting in 74% of the *C. rubella* reference genome remaining accessible to base calling in both species, with almost half (47%) of the masked sites in the repeat rich pericentromeric regions. After filtering, there were 5,784,607 SNPs and 883,837 indels. Unless otherwise stated, all subsequent analyses were performed using SNPs. Of these, only 27,852 were fixed between the two species, whereas 824,540 were found in both species (ts_{CgCr}SNPs), consistent with the expected sharing of variation between the two species. In addition, 4,291,959 SNPs segregated only in *C. grandiflora* (species-specific SNPs; ss_{Cg}SNPs), and 640,256 only in *C. rubella* (ss_{Cr}SNPs). Sample rarefaction by subsampling our sequenced accessions indicated that ss_{Cr}SNP and ts_{CgCr}SNP discovery was near saturation in our experiment (Figure 1D).

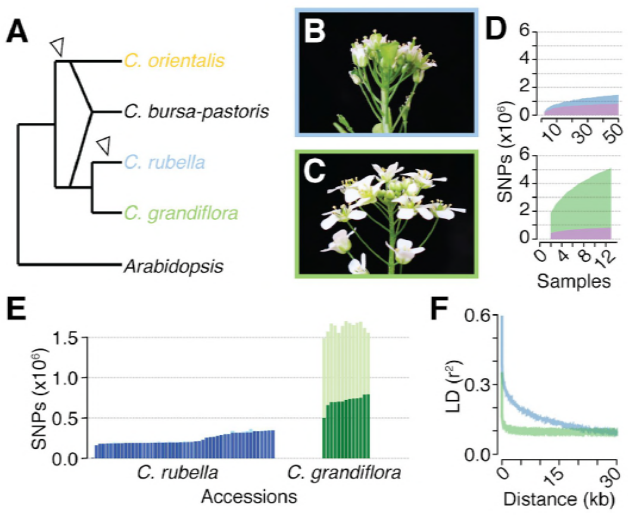


Figure 1. Polymorphism discovery in *Capsella*. (A) Diagram of the relationships between *Capsella* species. Arrowheads indicate transitions from outcrossing to self-fertilization. (B) Inflorescence of *C. rubella* with small flowers. (C) Inflorescence of *C. grandiflora* with large, showy flowers, to attract pollinators. (D) SNP discovery in *C. rubella* (top) and *C. grandiflora* (bottom). Samples were randomly downsampled ten times. Means of segregating transpecific (tsSNPs, purple), species specific in *C. rubella* (ss_{Cr}SNPs, blue), and species specific in *C. grandiflora* (ss_{Cg}SNPs, green) SNPs. (E) Number of heterozygous (light colors) and homozygous SNP calls (dark colors). (F) Average decay of linkage disequilibrium in *C. grandiflora* (green) and *C. rubella* (blue).

The consequences of selfing are easily seen as a dramatic reduction in genetic diversity in *C. rubella* (Figure 1 - figure supplement 3), consistent with the previously suggested genetic bottleneck (Foxe et al., 2009; Guo et al., 2009). As expected from a predominantly selfing species, SNPs segregating in *C. rubella* were much less likely to occur in the heterozygous state than those segregating in *C. grandiflora*, though evidence for occasional outcrossing in *C. rubella* is observed in the form of a variable number of heterozygous calls (Figure 1E). Selfing is also expected to reduce the effective rate of recombination between segregating polymorphisms. Linkage disequilibrium (LD) decayed, on average, to 0.1 within 5 kb in *C. grandiflora*, while it only reached this value at distances greater than 20 kb in *C. rubella* (Figure 1F). Though *C. rubella* is a relatively young species, it exhibits characteristics typical of a predominantly (but not exclusively) self-fertilizing species: reduced genetic diversity, reduced observed heterozygosity, and reduced effective recombination rate. This last effect could potentially increase the visibility of signals for balancing selection from linked sites (Wiuf et al., 2004b).

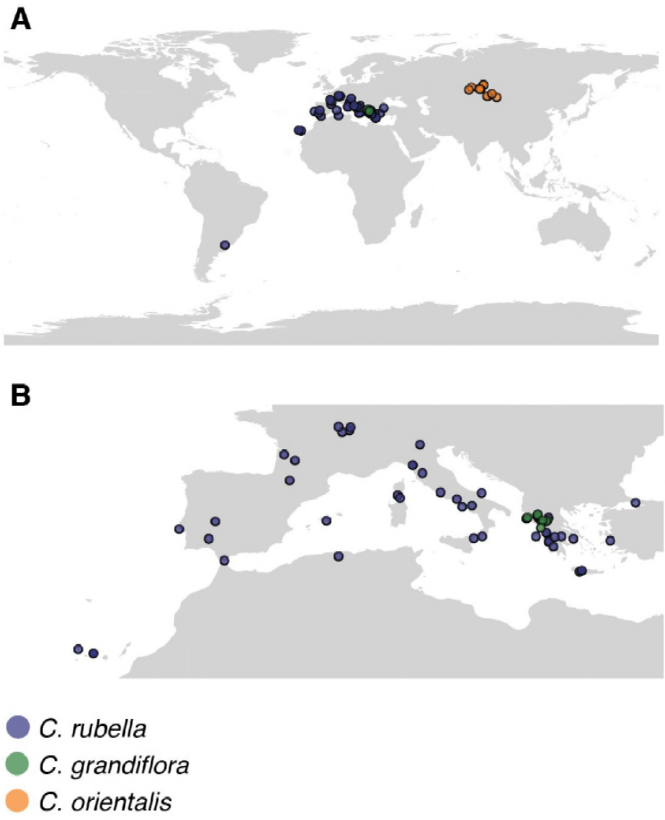


Figure 1 - figure supplement 1. Map of collections. (A) Global and (B) European locations of samples.

–SEE END OF DOCUMENT–

Figure 1 - figure supplement 2. Sample information.

Annotation	Diversity Cr	Diversity Cg	dCgCr
1	0.00121	0.00428	0.00462
2	0.00322	0.01303	0.01387
3	0.00441	0.01793	0.01941
4	0.00464	0.01991	0.02118
INTERGENIC	0.00360	0.01474	0.01582
INTRON	0.00354	0.01598	0.01700
UTR_3_PRIME	0.00269	0.01225	0.01300
UTR_5_PRIME	0.00233	0.01082	0.01139

Figure 1 - figure supplement 3. Diversity and divergence estimates for *C. grandiflora* and *C. rubella*.

Capsella rubella demography

The degree of trans-specific allele sharing is dependent upon the level of gene flow between species, the age of the speciation event, and the demographic history of each resultant species. We first sought to understand how these neutral processes have affected the pattern of extant polymorphism in *C. grandiflora* and *C. rubella*. We searched for evidence of population structure in our dataset by fitting individual ancestries to different numbers of genetic clusters with ADMIXTURE (Alexander et al., 2009) (Figure 2A and supplement 1 A-B; *k*-values from 1 to 6). The best fit as determined by the minimum cross-validation error was three clusters, with one including all *C. grandiflora* individuals, and *C. rubella* samples split into two clusters. Principal component (PC) analysis (Price et al., 2006) of genetic variation revealed a similar picture, with PC1 separating the two species and PC2 separating the *C. rubella* samples (Figure 2A).

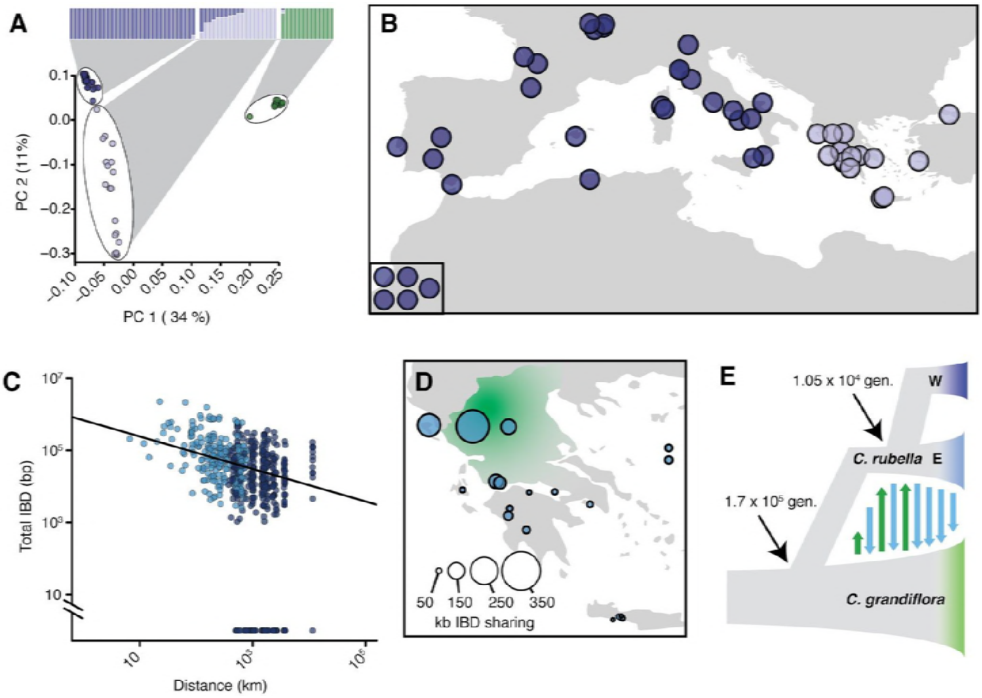


Figure 2. Demographic analysis of *C. rubella*. (A) Admixture bar graphs (top) and PCA of population structure in *C. grandiflora* (green) and *C. rubella* (blue). The *C. rubella* colors correspond to the sampling locations in (B). Inset shows lines from outside Eurasia (Canary Islands and Argentina). (C) Pairwise interspecific identity-by-descent (IBD) between *C. grandiflora* and *C. rubella* samples. Comparisons between West *C. rubella* and *C. grandiflora* are in dark blue and E *C. rubella* and *C. grandiflora* in light blue. The minimum segment length

threshold was 1 kb, and comparisons without IBD segments (all from the West *C. rubella* population) are at the bottom of the plot. (D) Total lengths of interspecific IBD sharing by sample site within the E *C. rubella* population. An approximate distribution of *C. grandiflora* is shown for comparison in green. (E) The most likely demographic model of *C. rubella* and *C. grandiflora* evolution as inferred from joint allele frequency spectra by fastsimcoal2. Arrows indicate gene flow.

C. rubella population structure was strongly associated with geography. Samples from western Europe and southeastern Greece were unambiguously assigned to separate groups, while samples from northern and western Greece, near the presumed site of speciation in the current range of *C. grandiflora* (Hurka and Neuffer, 1997), showed mixed ancestry (or intermediate assignment to these groups, Figure 2A-B). This pattern is consistent with the center of diversity for *C. rubella* being in northern Greece and a more recent rapid expansion into Western Europe, and agrees with predictions made based on previous, smaller datasets (Brandvain et al., 2013). Subsequently, we refer to these two distinct groups as the western (W) and eastern (E) populations.

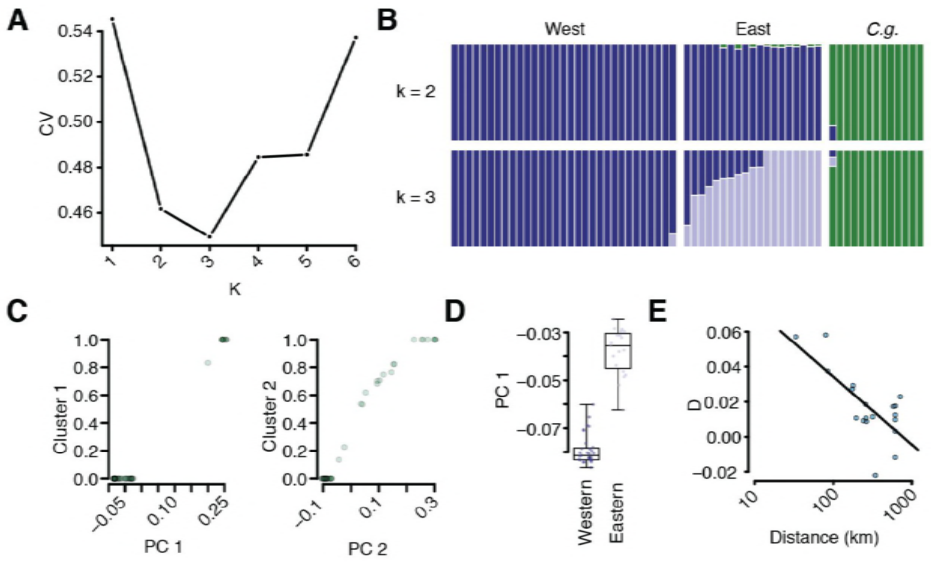


Figure 2 - figure supplement 1. Additional population structure and migration analyses. (A) Coefficient of variation for different numbers of *k* in ADMIXTURE runs. (B) Comparison of *k*=2 and *k*=3 results. (C) Correlation of sample position along PC1 and assignment to the *C. grandiflora* cluster and along PC2 and assignment to the East *C. rubella* cluster. (D) Position of East and West *C. rubella* samples along PC1. Dots show raw values. (E) *D* statistic in E *C. rubella* as a function of distance from northern Greece, where *C. grandiflora* occurs at the highest density. The line depicts a linear regression of this relationship ($p<<0.01$).

Because their current ranges overlap, ongoing gene flow between sympatric *C. rubella* and *C. grandiflora* could be a potentially important source of allele sharing between the two species. While a previous study had not found any evidence for such a scenario (Brandvain et al., 2013), one of our *C. grandiflora* samples was assigned partial ancestry to the otherwise *C. rubella*-specific clusters, and resided at an intermediate position along PC1 (Figure 2A). Furthermore, eastern *C. rubella* individuals, many of which grew in sympatry with *C. grandiflora*, were less differentiated from *C. grandiflora* compared to western *C. rubella* samples along PC1 (Figure 2A and Figure 2 - supplement figure 1C-D). Ongoing gene flow between eastern *C. rubella* and *C. grandiflora* was supported by significant genome-wide *D*-statistics for *C. rubella* samples from the *C. grandiflora* range (ABBA-BABA test; comparing each E individual with the W population) (Durand et al., 2011; Green et

al., 2010), with *D* decreasing as a function of distance from the center of *C. grandiflora*'s range (Figure 2 - supplement figure 1 and 2). Because *D* statistics can be sensitive to ancient population structure (Durand et al., 2011), we further relied on identity-by-descent (IBD) segments as detected by BEAGLE (Browning and Browning, 2013) to identify genome regions of very recent co-ancestry across these species. The proportion of the genome shared in IBD segments between *C. rubella* and *C. grandiflora* also decreased as a function of distance between samples, and the strongest evidence for recent ancestry was found between *C. grandiflora* individuals and sympatric northern Greek *C. rubella* lines (Figure 2 C-D). These results indicate that gene flow is ongoing between the species, consistent with interspecific crosses often producing fertile offspring (Sicard et al., 2011).

Accession	D statistic	Z-Score	p-value	
1GR1	0.01724	1.49	0.1362242359	
78.1	0.01066	0.72	0.4715249956	
83.1	0.02927	2.16	0.0307726696	*
72.12	0.00869	0.69	0.4901941873	
84.15	0.02726	1.98	0.0477035287	*
1407-8	0.01754	1.57	0.1164151113	
762	-0.02205	-1.01	0.3124952900	
844	-0.01185	-0.84	0.4009083865	
879	0.01238	1.2	0.2301393404	
907	0.05812	4.61	0.0000040267	*
100.8	0.03742	1.95	0.0511761190	
79.1	0.0184	1.52	0.1285109756	
75.2	0.00911	0.69	0.4901941873	
925	0.05708	3.78	0.0001568284	*
80TR1-TS1	0.02286	1.32	0.1868350180	
1408	0.00317	0.31	0.7565609564	
1411-3	0.00969	0.94	0.3472175607	
81.2	0.01137	1.03	0.3030100056	
77.16	0.01069	0.9	0.3681202507	

Figure 2 - figure supplement 2. *D* statistics comparing East and West *C. rubella* populations.

To estimate the magnitude and direction of gene flow and other demographic events that have shaped genetic variation in the two species we used fastsimcoal2 (Excoffier et al., 2013) to compare the likelihood of a large number of demographic models given the observed joint site frequency spectrum (Figure 2E and figure 2 - figure supplement 3 and 4). The best fitting model estimated the split between *C. rubella* and *C. grandiflora* to have occurred 170,000 generations ago, associated with a strong reduction in *C. rubella* population size (to only 2-14 effective chromosomes, or 1-7 individuals). Bidirectional gene flow at a relatively low rate apparently occurred until just over 10,000 generations ago, when *C. rubella* split into the W and E populations, after which gene flow continued only from E *C. rubella* to *C. grandiflora* (Figure 2E).

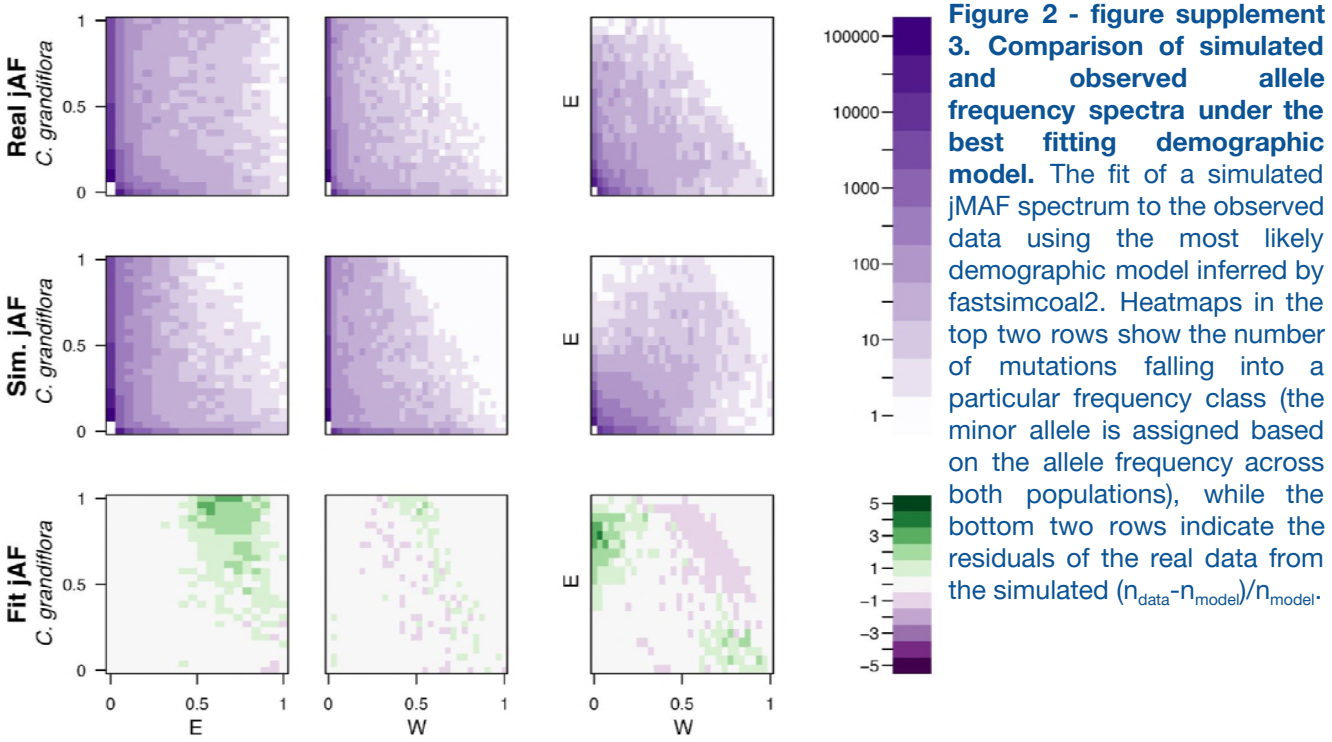


Figure 2 - figure supplement 3. Comparison of simulated and observed allele frequency spectra under the best fitting demographic model. The fit of a simulated jMAF spectrum to the observed data using the most likely demographic model inferred by fastsimcoal2. Heatmaps in the top two rows show the number of mutations falling into a particular frequency class (the minor allele is assigned based on the allele frequency across both populations), while the bottom two rows indicate the residuals of the real data from the simulated $(n_{data}-n_{model})/n_{model}$.

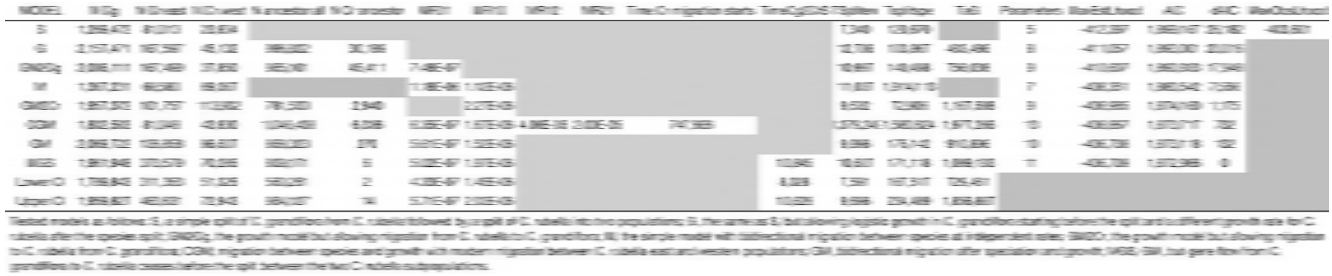


Figure 2 - figure supplement 4. Inferred demographic parameters from fastsimcoal2.

The close timing of the end of gene flow into *C. rubella* and the split into two populations suggests that westward expansion of the *C. rubella* range reduced the opportunity for gene flow from *C. grandiflora*, with potential genetic reinforcement by the development of hybrid incompatibilities (Sicard et al., 2015). If we assume an average of 1.3 years per generation as found in the close relative, *A. thaliana(Falahati-Anbaran et al., 2014)*, which has similar life history and ecology, the population split and the end of introgression from *C. grandiflora* occurred around 13,500 years ago. This date is similar to the spread of agriculture and the end of the last glaciation in Europe (Walker et al., 2009), suggesting that *C. rubella*'s success might have been facilitated by one or both of these events.

Non-random polymorphism sharing after a genetic bottleneck

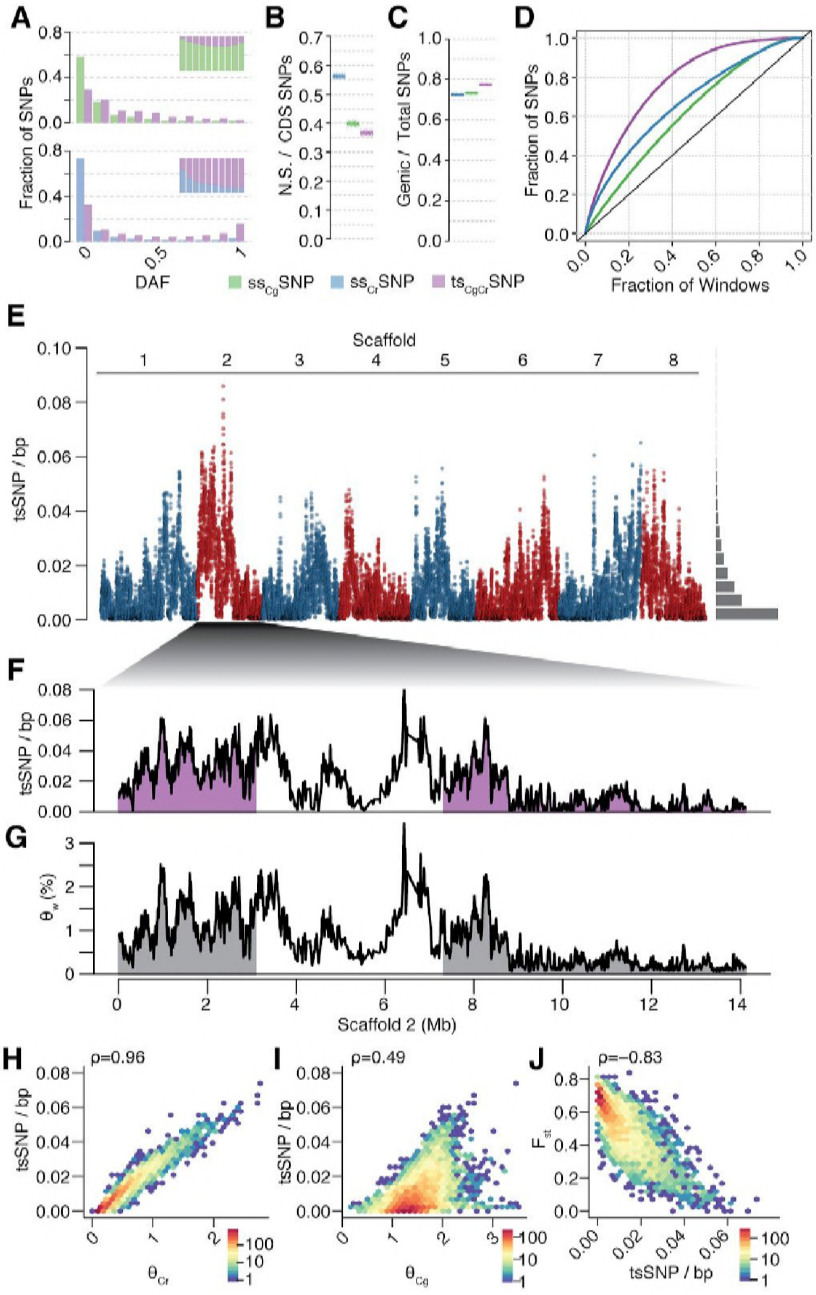


Figure 3. Unequal presence of ancestral variation in modern *C. rubella*. (A) Derived allele frequency spectra (DAF) of ss_{Cg}SNPs (green), ss_{Cr}SNPs (blue), and ts_{CgCr}SNPs (purple) in *C. grandiflora* (top) and *C. rubella* (bottom). The inset depicts the fraction of alleles that are species or transpecific as a function of derived allele frequency (DAF). (B) Fraction of coding (CDS) SNPs that result in non-synonymous changes as a function of SNP sharing. (C) Fraction of genic SNPs as a function of SNP sharing. Because SNPs in different classes (ssSNPs, tsSNPs) differ in allele frequency distributions, we normalized by downsampling to comparable frequency spectra. Each bar consists of 1,000 points depicting downsampling values. (D) 20 kb genomic windows required to cover different fractions of ssSNPs and tsSNPs. The black line corresponds to a completely even distribution of SNPs in the genome. tsSNPs deviate the most from this null distribution. (E) ts_{CgCr}SNP density in 20 kb windows (5 kb steps) along the eight *Capsella* chromosomes. Histogram on the right shows distribution of values across the entire genome. (F) ts_{CgCr}SNP density and (G) Watterson's estimator (Θ_w) of genetic diversity along scaffold 2. The repeat dense pericentromeric regions are not filled. (H-J) Correlation of ts_{CgCr}SNP density in 20 kb non-overlapping windows with genetic diversity in *C. rubella* (H), genetic diversity in *C. grandiflora* (I), and interspecific F_{st} (J). Spearman's rho is always given on the top left. Only windows with at least 5,000 accessible sites in both species were considered.

Our analyses provide dates for the bottleneck and rapid colonization events that have led to dramatically reduced genetic variation in *C. rubella*. Yet, over half of the segregating variants in *C. rubella* were also found in *C. grandiflora* (Figure 1D). Such ts_{CgCr}SNPs could originate from independent mutation in each species (identity by state, IBS). Alternatively, they could be the result of introgression after speciation or they could reflect retention of the same alleles since the species split (identity by descent, IBD). Older retained alleles are expected to be found at elevated frequencies relative to the genome-wide average, while younger, recurrent mutations are expected to

be rare. We therefore identified ancestral and derived alleles by comparison with the related genus *Arabidopsis*, and then compared the derived allele frequency spectra of ts_{CgCr}SNPs and ssSNPs in *Capsella* as a proxy for allele age. We found that ts_{CgCr}SNPs are strongly enriched among high-frequency alleles in both *Capsella* species (Figure 3A, p-value << 0.0001 in *C. grandiflora* and *C. rubella*, Mann-Whitney U-test). At allele frequencies greater than 0.25 in *C. rubella*, ts_{CgCr}SNPs accounted for more than 80% of all variation. These results indicate that ts_{CgCr}SNPs are predominantly older alleles that were already present in the common ancestral population of *C. rubella* and *C. grandiflora* or that were introgressed from *C. grandiflora* prior to its expansion into western Europe.

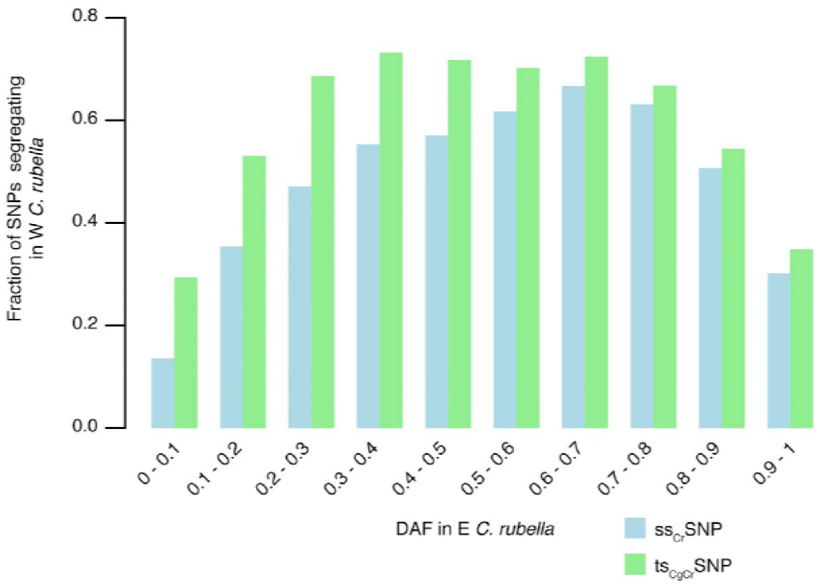


Figure 3 - figure supplement 1. Sharing of ts_{CgCr}SNPs and ss_{Cr}SNPs alleles after colonization.

The distribution of ts_{CgCr}SNPs was uneven across the genome. When compared to ss_{Cr}SNPs drawn from the same allele frequency distribution, ts_{CgCr}SNPs were less likely to result in nonsynonymous changes (Figure 3B, p-value < 0.001, from 1,000 jackknife resamples from the same allele frequency distribution), but they were more likely to be in genes (Figure 3C). Eighty-three percent of all ts_{CgCr}SNPs were in complete LD with at least one other ts_{CgCr}SNP in *C. rubella*, and the density of tsSNPs along the genome was highly variable (Figure 3D-G). ts_{CgCr}SNP density was positively correlated with local genetic diversity in *C. rubella* (and less strongly so with genetic diversity in *C. grandiflora*; Figure 3F-I and figure supplement 2), and negatively correlated with differentiation between the species as measured by F_{st} (Figure 3J and figure supplement 2). The uneven pattern of diversity was similar in each *C. rubella* subpopulation (Figure 3 - figure supplement 3), indicating that most of the retained polymorphism already segregated prior to colonization. Thus, most common genetic variation in *C. rubella* is also retained in its outcrossing ancestor, and the rate of retention varies dramatically between genomic regions.

–SEE END OF DOCUMENT–

Figure 3 - figure supplement 2. Diversity in *C. rubella* and *C. grandiflora* along the eight major scaffolds (chromosomes). Data calculated in 20 kb windows, 5 kb steps. Windows with fewer than 5,000 covered sites were masked for all calculations. Dotted lines indicate the pericentromeric regions excluded from our analysis, and shaded regions highlight the regions with significant evidence for balancing selection.

–SEE END OF DOCUMENT–

Figure 3 - figure supplement 3. Distribution of diversity in East and West *C. rubella* populations along the eight major scaffolds (chromosomes). As in S5, with diversity calculated within and F_{st} between populations. In general, regions of high diversity species-wide show high diversity and low F_{st}.

High density of tsSNPs around immunity-related loci

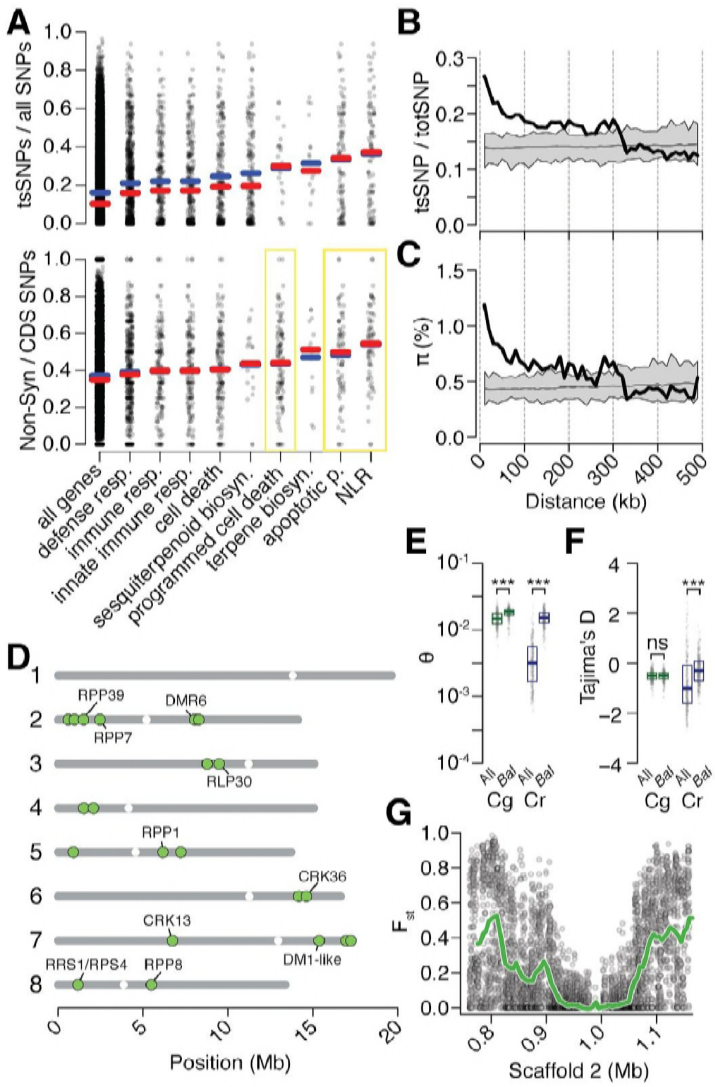


Figure 4. Preferential sharing of alleles near immunity genes. (A) Enrichment of ts_{CgCr}SNPs and non-synonymous ts_{CgCr}SNPs for genes associated with significant GO terms (means, blue; medians, red). GO terms with a significantly increased ratio of nonsynonymous changes are highlighted with a yellow box. (B) ts_{CgCr}SNP frequency as a function of distance to the closest NLR cluster. (C) Pairwise genetic diversity at neutral (four-fold degenerate) sites as a function of distance to the closest NLR cluster. Thick black lines, mean values calculated in 500 bp windows as a function of distance from a NLR gene. Thin black lines, mean values from 100 random gene sets of equivalent size. Grey polygon, range of values across all 100 random gene sets. (D) Chromosomal locations of *Bal* regions with the strongest evidence for balancing selection. Immunity genes with known function in *A. thaliana* in each region indicated. (E) Values for Watterson's estimator (Θ_w) of diversity in *Bal* regions, calculated from 20 kb windows. (F) Tajima's D. Dots in (E, F) are a random sample of 1,000 windows for non candidate windows. Boxplots report the median 1st and 3rd quantiles of all windows in each class. (G) An example of the site level (dots) and windowed (green) decrease in F_{st} at the first region on chromosome 2. The subregion without data near 1 Mb is a CC-NLR cluster, which was largely masked for variant calling.

The observed heterogeneity in shared diversity across the *C. rubella* genome could be a simple consequence of a bottleneck during the transition to selfing. In the simplest scenario, *C. rubella* was

founded by a small number of closely related individuals, and stochastic processes during subsequent inbreeding caused random losses of population heterozygosity. A study of genetic variation in bottlenecked populations of the Catalina fox found this exact pattern (Robinson et al., 2016). Alternatively, there may be selective maintenance of diversity in specific regions of the genome due to balanced polymorphisms, with contrasting activities of the different alleles. To explore this latter possibility, we tested whether the likelihood of allele sharing was dependent on annotated function of the affected genes. We found that ts_{CgCr}SNPs were strongly biased towards genes involved in plant biotic interactions, including defense and immune responses, and also toward pollen-pistil interactions, though less strongly (Supplementary file 1, Figure 4A). Amongst the top ten enriched Gene Ontology (GO) categories for biological processes were apoptotic process, defense response, innate immune response, programmed cell death, and defensive secondary metabolite production (specifically associated with terpenoids). Of genes annotated with apoptotic process, 87% were homologs of *A. thaliana* NLR genes, a class of genes best known for its involvement in perception and response to pathogen attack (Jones and Dangl, 2006). An even higher enrichment for ts_{CgCr}SNPs was found when testing this class of genes specifically, with ts_{CgCr}SNPs falling in NLR genes being more likely than those in other types of genes to result in nonsynonymous changes that could potentially have functional consequences (Figure 4A-B). These results indicate that despite a severe global loss of genetic diversity, genes involved in plant-pathogen interactions have maintained high levels of genetic variation in *C. rubella*.

While the high density of ts_{CgCr}SNPs near immunity genes was intriguing, NLR genes frequently occur in complex clusters, which could elevate error rates during SNP calling and thus potentially influence our analyses. Of particular concern is that sequencing reads derived from paralogs not found in the reference, but present in some accessions, could be mismapped against the reference, leading to false positive ts_{CgCr}SNPs calls. We therefore examined whether ts_{CgCr}SNPs showed more evidence of such errors than other SNPs. Mismapping should increase coverage and reduce concordance (the fraction of reads supporting a particular call) at a site. That the distributions of these two metrics were nearly identical at ts_{CgCr}SNPs and ssSNP sites indicates, however, that mismapping is unlikely to have affected our SNP calls (Figure 4 - figure supplement 1). Mismapping is also expected to cause pseudo-heterozygous calls, due to reads from different positions in the focal genome being mapped to the same target in the reference genome. However, ts_{CgCr}SNPs were not more likely to be found in the heterozygous state as compared to ssSNPs (Figure 4 - figure supplement 1). In addition, we asked whether the signal of increased ts_{CgCr}SNPs density extended into sequences adjacent to NLRs and is detectable even when masking the NLR clusters themselves. For this purpose we collapsed NLR genes within 10 kb of one another into a single region, and calculated ts_{CgCr}SNPs rates and genetic diversity as a function of distance from these collapsed regions, ignoring SNPs

within the focal cluster. We found that elevated ts_{CgCr}SNPs sharing and genetic diversity extended over 100 kb from NLR genes. Thus, increased sharing is not an artifact of the internal structure of NLR clusters (Figure 4B-C).

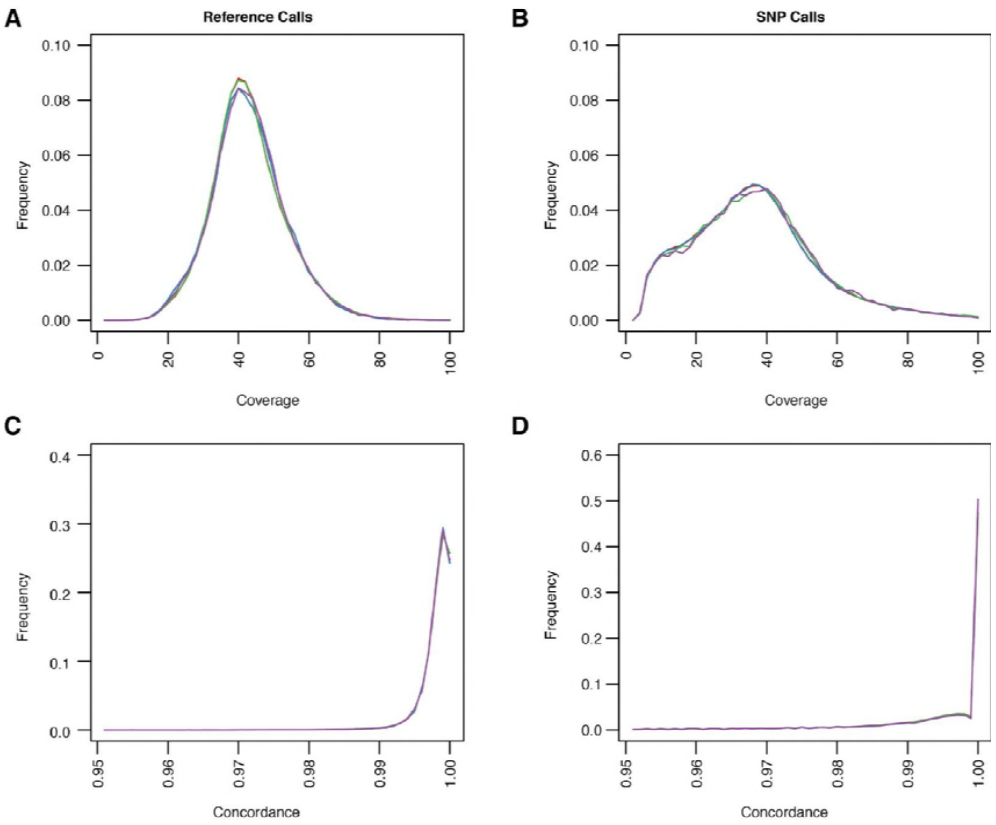


Figure 4 - figure supplement 1. Quality metrics for ssSNPs and tsSNPs in *C. rubella*. (A-D) Average coverage and concordance for reference calls (A and C) or non-reference calls (B and D). ssSNPs or tsSNPs outside of regions with significant evidence for balancing selection (red and blue) or inside of these regions (green and purple).

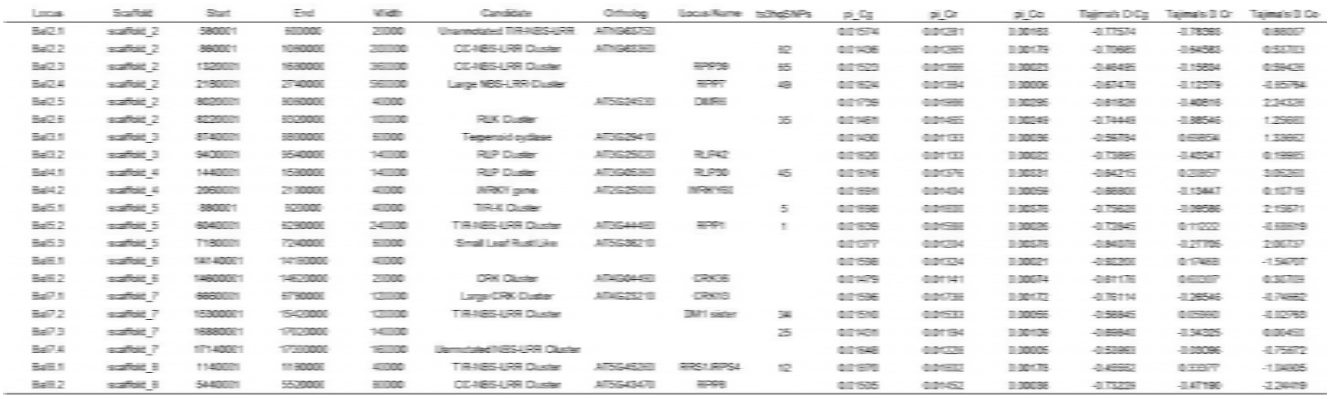


Figure 4 - figure supplement 2. Regions with evidence for balancing selection.

Increased retention of genetic diversity near immunity loci suggests that these genes might be the targets of balancing selection in either *C. rubella*, *C. grandiflora*, or both species. However, neutral processes including random introgression and stochastic allele fixation can give rise to uneven distributions of genetic variation across the genome after genetic bottlenecks (Robinson et al., 2016). We sought to identify regions that showed a pattern of allele sharing that was unlikely to have occured neutrally, as indicated by low values of the fixation index F_{st}, which quantifies genetic

differentiation between populations. We compared the observed values of F_{st} between *C. rubella* and *C. grandiflora* to a distribution calculated from simulated sequences under our previously inferred neutral demographic model, which included gene flow between *C. rubella* and *C. grandiflora*. We simulated one million 20 kb DNA segments, or just over 7,000 *C. rubella* genome equivalents, under the neutral model and calculated the expected distribution of F_{st} values. Using this distribution, we assigned the probability of observing the F_{st} value for each non-overlapping 20 kb window throughout the genome. After Bonferroni correction and joining of adjacent significant segments, we identified 21 genomic regions that we designated as candidate targets of balancing selection (*Bal*, Figure 4D and figure supplement 2). *Bal* regions showed several classical indications of balancing selection including substantially higher Tajima's *D* and within-*C. Rubella* genetic diversity relative to the remainder of the genome (Figure 4 E-F; p<< 0.001 Mann-Whitney U-test for both statistics). ts_{CgCr}SNPs in *Bal* regions were also less likely to have been lost during colonization of Western Europe then ss_{Cr}SNPs or ts_{CgCr}SNPs in other parts of the genome, and allele frequencies in *Bal* regions showed elevated correlation across populations (Figure 4 - figure supplement 3). Like tsSNPs in general, *Bal* regions did not show evidence for increased heterozygosity that might indicate increased error rates in SNP calling (Median Observed - Expected Heterozygosity in 20 kb windows was -0.021 inside of *Bal* regions, and -0.020 outside of these regions).

ts CgCr SNPs		
	C. rubella-East	*C. rubella*-West
C. grandiflora	0.166998	0.07838415
C. rubella-East		0.36955393
ts CgCr SNPs in Balanced Regions		
	C. rubella-East	*C. rubella*-West
C. grandiflora	0.4743708	0.4300125
C. rubella-East		0.668121
ts3way-HQ SNPs		
	C. rubella-East	*C. rubella*-West
C. grandiflora	0.458713	0.2692516
C. rubella-East		0.6802083
ts 3way-HQ SNPs in Balanced Regions		
	C. rubella-East	*C. rubella*-West
C. grandiflora	0.6503384	0.5241294
C. rubella-East		0.7869077

Figure 4 - figure supplement 3. Spearman's correlations of allele frequencies for different classes of tsSNPs. Only SNPs with a minor allele frequency greater than 0.05 in all three populations were considered.

Estimates of F_{st} were reduced in large segments surrounding NLR and other immune gene candidate clusters (Figure 4G), consistent with allele sharing being the result of linkage to a nearby balanced polymorphism. Of the 21 candidate regions, nine overlapped with clusters of NLR genes, and five with clusters of RLK/RLP or CRK genes, two classes of genes with broad roles in innate immunity (Yeh et al., 2015; Zipfel, 2008). Many of the specific regions we identified in *Capsella* have been directly demonstrated to function in disease resistance in *A. thaliana* (Gassmann et al., 1999; Goritschnig et al., 2012; Holub et al., 1994; McDowell et al., 2000, 1998; Xu et al., 2006; Yeh et al.,

2015; Zhang et al., 2014, 2013). *RPP1* and *RPP8* have been previously suggested as candidate targets of balancing selection, and trans-specific polymorphism has been reported at the *RPP8* locus in the genus *Arabidopsis* (Bergelson et al., 2001; Wang et al., 2011). It should be noted, however, that these genes are often members of larger linked NLR gene superclusters, with some of the regions our approach identified being sizeable and thus making it difficult to pinpoint a single focal gene. Indeed the strong signal found in these regions could result from multiple linked balanced sites. Furthermore, the strongest signals of balancing selection are mostly derived from linked sites, rather than the clusters themselves, because confident SNP calling is very difficult, if not impossible, with short reads in the most complex genomic regions (Figure 4G).

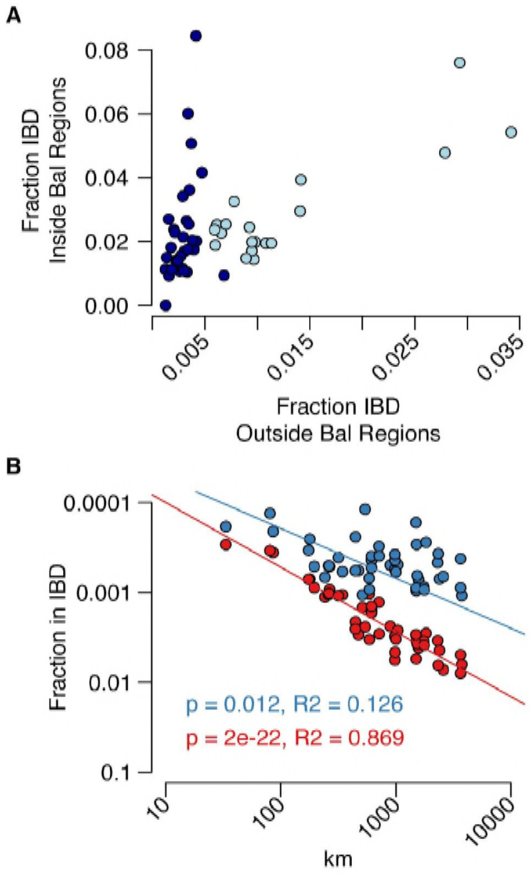


Figure 4 - figure supplement 4. Analysis of IBD in balanced regions and genome-wide. (A) The relationship between genome-wide and balanced region IBD. Dark blue, west population accessions, Light blue, east population accessions. (B) The decay of IBD genome-wide (red) and within balanced regions (blue) as a function of distance from the center of *C. grandiflora*'s range.

It is formally possible that the unusual pattern of diversity that we observe near *Bal* loci could result from historical balancing selection in the outcrossing ancestor *C. grandiflora* rather than ongoing selection in the selfing *C. rubella*. Population genetic indices such as F_{st}, nucleotide diversity pi, Tajima's D, and allele sharing are not fully independent, and elevated diversity in the *C. rubella* founding population, driven by historical balancing selection, could also generate the observed patterns. ,Since genetic diversity was only modestly elevated in these regions in *C. grandiflora* ($p << 0.001$ Mann-Whitney U-test, Figure 4E), and Tajima's *D* was not significantly different from other windows (Figure 4F), suggest that this is not very likely. If balancing selection is acting at these loci in the outcrosser, it is clear that its genomic footprint is small, perhaps due to the rapid decay of LD in

this species relative to the selfing *C. rubella*. Still, it is possible that even a small elevation of genetic diversity in *Bal* regions in the founding populations might have considerable impact on subsequent *C. rubella* diversity. We approximated this situation using our simulated genetic data. We subsetted simulations by the level of genetic diversity in *C. grandiflora*, choosing the top 1% of simulated values. Even in the case of elevated founder diversity in these data, the observed F_{st} values in *Bal* regions remain exceptionally unlikely ($p<0.0001$). These observations point to ongoing balancing selection within *C. rubella* maintaining diversity in *Bal* regions.

Adaptive retention of *C. grandiflora* diversity in *Bal* regions could be explained by two non-exclusive models. Allelic variation might have been present in the *C. rubella* founding population and maintained by balancing selection until the present. Alternatively, beneficial alleles may have been introgressed from *C. grandiflora* after the evolution of selfing, and retained by balancing selection. We searched for evidence of recent ancestry between the two species in *Bal* regions. A larger fraction of *Bal* region sequence was found to be IBD when compared to the genome-wide average (Figure 4 - figure supplement 4), consistent with elevated rates of gene flow. Shared segments in *Bal* regions were on average shorter than those found in other parts of the genome, suggesting that they are older and have been subjected to longer periods of recombination since the introgression event (median within 3,503 bp, median outside 6,661 bp, Wilcoxon-rank sum test, p = 1e-54), although we cannot exclude the influence of differing patterns of recombination in these regions as a contributing factor to this observation.

Elevated IBD rates in *Bal* regions might result from gene flow between the species in either direction, and our previous results suggested that most modern gene flow occurs through introgression of *C. rubella* alleles into *C. grandiflora*. We explored the geographic pattern of IBD between *C. rubella* and *C. grandiflora* in *Bal* regions to determine whether it differs from that of the genome-wide pattern. Within the East population, IBD decayed as a function of distance from the *C. grandiflora* range in a manner comparable to the observed genome-wide pattern, albeit with a more shallow slope (Figure 4 - figure supplement 4). In contrast to the genome-wide pattern, high levels of IBD were observed between *C. grandiflora* and West population accessions. Thus, *Bal* regions are influenced by neutral gene flow much like the rest of the genome, perhaps dominated by *C. rubella* to *C. grandiflora* introgression as indicated by our demographic simulations. However, allele sharing appears to be older in *Bal* regions and introgressed alleles have been retained for longer periods even after colonization of Western Europe. This latter observation is consistent with the hypothesis that alleles were introgressed prior to the most recent range expansions in *C. rubella*, and that variation was subsequently maintained by selection in *Bal* regions.

Balancing selection over millions of years

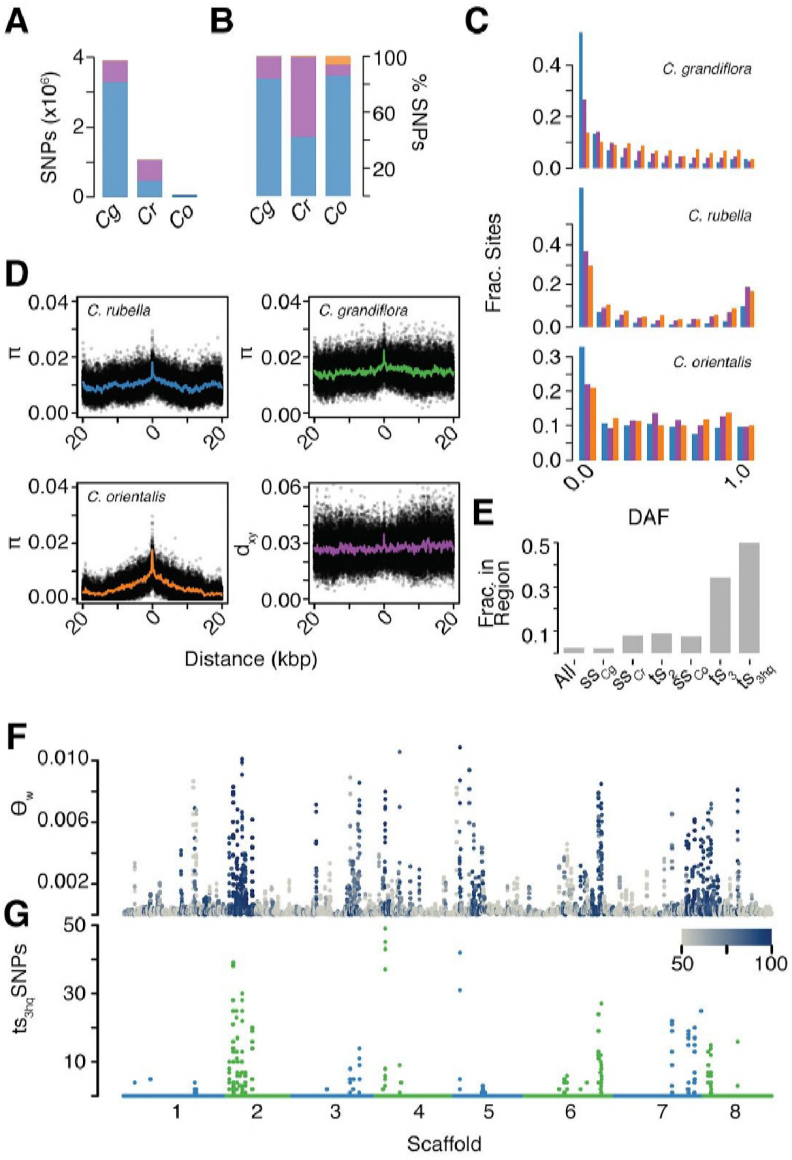


Figure 5. The signal of ancient balancing selection. (A) Absolute number and (B) fraction of ssSNPs (blue), $ts_{2\text{-way}}$SNPs (purple), and $ts_{3\text{-way}}$SNPs (orange) for *Capsella*. Only sites accessible to read mapping in all three species were considered. (C) Derived allele frequency (using *A. lyrata* and *A. thaliana* as outgroup) of ssSNPs (blue), $ts_{2\text{-way}}$SNPs (purple) and $ts_{3\text{-way}}$SNPs (orange). (D) Pairwise diversity (π) as a function of distance from a $ts_{3\text{-way}}$SNP in *C. grandiflora*, *C. rubella*, and *C. orientalis*. Bottom right, Divergence between the *C. rubella*/*C.grandiflora* and *C. orientalis* lineages as a function of distance from a $ts_{3\text{-way}}$SNP. Black dots, means over all sites at a particular distance, and colored lines, means over bins of 50 bp. (E) Enrichment of $ts_{3\text{-way}}$SNPs and $ts_{3\text{-wayhq}}$SNPs in candidate balanced regions from the *C. rubella*/*C. grandiflora* comparison. (F) Watterson's estimator (Θ_w) of genetic diversity in *C. orientalis*, in 20 kb windows. Grey-to-blue scale indicates the genome-wide percentile of the same window for Θ_w) in *C. rubella*. (G) ts_{3hq}SNP number in each window.

Although evidence for balancing selection at immunity-related loci in *C. rubella* is much stronger than in *C. grandiflora*, it is difficult to completely exclude the effect of founder diversity at these loci on the observed patterns. We therefore sought to validate our findings in a related species that has been separated from *C. grandiflora* and *C. rubella* for a long time. The genus *Capsella* offers a unique opportunity to test the longevity of balancing selection, because selfing has evolved independently in *C. orientalis*, which diverged from *C. grandiflora* and *C. rubella* more than one million years ago and whose modern range no longer overlaps with the two other species, preventing ongoing introgression (Douglas et al., 2015; Hurka et al., 2012). We expected the evolution of selfing to have generated a similar bottleneck as in *C. rubella* (Bachmann et al., 2018; Douglas et al., 2015), and we

therefore resequenced 16 *C. orientalis* genomes, to test whether there is evidence of balancing selection at similar types of loci.

After alignment, SNP calling, and filtering, we identified a mere 71,454 segregating SNPs in *C. orientalis*. This is a surprisingly small amount of variation, corresponding to an almost 50-fold reduction in diversity relative to the outcrossing *C. grandiflora* (Figure 5 - figure supplement 1). Using our divergence and diversity measures, we estimated that *C. orientalis* diverged from *C. grandiflora* over 1.8 million generations ago (calculated as in ref. (Brandvain et al., 2013)). The combination of long divergence times and low variability in *C. orientalis* makes it unlikely that alleles will have been maintained by random chance. (Using the methodology of ref. (Wiuf et al., 2004a), which assumes constant N_e in *C. orientalis* and *C. grandiflora*, the probability of finding a single tsSNP is $< 7 \times 10^{-40}$.) It was therefore surprising that 8,408 *C. orientalis* variants were shared with either *C. rubella* or *C. grandiflora* ($ts_{2\text{-way}}$SNPs), and 3,992 with both ($ts_{3\text{-way}}$SNPs, Figure 5A-B). In each of the three species, $ts_{3\text{-way}}$SNPs were enriched at higher derived allele frequencies relative to ssSNPs and $ts_{2\text{-way}}$SNPs, suggesting that they are on average the oldest SNPs (Figure 5C).

Annotation	Diversity Cr	Diversity Cg	Diversity Co	dCgCr	dCrCo	dCgCo
1	0.00114	0.00405	0.00018	0.00438	0.01008	0.01005
2	0.00312	0.01272	0.00032	0.01355	0.02967	0.02962
3	0.00426	0.01764	0.00039	0.01900	0.04151	0.04115
4	0.00449	0.01945	0.00041	0.02068	0.04508	0.04514
INTERGENIC	0.00311	0.01360	0.00034	0.01433	0.03084	0.03168
INTRON	0.00336	0.01561	0.00032	0.01647	0.03573	0.03635
UTR_3_PRIME	0.00252	0.01180	0.00028	0.01243	0.02745	0.02793
UTR_5_PRIME	0.00220	0.01042	0.00028	0.01092	0.02454	0.02506

Figure 5 - figure supplement 1. Three species diversity and divergence.

Because this large amount of trans-specific polymorphism was unexpected, we wanted to ensure that this was not due to more error-prone read mapping to a distant reference. We therefore also used an additional set of more stringent filters to identify high confidence $ts_{3\text{-way}}$SNPs ($ts_{3\text{-wayhq}}$SNPs; see methods). Importantly, we required $ts_{3\text{-wayhq}}$SNPs to be in LD with at least one other $ts_{3\text{-wayhq}}$SNP in all three species (r^2>0.2 in the same phase), to provide evidence that they represented the same ancestral haplotype. The aim was to improve the likelihood that such SNPs were true examples of identity by descent. Furthermore, we generated a draft assembly of the *C. orientalis* genome using Pacific Biosciences SMRT cell technology, and recalled $ts_{3\text{-way}}$SNP sites to remove potential reference bias. We identified 812 high quality transpecific SNPs segregating in all three species ($ts_{3\text{-wayhq}}$SNPs).

As discussed earlier, the presence of trans-specific polymorphism in diverged species could be driven by stable balancing selection or it could result from gene flow between the species. While *C.*

grandiflora and *C. rubella* occur around the Mediterranean, *C. orientalis* is restricted to Central Asia(Hurka et al., 2012) and its current distribution is far from that of *C. grandiflora* and *C. rubella*. Modern gene flow between the *C. orientalis* and *C. rubella* / *C. grandiflora* lineages is therefore unlikely, but it is possible that the ranges of these species overlapped in the past. If alleles have been maintained since the split between the lineages, then the divergence between maintained alleles should meet or exceed the divergence between the species. On the other hand, if $ts_{3\text{-}wayhq}$SNPs are the result of recent gene flow between the lineages, then divergence between species near these SNPs should be reduced. We examined diversity and divergence at neutral (four-fold degenerate) sites surrounding $ts_{3\text{-}wayhq}$SNPs (Figure 5D). In all three species, diversity was high directly adjacent to $ts_{3\text{-}wayhq}$SNPs, close to average levels for genome-wide divergence between the two *Capsella* lineages. This footprint of elevated diversity is much more discernible in the two selfing species than in *C. grandiflora*. No obvious reduction in divergence was observed near $ts_{3\text{-}wayhq}$SNPs (Figure 5D). We conclude that $ts_{3\text{-}wayhq}$SNPs correspond predominantly to long-term maintained alleles that diverged on ancient time scales and that they are not the result of recent introgression.

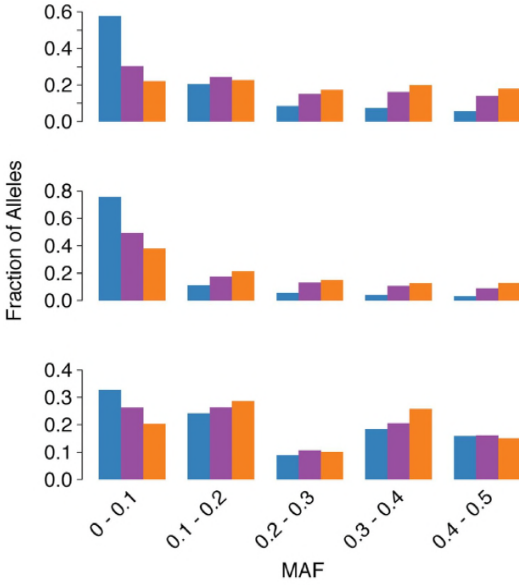


Figure 5 - figure supplement 2. $tr_{3\text{-}wayhq}$SNPs MAF. (A) MAF of ssSNPs (blue), $ts_{2\text{-}way}$SNPs (purple), and $ts_{3\text{-}way}$SNPs (orange) in each species (top: *C. grandiflora*, middle: *C. rubella*, bottom: *C. orientalis*).

The finding of tsSNPs shared between two independent lineages, *C. grandiflora*/*C. rubella* and *C. orientalis*, for over a million generations in spite of strong geographic barriers suggests that they are targets of stable long-term balancing selection. If this selection pressure remains constant across species, ancient alleles are expected to evolve towards similar equilibrium intermediate frequencies. In comparison to $ts_{2\text{-}way}$SNPs, the minor alleles at $ts_{3\text{-}wayhq}$SNP sites are closer to intermediate frequencies in all three species (Figure 5 - figure supplement 2). Furthermore, $ts_{3\text{-}wayhq}$SNPs segregate at more similar allele frequencies in *C. rubella* and *C. grandiflora* than other two-way tsSNPs, as measured by F_{st} median values: 0.03 for $ts_{3\text{-}wayhq}$SNPs and 0.16 for $ts_{2\text{-}way}$SNPs, $p << 0.001$ Mann-Whitney test) and correlation of derived allele frequencies (Figure 4 - figure supplement 3).

These results suggest a conserved equilibrium maintained since the isolation of *C. rubella* and *C. grandiflora* over 10,000 generations ago. Derived allele frequencies for $ts_{3\text{-}wayhq}$SNPs are not correlated between the two ancient *Capsella* lineages (Spearman's rho -0.08 *C. orientalis* to *C. grandiflora* and -0.04 to *C. rubella*). It is possible that demographic reduction or habitat shift in *C. orientalis* has disturbed this equilibrium.

Like ts_{CgCr}SNPs, $ts_{3\text{-}way}$SNPs are strongly enriched in GO categories associated with immunity (Supplementary file 2). Our previously identified balanced regions strongly predicted the genomic distribution of $ts_{3\text{-}way}$SNPs; 50% of $ts_{3\text{-}wayhq}$SNPs fell into these regions, even though they encompass fewer than 10% of ts_{CgCr}SNPs and fewer than 3% of all SNPs , resulting in an even more skewed and uneven distribution of genetic diversity along the genome (Figure 5F-G). At least one $ts_{3\text{-}wayhq}$SNP was found in each of 10 of the 21 original candidate regions under balanced selection. Six of these corresponded to NLR clusters, two to RLK / RLP clusters, and one to a TIR-X cluster. Only one region did not contain a clear immunity candidate, with the caveat that this conclusion is based on the single annotated *C. rubella* reference genome (Slotte et al., 2013). Thus, even in a situation where a recent genetic bottleneck has wiped out almost all genetic diversity, there is very strong selection to maintain allelic diversity at specific immunity-related loci, consistent with these alleles having persisted already for very long evolutionary times.

Insights into balancing selection from *de novo* assembly of *MLO2*

The balanced regions we identified contained very old tsSNPs, yet as mentioned, the immunity genes themselves are often not accessible to variant discovery based on mapping short reads to a single reference genome. Furthermore, it is possible, or even likely, that the strongest evidence for balancing selection comes from loci that include several linked targets of balancing selection. This combination of factors makes it difficult to pinpoint potential functional changes maintained by balancing selection in these regions. To discover functional changes, we therefore focused on $ts_{3\text{-}wayhq}$SNPs that did not fall in our large balanced regions but were clustered in regions of the genome that were likely less complex. We selected genes that were well covered by reads in all three species (>80% sites), contained at least six high quality tsSNPs, at least one non-synonymous $ts_{3\text{-}wayhq}$SNP, were at least 100 kb from any of our candidate balanced regions, and had been functionally characterized in *A. thaliana*. These filters singled out a homolog of the *A. thaliana MLO2* gene as a particularly good candidate for more detailed analysis (Supplementary file 3 and Figure 6).

MLO2 encodes a seven-transmembrane domain protein with a conserved role in plant disease susceptibility (Figure 6A) (Consonni et al., 2006). The *C. rubella MLO2* locus has experienced a tandem duplication, resulting in two genes, *MLO2a* and *MLO2b*. Although both homologs are

sufficiently diverged to be accessible to unambiguous read mapping, all 6 $ts_{3\text{-}wayhq}$SNPs were in *MLO2b* (Figure 6B-C). In *C. rubella* and *C. orientalis*, the $ts_{3\text{-}wayhq}$SNPs were arranged in five different haplotypes, which we collapsed into three related haplogroups, A, B and C (Figure 6B). The reference haplogroup A was most frequent in both species.

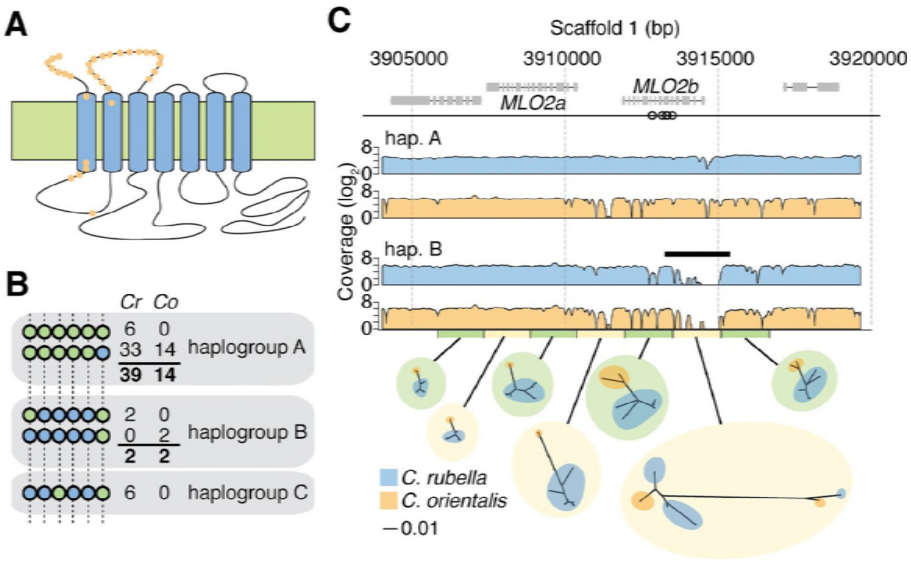


Figure 6. Evidence for long term balancing selection at *MLO2*. (A) Diagram of MLO2 protein in the cell membrane. Blue ovals, transmembrane domains. Top is the extracellular space. Orange dots represent amino acid differences between proteins encoded by haplogroups A and B. (B) Haplogroup identification with reference based SNP calls. Circles represent $ts_{3\text{-}wayhq}$SNPs and colors represent the reference (green) and alternative (blue) SNP calls. Numbers indicate haplotype frequencies in each species. (C) Top: A diagram of the *MLO2* region on scaffold 1. Grey boxes represent coding regions. Empty circles show the positions of the seven initially identified $ts_{3\text{-}wayhq}$SNPs shown in (B). *MLO2b* gene is drawn based on the reference annotation, but alignment with orthologous genes suggested a misannotation of the last splice site acceptor leading to truncation of the annotated gene. For final alignments, the corrected annotation was used. Middle: Average read coverage by haplogroup and species (blue is *C. rubella* and orange is *C. orientalis*). Region of poor coverage in haplogroup B is highlighted with a black bar. The green and yellow bars below the coverage plots highlight the *de novo* assembled region and the windows from which the neighbor joining trees were built (excluding indels, each window is 1 kb). The blue and orange circles on the tree indicate samples from each species. Black scale indicates substitutions in trees.

Because several known targets of balancing selection in *A. thaliana* are the result of structural variation (Mauricio et al., 2003; Stahl et al., 1999), we examined coverage patterns around the *MLO2* locus to identify potential linked indels. We found that haplogroup B in both *C. rubella* and *C. orientalis* exhibited similar patterns of low read coverage at the 5' end of *MLO2b*, suggesting a possible indel (Figure 6C). To examine the exact sequence of each allele, we took advantage of the homozygous nature of sequence data from these two selfing species and performed local *de novo* assembly of the *MLO2* locus from read pairs mapping to this region. We were able to reconstruct the locus for 15 *C. orientalis* samples (13 haplogroup A and 2 haplogroup B) and 40 *C. rubella* samples (31 A, 2 B, and 7 C). Surprisingly, a comparison of the different haplotypes revealed that the pattern of low coverage observed for haplogroup B was not due to structural variation, but instead to extremely high divergence from the reference haplogroup A (Figure 6 - figure supplement 1). Divergence between alleles within species was greater than 0.15 differences per bp, over 3 times higher than the genome-wide divergence between the species (Figure 6 - figure supplement 1 and Figure 5 - figure supplement 1). This highly diverged region had therefore been originally inaccessible

to reference-based read mapping in haplogroup B samples. *De novo* assembly allowed us to identify a total of 204 additional tsSNPs, nearly all of which mapped to the 5' end of *MLO2b* (Figure 6 - figure supplement 1). Neighbor joining trees revealed the expected clustering of samples by species in regions adjacent to *MLO2b*, but clear clustering by haplogroup within the 5' region (Figure 6C). Importantly, divergence within haplogroup across species was similar for both A and B, demonstrating that recent introgression did not give rise to allele sharing (Figure 6 - figure supplement 1).

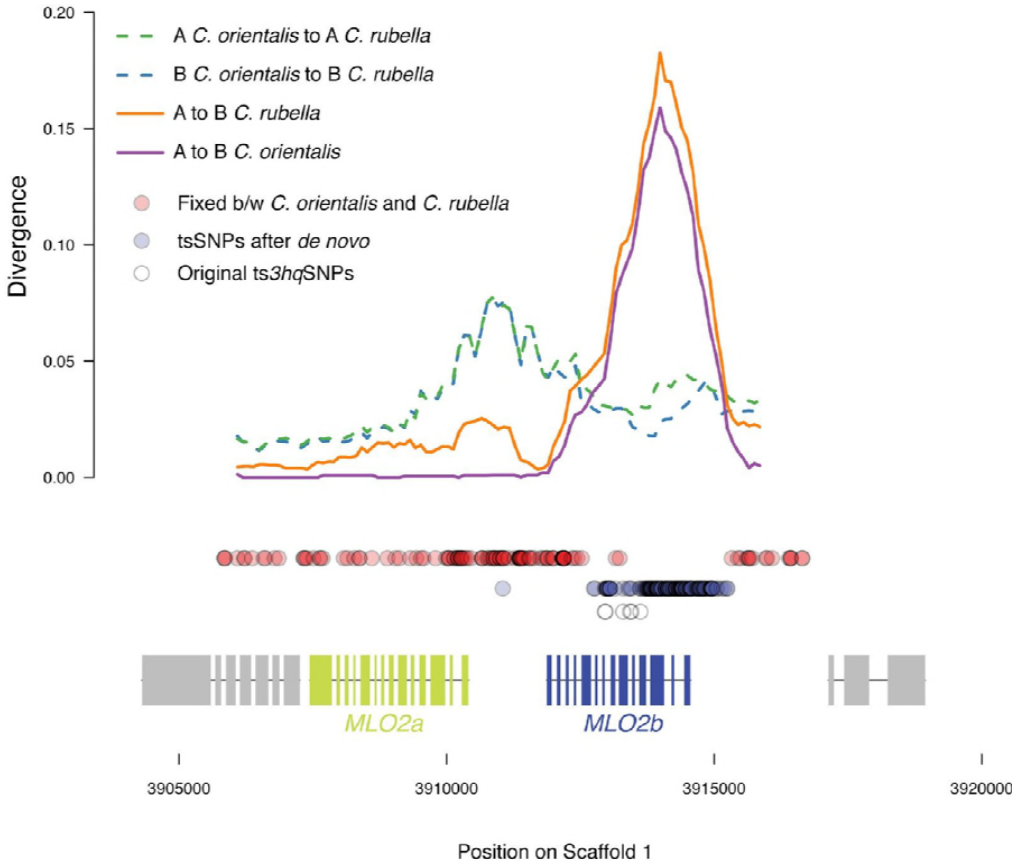


Figure 6 - figure supplement 1. Sliding windows of allelic divergence and positions of tsSNPs. Divergence between haplotype A and B within species, or between the same haplotype across species at the *MLO2* locus. Circles indicate positions of SNPs by class.

The high nucleotide divergence between haplogroups A and B translates into numerous amino acid differences in the N terminal half of the encoded proteins. In a 157 amino acid stretch, 31 amino acid differences are found in both species (Figure 6A and figure supplement 2), with an indel polymorphism accounting for another seven amino acid differences. The large number of differences between the two haplogroups makes it difficult to point to any specific change as the target of balancing selection, but it seems likely that the two alleles differ functionally, perhaps reinforced by additional differences in the promoter. In summary, the nucleotide divergence in this region suggests that the *MLO2b* haplogroups are much older than the split between the two species.

Figure 6 - figure supplement 2. Alignment of amino acid sequences at the MLO2 N-terminus. Examples of alleles from each haplogroup in each species are shown. Triangles show fixed amino acid substitutions between the two alleles. The inferred domain structure is shown above the alignments.

Discussion

While balancing selection has long been recognized as an important evolutionary force, its relevance as a major factor shaping genomic variation has remained unclear (Asthana et al., 2005; Charlesworth, 2006; Wiuf et al., 2004b). We have taken advantage of unique demographic situations in two *Capsella* lineages to demonstrate not only that there is pervasive balancing selection at immunity-related loci in this genus, but also that the same alleles are maintained in species that are likely experiencing quite different pathogen pressures. We expect that balancing selection plays a similar role in other taxa, but that its effects are masked by a background of higher neutral genetic diversity and more frequent recombination between balanced sites and linked variants (Charlesworth, 2006; Wiuf et al., 2004b). In addition, the detection of long-term balancing selection is further compounded by very old alleles being less accessible to short read re-sequencing, the dominant mode of variant discovery today. In the two selfing *Capsella* species, the footprints of balancing selection extend for tens of kilobases, greatly impacting diversity of many other genes. While this makes it more difficult to pinpoint the actual selected variants, it greatly improves statistical power to identify regions under balancing selection. This is reminiscent of genome-wide association studies, where extended LD improves statistical power to detect causal regions of the genome but reduces the ability to identify the specific causal variants (Atwell et al., 2010).

The nature of balancing selection acting on the regions we have identified remains to be clarified. Stable balancing selection in self fertilizing species is unlikely to derive from heterozygous advantage, pointing to negative frequency-dependent selection or fluctuating selection from variable pathogen pressures as possible factors. While the mode of selection cannot be determined from these static data, the strong signal that we observe in highly selfing lineages points to environmental heterogeneity or negative frequency dependent selection over heterozygote advantage. Based on the enrichment of immunity-related genes, it appears that biotic factors are the dominant drivers of long-term maintenance of polymorphism. This observation is consistent with a large body of work on intraspecific variation in *A. thaliana*. The signal of balancing selection has been observed for specific pairs of disease resistance alleles in *A. thaliana* (Bakker et al., 2006; Mauricio et al., 2003; Stahl et al., 1999; Tian et al., 2003, 2002), and in the case of the resistance gene *RPS5*, alternative alleles have been shown to affect fitness in the field (Karasov et al., 2014). It is possible, or perhaps even likely,

that the signal of balancing selection is amplified by the fact that immunity-related loci occur in clusters (Meyers et al., 2003) and that our strongest signal is the result of simultaneous selection on several genes in these regions in a situation analogous to the MHC in animals (Hedrick, 1998). Thus, biotic factors might not be quite as important as our analyses make them appear. On the other hand, it is also possible that the clustering of disease resistance genes itself is a product of selection, if selection was more effective when acting on groups of genes (Charlesworth and Charlesworth, 1975), or if evolution under a balanced regime was deleterious at other types of loci. Even if we accept that biotic factors predominate, the nature of the potential trade-offs that prevent individual alleles from becoming fixed is still a mystery, but it might involve conflicts between growth and defense (Coley et al., 1985; Herms and Mattson, 1992; Walling, 2009), beneficial and harmful microbe interactions (Walters and Heil, 2007), or defense against different types of pathogens (Kliebenstein and Rowe, 2008). What is clear is that the trade-offs must be stable over very long periods of evolution.

Our findings suggest a model in which the success of self fertilizing populations may be buoyed by gene flow from outcrossing relatives in a situation analogous to evolutionary rescue strategies in conservation biology (Whiteley et al., 2015). This model is a variation on the theme of adaptive introgressions, which have recently emerged as a major evolutionary force in a wide range of taxa (Castric et al., 2008; Hedrick, 2013; Heliconius Genome Consortium, 2012; Henning and Meyer, 2014; Huerta-Sánchez et al., 2014; Pease et al., 2016; Racimo et al., 2015; Whitney et al., 2006)(Castric et al., 2008; Hedrick, 2013; Heliconius Genome Consortium, 2012; Huerta-Sánchez et al., 2014; Pease et al., 2016; Whitney et al., 2006). The unique feature of self-fertilization in comparison to these examples is that the amplified effects of linked selection and genetic drift lead to a steady loss of genetic variation over time. Constant replenishment via adaptive introgression from an outcrossing relative counters the loss of diversity at immunity-related loci, thereby preventing decreased fitness in competition with pathogens. Whether this model generally applies will require independent study of other lineages of related self-fertilizing and outcrossing populations at various stages of speciation.

Finally, we note that maintenance of ancient variants is most easily detectable in a background of low variation. Therefore, it could potentially be used to rapidly identify loci with meaningful functional variation. Typically, agricultural breeding panels seek to maximize surveyed diversity, but our results indicate that identification of useful immunity-related polymorphism with genomic data might be facilitated in otherwise homogeneous wild populations.

MATERIALS AND METHODS

Plant Material and DNA extraction

Seeds were stratified for two weeks at 4°C and germinated in controlled environment chambers. Four to six rosette leaves were collected from each accession and frozen in liquid nitrogen for gDNA extraction. The methods available for extraction and sequencing varied as the project progressed, and 24 of the *C. rubella* and the 13 *C. grandiflora* samples were analyzed independently in previous studies (Agren et al., 2014; Williamson et al., 2014). See Table S1 for a listing of DNA preparation, library construction, and sequencing technology by sample. In brief, DNA was extracted following an abbreviated nuclei enrichment protocol (Becker et al., 2011) or using the Qiagen Plant DNeasy Extraction kit. The recovered DNA was sheared to the desired length using a Covaris S220 instrument, and Illumina sequencing libraries were prepared using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs) or the Illumina TruSeq DNA Library Preparation Kit and sequenced on the instrument as listed in Table S1. We aimed for a minimum genome coverage of 40x. We mapped reads to the *C. rubella* reference genome (Slotte et al., 2013) resulting in realized coverages of 30 – 126x.

Sequence handling and variant calling

Initial sequence read processing, alignment, and variant calling were carried out using the SHORE (v0.8) software package (Ossowski et al., 2008). Read filtering, de-multiplexing, and trimming were accomplished using the import command discarding reads that had low complexity, contained more than 10% ambiguous bases, or were shorter than 75 bp after trimming. Reads were mapped to the *C. rubella* reference genome (Phytozome v.1.0) using the GenomeMapper aligner (Schneeberger et al., 2009) with a maximum edit distance (gaps or mismatches) of 10%. Alignments from each sample were then processed to generate raw whole genome reference and variant calls with qualities computed using an empirical scoring matrix approach (Cao et al., 2011) allowing heterozygous positions. Of the initial 53 *C. rubella* samples, two were removed because of low or uneven coverage, and one was removed as a misidentified *C. bursa-pastoris* sample (*C. rubella* and its polyploid relative *C. bursa-pastoris* are not easily identified phenotypically, but they can be distinguished by the extreme number of pseudo-heterozygous calls in the latter).

The per-sample raw consensus calls produced by SHORE were used to construct a whole genome matrix of finalized genotype calls for each species. Positions were considered only if covered by at

least four reads and if overlapping reads mapped uniquely (GenomeMapper applies a "best match" approach, so unique means that only one best match exists) (Schneeberger et al., 2009). We simultaneously considered information from all samples within a species to make base pair calls. If no variant was called in any sample then the site was treated as reference. Individual sample calls were made if four reads supported the reference base, the computed quality was above 24, and at least 80% of reads supported a reference call. A site was excluded if more than 30% of the samples from that species did not meet these criteria.

If at least one sample reported a difference from the reference in the raw consensus, then variant (indel or SNP) or reference calls were considered. The SNP calling parameters were slightly different for the two selfing species as compared to the outcrossing *C. grandiflora* because variants should only rarely be found in the heterozygous state in the former (and the frequency of heterozygous calls in a selfing species is a powerful filter to detect problems with mismapped reads). The general approach was to require at least one high quality variant call at a site and then to call genotypes in other samples with slightly reduced stringency. If no variant call met the more stringent threshold, then the site was reconsidered using the above reference criteria. Finally, the calls from each of the three species were combined into a master matrix. If a position was not called biallelic or invariant across the compared species, then it was not considered. To facilitate further analyses in PLINK (v1.9) (Chang et al., 2015) and vcftools (v0.1.12a) (Danecek et al., 2011), the genome matrix at biallelic SNP sites was also converted into a minimal vcf format.

Defining pericentromeric regions

Regions of high repeat density near the centromeres of all chromosomes as well as two large, repeat-rich regions in chromosomes 1 and 7 were removed from genome scans. Coordinates for these regions are listed in Supplementary file 4.

Site annotations

We used the SnpEff (v.3.2a) (Cingolani et al., 2012) software package to annotate variant and invariant sites for the whole genome. The annotation database was built using the *C. rubella* v1.0 Phytozome gff file. Sites were annotated using the table input function that includes annotation of fold degeneracy for each site in coding regions. Invariant sites were annotated using a table with dummy SNPs at each position. The SnpEff program outputs several annotations for some sites, and a primary annotation was selected by ranking the strength of effect of each annotation and reporting the annotation with the strongest effect (the rankings are listed in Supplementary file 5).

Ancestral state assignment

To calculate derived allele frequency spectra we assigned ancestral state to each polymorphic site using three-way whole genome alignments between *C. rubella*, *A. thaliana*, and *A. lyrata* (Slotte et al., 2013). Only biallelic sites identical between *A. lyrata* and *A. thaliana* (indels were ignored) were considered. For the two species analysis, only sites also fixed for the ancestral allele in *C. orientalis* were considered.

Trans-specific SNP annotation comparisons

To compare tsSNP and ssSNP annotations from similar allele frequency spectra, we binned 20,000 tsSNPs randomly drawn from throughout the genome by derived allele frequency (10 bins). We then drew an equivalent number of ssSNPs from each allele frequency bin and calculated the fraction of CDS SNPs that caused nonsynonymous changes and the fraction that fell in genes. This process was repeated 1,000 times for both species to generate the plots shown in Figure 3B.

Analysis of population structure and demographic modeling

Genotypes at four-fold degenerate SNP sites called in *C. grandiflora* and *C. rubella* were pruned in PLINK (50 kb windows, 5 kb step, and 0.2 r2 LD threshold) and used as input for ADMIXTURE (v.1.23) (Alexander et al., 2009) and EIGENSTRAT (v6.0 beta) (Price et al., 2006). For demographic modelling in Fastsimcoal (v2.5.2.11) (Excoffier et al., 2013), joint minor allele frequency spectra were generated at four-fold degenerate sites with complete information and ignoring heterozygous calls in selfing lineages (counting only one allele from each individual). Demographic parameters for each tested model were then inferred in 50 runs of Fastsimcoal (parameters: -l40 -L40 -n100000 -N100000 -M0.001 -C5). The global maximum likelihood model was selected after correcting for number of estimated parameters using Akaike Information Criterion. Confidence intervals were set for estimated parameters using 100 bootstraps of identical inference runs on simulated data under the most likely model. To reduce computational times, global maximum likelihoods were calculated for bootstraps after 13 runs rather than 50. The mutation rate assumed for this and other analyses was 7×10^{-9} mutations / generation / bp based on mutation rate measurements in *Arabidopsis thaliana (Ossowski et al., 2010)*.

Segments of recent ancestry and interspecific introgression

Segments of IBD were identified using the phasing and segment identification in Beagle (r1339) (Browning and Browning, 2013). For the analysis presented here, we considered only the first haplotype from each *C. rubella* sample and both haplotypes from each *C. grandiflora* sample. D statistics were calculated as in (Green et al., 2010; Heliconius Genome Consortium, 2012; Patterson et al., 2012) comparing each individual genotype from the eastern *C. rubella* population to allele frequencies from western *C. rubella* and *C. grandiflora*.

Sliding window analysis of genetic diversity

Population genetic diversity statistics for genome scans were calculated for each species by transforming variant calls from the genome matrix into FASTA files and inputting these files into the compute function from the libsequence analysis package (Thornton, 2003). Heterozygous bases were randomly assigned as reference or variant to generate a single haplotype for each sample. Weir and Cockerham's F_{st} was calculated using vcftools (v.0.1.12a) on biallelic SNP sites.

Identification of balanced regions

To identify regions of the genome with unusually low F_{st} after speciation, we generated a null distribution of F_{st} values by simulating one million 20 kb segments under our inferred best demographic model using Fastsimcoal2. The output of each simulation was transformed to vcf format and F_{st} between *C. grandiflora* and each *C. rubella* subpopulation was calculated using vcftools. The probability of a particular F_{st} value in the observed data was then assigned based on its rank in these simulations (independently for the two subpopulations; one sided test). Multiple testing was accounted for using Bonferroni correction. Significant outlier windows (adjusted p-value < 0.05) identified for each subpopulation were collapsed into regions using a two state hidden markov-model as implemented in the Rhmm package. The HMM approach has the advantage of joining windows of high coverage separated by a low coverage window. Only regions significant in both subpopulations were considered for further analysis. Windows overlapping the pericentromeric regions were removed from the analysis.

Linkage disequilibrium

LD was calculated in 30 kb windows in *C. grandiflora* and *C. rubella* using PLINK (v.1.9). The decay of LD is the mean value at each position up to 30 kb from a focal SNP.

Gene ontology (GO) enrichment

Because the *C. rubella* annotation is sparse, we used annotations from nucleotide blast best hit matches (e < 1e-10) to CDS sequences from its close relative, *A. thaliana,* for our GO analysis. Enrichment tests were performed with the SNP2GO R library (Szkiba et al., 2014) using ts_{CgCr}SNPs as the test set and all SNP sites called in either *C. rubella* or *C. grandiflora* as the background set. We chose this approach because it is less sensitive to gene length (which should similarly affect tsSNP and non-tsSNP distributions across genes). A corresponding analysis was performed in the three-way comparisons using a background set of all SNP sites called in all three species. Significant enrichments were considered at a q-value threshold of $q < 0.01$ after false discovery correction. A gene was considered as belonging to the NLR family in *C. rubella* if its best blast hit in *A. thaliana* was annotated as such (Supplementary file 6).

Identification of high quality three-way tsSNPs

To generate a list of high quality $ts_{3\text{-way}}$SNPs, we applied a series of empirical filters. First, all $ts_{3\text{-way}}$SNPs were required to have an $r^2 > 0.2$ with another $ts_{3\text{-wayhq}}$SNP in the same phase in all three species. We excluded SNPs overlapping pericentromeric or annotated repeat sequences (Slotte et al., 2013). We also required that the coverage of SNPs was no more than two standard deviations above the mean coverage of all SNPs for that species, to have an average concordance greater than 0.98, and to be identified in more than one individual. These criteria were selected to increase our confidence in identified tsSNPs; it is likely that our inferences are conservative.

To validate our trans-specific SNPs we aligned the *C. orientalis* samples against the draft *C. orientalis* assembly using the bwa (v.0.7.12) mem command with default parameters. The output bam format file was sorted using samtools (v.1.6) and multisample variant calls were made with freebayes (v.1.1.0) using the parameter settings -z .1 -0 -w. The resulting vcf file was filtered using vcftools (v.0.1.13) using the settings --remove-indels --minQ 50 --max-missing 0.8 --max-alleles 2 and further filtered to remove sites that were called as heterozygous in more than 5% of the samples. The sites overlapping with the original call set were extracted from this vcf and used for validation.

Coordinate transforms between the two genomes were necessary to validate tsSNPs. The draft assembly of *C. orientalis* and the *C. rubella* reference genome were aligned using the LAST (v.923) aligner. The *C. rubella* reference database was built with the lastdb command with the parameter settings -uMAM8 -cR11, and then the two genomes were aligned with the lastal command with the settings -m50 -E0.05. Equivalent sites were considered if they were present in alignments at least 500 bp long and contained only one *C. orientalis* and one *C. rubella* sequence.

Local *de novo* assembly and analysis of *MLO2*

To reconstruct alleles from the *MLO2* locus, we used an iterative assembly approach. Reads were first mapped to the entire reference genome using bwa (v.0.7.8) (Li and Durbin, 2009) using the bwa-mem alignment algorithm for each sample. Reads that mapped to the *MLO2* locus were then extracted and assembled de novo using SPAdes (v.3.5.0) (Nurk et al., 2013). Assemblies were filtered to be longer than 2,000 bp with a coverage greater than 5, and then used to create an index for a second round of read mapping. Reads that mapped to the assembly without mismatches were collected together with their mates (regardless of the mate's mapping quality), and were again *de novo* assembled. This process was iterated six times until scaffolds covering both coding regions were achieved. Format conversions and file handling made use of the software samtools (v.0.1.19) (Li et al., 2009) and bamutil (v.1.0.13) .

Assemblies were filtered for appropriate length, and aligned using MAFFT (Katoh and Standley, 2013). Alignments were visualized using AliView (Larsson, 2014), and manually edited where appropriate. The protein encoded by *MLO2b* annotated in the *C. rubella* reference was truncated relative to *A. thaliana MLO2*. We aligned the genomic and coding regions from both species and found that the premature stop in *MLO2b* is likely due to a mis-annotated splice junction. The *A. thaliana* junction is conserved in *C. rubella* and alternative annotations on phytozome identify the *A. thaliana*-like splice variant. We therefore used the full-length version derived from manual alignments for our analysis. To determine where amino acid substitutions had occurred, we aligned the proteins encoded by each allele against the barley *mlo* protein and annotated domains (UniProtKB P3766).

Draft assembly of the *C. orientalis* genome

The draft genome from the *C. orientalis* accession 2007-03 (Figure 1 - figure supplement 2) was assembled from long reads generated by PacBio single-molecule real-time sequencing. Long reads were assembled with Falcon (Chin et al., 2016) (version 0.5.4, max_diff = 150, max_cov = 150, min_cov = 2). The resulting primary contig set was iteratively polished with Quiver again using long

reads (Chin et al., 2013) (version 2.0.0) and with Pilon (Walker et al., 2014) (version 1.16) using short reads from a single Illumina TruSeq DNA PCR-free library. The draft genome of *C. orientalis* comprises 135 Mb distributed over 423 gap-free contigs and covers 60% of the *C. rubella* reference with non-ambiguous 1-to-1 whole genome alignments. Its completeness is comparable to that of the *C. rubella* reference.

ACKNOWLEDGEMENTS

We thank Christa Lanz for expert assistance with Illumina sequencing. We thank Danelle Seymour, Rebecca Schwab, Beth Rowan, Derek Lundberg, Wangsheng Zhu, Efthymia Symeonidi, Gautam Shirsekar, Rui Wu, Patricia Lang, Talia Karasov, Hernán Burbano, Moisés Exposito Alonso, Maricris Zaidem, Rafal Gutaker, Eunyoung Chae, and Diep Tran for reading of the manuscript and insightful comments. Thank you to Dmitry German for his identification of *C. orientalis* from herbarium samples, making this study possible in the first place. This work was supported by a Human Frontiers Science Program Long-Term Fellowship to D.K. (LT000783/2010-L) and by DFG-SPP1529 ADAPTOMICS (WE 2897/4-2), ERC Advanced Grant IMMUNEMESIS and the Max Planck Society (D.W.).

REFERENCES

1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell* **166**:481–491.

Agren JÅ, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI. 2014. Mating system shifts and transposable element evolution in the plant genus *Capsella*. *BMC Genomics* **15**:602.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**:1655–1664.

Asthana S, Schmidt S, Sunyaev S. 2005. A limited role for balancing selection. *Trends Genet* **21**:30–32.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**:627–631.

Bachmann JA, Tedder A, Laenen B, Fracassetti M, Désamoré A, Lafon-Placette C, Steige KA, Callot C, Marande W, Neuffer B, Bergès H, Köhler C, Castric V, Slotte T. 2018. Genetic basis and timing of a major mating system shift in Capsella. *bioRxiv*. doi:10.1101/425389

Bakker EG, Toomajian C, Kreitman M, Bergelson J. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* **18**:1803–1818.

Bechsgaard J, Jorgensen TH, Schierup MH. 2017. Evidence for Adaptive Introgression of Disease Resistance Genes Among Closely Related Arabidopsis Species. *G3* . doi:10.1534/g3.117.043984

Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* **480**:245–249.

Bergelson J, Kreitman M, Stahl EA, Tian D. 2001. Evolutionary dynamics of plant R-genes. *Science* **292**:2281–2285.

Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, Beynon JL, Daniels MJ, Holub EB, Jones JD. 1998. Three genes of the Arabidopsis RPP1 complex resistance locus recognize distinct Peronospora parasitica avirulence determinants. *Plant Cell* **10**:1847–60.

Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G. 2013. Genomic identification of founding haplotypes reveals the history of the selfing species *Capsella rubella*. *PLoS Genet* **9**:e1003754.

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**:459–471.

Caicedo AL, Schaal BA, Kunkel BN. 1999. Diversity and molecular evolution of the RPS2 resistance gene in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **96**:302–306.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**:956–963.

Castric V, Bechsgaard J, Schierup MH, Vekemans X. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet* **4**:e1000168.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**:7.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* **2**:e64.

Charlesworth D, Charlesworth B. 1975. Theoretical genetics of Batesian mimicry II. Evolution of supergenes. *J Theor Biol* **55**:305–324.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**:563–569.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**:1050–1054.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly* **6**:80–92.

Coley PD, Bryant JP, Chapin FS 3rd. 1985. Resource availability and plant antiherbivore defense. *Science* **230**:895–899.

Consonni C, Humphry ME, Hartmann HA, Livaja M, Durner J, Westphal L, Vogel J, Lipka V, Kemmerling B, Schulze-Lefert P, Somerville SC, Panstruga R. 2006. Conserved requirement for a plant host cell protein in powdery mildew pathogenesis. *Nat Genet* **38**:716–720.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158.

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genet* **10**:e1004561.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Ågren JA, Hazzouri KM, Wang W, Platts AE, Williamson RJ, Neuffer B, Lascoux M, Slotte T, Wright SI. 2015. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc Natl Acad Sci U S A* **112**:2806–2811.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28**:2239–2252.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**:e1003905.

Falahati-Anbaran M, Lundemo S, Stenøien HK. 2014. Seed dispersal in time can counteract the effect of gene flow between natural populations of *Arabidopsis thaliana*. *New Phytol* **202**:1043–1054.

Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Mol Ecol* **24**:3529–3545.

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. *Proc Natl Acad Sci U S A* **106**:5241–5245.

Gassmann W, Hinsch ME, Staskawicz BJ. 1999. The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J* **20**:265–277.

Goritschnig S, Krasileva KV, Dahlbeck D, Staskawicz BJ. 2012. Computational prediction and molecular characterization of an oomycete effector and the cognate *Arabidopsis* resistance gene. *PLoS Genet* **8**:e1002502.

Gos G, Slotte T, Wright SI. 2012. Signatures of balancing selection are maintained at disease resistance loci following mating system evolution and a population bottleneck in the genus *Capsella*. *BMC Evol Biol* **12**:152.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S. 2010. A Draft Sequence of the Neandertal Genome. *Science* **328**:710–722.

Guo Y-L, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proc Natl Acad Sci U S A* **106**:5246–5251.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol* **22**:4606–4618.

Hedrick PW. 1998. Balancing selection and MHC. *Genetica* **104**:207–214.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**:94–98.

Henning F, Meyer A. 2014. The evolutionary genomics of cichlid fishes: explosive speciation and adaptation in the postgenomic era. *Annu Rev Genomics Hum Genet* **15**:417–441.

Herms DA, Mattson WJ. 1992. The dilemma of plants: to grow or defend. *Q Rev Biol* **67**:283–335.

Holub EB, Beynon JL, Crute IR. 1994. Phenotypic and genotypic characterization of interactions between isolates of *Peronospora parasitica* and accessions of *Arabidopsis thaliana*. *MPMI-Molecular Plant Microbe Interactions* **7**:223–239.

Huard-Chauveau C, Perchepied L, Debieu M, Rivas S, Kroj T, Kars I, Bergelson J, Roux F, Roby D. 2013. An atypical kinase under balancing selection confers broad-spectrum disease resistance in Arabidopsis. *PLoS Genet* **9**:e1003766.

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang, Luosang J, Cuo ZXP, Li K, Gao G, Yin Y, Wang W, Zhang X, Xu X, Yang H, Li Y, Wang J, Wang J, Nielsen R. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**:194–197.

Hurka H, Friesen N, German DA, Franzke A, Neuffer B. 2012. “Missing link”species *Capsella orientalis* and *Capsella thracica* elucidate evolution of model plant genus *Capsella* (Brassicaceae). *Mol Ecol* **21**:1223–1238.

Hurka H, Neuffer B. 1997. Evolutionary processes in the genus Capsella (Brassicaceae). *Plant Syst Evol* **206**:295–316.

Jones JDG, Dangl JL. 2006. The plant immune system. *Nature* **444**:323–329.

Karasov TL, Kniskern JM, Gao L, DeYoung BJ, Ding J, Dubiella U, Lastra RO, Nallu S, Roux F, Innes RW, Barrett LG, Hudson RR, Bergelson J. 2014. The long-term maintenance of a resistance

polymorphism through diffuse interactions. *Nature* **512**:436–440.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772–780.
Kliebenstein DJ, Rowe HC. 2008. Ecological costs of biotrophic versus necrotrophic pathogen resistance, the hypersensitive response and signal transduction. *Plant Sci* **174**:551–556.
Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large data sets. Bioinformatics btu531.
Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P. 1988. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature* **335**:268–271.
Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, Donnelly P, McVean G, Przeworski M. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**:1578–1582.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.
Mauricio R, Stahl EA, Korves T, Tian D, Kreitman M, Bergelson J. 2003. Natural selection for polymorphism in the disease resistance gene *RPS2* of *Arabidopsis thaliana*. *Genetics* **163**:735–746.
Mayer WE, Jonker M, Klein D, Ivanyi P, van Seventer G, Klein J. 1988. Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. *EMBO J* **7**:2765–2774.
McConnell TJ, Talbot WS, McIndoe RA, Wakeland EK. 1988. The origin of MHC class II gene polymorphism within the genus Mus. *Nature* **332**:651–654.
McDowell JM, Cuzick A, Can C, Beynon J, Dangl JL, Holub EB. 2000. Downy mildew (*Peronospora parasitica*) resistance genes in *Arabidopsis* vary in functional requirements for NDR1, EDS1, NPR1 and salicylic acid accumulation. *Plant J* **22**:523–529.
McDowell JM, Dhandaydham M, Long TA, Aarts MG, Goff S, Holub EB, Dangl JL. 1998. Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of *Arabidopsis*. *Plant Cell* **10**:1861–1874.
Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. 2003. Genome-wide analysis of NBS-LRR encoding genes in *Arabidopsis*. *Plant Cell* **15**:809–834.
Noël L, Moores TL, van Der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JD. 1999. Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. *Plant Cell* **11**:2099–2112.
Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, Holm S, Säll T, Schlötterer C, Marhold K, Widmer A, Sese J, Shimizu KK, Weigel D, Krämer U, Koch MA, Nordborg M. 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet*. doi:10.1038/ng.3617
Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Others. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads, p 158--170. *Research in computational molecular biology Springer Verlag, Berlin, Germany*.
Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18**:2024–2033.
Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**:92–94.
Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* **192**:1065–1093.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS Biol* **14**:e1002379.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**:904–909.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet* **16**:359–371.

Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, vonHoldt BM, Marsden CD, Lohmueller KE, Wayne RK. 2016. Genomic Flatlining in the Endangered Island Fox. *Curr Biol* **26**:1183–1189.

Rose LE, Bittner-Eddy PD, Langley CH, Holub EB, Michelmore RW, Beynon JL. 2004. The maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in Arabidopsis thaliana. *Genetics* **166**:1517–1527.

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D. 2009. Simultaneous alignment of short reads against multiple genomes. *Genome Biol* **10**:R98.

Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, Moyse J, Ross S, Gamble K, Sella G, Ober C, Przeworski M. 2012. The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences* **109**:18493–18498.

Sicard A, Kappel C, Josephs EB, Lee YW, Marona C, Stinchcombe JR, Wright SI, Lenhard M. 2015. Divergent sorting of a balanced ancestral polymorphism underlies the establishment of gene-flow barriers in *Capsella*. *Nat Commun* **6**:7960.

Sicard A, Stacey N, Hermann K, Dessoly J, Neuffer B, Bäurle I, Lenhard M. 2011. Genetics, evolution, and adaptive significance of the selfing syndrome in the genus *Capsella*. *Plant Cell* **23**:3156–3171.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, Wang W, Mandáková T, Vello E, Smith LM, Henz SR, Steffen J, Takuno S, Brandvain Y, Coop G, Andolfatto P, Hu TT, Blanchette M, Clark RM, Quesneville H, Nordborg M, Gaut BS, Lysak MA, Jenkins J, Grimwood J, Chapman J, Prochnik S, Shu S, Rokhsar D, Schmutz J, Weigel D, Wright SI. 2013. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45**:831–835.

Slotte T, Hazzouri KM, Stern D, Andolfatto P, Wright SI. 2012. Genetic architecture and adaptive significance of the selfing syndrome in *Capsella*. *Evolution* **66**:1360–1374.

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.

St. Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE. 2011. Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. *Mol Ecol* **20**:3306–3320.

Szkiba D, Kapun M, von Haeseler A, Gallach M. 2014. SNP2GO: functional analysis of genome-wide association studies. *Genetics* **197**:285–289.

Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, Nickel B, Fischer A, Halbwax M, Andre C, Atencia R, Meyer M, Parra G, Pääbo S, Andrés AM. 2015. Long-Term Balancing Selection in *LAD1* Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. *Mol Biol Evol* **32**:1186–1196.

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.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**:2325–2327.

Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002. Signature of balancing selection in *Arabidopsis*. *Proc Natl Acad Sci U S A* **99**:11525–11530.

Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J. 2003. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. *Nature* **423**:74–77.

Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, Kuhns C, Sureshkumar S, Schwartz C, Lanz C, Laitinen RA, Huang Y, Chory J, Lipka V, Borevitz JO, Dangl JL, Bergelson

J, Nordborg M, Weigel D. 2010. Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature* **465**:632–636.

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**:1157–1165.

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.

Walker M, Johnsen S, Rasmussen SO, Popp T, Steffensen J-P, Gibbard P, Hoek W, Lowe J, Andrews J, Björck S, Cwynar LC, Hughen K, Kershaw P, Kromer B, Litt T, Lowe DJ, Nakagawa T, Newnham R, Schwander J. 2009. Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxiliary records. *J Quat Sci* **24**:3–17.

Walling LL. 2009. Chapter 13 Adaptive Defense Responses to Pathogens and InsectsAdvances in Botanical Research. Academic Press. pp. 551–612.

Walters D, Heil M. 2007. Costs and trade-offs associated with induced resistance. *Physiol Mol Plant Pathol* **71**:3–17.

Wang J, Zhang L, Li J, Lawton-Rauh A, Tian D. 2011. Unusual signatures of highly adaptable R-loci in closely-related *Arabidopsis* species. *Gene* **482**:24–33.

Watkins DI, Chen ZW, Hughes AL, Evans MG, Tedder TF, Letvin NL. 1990. Evolution of the MHC class I genes of a New World primate from ancestral homologues of human non-classical genes. *Nature* **346**:60–63.

Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA. 2015. Genetic rescue to the rescue. *Trends Ecol Evol* **30**:42–49.

Whitney KD, Randell RA, Rieseberg LH. 2006. Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *Am Nat* **167**:794–807.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet* **10**:e1004622.

Wiuf C, Zhao K, Innan H, Nordborg M. 2004a. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* **168**:2363–2372.

Wiuf C, Zhao K, Innan H, Nordborg M. 2004b. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* **168**:2363–2372.

Wright S i., Ness R w., Foxe JP, Barrett S c. h. 2008. Genomic consequences of outcrossing and selfing in plants. *Int J Plant Sci* **169**:105–118.

Wu J, Saupe SJ, Glass NL. 1998. Evidence for balancing selection operating at the *het-c* heterokaryon incompatibility locus in a group of filamentous fungi. *Proc Natl Acad Sci U S A* **95**:12398–12403.

Xu X, Chen C, Fan B, Chen Z. 2006. Physical and functional interactions between pathogen-induced *Arabidopsis* WRKY18, WRKY40, and WRKY60 transcription factors. *Plant Cell* **18**:1310–1326.

Yeh Y-H, Chang Y-H, Huang P-Y, Huang J-B, Zimmerli L. 2015. Enhanced *Arabidopsis* pattern-triggered immunity by overexpression of cysteine-rich receptor-like kinases. *Front Plant Sci* **6**:322.

Zhang L, Kars I, Essenstam B, Liebrand TWH, Wagemakers L, Elberse J, Tagkalaki P, Tjoitang D, van den Ackerveken G, van Kan JAL. 2014. Fungal endopolygalacturonases are recognized as microbe-associated molecular patterns by the *arabidopsis* receptor-like protein RESPONSIVENESS TO BOTRYTIS POLYGALACTURONASES1. *Plant Physiol* **164**:352–364.

Zhang W, Fraiture M, Kolb D, Löffelhardt B, Desaki Y, Boutrot FFG, Tör M, Zipfel C, Gust AA, Brunner F. 2013. Arabidopsis receptor-like protein30 and receptor-like kinase suppressor of BIR1-1/EVERSHED mediate innate immunity to necrotrophic fungi. *Plant Cell* **25**:4227–4241.

Zipfel C. 2008. Pattern-recognition receptors in plant innate immunity. *Curr Opin Immunol* **20**:10–16.

SUPPLEMENTARY FILES

Supplementary file 1. GO enrichment analysis of tsSNPs.

Supplementary file 2. GO enrichment for $tr_{3\text{-}way}$SNPs.

Supplementary file 3. List of well covered genes for targeted analysis of potential balancing selection.

Supplementary file 4. Pericentromeric or repeat dense genomic regions filtered in genome scans.

Supplementary file 5. Annotation hierarchies for SNPs with multiple annotations.

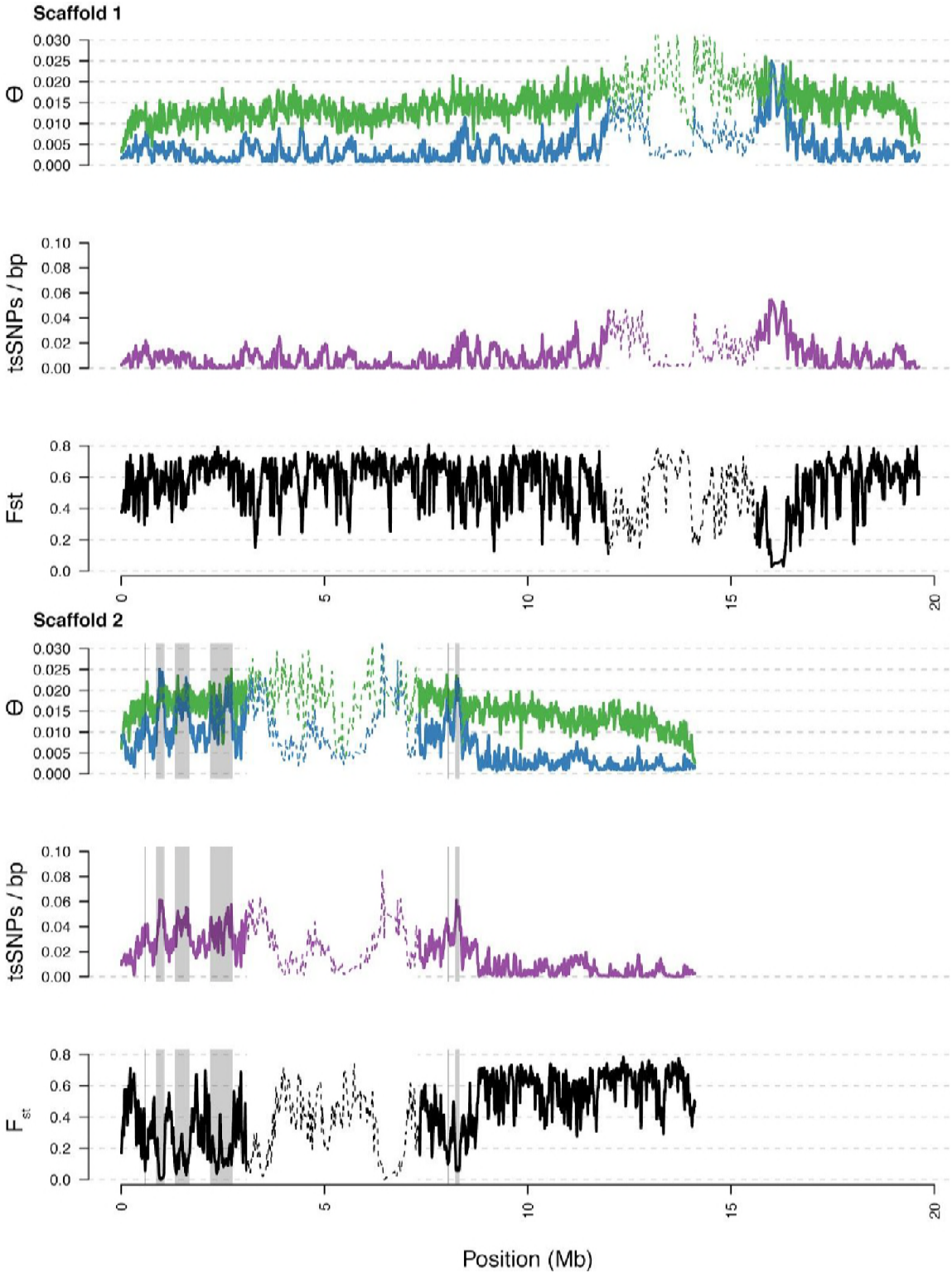
Supplementary file 6. List of *A. thaliana* NLR genes used for ortholog identification.

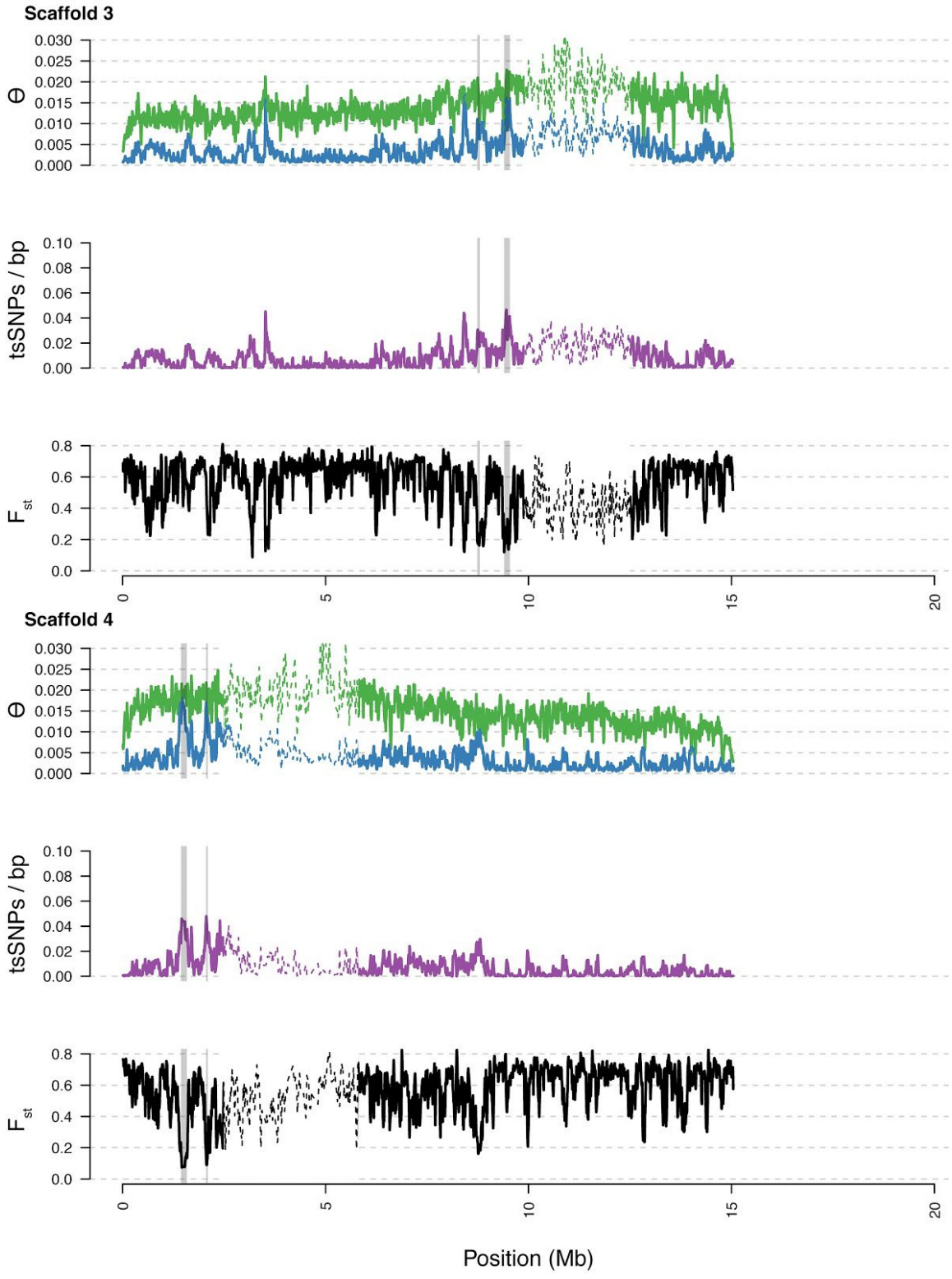
SUPPLEMENTS NOT SHOWN IN MAIN TEXT

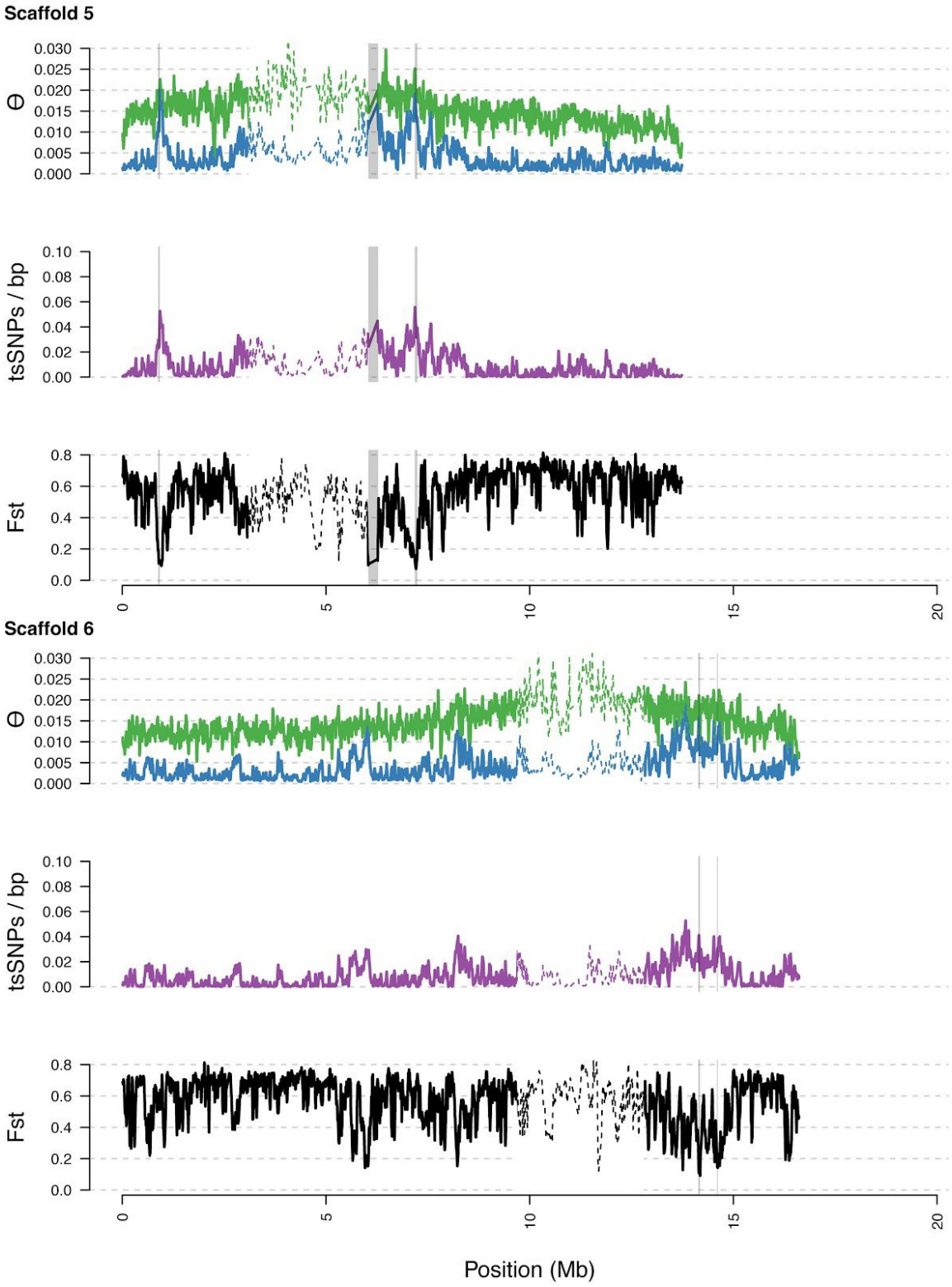
Species	Accession	Latitude	Longitude	Covered Sites	Fraction of genome	Average Coverage	Machine	Read length	Run type	DNA extraction method
Capsella rubella	690	36.15	-5.58	126001336	96.43%	45.52	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	697	-	-	123161423	94.25%	15.34	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	698	-	-	124962254	95.63%	34.85	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	762	37.966667	23.716667	122513726	93.76%	53.27	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	844	35.202554	24.233091	121329288	92.85%	81.83	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	879	35.29	24.42	123690648	94.66%	53.44	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	907	39.666667	19.8	121421409	92.92%	68.71	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	925	39.666667	20.85	123737895	94.70%	51.14	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	984	39.5	3	124755181	95.47%	38.65	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1207	28.316667	-16.566667	124326589	95.15%	114.33	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1208	28.316667	-16.566667	124706029	95.44%	58.39	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1209	28.316667	-16.566667	124491196	95.27%	79.04	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1311	42.8833333	-0.1	124745870	95.47%	48.70	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1377	-34.6666667	-58.5	123919991	94.83%	61.13	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1453	43.466667	11.033333	124592206	95.35%	43.84	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	774	41.8333333	16	129582019	99.17%	54.77	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	86IT1	40.62	14.37	124745454	95.47%	31.91	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1GR1	-	-	121591883	93.05%	33.64	Illumina GAII	100	Paired end	CTAB
Capsella rubella	1574-1	41.6	8.983333	124175014	95.03%	32.41	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1407-8	35.183333	24.233333	123498038	94.51%	73.48	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1504-11	28.666667	-17.866667	124832658	95.53%	60.99	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1575-1	41.383333	9.166667	125029339	95.68%	39.96	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1249-11	38.7833333	-9.3833333	124926266	95.60%	40.99	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1267-15	37.933333	-6.883333	125370742	95.94%	51.95	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	987-25	45.033333	-0.583333	124718548	95.45%	46.53	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1245-12	39.433333	-6.333333	124991085	95.65%	56.82	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1354-12	45.883333	10.833333	124742295	95.46%	57.42	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1316-5	47.1	4.933333	124955953	95.63%	51.61	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	1321-1	46.95	4.3	124804315	95.51%	56.18	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	100.8	-	-	122812639	93.99%	47.03	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	79.1	37.7295	21.687317	123137159	94.24%	41.18	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	75.2	38.090667	22.171933	122964219	94.10%	52.71	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	80TR1-TS1	41.02	28.96	123019989	94.15%	33.15	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	TAAL-1	-	-	125232822	95.84%	36.02	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	39.1	37.989117	15.3369	124261563	95.10%	50.04	Illumina GAII	150	Paired end	Nuclei Extraction
Capsella rubella	78.1	37.69185	21.62535	123261469	94.33%	93.69	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	83.1	38.4379	21.4243167	123025981	94.15%	63.40	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	72.12	38.1557667	22.7215	123524414	94.53%	94.54	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	34.11	38.1605	16.051867	123470597	94.49%	66.10	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	84.15	38.476433	21.429833	122370981	93.65%	94.18	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	6.26	44.56205	0.353967	125920620	96.37%	126.84	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	23.9	40.723	15.1951167	124780722	95.49%	73.00	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	22.13	41.289667	13.951083	125013935	95.67%	55.61	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1408	35.29	24.42	122692159	93.90%	70.16	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	RIAH	-	-	125174646	95.79%	123.99	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1411-3	35.29	24.42	123372390	94.42%	60.38	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1314-10	47.316667	5.016667	125083806	95.73%	74.49	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	1319-3	47.366667	3.983334	125166467	95.79%	71.63	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	81.2	37.295517	22.060517	123950653	94.86%	92.15	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella rubella	77.16	38.1753675	20.5692179	124523820	95.30%	96.59	Illumina HiSeq 2000	100	Paired end multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella grandiflora	103.173	39.51838	21.56092	119484548	91.44%	48.51	GAII	100	Single end	CTAB
Capsella grandiflora	5a	39.705025	19.757344	115849190	88.66%	50.45	GAII	100	Single end	CTAB
Capsella grandiflora	83.17	39.4379	21.42983	116740093	89.34%	36.01	GAII	100	Single end	CTAB
Capsella grandiflora	85.33	39.55522	20.91642	115742135	88.58%	31.19	GAII	100	Single end	CTAB
Capsella grandiflora	86.8	39.51723	20.96527	116854539	89.43%	53.04	GAII	100	Single end	CTAB
Capsella grandiflora	88.56	39.8845167	20.7333	116216438	88.94%	35.42	GAII	100	Single end	CTAB
Capsella grandiflora	91.23	39.86715	20.70708	116282819	88.99%	53.26	GAII	100	Single end	CTAB
Capsella grandiflora	93.23	39.96448	20.71075	116236776	88.95%	33.78	GAII	100	Single end	CTAB
Capsella grandiflora	94.12	39.95974	20.72393	116661288	89.28%	57.92	GAII	100	Single end	CTAB
Capsella grandiflora	95.15	39.97875	20.72483	117323055	89.79%	35.76	GAII	100	Single end	CTAB
Capsella grandiflora	97.26	39.51727	21.15622	117862965	90.20%	50.97	GAII	100	Single end	CTAB
Capsella grandiflora	AxE	-	-	116573989	89.21%	43.71	GAII	100	Single end	CTAB
Capsella grandiflora	98	38.86465	20.9873	115864671	88.67%	33.12	GAII	100	Single end	CTAB
Capsella grandiflora	918-8	39.75	19.86667	-	-	-	-	-	-	-
Capsella grandiflora	Cg2e	39.67175	19.70102	-	-	-	-	-	-	-
Capsella orientalis	1718-9	48.5833333	88.4333333	99824130	76.39%	107.08	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1719-1	48.5833333	88.4333333	99746268	76.33%	78.90	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1938-1_1	53.3333333	83.75	99458682	76.11%	56.11	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1939-1_6	52.2666667	76.95	99180784	75.90%	77.19	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1940-1_1	50.7833333	75.6833333	98311522	75.24%	51.02	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1978-6	51.1	81.9	100377405	76.82%	101.67	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1979-02	51.1166667	81.8	98087502	75.07%	47.74	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1980-1	51.1666667	81.6666667	100019324	76.54%	113.49	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1981-3	51.1333333	81.6	98950978	75.73%	64.96	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1982-9	51.5	81.2166667	99425987	76.09%	71.79	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1983-6	51.3666667	82.2	99275895	75.97%	69.01	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1984-2	53.35	83.7333333	101010870	77.30%	72.46	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	1985-11	53.35	83.7333333	99383443	76.06%	73.92	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	2006-01	47.15	86.1166667	99034982	75.79%	76.16	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	2007-03	47.2333333	85.7166667	99771592	76.35%	81.69	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit
Capsella orientalis	2008-01	46.6166667	90.8666667	98946884	75.72%	67.74	Illumina HiSeq 2000	100	Paired End Multiplexed	Qiagen Dneasy Plant Mini Kit

ındiflora sample is a hybrid between samples from populations 918-8 and Cg2e

Figure 1 - figure supplement 2. Sample information.







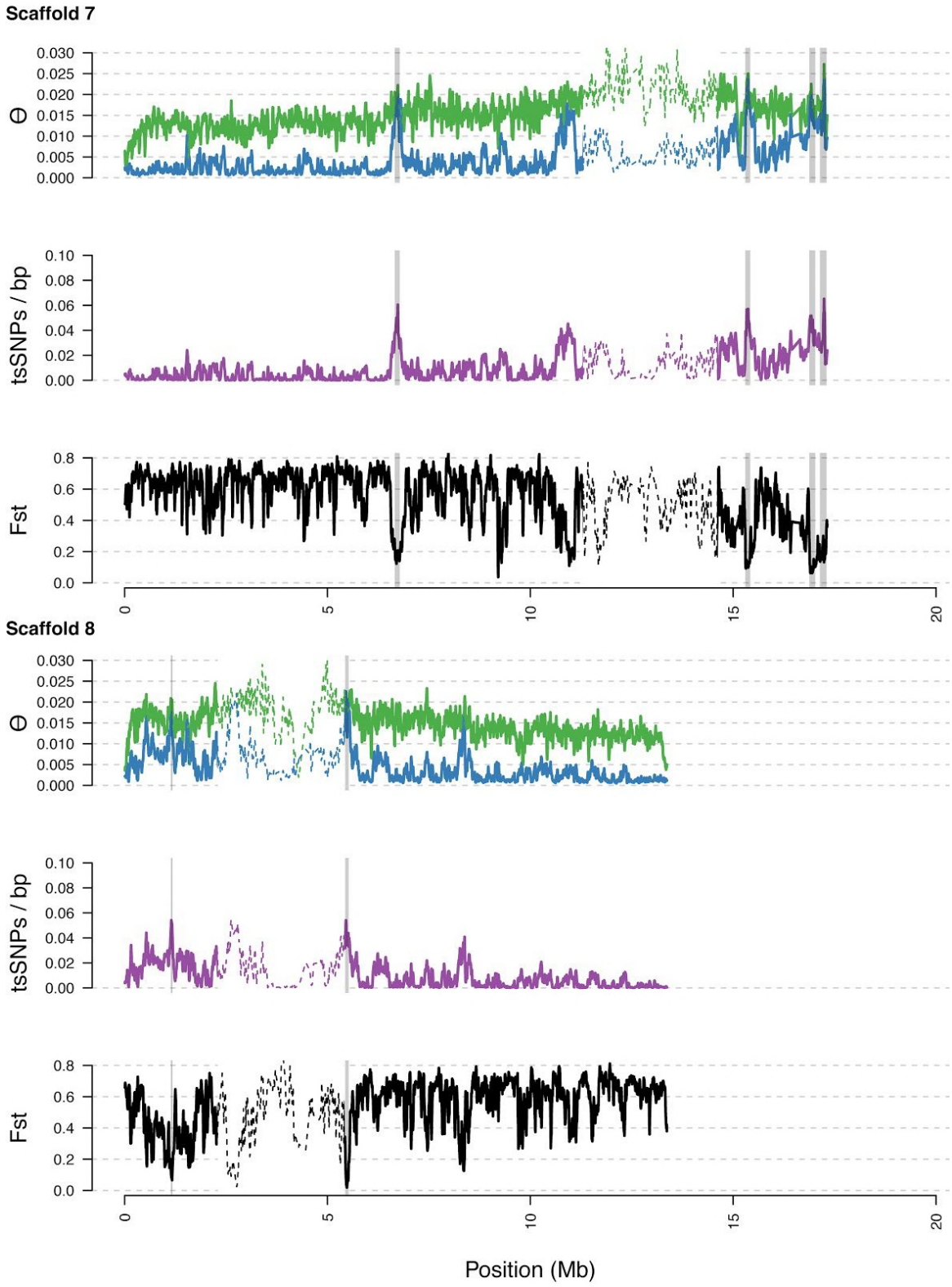
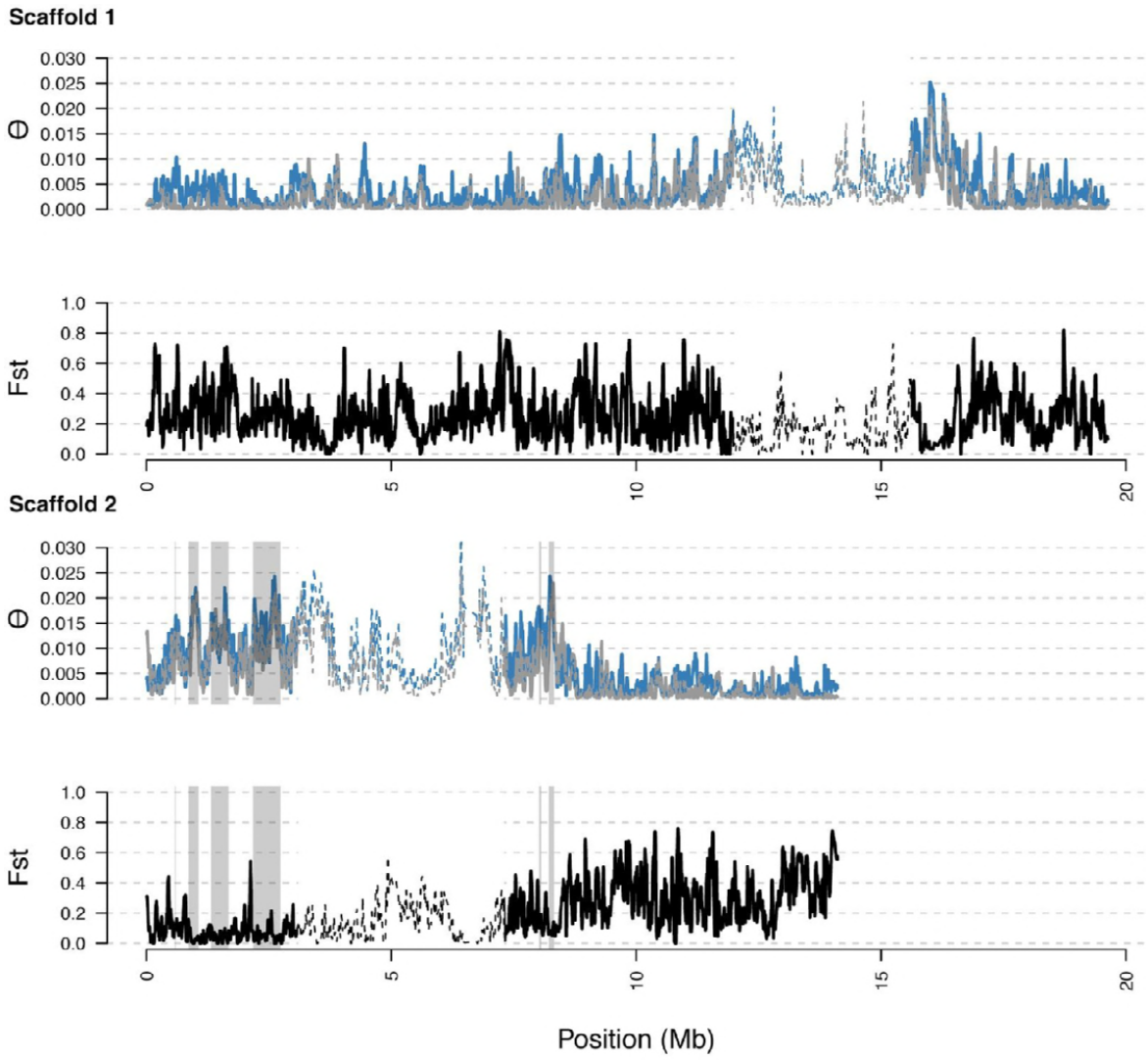
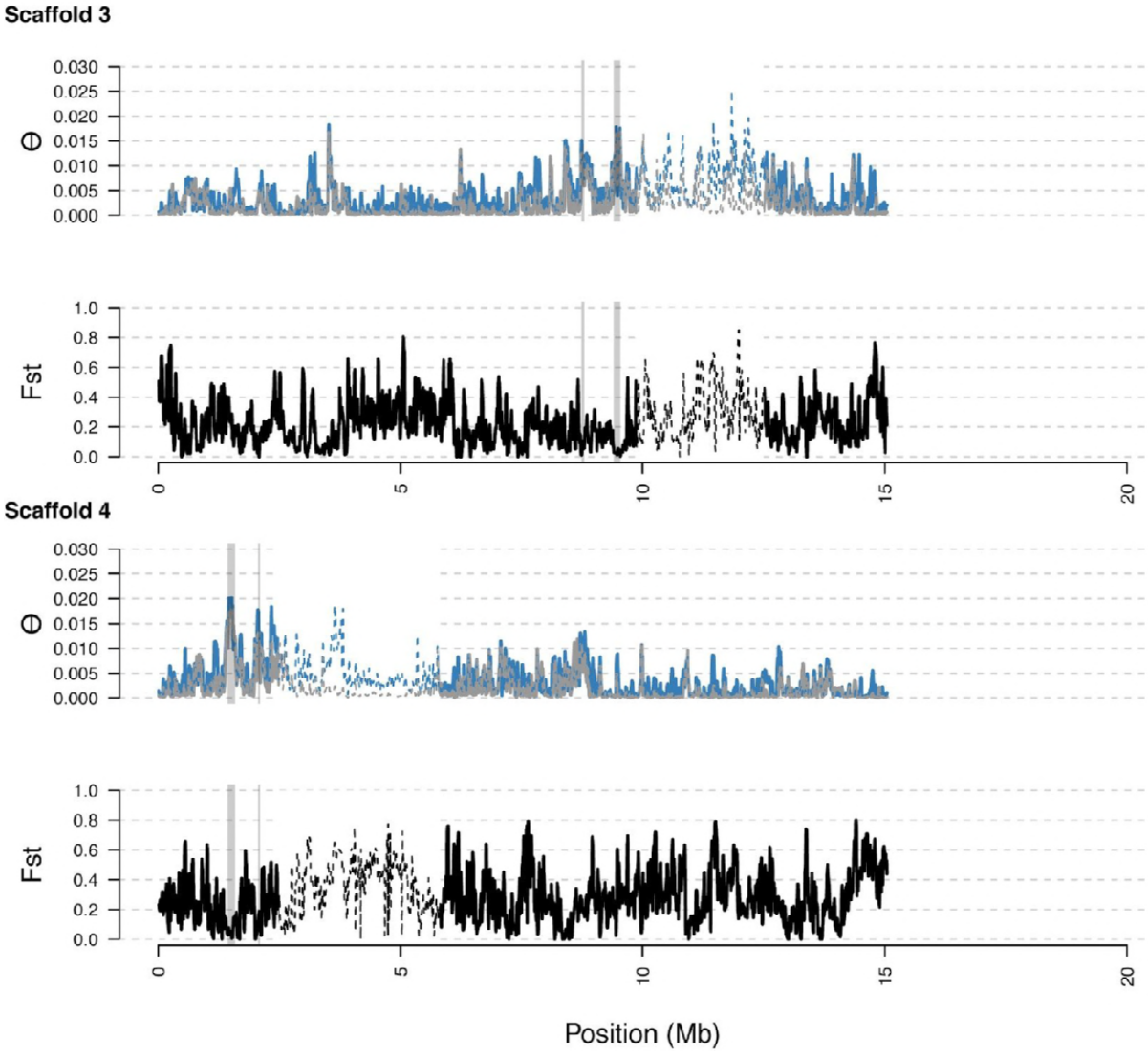
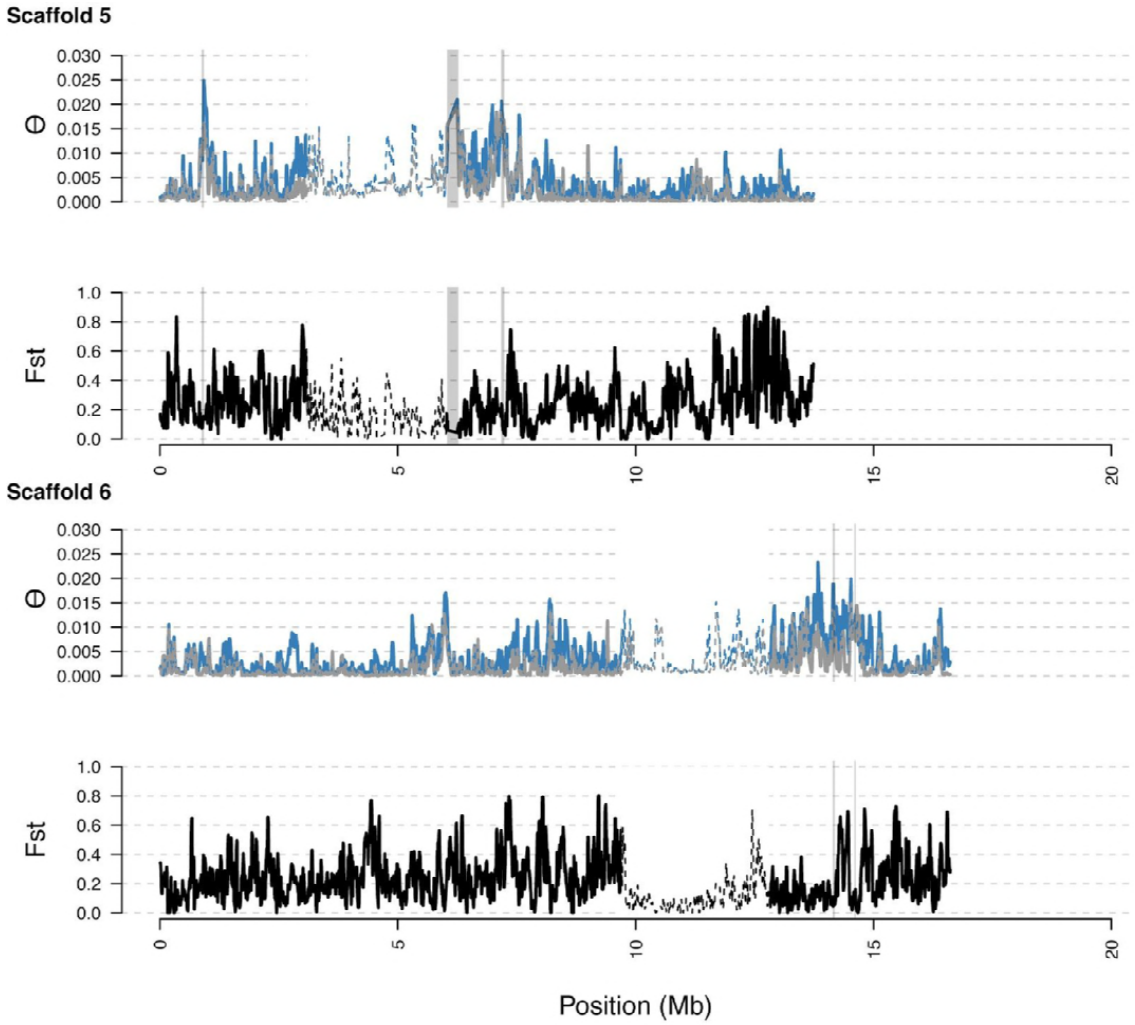


Figure 3 - figure supplement 2. Diversity in *C. rubella* and *C. grandiflora* along the eight major scaffolds (chromosomes). Data calculated in 20 kb windows, 5 kb steps. Windows with fewer than 5,000 covered sites were masked for all calculations. Dotted lines indicate the pericentromeric regions excluded from our analysis, and shaded regions highlight the regions with significant evidence for balancing selection.







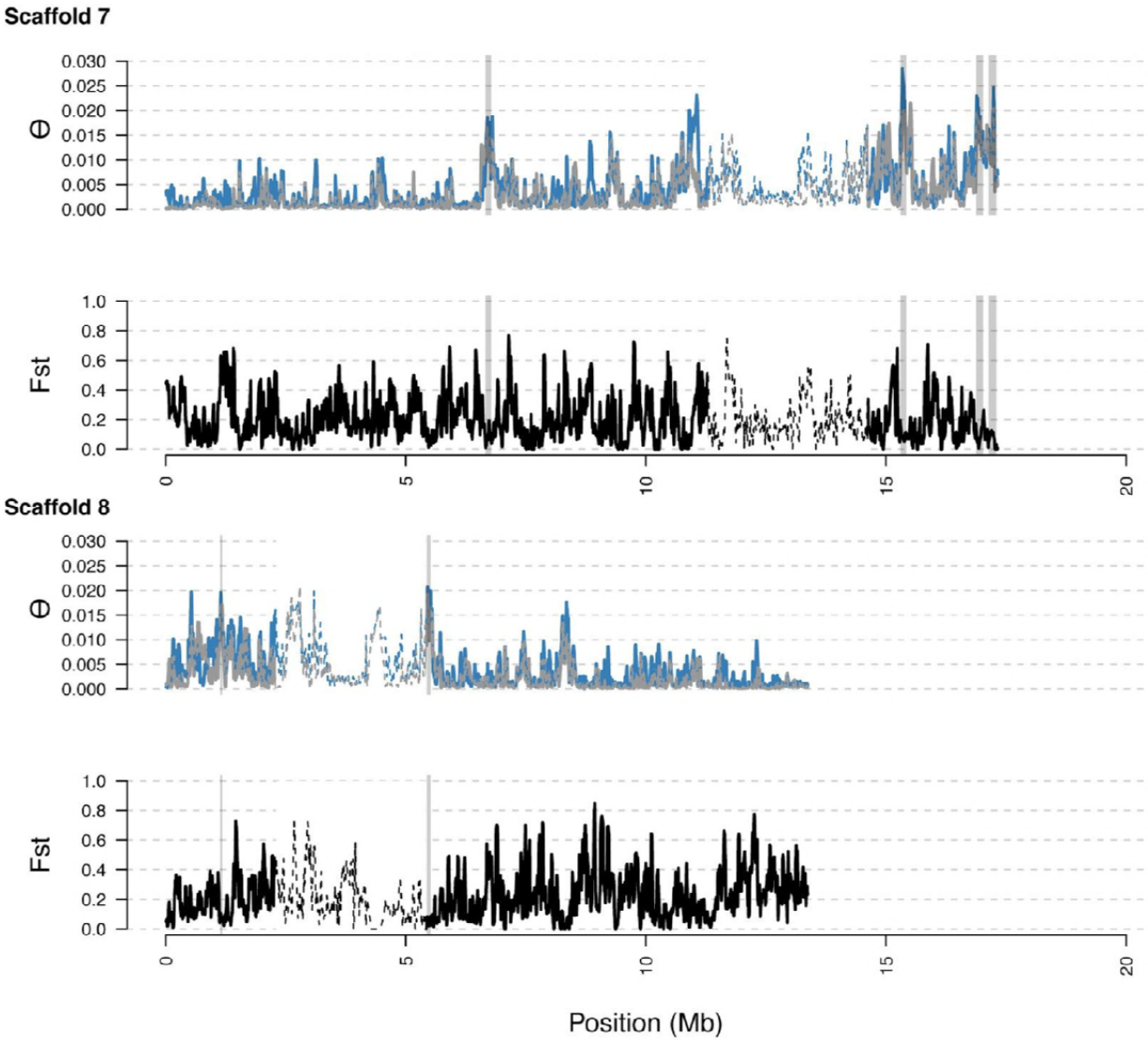


Figure 3 - figure supplement 3. Distribution of diversity in East and West *C. rubella* populations along the eight major scaffolds (chromosomes). As in S5, with diversity calculated within and F_{st} between populations. In general, regions of high diversity species-wide show high diversity and low F_{st}.