eLIFE elifesciences.org

RESEARCH ARTICLE

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

Daniel Koenig[1†*], Jörg Hagmann[1‡], Rachel Li[1§], Felix Bemm[1#], Tanja Slotte[2], Barbara Neuffer[3], Stephen I Wright[4], 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, Canada

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. By reconstructing the evolution of the disease-related locus *MLO2b*, we find that divergence between ancient haplotypes can be obscured by referenced based re-sequencing methods, and that trans-specific alleles can encode substantially diverged protein sequences. 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.
DOI: https://doi.org/10.7554/eLife.43606.001

***For correspondence:** dkoenig@ucr.edu (DK); weigel.elife@gmail.com (DW)

Present address: †Department of Botany and Plant Sciences, University of California, Riverside, United States; ‡Computomics GmbH, Tübingen, Germany; §Berkeley Brewing Science, Oakland, United States; #KWS SE, Einbeck, Germany

Competing interest: See page 21

Funding: See page 21

Received: 13 November 2018
Accepted: 26 February 2019
Published: 26 February 2019

Reviewing editor: Molly Przeworski, Columbia University, United States

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 (***Tellier et al., 2014***), and many candidates have been identified using diverse methodology (***Fijarczyk and Babik, 2015***). However, balanced alleles will be stochastically lost over long time spans, suggesting that most balanced polymorphism is short lived (***Fijarczyk and Babik, 2015***).

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-type loci of fungi (***Wu et al., 1998***), and the major histocompatibility complex (MHC) and ABO blood group loci in vertebrates (***McConnell et al., 1988***; ***Mayer et al., 1988***; ***Lawlor et al., 1988***; ***Watkins et al., 1990***; ***Ségurel et al., 2012***). 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

eLife digest *Capsella rubella* is a small plant that is found in southern and western Europe. This plant is young in evolutionary terms: it is thought to have emerged less than 200,000 years ago from a small group of plants belonging to an older species known as *Capsella grandiflora*.

Individuals of the same species may carry alternative versions of the same genes – known as alleles – and the total number of alleles present in a population is referred to as genetic diversity. When a few individuals form a new species, the gene pool and the genetic diversity in the new species is initially much lower than in the ancestral species, which may make the new species less robust to fluctuations in the environment. For example, alternative versions of a gene might be preferable in hot or cold climates, and loss of one of these versions would limit the species' ability to survive in both climates.

A mechanism known as balancing selection can maintain various alleles in a species, even if the population is very small. However, it was not clear how common long-lasting balancing selection was after a species had split. To address this question, Koenig et al. assembled collections of wild *C. rubella* and *C. grandiflora* plants and sequenced their genomes in search of alleles that were shared between individuals of the two species.

The analysis found not just a few, but thousands of examples where the same genetic differences had been maintained in both *C. rubella* and *C. grandiflora*. Some of these allele pairs were also shared with individuals of a third species of *Capsella* that had split from *C. rubella* and *C. grandiflora* over a million years ago. The shared alleles did not occur randomly in the genome; genes involved in immune responses were far more likely to be targets of balancing selection than other types of genes.

These findings indicate that there is strong balancing selection to maintain different alleles of immunity genes in wild populations of plants, and that some of this diversity can be maintained over hundreds of thousands, if not millions of years. The strategy developed by Koenig et al. may help to identify new versions of immunity genes from wild relatives of crop plants that could be used to combat crop diseases.

DOI: https://doi.org/10.7554/eLife.43606.002

loci in primates (**Leffler et al., 2013b**; **Teixeira et al., 2015**) and up to 129 loci, that were identified by at least two shared SNPs each, in *Arabidopsis* (**Novikova et al., 2016**; **Bechsgaard et al., 2017**), 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 (**Huard-Chauveau et al., 2013**; **Botella, 1998**; **Caicedo et al., 1999**; **Noel, 1999**; **Stahl et al., 1999**; **Tian et al., 2002**; **Bakker, 2006**; **Rose et al., 2004**; **Todesco et al., 2010**). However, even with the ability to conduct whole-genome scans for balancing selection in *A. thaliana*, the total number of examples with robust evidence across species remains small (**Cao et al., 2011**; **1001 Genomes Consortium, 2016**).

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 (**Fijarczyk and Babik, 2015**; **DeGiorgio et al., 2014**). 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 hypothesised that self-fertilizing species provide increased sensitivity to detect balancing selection based on two observations (**Wiuf et al., 2004**; **Wright et al., 2008**). First, self fertilisation greatly reduces the effective rate of recombination, thus potentially expanding the footprint of balancing selection. In addition, the transition to self fertilisation is generally associated with dramatic genome-wide reductions in polymorphism, potentially making it easier to detect outlier 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 fertilisation 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***)(***Foxe et al., 2009***; ***Guo et al., 2009***; ***Bachmann et al., 2018***). The genomes of both species exhibit the drastic loss of genetic diversity typical for many selfers (***Figure 1B–C***) (***Guo et al., 2009***; ***Foxe et al., 2009***; ***St Onge et al., 2011***; ***Slotte et al., 2013***; ***Brandvain et al., 2013***; ***Slotte et al., 2012***). 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 (***Gos et al., 2012***; ***Brandvain et al., 2013***), motivating the present study.

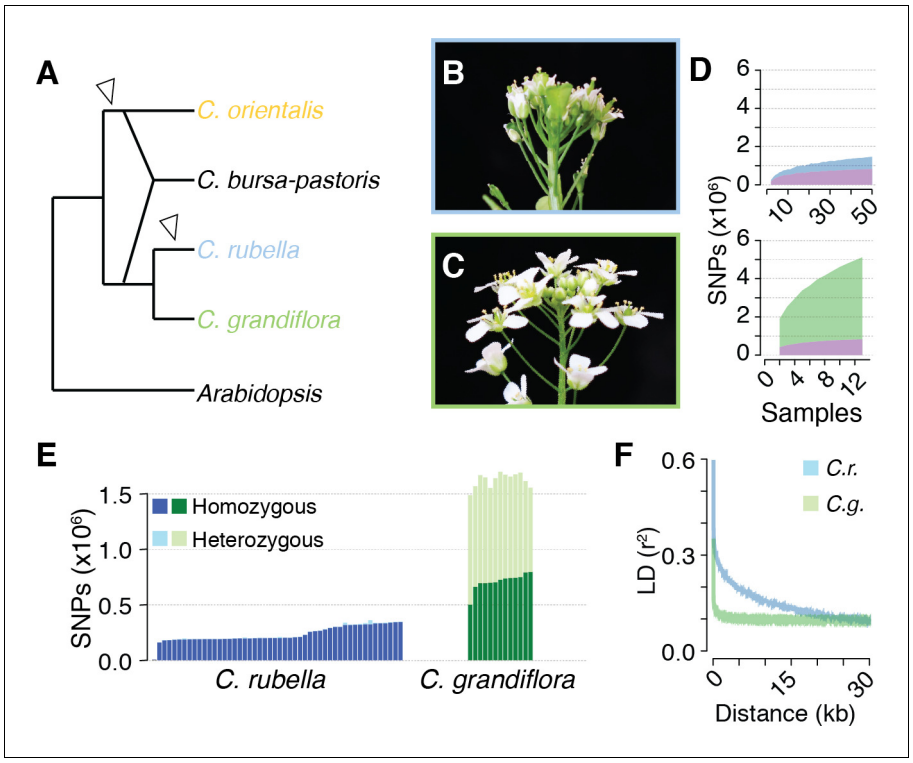

Figure 1. Polymorphism discovery in *Capsella*. (**A**) Diagram of the relationships between *Capsella* species. Arrowheads indicate transitions from outcrossing to self-fertilisation. (**B**) Inflorescence of *C. rubella* with small flowers. (**C**) Inflorescence of *C. grandiflora* with large, showy 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 colours) and homozygous SNP calls (dark colours). (**F**) Average decay of linkage disequilibrium in *C. grandiflora* (green) and *C. rubella* (blue).

DOI: https://doi.org/10.7554/eLife.43606.003

The following source data and figure supplement are available for figure 1:

Source data 1. Sample information.

DOI: https://doi.org/10.7554/eLife.43606.005

Source data 2. Diversity and divergence estimates for C.grandiflora and C. rubella.

DOI: https://doi.org/10.7554/eLife.43606.006

Figure supplement 1. Map of collections.

DOI: https://doi.org/10.7554/eLife.43606.004

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 retention of *C. grandiflora* alleles across the *C. rubella* genome (**Gos et al., 2012**; **Brandvain et al., 2013**), leading us to analyse 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—source data 1**). 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 common ss_{Cr}SNP and ts_{CgCr}SNP discovery was near saturation in our experiment, though additional sampling will continue to uncover rare alleles (**Figure 1D**).

The consequences of selfing are easily seen as a dramatic reduction in genetic diversity in *C. rubella* (**Figure 1—source data 2**), 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 be heterozygous 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-fertilising 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., 2004**).

Capsella rubella demography

The degree of trans-specific allele sharing depends 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 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 **Figure 2—figure supplement 1A-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**).

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**). A single *C. rubella* sample from western Europe showed some mixed ancestry. This sample was collected near Gargano National Park on the eastern coast of Italy. The source of its mixed ancestry is unclear, but its proximity to Greece suggests that it may result from ongoing migration across the Adriatic Sea. The general pattern of population structure is consistent with the centre 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

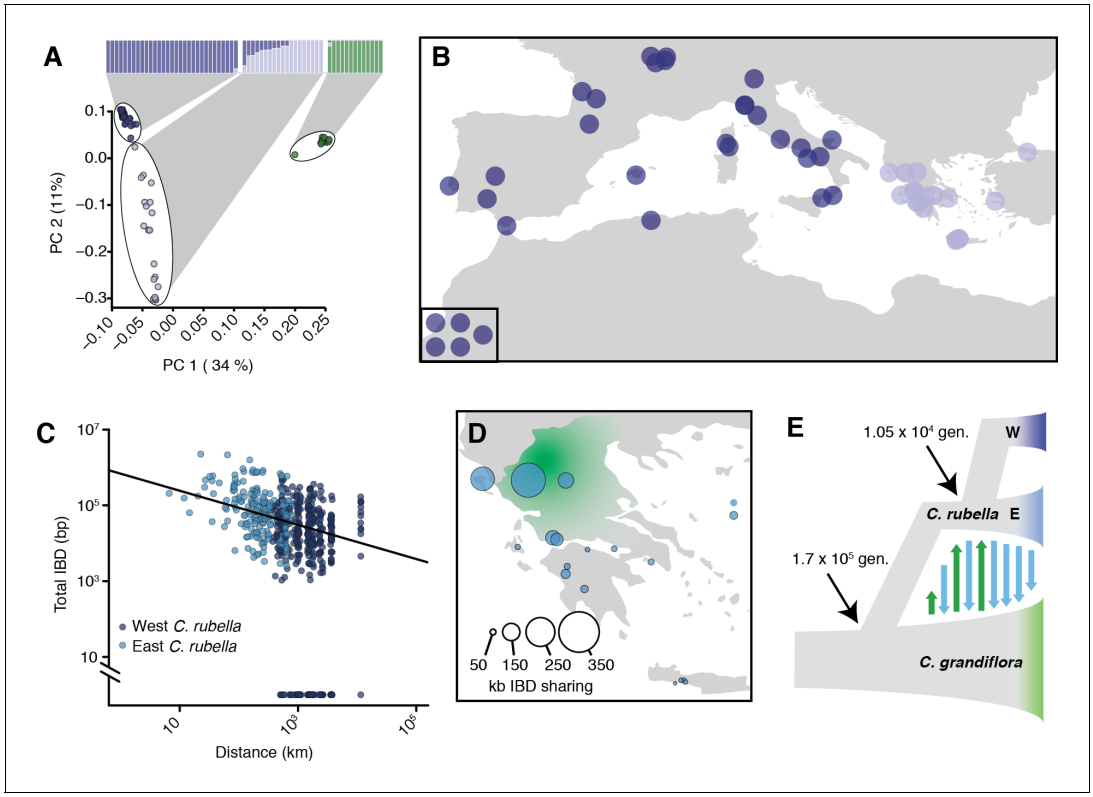

Figure 2. Demographic analysis of *C.rubella*. (**A**) Admixture bar graphs (top) and PCA of population structure in *C. grandiflora* (green) and *C. rubella* (blue). The *C. rubella* colours correspond to the sampling locations in (**B**). Inset shows lines from outside Eurasia (Canary Islands and Argentina). (**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.

DOI: https://doi.org/10.7554/eLife.43606.007

The following source data and figure supplements are available for figure 2:

Source data 1. D statistics comparing East and West C.rubella populations.

DOI: https://doi.org/10.7554/eLife.43606.010

Source data 2. Inferred demographic parameters from fastsimcoal2.

DOI: https://doi.org/10.7554/eLife.43606.011

Figure supplement 1. Additional population structure and migration analyses.

DOI: https://doi.org/10.7554/eLife.43606.008

Figure supplement 2. Comparison of simulated and observed allele frequency spectra under the best fitting demographic model.

DOI: https://doi.org/10.7554/eLife.43606.009

(***Brandvain et al., 2013***). The observed structure is principally organised by a major geographic barrier, the Adriatic Sea. We therefore separated our samples into into two distinct groups to the west (W) and east (E) of the Adriatic Sea for subsequent analyses.

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—figure supplement 1C-D***). 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) (**Green et al., 2010**; **Durand et al., 2011**), with *D* decreasing as a function of distance from the centre of *C. grandiflora*'s range (**Figure 2—figure supplement 1** and **Figure 2—source data 1**). 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 genomic regions of more 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 2C–D**). These results indicate that gene flow is ongoing between the species, consistent with interspecific crosses often producing fertile offspring, specifically with *C. rubella* as the paternal parent (**Sicard et al., 2011**; **Rebernig et al., 2015**).

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**, **Figure 2—figure supplement 2** and **Figure 2—source data 2**). 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**).

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

Our analyses provide dates for the bottleneck and rapid colonisation 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* to *C. rubella prior to its expansion into western Europe.*

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 1000 jackknife resamples from the same allele frequency distribution), but they were more likely to be in genes (**Figure 3C**). As expected for transpecific haplotype sharing, 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 3—figure supplements 2–5**), and negatively correlated with differentiation between the species as measured by F_{st} (**Figure 3J** and **Figure 3—figure supplements 2–5**). The uneven pattern of diversity

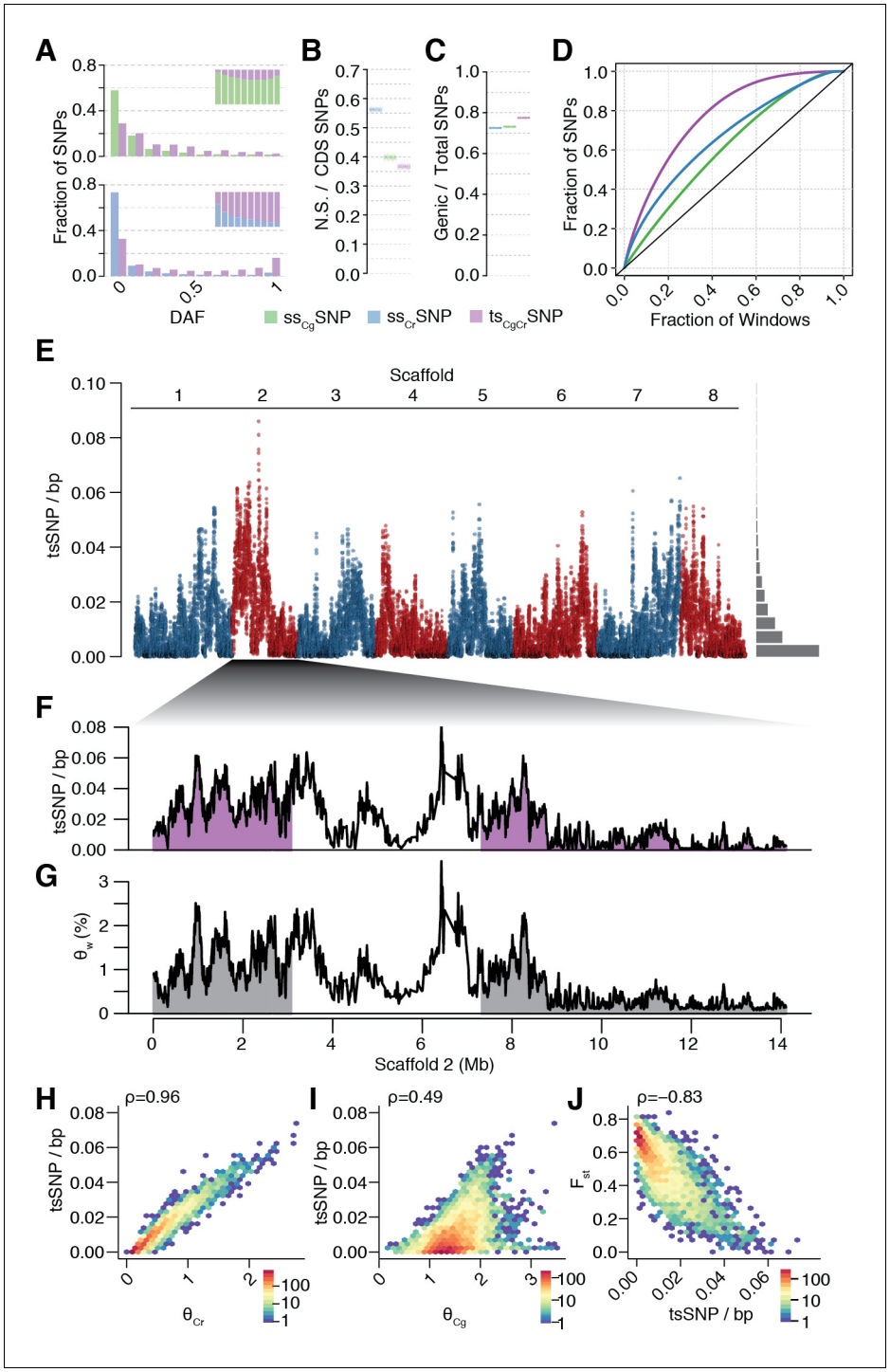

Figure 3. Unequal presence of ancestral variation in modern *C.rubella*. (**A**) Derived allele frequency spectra (DAF) of ss_{Cg}SNPs (green), ss_{Cr}SNPs (blue), and ts_{CgCr}SNPs (purple) in *C. grandiflora* (top) and *C. rubella* (bottom). The inset depicts the fraction of alleles that are species 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 normalised by downsampling to comparable frequency spectra. Each bar consists of 1000 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

Figure 3 continued on next page

Figure 3 continued

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 5000 accessible sites in both species were considered.

DOI: https://doi.org/10.7554/eLife.43606.012

The following figure supplements are available for figure 3:

Figure supplement 1. Sharing of ts_{CgCr}SNPs and ss_{Cr}SNPs alleles after colonisation.

DOI: https://doi.org/10.7554/eLife.43606.013

Figure supplement 2. Diversity in *C. rubella* and *C. grandiflora* along scaffolds (chromosomes) 1 and 2.

DOI: https://doi.org/10.7554/eLife.43606.014

Figure supplement 3. Diversity in *C. rubella* and *C. grandiflora* along scaffolds (chromosomes) 3 and 4.

DOI: https://doi.org/10.7554/eLife.43606.015

Figure supplement 4. Diversity in *C. rubella* and *C. grandiflora* along scaffolds (chromosomes) 5 and 6.

DOI: https://doi.org/10.7554/eLife.43606.016

Figure supplement 5. Diversity in *C. rubella* and *C. grandiflora* along scaffolds (chromosomes) 7 and 8.

DOI: https://doi.org/10.7554/eLife.43606.017

Figure supplement 6. Distribution of diversity in East and West *C. rubella* populations along scaffolds (chromosomes) 1 and 2.

DOI: https://doi.org/10.7554/eLife.43606.018

Figure supplement 7. Distribution of diversity in East and West *C. rubella* populations along scaffolds (chromosomes) 3 and 4.

DOI: https://doi.org/10.7554/eLife.43606.019

Figure supplement 8 Distribution of diversity in East and West *C. rubella* populations along scaffolds (chromosomes) 5 and 6.

DOI: https://doi.org/10.7554/eLife.43606.020

Figure supplement 9. Distribution of diversity in East and West *C. rubella* populations along scaffolds (chromosomes) 7 and 8.

DOI: https://doi.org/10.7554/eLife.43606.021

was similar in each *C. rubella* subpopulation (***Figure 3—figure supplements 6–9***), indicating that most of the retained polymorphism already segregated prior to colonisation. 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.

High density of tsSNPs around immunity-related loci

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

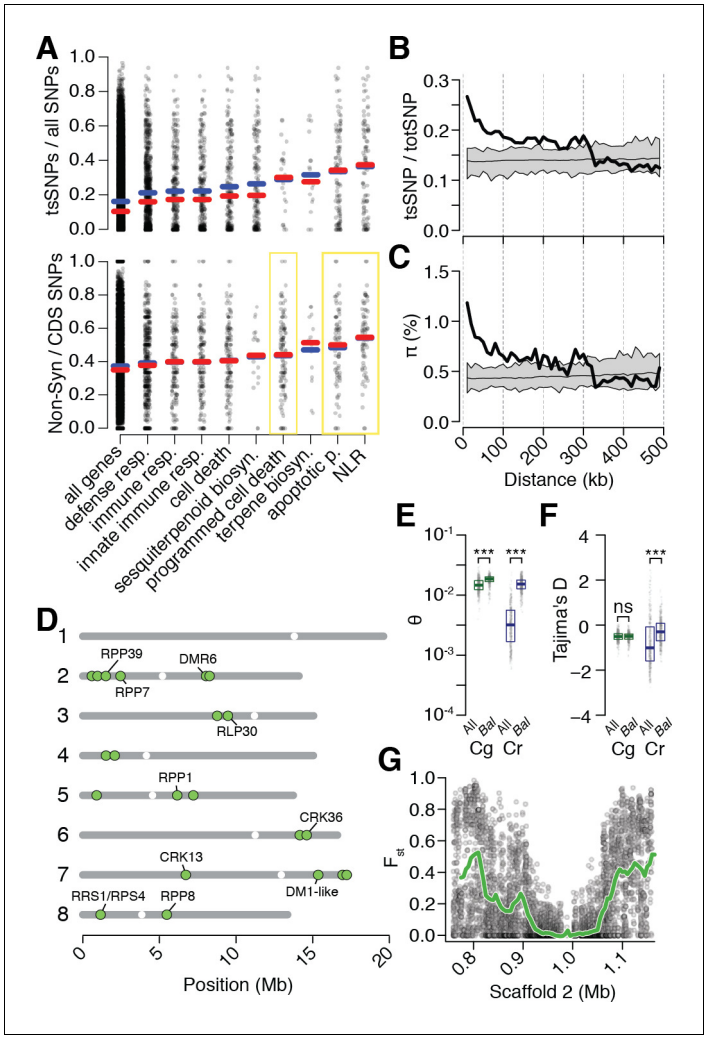

Figure 4. Preferential sharing of alleles near immunity genes. (**A**) Enrichment of ts_{CgCr}SNPs and non-synonymous ts_{CgCr}SNPs for genes associated with significant GO terms (means, blue; medians, red). Each point represents values calculated for an individual gene. For example, in the upper subplot each point is the number of tsSNPs identified in a gene divided by the total number of SNPs identified for a gene. 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. For (**B–C**) the thick black lines are mean values calculated in 500 bp windows as a function of distance from a NLR gene. The thin black lines are mean values from 100 random gene sets of equivalent size. The grey polygons are the 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 1000 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.

DOI: https://doi.org/10.7554/eLife.43606.022

The following source data and figure supplements are available for figure 4:

Source data 1. Regions with evidence for balancing selection.

DOI: https://doi.org/10.7554/eLife.43606.025

Source data 2. Spearman's correlations of allele frequencies for different classes of tsSNPs.

DOI: https://doi.org/10.7554/eLife.43606.026

Figure supplement 1. Quality metrics for ssSNPs and tsSNPs in *C. rubella*.

DOI: https://doi.org/10.7554/eLife.43606.023

Figure 4 continued on next page

Figure 4 continued

Figure supplement 2. Analysis of IBS in balanced regions and genome-wide.
DOI: https://doi.org/10.7554/eLife.43606.024

(***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 artefact of the internal structure of NLR clusters (***Figure 4B–C***).

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 4—source data 1***). *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 4E–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 colonisation of Western Europe than 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—source data 2***). 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).

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* (***McDowell et al., 2000***; ***McDowell, 1998***; ***Goritschnig et al., 2012***; ***Holub, 1994***; ***Gassmann et al., 1999***; ***Zhang et al., 2014***;

Zhang et al., 2013; *Xu, 2006*; *Yeh et al., 2015*). *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*).

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. 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*), suggesting 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 2*), consistent with elevated retention of introgressed alleles in these regions. 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 2*). In contrast to the genome-wide pattern, high levels of IBD were observed between *C. grandiflora* and West population accessions. Thus, we find evidence for neutral gene flow throughout 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 colonisation 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

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 (**Hurka et al., 2012**; **Douglas et al., 2015**). We expected the evolution of selfing to have generated a similar bottleneck as in *C. rubella* (**Douglas et al., 2015**; **Bachmann et al., 2018**), 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—source data 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 estimates of N_e from nucleotide diversity at four-fold degenerate sites (*C. orientalis* [14,643] and *C. grandiflora* [694,643]), the divergence time above, and the genome assembly size of 134.8 Mb, the probability of finding a single tsSNP is $<4\times10^{-19}$ using the methodology of Leffler and colleagues and Wiuf and colleagues (**Leffler et al., 2013a**; **Wiuf et al., 2004**), which assumes constant population size. 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 3992 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**).

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 Materials and 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 re-called $ts_{3\text{-way}}$SNP sites. We identified 812 high quality transpecific SNPs segregating in all three species ($ts_{3\text{-wayhq}}$SNPs). The distributions of coverage and concordance values in this dataset were similar between $ts_{3\text{-way}}$SNP sites and other *C. orientalis* sites, further supporting their authenticity (**Figure 5—figure supplement 1**).

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 compared to the genome-wide average divergence. 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.

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.

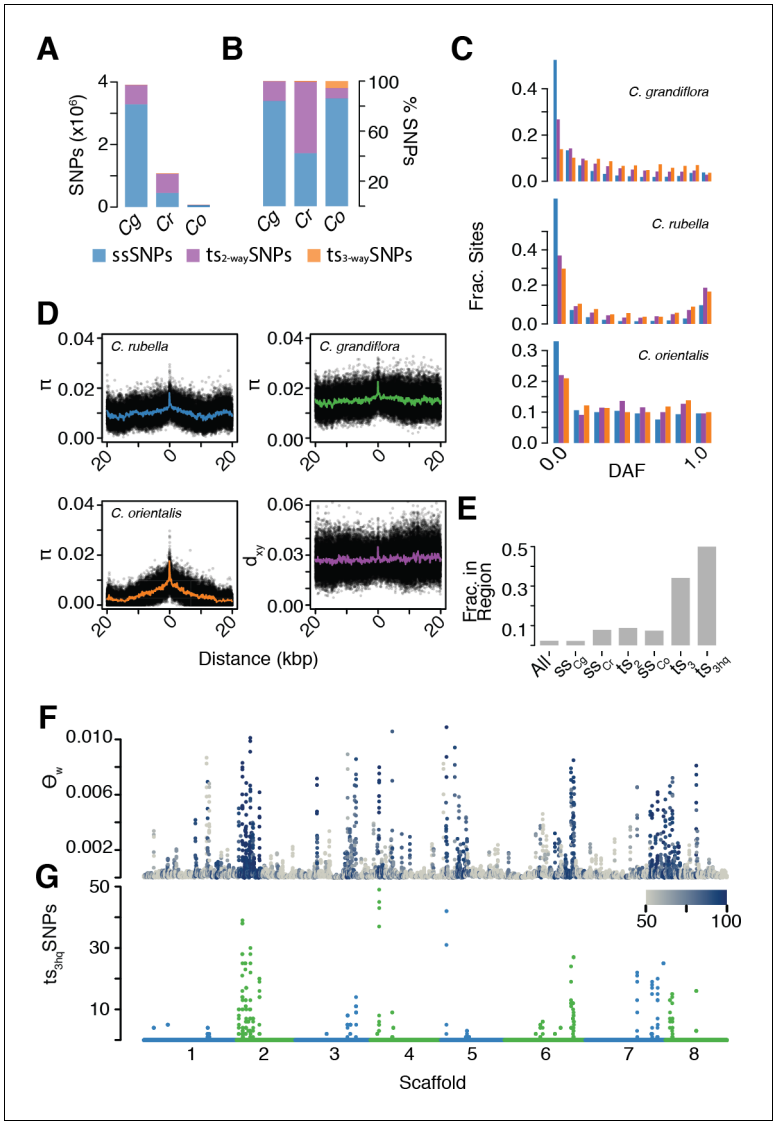

Figure 5. The signal of ancient balancing selection. (**A**) Absolute number and (**B**) fraction of ssSNPs (blue), $ts_{2\text{-}way}$SNPs (purple), and $ts_{3\text{-}way}$SNPs (orange) for *Capsella*. Only sites accessible to read mapping in all three species were considered. (**C**) Derived allele 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 coloured 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.

DOI: https://doi.org/10.7554/eLife.43606.027

The following source data and figure supplements are available for figure 5:

Source data 1. Three species diversity and divergence.

DOI: https://doi.org/10.7554/eLife.43606.030

Figure supplement 1. Distributions of concordance and coverage values for different SNP classes in *C. orientalis*.

DOI: https://doi.org/10.7554/eLife.43606.028

Figure supplement 2. $tr_{3\text{-}wayhq}$SNPs MAF.

DOI: https://doi.org/10.7554/eLife.43606.029

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—source data 2***). 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 characterised 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 six $ts_{3\text{-}wayhq}$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.

Because several known targets of balancing selection in *A. thaliana* are the result of structural variation, or lesions larger than 1 kb (***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 two haplogroup B) and 43 *C. rubella* samples (34 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,

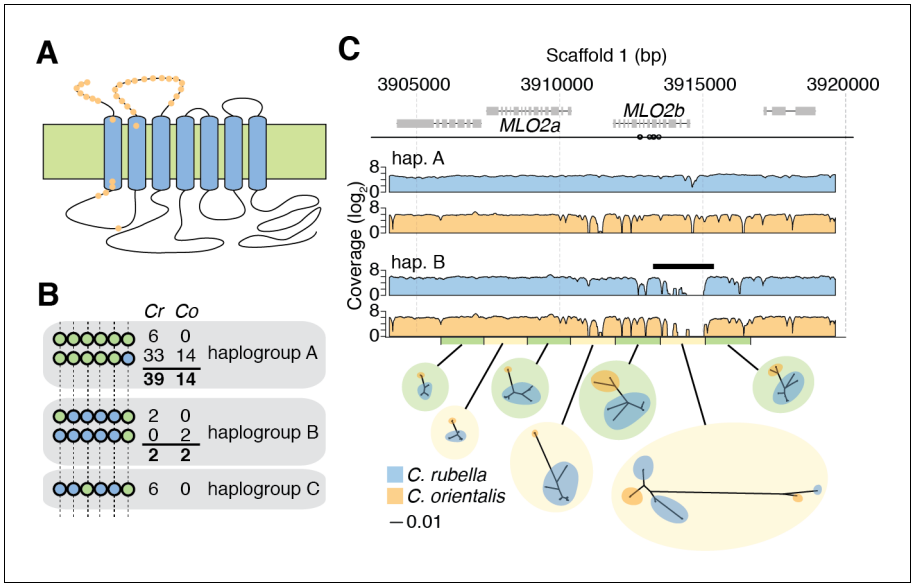

Figure 6. Evidence for long-term balancing selection at *MLO2*. (**A**) Diagram of MLO2 protein in the cell membrane. Blue ovals, transmembrane domains. Top is the extracellular space. Orange dots represent amino acid differences between proteins encoded by haplogroups A and B. (**B**) Haplogroup identification with reference based SNP calls. Circles represent $ts_{3\text{-}wayhq}$SNPs and colours 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 neighbour-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.

DOI: https://doi.org/10.7554/eLife.43606.031

The following figure supplements are available for figure 6:

Figure supplement 1. Sliding windows of allelic divergence and positions of tsSNPs.

DOI: https://doi.org/10.7554/eLife.43606.032

Figure supplement 2. Phylogenetic analysis of the MLO2 CDS.

DOI: https://doi.org/10.7554/eLife.43606.033

Figure supplement 3. Alignment of amino acid sequences at the MLO2 N-terminus.

DOI: https://doi.org/10.7554/eLife.43606.034

over three times higher than the genome-wide divergence between the species (***Figure 6—figure supplement 1*** and ***Figure 5—source data 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***). Neighbour-joining trees revealed the expected clustering of samples by species in regions adjacent to *MLO2b*, but clear clustering by haplogroup within the 5' region, a pattern that is reproduced in phylogenetic analysis of the entire CDS (***Figure 6C*** and ***Figure 6—figure supplement 2***). Importantly, divergence within haplogroup across species was greater than, or similar to genome-wide averages for both A and B, demonstrating that recent introgression did not give rise to allele sharing (***Figure 6—figure supplement 1***).

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 6—figure supplement 3***), 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.

Discussion

While balancing selection has long been recognised as an important evolutionary force, its relevance as a major factor shaping genomic variation has remained unclear (**Charlesworth, 2006**; **Wiuf et al., 2004**; **Asthana et al., 2005**). 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 (**Wiuf et al., 2004**; **Charlesworth, 2006**). 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 fertilising 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* (**Stahl et al., 1999**; **Tian et al., 2002**; **Tian et al., 2003**; **Mauricio et al., 2003**; **Bakker, 2006**), 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, 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**; **Walling, 2009**; **Herms and Mattson, 1992**), 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 fertilising 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 (**Whitney et al., 2006**; **Castric et al., 2008**; **Pease et al., 2016**; **Dasmahapatra et al., 2012**; **Henning and Meyer, 2014**; **Hedrick, 2013**; **Huerta-Sánchez et al., 2014**; **Racimo et al., 2015**; **Castric et al., 2008**; **Pease et al., 2016**; **Dasmahapatra et al., 2012**; **Whitney et al., 2006**;

Huerta-Sánchez et al., 2014; *Hedrick, 2013*). The unique feature of self-fertilisation 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-fertilising 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 maximise 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 analysed independently in previous studies (*Agren et al., 2014*; *Williamson et al., 2014*). See *Figure 1—source data 1* 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 *Figure 1—source data 1*. We aimed for a minimum genome coverage of 40x. We mapped reads to the *C. rubella* reference genome (*Slotte et al., 2013*) resulting in realised 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 finalised 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 1000 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. Segments were required to be larger than 1 kb to be considered in the analyses. D statistics were calculated as in **Green et al. (2010)**; **Patterson et al. (2012)**; **Dasmahapatra et al. (2012)** comparing each individual genotype from the eastern *C. rubella* population to allele frequencies from western *C. rubella* and *C. grandiflora*. The outgroup species for these analyses was *C. orientalis*.

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 two 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 visualised 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. The phylogeny of *Capsella MLO2* CDS sequences was produced using the optim.pml command from the R package phangorn using Jukes-Cantor distances. 1000 bootstrap iterations were run to estimate support for nodes in the tree. 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—source data 1**) 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 DK (LT000783/2010 L) and by DFG-SPP1529 ADAPTOMICS (WE 2897/4–2), ERC Advanced Grant IMMUNEMESIS and the Max Planck Society (DW).

Additional information

Competing interests

Detlef Weigel: Deputy editor, *eLife*. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
European Research Council	IMMUNEMESIS	Detlef Weigel
Human Frontier Science Program	LT000783/2010-L	Daniel Koenig
Deutsche Forschungsgemeinschaft	WE 2897/4-2	Detlef Weigel
Max-Planck-Gesellschaft	Open-access funding	Detlef Weigel

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Daniel Koenig, Conceptualization, Resources, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing; Jörg Hagmann, Data curation, Writing—review and editing; Rachel Li, Felix Bemm, Investigation, Writing—review and editing; Tanja Slotte, Stephen I Wright, Resources, Supervision, Writing—review and editing; Barbara Neuffer, Resources, Writing—review and editing; Detlef Weigel, Conceptualization, Supervision, Funding acquisition, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Daniel Koenig http://orcid.org/0000-0002-1037-5346
Rachel Li https://orcid.org/0000-0002-8112-4237
Stephen I Wright http://orcid.org/0000-0001-9973-9697
Detlef Weigel http://orcid.org/0000-0002-2114-7963

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.43606.047
Author response https://doi.org/10.7554/eLife.43606.048

Additional files

Supplementary files

- Supplementary file 1. GO enrichment analysis of tsSNPs.
DOI: https://doi.org/10.7554/eLife.43606.035
- Supplementary file 2. GO enrichment for $tr_{3\text{-}way}$SNPs.
DOI: https://doi.org/10.7554/eLife.43606.036
- Supplementary file 3. List of well covered genes for targeted analysis of potential balancing selection.
DOI: https://doi.org/10.7554/eLife.43606.037
- Supplementary file 4. Pericentromeric or repeat dense genomic regions filtered in genome scans.
DOI: https://doi.org/10.7554/eLife.43606.038

• Supplementary file 5. Annotation hierarchies for SNPs with multiple annotations.
DOI: https://doi.org/10.7554/eLife.43606.039

• Supplementary file 6. List of *A. thaliana* NLR genes used for ortholog identification.
DOI: https://doi.org/10.7554/eLife.43606.040

• Transparent reporting form
DOI: https://doi.org/10.7554/eLife.43606.041

Data availability

All raw sequencing data are depsoited under the accession codes PRJEB6689.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Koenig D, Hagmann J, Li R, Bemm F, Slotte T, Neuffer B, Wright SI, Detlef Weigel	2018	Whole genome resequencing of Capsella species	https://www.ebi.ac.uk/ena/data/view/PRJEB6689	European Nucleotide Archive, PRJEB6689

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Williamson R, Josephs EB, Platts AE	2014	Capsella grandiflora WGS	https://www.ebi.ac.uk/ena/data/view/PRJEB6689	European Nucleotide Archive, PRJEB6689

References

1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell* **166**:481–491. DOI: https://doi.org/10.1016/j.cell.2016.05.063, PMID: 27293186

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**. DOI: https://doi.org/10.1186/1471-2164-15-602, PMID: 25030755

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655–1664. DOI: https://doi.org/10.1101/gr.094052.109, PMID: 19648217

Asthana S, Schmidt S, Sunyaev S. 2005. A limited role for balancing selection. *Trends in Genetics* **21**:30–32. DOI: https://doi.org/10.1016/j.tig.2004.11.001, PMID: 15680511

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, et al. 2010. Genome-wide association study of 107 phenotypes in *arabidopsis* thaliana inbred lines. *Nature* **465**:627–631. DOI: https://doi.org/10.1038/nature08800, PMID: 20336072

Bachmann JA, Tedder A, Laenen B, Fracassetti M, Désamoré A, Lafon-Placette C, Kim A. 2018. Genetic basis and timing of a major mating system shift in capsella. *bioRxiv*. DOI: https://doi.org/10.1101/425389

Bakker EG. 2006. A Genome-Wide survey of R gene polymorphisms in *Arabidopsis*. *The Plant Cell Online* **18**:1803–1818. DOI: https://doi.org/10.1105/tpc.106.042614

Bechsgaard J, Jorgensen TH, Schierup MH. 2017. Evidence for Adaptive Introgression of Disease Resistance Genes Among Closely Related *Arabidopsis* Species. *G3: Genes|Genomes|Genetics* **7**:2677–2683. DOI: https://doi.org/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. DOI: https://doi.org/10.1038/nature10555, PMID: 22057020

Bergelson J, Kreitman M, Stahl EA, Tian D. 2001. Evolutionary dynamics of plant R-genes. *Science* **292**:2281–2285. DOI: https://doi.org/10.1126/science.1061337, PMID: 11423651

Botella MA. 1998. Three genes of the arabidopsis RPP1 complex resistance locus recognize distinct peronospora parasitica avirulence determinants. *The Plant Cell Online* **10**:1847–1860. DOI: https://doi.org/10.1105/tpc.10.11.1847

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 Genetics* **9**:e1003754. DOI: https://doi.org/10.1371/journal.pgen.1003754, PMID: 24068948

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**:459–471. DOI: https://doi.org/10.1534/genetics.113.150029, PMID: 23535385

Caicedo AL, Schaal BA, Kunkel BN. 1999. Diversity and molecular evolution of the RPS2 resistance gene in arabidopsis thaliana. *PNAS* **96**:302–306. DOI: https://doi.org/10.1073/pnas.96.1.302, PMID: 9874813

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. *Nature Genetics* **43**:956–963. DOI: https://doi.org/10.1038/ng.911, PMID: 21874002

Castric V, Bechsgaard J, Schierup MH, Vekemans X. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genetics* **4**:e1000168. DOI: https://doi.org/10.1371/journal.pgen.1000168, PMID: 18769722

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**. DOI: https://doi.org/10.1186/s13742-015-0047-8, PMID: 25722852

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**:e64. DOI: https://doi.org/10.1371/journal.pgen.0020064, PMID: 16683038

Charlesworth D, Charlesworth B. 1975. Theoretical genetics of batesian mimicry II. evolution of supergenes. *Journal of Theoretical Biology* **55**:305–324. DOI: https://doi.org/10.1016/S0022-5193(75)80082-8, PMID: 1207161

Chin CS, 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. *Nature Methods* **10**:563–569. DOI: https://doi.org/10.1038/nmeth.2474, PMID: 23644548

Chin CS, 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. *Nature Methods* **13**:1050–1054. DOI: https://doi.org/10.1038/nmeth.4035, PMID: 27749838

Cingolani P, Platts A, Wang leL, 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 w1118; *iso-2*; *iso-3*. *Fly* **6**:80–92. DOI: https://doi.org/10.4161/fly.19695, PMID: 22728672

Coley PD, Bryant JP, Chapin FS. 1985. Resource availability and plant antiherbivore defense. *Science* **230**:895–899. DOI: https://doi.org/10.1126/science.230.4728.895, PMID: 17739203

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. *Nature Genetics* **38**:716–720. DOI: https://doi.org/10.1038/ng1806, PMID: 16732289

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. DOI: https://doi.org/10.1093/bioinformatics/btr330, PMID: 21653522

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin AV, Hughes DST, Ferguson LC, Martin SH, Salazar C, Lewis JJ, Adler S, Ahn S-J, Baker DA, Baxter SW, Chamberlain NL, Chauhan R, Counterman BA, Dalmay T, et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**:94–98. DOI: https://doi.org/10.1038/nature11041, PMID: 22722851

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics* **10**:e1004561. DOI: https://doi.org/10.1371/journal.pgen.1004561, PMID: 25144706

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*. *PNAS* **112**:2806–2811. DOI: https://doi.org/10.1073/pnas.1412277112, PMID: 25691747

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**:2239–2252. DOI: https://doi.org/10.1093/molbev/msr048, PMID: 21325092

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics* **9**:e1003905. DOI: https://doi.org/10.1371/journal.pgen.1003905, PMID: 24204310

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 Phytologist* **202**:1043–1054. DOI: https://doi.org/10.1111/nph.12702, PMID: 24471774

Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Molecular Ecology* **24**:3529–3545. DOI: https://doi.org/10.1111/mec.13226, PMID: 25943689

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in *capsella*. *PNAS* **106**:5241–5245. DOI: https://doi.org/10.1073/pnas.0807679106, PMID: 19228944

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. *The Plant Journal* **20**:265–277. DOI: https://doi.org/10.1046/j.1365-313X.1999.t01-1-00600.x, PMID: 10571887

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 Genetics* **8**:e1002502. DOI: https://doi.org/10.1371/journal.pgen.1002502, PMID: 22359513

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 Evolutionary Biology* **12**:152. DOI: https://doi.org/10.1186/1471-2148-12-152, PMID: 22909344

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, et al. 2010. A draft sequence of the neandertal genome. *Science* **328**:710–722. DOI: https://doi.org/10.1126/science.1188021, PMID: 20448178

Guo YL, 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. *PNAS* **106**:5246–5251. DOI: https://doi.org/10.1073/pnas.0808012106, PMID: 19307580

Hedrick PW. 1998. Balancing selection and MHC. *Genetica* **104**:207–214. DOI: https://doi.org/10.1023/A:1026494212540, PMID: 10386384

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology* **22**:4606–4618. DOI: https://doi.org/10.1111/mec.12415, PMID: 23906376

Henning F, Meyer A. 2014. The evolutionary genomics of cichlid fishes: explosive speciation and adaptation in the postgenomic era. *Annual Review of Genomics and Human Genetics* **15**:417–441. DOI: https://doi.org/10.1146/annurev-genom-090413-025412, PMID: 24898042

Herms DA, Mattson WJ. 1992. The dilemma of plants: to grow or defend. *The Quarterly Review of Biology* **67**:283–335. DOI: https://doi.org/10.1086/417659

Holub EB. 1994. Phenotypic and Genotypic Characterization of Interactions Between Isolates of *Peronospora parasitica* and Accessions of *Arabidopsis thaliana* . *Molecular Plant-Microbe Interactions* **7**:223–239. DOI: https://doi.org/10.1094/MPMI-7-0223

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 Genetics* **9**:e1003766. DOI: https://doi.org/10.1371/journal.pgen.1003766, PMID: 24068949

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 ZX, Li K, Gao G, Yin Y, Wang W, et al. 2014. Altitude adaptation in tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**:194–197. DOI: https://doi.org/10.1038/nature13408, PMID: 25043035

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). *Molecular Ecology* **21**:1223–1238. DOI: https://doi.org/10.1111/j.1365-294X.2012.05460.x, PMID: 22288429

Hurka H, Neuffer B. 1997. Evolutionary processes in the genusCapsella (Brassicaceae). *Plant Systematics and Evolution* **206**:295–316. DOI: https://doi.org/10.1007/BF00987954

Jones JD, Dangl JL. 2006. The plant immune system. *Nature* **444**:323–329. DOI: https://doi.org/10.1038/nature05286, PMID: 17108957

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. DOI: https://doi.org/10.1038/nature13439, PMID: 25043057

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780. DOI: https://doi.org/10.1093/molbev/mst010, PMID: 23329690

Kliebenstein DJ, Rowe HC. 2008. Ecological costs of biotrophic versus necrotrophic pathogen resistance, the hypersensitive response and signal transduction. *Plant Science* **174**:551–556. DOI: https://doi.org/10.1016/j.plantsci.2008.03.005

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**:3276–3278. DOI: https://doi.org/10.1093/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. DOI: https://doi.org/10.1038/335268a0, PMID: 3412487

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. 2013a. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**:1578–1582. DOI: https://doi.org/10.1126/science.1234070, PMID: 23413192

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. 2013b. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**:1578–1582. DOI: https://doi.org/10.1126/science.1234070, PMID: 23413192

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. DOI: https://doi.org/10.1093/bioinformatics/btp352, PMID: 19505943

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI: https://doi.org/10.1093/bioinformatics/btp324, PMID: 19451168

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. PMID: 12618410

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. *The EMBO Journal* **7**:2765–2774. DOI: https://doi.org/10.1002/j.1460-2075.1988.tb03131.x, PMID: 2460344

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. DOI: https://doi.org/10.1038/332651a0, PMID: 2895893

McDowell JM. 1998. Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* Locus of *Arabidopsis*. *The Plant Cell Online* **10**:1861–1874. DOI: https://doi.org/10.1105/tpc.10.11.1861

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. *The Plant Journal* **22**:523–529. DOI: https://doi.org/10.1046/j.1365-313x.2000.00771.x, PMID: 10886772

Meyers BC. 2003. Genome-Wide analysis of NBS-LRR-Encoding genes in *Arabidopsis*. *The Plant Cell Online* **15**:809–834. DOI: https://doi.org/10.1105/tpc.009308

Noel L. 1999. Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of arabidopsis. *The Plant Cell Online* **11**:2099–2112. DOI: https://doi.org/10.1105/tpc.11.11.2099

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, et al. 2016. Sequencing of the genus arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics* **48**:1077–1082. DOI: https://doi.org/10.1038/ng.3617, PMID: 27428747

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A. 2013. Assembling Genomes and Mini-Metagenomes from Highly Chimeric Reads. In: *Research in Computational Molecular Biology*. Berlin: Springer Verlag. p. 158–170.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *arabidopsis* thaliana with short reads. *Genome Research* **18**:2024–2033. DOI: https://doi.org/10.1101/gr.080200.108, PMID: 18818371

Ossowski S, Schneeberger K, Lucas-Lledó 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. DOI: https://doi.org/10.1126/science.1180677, PMID: 20044577

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. DOI: https://doi.org/10.1534/genetics.112.145037, PMID: 22960212

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLOS Biology* **14**:e1002379. DOI: https://doi.org/10.1371/journal.pbio.1002379, PMID: 26871574

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**:904–909. DOI: https://doi.org/10.1038/ng1847, PMID: 16862161

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**:359–371. DOI: https://doi.org/10.1038/nrg3936, PMID: 25963373

Rebernig CA, Lafon-Placette C, Hatorangan MR, Slotte T, Köhler C. 2015. Non-reciprocal interspecies hybridization barriers in the capsella genus are established in the endosperm. *PLOS Genetics* **11**:e1005295. DOI: https://doi.org/10.1371/journal.pgen.1005295, PMID: 26086217

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. *Current Biology* **26**:1183–1189. DOI: https://doi.org/10.1016/j.cub.2016.02.062, PMID: 27112291

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. DOI: https://doi.org/10.1534/genetics.166.3.1517, PMID: 15082565

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D. 2009. Simultaneous alignment of short reads against multiple genomes. *Genome Biology* **10**:R98. DOI: https://doi.org/10.1186/gb-2009-10-9-r98, PMID: 19761611

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. *PNAS* **109**:18493–18498. DOI: https://doi.org/10.1073/pnas.1210603109, PMID: 23091028

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. *The Plant Cell* **23**:3156–3171. DOI: https://doi.org/10.1105/tpc.111.088237, PMID: 21954462

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*. *Nature Communications* **6**:7960. DOI: https://doi.org/10.1038/ncomms8960, PMID: 26268845

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. DOI: https://doi.org/10.1111/j.1558-5646.2011.01540.x

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo YL, 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, et al. 2013. The capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nature Genetics* **45**:831–835. DOI: https://doi.org/10.1038/ng.2669, PMID: 23749190

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. *Molecular Ecology* **20**:3306–3320. DOI: https://doi.org/10.1111/j.1365-294X.2011.05189.x, PMID: 21777317

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. DOI: https://doi.org/10.1038/23260

Szkiba D, Kapun M, von Haeseler A, Gallach M. 2014. SNP2GO: functional analysis of genome-wide association studies. *Genetics* **197**:285–289. DOI: https://doi.org/10.1534/genetics.113.160341, PMID: 24561481

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. *Molecular Biology and Evolution* **32**:1186–1196. DOI: https://doi.org/10.1093/molbev/msv007, PMID: 25605789

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; International Journal of Organic Evolution* **68**:2211–2224. DOI: https://doi.org/10.1111/evo.12427, PMID: 24749791

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**:2325–2327. DOI: https://doi.org/10.1093/bioinformatics/btg316, PMID: 14630667

Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002. Signature of balancing selection in *arabidopsis*. *PNAS* **99**:11525–11530. DOI: https://doi.org/10.1073/pnas.172203599, PMID: 12172007

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. DOI: https://doi.org/10.1038/nature01588, PMID: 12721627

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. DOI: https://doi.org/10.1038/nature09083, PMID: 20520716

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**:1157–1165. PMID: 7982569

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. *Journal of Quaternary Science* **24**:3–17. DOI: https://doi.org/10.1002/jqs.1227

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**:e112963. DOI: https://doi.org/10.1371/journal.pone.0112963, PMID: 25409509

Walling LL. 2009. Adaptive Defense Responses to Pathogens and Insects. In: *In Advances in Botanical Research*. Academic Press. p. 551–612.

Walters D, Heil M. 2007. Costs and trade-offs associated with induced resistance. *Physiological and Molecular Plant Pathology* **71**:3–17. DOI: https://doi.org/10.1016/j.pmpp.2007.09.008

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. DOI: https://doi.org/10.1016/j.gene.2011.05.012, PMID: 21664259

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. DOI: https://doi.org/10.1038/346060a0, PMID: 2114550

Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA. 2015. Genetic rescue to the rescue. *Trends in Ecology & Evolution* **30**:42–49. DOI: https://doi.org/10.1016/j.tree.2014.10.009, PMID: 25435267

Whitney KD, Randell RA, Rieseberg LH. 2006. Adaptive introgression of herbivore resistance traits in the weedy sunflower *helianthus annuus*. *The American Naturalist* **167**:794–807. DOI: https://doi.org/10.1086/504606, PMID: 16649157

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 Genetics* **10**:e1004622. DOI: https://doi.org/10.1371/journal.pgen.1004622, PMID: 25255320

Wiuf C, Zhao K, Innan H, Nordborg M. 2004. The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics* **168**:2363–2372. DOI: https://doi.org/10.1534/genetics.104.029488, PMID: 15371365

Wright SI, Ness RW, Foxe JP, Barrett SCH. 2008. Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences* **169**:105–118. DOI: https://doi.org/10.1086/523366

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. *PNAS* **95**:12398–12403. DOI: https://doi.org/10.1073/pnas.95.21.12398, PMID: 9770498

Xu X. 2006. Physical and functional interactions between Pathogen-Induced *Arabidopsis* WRKY18, WRKY40, and WRKY60 Transcription Factors. *The Plant Cell Online* **18**:1310–1326. DOI: https://doi.org/10.1105/tpc.105.037523

Yeh YH, Chang YH, Huang PY, Huang JB, Zimmerli L. 2015. Enhanced *arabidopsis* pattern-triggered immunity by overexpression of cysteine-rich receptor-like kinases. *Frontiers in Plant Science* **6**. DOI: https://doi.org/10.3389/fpls.2015.00322, PMID: 26029224

Zhang W, Fraiture M, Kolb D, Löffelhardt B, Desaki Y, Boutrot FF, 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. *The Plant Cell* **25**:4227–4241. DOI: https://doi.org/10.1105/tpc.113.117010, PMID: 24104566

Zhang L, Kars I, Essenstam B, Liebrand TW, Wagemakers L, Elberse J, Tagkalaki P, Tjoitang D, van den Ackerveken G, van Kan JA. 2014. Fungal endopolygalacturonases are recognized as microbe-associated molecular patterns by the *arabidopsis* receptor-like protein responsiveness to botrytis polygalacturonases1. *Plant Physiology* **164**:352–364. DOI: https://doi.org/10.1104/pp.113.230698, PMID: 24259685

Zipfel C. 2008. Pattern-recognition receptors in plant innate immunity. *Current Opinion in Immunology* **20**:10–16. DOI: https://doi.org/10.1016/j.coi.2007.11.003, PMID: 18206360