Genome of selfing Siberian *Arabidopsis lyrata* explains establishment of allopolyploid *Arabidopsis kamchatica*

Uliana K. Kolesnikova[1,*], Alison Dawn Scott[1,*], Jozefien D. Van de Velde[1], Robin Burns[2], Nikita P. Tikhomirov[3,4], Ursula Pfordt[1], Andrew C. Clarke[5], Levi Yant[6], Xavier Vekemans[7], Stefan Laurent[8], Polina Yu. Novikova[1,§]

[1] Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, Cologne, 50829, Germany

[2] Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge CB2 3EA, UK

[3] Faculty of Biology, Lomonosov Moscow State University, Leninskie Gory Str. 1–12, Moscow, 119991, Russia

[4] Papanin Institute for Biology of Inland Waters, Russian Academy of Sciences, Borok, Yaroslavl region, 152742, Russia

[5] Future Food Beacon of Excellence and School of Biosciences, University of Nottingham, Sutton Bonington, LE12 5RD, UK

[6] Future Food Beacon of Excellence and School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, UK

[7] Univ. Lille, CNRS, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France

[8] Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, Cologne, 50829, Germany

[§] corresponding author, *co-first authors

Abstract

In outcrossing plants a lack of mating partners at range edges or resulting from karyotypic changes can be alleviated by a transition to self-compatibility. Here we determine the genetics of this transition in diploid Siberian *Arabidopsis lyrata* and allotetraploid *A. kamchatica*. We first provide chromosome-level genome assemblies for one Siberian and one North American selfing *A. lyrata* accession, and then give the structure of a fully assembled *S*-locus in the Siberian accession. We then reconstruct the events leading to the loss of self-incompatibility in Siberian *A. lyrata* and date it to ~140 Kya, showing it was an independent transition to selfing from that of North American *A. lyrata*. Moreover, using both *A. lyrata* genomes we detect segregating structural variants up to ~2.4 Mb in size. Finally, we show that this selfing Siberian *A. lyrata* lineage is parental to allotetraploid *A. kamchatica* and explain the selfing of the latter by the dominant loss-of-function mutation in the *S*-locus inherited from *A. lyrata*. This suggests that transition to self-compatibility in one of the parental lineages promoted the establishment of the new allotetraploid *A. kamchatica*.

Introduction

Most angiosperms are hermaphroditic, with both female and male organs in the same flower, and can thus at least potentially self-fertilize. Although a transition to selfing provides an immediate advantage in the face of low population density, usually at the edges of the species distribution range (Levin 2012), it is irreversible and considered an evolutionary dead-end (Stebbins 1957; Wyatt 1988). To avoid inbreeding, different recognition systems based on pollen–pistil interactions evolved repeatedly (Charlesworth et al. 2005; Zhao et al. 2022). Additionally, several independent transitions from outcrossing to self-pollination have occurred through the degradation of such recognition systems (Shimizu and Tsuchimatsu 2015).

In Brassicaceae, a sporophytic self-incompatibility (SI) system is determined by the *S*-locus, where two main genes are linked: The *SCR* gene is expressed on the surface of pollen and serves as a ligand for the receptor kinase *SRK* gene, which is expressed on the surface of stigma (Takayama and Isogai 2005; Nasrallah 2019). A breakdown of SI and transition to selfing occurs when recognition between *SCR* and *SRK* is impaired. In outcrossing *Arabidopsis* species (e.g., *A. lyrata*, *A. halleri*, *A. arenosa*) more than 50 different S haplotypes can segregate in a population (Castric et al. 2008). This diversity is essential for a SI system to function and has been maintained by frequency-dependent balancing selection for over 8 My (Mable et al. 2003; Castric and Vekemans 2004; Mable et al. 2004; Castric et al. 2008; Llaurens et al. 2008; Le Veve et al. 2022). A diploid outcrossing individual can possess two different S haplotypes but, to increase the chances of reproduction, only one of them will be expressed. This is achieved by a strong dominance hierarchy among S haplotypes. Such a hierarchy is conditioned by different trans-acting microRNA precursors and their targets on recessive S haplotypes. MicroRNAs produced on dominant S haplotypes silence the expression of the SCR gene on recessive S haplotypes (Tarutani et al. 2010; Durand et al. 2014; Fujii and Takayama 2018), increasing the chances for successful mating in heterozygous outcrossers.

The ancestral state in the *Arabidopsis* genus is outcrossing. However, there are several selfing species: the model species *A. thaliana*, and both allotetraploids *A. suecica* and *A. kamchatica*. After polyploid formation, a key early challenge is the scarcity of compatible karyotypes for mating, and competition with established nearby diploids (Levin 1975). Although autopolyploid individuals will be compatible with the unreduced (2n) gametes of their diploid progenitors, most allopolyploids would need to become self-compatible to propagate sexually. In *A. suecica,* the transition to selfing was immediate after the cross between an *A. thaliana* with non-functional but dominant S haplotypes (Tsuchimatsu et al. 2010) and an outcrossing *A. arenosa* (Novikova et al. 2017). However, the status of *A. kamchatica* is less clear: it originated from crosses between *A. lyrata* and *A. halleri* in East Asia (Shimizu et al. 2005; Shimizu-Inatsugi et al. 2009; Tsuchimatsu et al. 2012; Paape et al. 2018), but the self-compatibility status of the progenitor populations is unknown. *A. halleri* is an obligate outcrosser, whereas *A. lyrata* is predominantly outcrossing with the only known selfing populations restricted to the Great Lakes region of North America (Mable et al. 2005; Foxe et al. 2010; Griffin and Willi 2014). Selfing in *A. kamchatica* is most probably driven by loss of function of the *SCR* gene on the *A. lyrata* subgenome with the dominant S haplotype (Tsuchimatsu et al. 2012). However, since the *A. lyrata* subgenome of *A. kamchatica* and selfing populations of *A. lyrata* in North America do not appear closely related and bear different S haplotypes, the origin of *A. kamchatica* and the mechanism of its transition to selfing has remained an enigma.

We discovered a previously undescribed selfing lineage of *A. lyrata* in Siberia ranging between Lake Taymyr and Chukotka, across north-central and eastern Russia. Here we first present chromosome level assemblies of a Siberian selfing *A. lyrata* and the reference North American selfing accession (Hu et al. 2011), characterize the genomic and structural differences between them, and describe the *S*-locus structure and the mechanism of the failure of self-incompatibility in the Siberian selfing populations. Then, exploring overall genetic relatedness and phylogeny at the *S*-locus, we show that the Siberian selfing *A. lyrata* was likely a progenitor of the allotetraploid *A. kamchatica*. Together, our results explain the origin and transition to selfing in the allotetraploid *A. kamchatica*.

Results

Genome assembly of the selfing Siberian NT1 accession

We collected seeds from three *A. lyrata* populations during an expedition to the Yakutia region in Russia in the summer of 2019 (Supplementary Data 1), one of which (NT1) appeared to be selfing. NT1 samples were collected on a sandy island in the course of the Lena river (GPS coordinates 66.80449, 123.46546; Supplementary Figure 1a shows a picture of the collection site). We noticed that these NT1 plants formed long fruits when grown in the greenhouse. We confirmed that pollen successfully germinated in a selfed NT1 accession and made pollen tubes, whereas attempted self-pollination of an outcrossing plant from the NT8 population did not result in pollen tube growth (Supplementary Figure 2). We also noticed that flowers of the selfing NT1 accession are substantially smaller compared to flowers of outcrossing plants and another selfing accession MN47 from North America (Supplementary Fig 1b,c).

We extracted DNA from NT1 leaf tissue and obtained 1,100,878 high-fidelity (HiFi) PacBio reads with N50 read length of 14,161bp (total length of raw read sequences is ~15,9Gbp) from one SMRT cell on a PacBio Sequel II. We assembled those reads using the Hifiasm (Cheng et al. 2021) into 1,070 contigs with N50 of 5.508 MB. We scaffolded these contigs further along the MN47 *A. lyrata* assembly (Hu et al. 2011) with RagTag (Alonge et al. 2019) reaching chromosome-level with an N50 of 24.641Mb. We then assessed completeness of the NT1 *A. lyrata* genome assembly using BUSCO and found 4,463 complete and single-copy (97.1%), 88 complete and duplicated (1.9%), 7 fragmented (0.2%), and 38 missing genes (0.8%) from the Brassicales_odb10 set. Repeated sequences composed about 49.9% of the assembly. We annotated 28,596 genes by transferring gene annotation from the reference *A. lyrata* genome (Rawat et al. 2015) using Liftoff (Shumate and Salzberg 2020).

Various papers (Long et al. 2013; Slotte et al. 2013; Henry et al. 2014; Burns et al. 2021; Dukić and Bomblies 2022) have reported potential artifacts in the reference *A. lyrata* MN47 (version 1 or v1) genome assembly (Hu et al. 2011). Our comparison of the Siberian NT1 with the MN47 v1 *A. lyrata* reference genome indicated multiple structural variants in the same genomic regions as those between the genomes of MN47 v1 and the *A. arenosa* subgenome of

A. suecica (Fig.1A), MN47 v1 and *Capsella rubella,* and MN47 v1 and a diploid *A. arenosa* (Supplementary Data 2) (Long et al. 2013; Slotte et al. 2013; Burns et al. 2021; Dukić and Bomblies 2022). We confirmed the existence of such artifacts and corrected them through long read DNA sequencing. Specifically, we sequenced the MN47 accession on one SMRT cell on a PacBio Sequel II, obtaining 868,563 HiFi reads with N50 length of 20,206 bp (total length of raw read sequences is ~17,6Gbp). We then reassembled the reference genome of MN47 *A. lyrata* using ~80x coverage of HiFi reads. Contigs totalled 820 with an N50 of 23,506,252 bp, indicating that full chromosome arms were assembled as single contigs. The assembled contigs summed to ~244Mb. Contigs were scaffolded into eight chromosomes using the genomes of MN47 v1 and NT1 as guides. The scaffolded contigs amount to ~209Mb. Completeness of the new MN47 v2 *A. lyrata* genome assembly by BUSCO was 4,544 complete and single-copy (97.1%), 83 complete and duplicated (1.8%), 8 fragmented (0.2%), and 44 missing genes (0.9%) from the Brassicales_odb10 set. The placement and orientation of contigs in the scaffolds were corrected using previously published Hi-C data (Zhu et al. 2017) and by manual examination of the long reads (*see Methods*, Supplementary Figs. 2-6).

Our re-assembled long read-based MN47 v2 genome confirmed the existence of the expected artifacts in the MN47 v1 genome (Figure 1, Supplementary Figures 3-7). Interestingly, additional segregating inversions in *A. lyrata* were confirmed by comparison of the MN47 v2 genome and that of NT1. Of these inversions, three were not observed using the previous version of the MN47 genome, the most notable of which is on chromosome 1 and is ~2,4Mb in size. All the identified inversions between the genomes comparisons are listed in Supplementary Data 2. Inversions between *A. lyrata* MN47 and NT1 accessions are listed in Supplementary Table 1. In each genome comparison we identified multiple alleles of structural variants at the end of chromosome 3. This may be explained by the fact that one of the nucleolar organizer regions (NORs) of *A. lyrata* is located at the end of chromosome 3 (Lysak et al. 2006). We confirmed that chromosome 3 contains a partially assembled NOR using BLAST. Overall, we have assembled high quality chromosome-level genomes for two *A. lyrata* accessions, one of which being the reference *A. lyrata* genome, and through pairwise genome alignment we identified several inversions up to 2.4 Mb long segregating in the species.

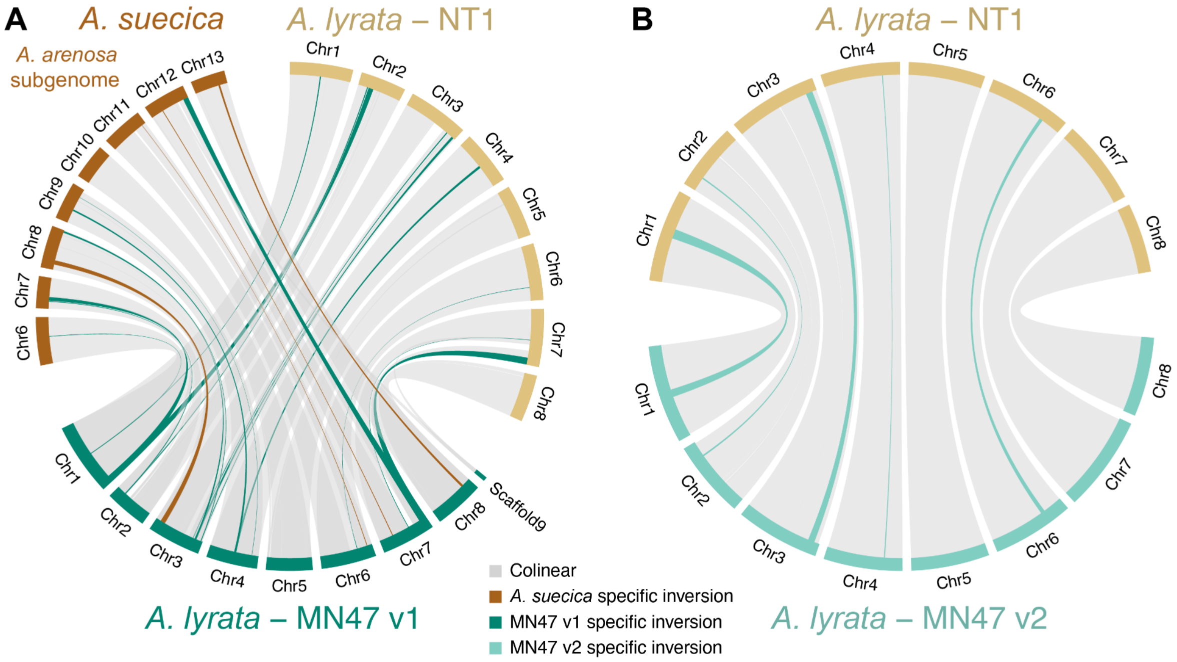


Figure 1. Segregating large structural variants in *A. lyrata*. (A) Eleven large inversions (dark green) between North American MN47 (dark green) and Siberian NT1 (dark yellow) *A. lyrata* are also observed between MN47 and the *A. arenosa* subgenome of *A. suecica* (brown), but are not observed in the comparison between NT1 and *A. suecica* (Supplementary Data 2), which suggests these inversions are likely artifacts in the original MN47 assembly. (B) Segregating inversions in *A. lyrata* observed following re-assembly by long reads and manual curation using Hi-C data of the North American MN47 genome (turquoise) and its alignment to the Siberian NT1 (dark yellow) *A. lyrata* genome. Five inversions that are unique to MN47 (the longest being ~2.4 Mb in size) are highlighted in turquoise.

Breakdown of the SI system in Siberian *A. lyrata* NT1

Both genes flanking *S*-locus (*U-box* and *ARK3*) were assembled in a single contig in the HiFi assembly before any scaffolding, which suggests that the entire ~44.5kb *S*-locus of the NT1 accession was fully assembled. We confirmed the completeness of the *S*-locus by mapping the PacBio reads back to the assembly; coverage was even along the *S*-locus with no gaps (Supplementary Fig. 8). BLAST analysis of *SRK* and *SCR* sequences from the known S haplotypes (Boggs et al. 2009; Tsuchimatsu et al. 2010; Guo et al. 2011; Goubet et al. 2012; Tsuchimatsu et al. 2012) revealed no hits for *SRK*, and one hit for *SCR* from the *A. halleri* S12 haplogroup (Figure 2b). Due to long-term frequency-dependent balancing selection on the *S*-locus in Brassicaceae, relatedness between S haplotypes does not follow species relatedness, such that the closest sequences to *A. halleri* S12 (AhS12) are not other *A. halleri* S haplotypes but rather specific S haplotypes from *A. lyrata* S42 (AlS42) and *A. kamchatica* D (Ak-D) (Tsuchimatsu et al. 2012). We performed estimated a phylogeny of the known SCR protein sequences (Guo et al. 2011; Goubet et al. 2012) and the manually annotated NT1 *A. lyrata* SCR sequence from the BLAST results (Figure 2a). As expected, the SCR phylogeny has a different topology than the species phylogeny, as S haplotypes are trans-specifically shared across *Arabidopsis*. The SCR phylogeny confirms that the closest haplotype to the NT1 *A. lyrata* *S*-locus is the S12 haplotype from *A. halleri* (AhS12).

We compared the structures of the AhS12 and NT1 *S*-loci (Figure 2b) and confirmed the absence of *SRK*, (i.e. the female component of the self-incompatibility system) which is sufficient to explain the selfing nature of the NT1 accession. We also mapped short reads of the NT1 accession to the NT1 genome assembly plus sequence of the intact AhS12 containing *SRK*; and found no reads mapped to *SRK* (Supplementary Fig. 9C). This provides additional confirmation of a complete loss of *SRK* from the NT1 *S*-locus. Analyzing the SCR protein sequences more closely, we also noticed that the NT1 SCR protein has lost one of the eight conserved cysteines, which are important in protein-folding and the recognition of the SCR ligand by the SRK receptor (Kusaba et al. 2001; Mishima et al. 2003; Tsuchimatsu et al. 2010). This suggests that the *SCR* is non-functional in the NT1 *A. lyrata* accession. Also, we tested for expression of *SCR* in the anther of NT1 using RNAseq data from flower buds and did not detect any transcript of the AhS12 *SCR* gene (Supplementary Fig. 9A,B). However, sequence comparison of the *SCR* region between AhS12 and NT1 showed high similarity in the promoter region (Supplementary Fig. 9D) suggesting absence of major re-arrangements that could have caused this loss of expression, yet nucleotide substitutions at critical sites cannot be excluded. To check whether *SCR* is indeed non-functional and/or not expressed in NT1 accession we performed a cross between an outcrossing *A. lyrata* accession expressing the same AhS12 haplogroup as maternal plant and NT1 pollen, which resulted in compatible reaction with successful pollen tube growth (Supplementary Figure 10).

According to the classification of S haplotypes, AhS12 belongs to the most dominant class and has an sRNA precursor which can silence the expression of *SCR* genes on recessive S haplotypes (Durand et al. 2014). Indeed, by BLAST analysis, we found an sRNA precursor sequence in the NT1 S-locus assembly similar to the mirS3 precursor of *A. halleri* S12 haplotype (Durand et al. 2014), suggesting conserved dominance class of *A. lyrata* S12.

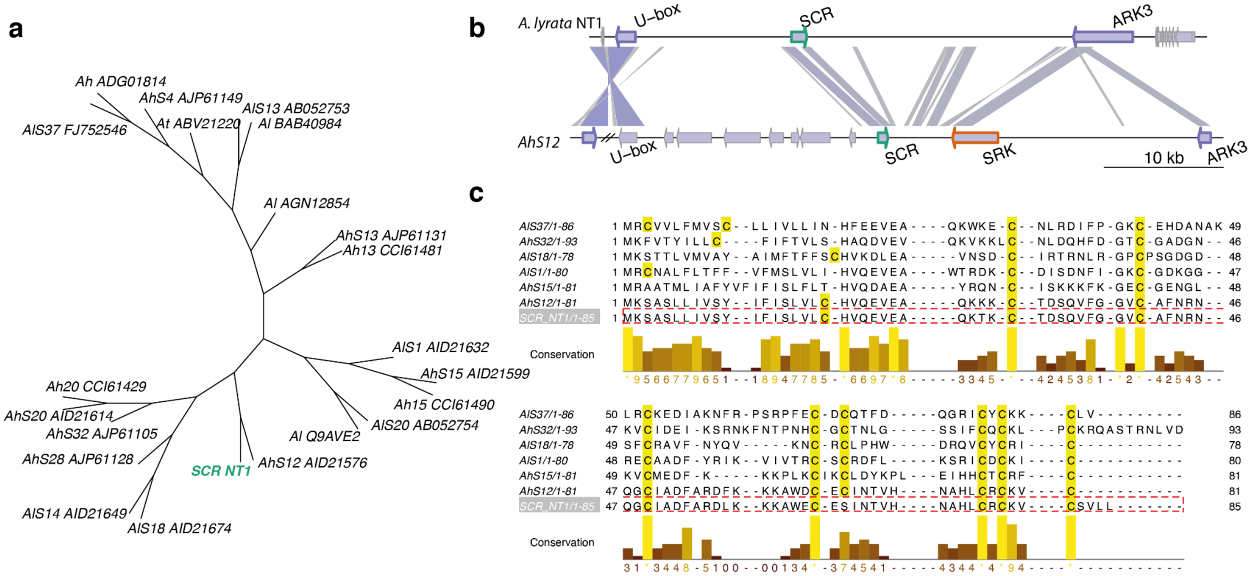

a
Ah ADG01814
AhS4 AJP61149
AlS13 AB052753
At ABV21220
Al BAB40984
AlS37 FJ752546
Al AGN12854
AhS13 AJP61131
Ah13 CCI61481
AlS1 AID21632
AhS15 AID21599
Ah15 CCI61490
AlS20 AB052754
Al Q9AVE2
AhS12 AID21576
SCR NT1
Ah20 CCI61429
AhS20 AID21614
AhS32 AJP61105
AhS28 AJP61128
AlS14 AID21649
AlS18 AID21674
b
A. lyrata NT1
U–box
SCR
ARK3
AhS12
U–box
SCR
SRK
10 kb
ARK3
c
AlS37/1-86 1 MRCVVLFMVSC--LLIVLLIN-HFEEVEA----QKWKE-C--NLRDIFP-GKC-EHDANAK 49
AhS32/1-93 1 MKFVTYILLC----FIFTVLS-HAQDVEV----QKVKKLC--NLDQHFD-GTC-GADGN-- 46
AlS18/1-78 1 MKSTTLVMVAY--AIMFTFFSCHVKDLEA-----VNSD-C--IRTRNLR-GPCPSGDGD-- 48
AlS1/1-80 1 MRCNALFLTFF--VFMSLVLI-HVQEVEA----WTRDK-C--DISDNFI-GKC-GDKGG-- 47
AhS15/1-81 1 MRAATMLIAFYVFIFISLFLT-HVQDAEA-----YRQN-C--ISKKKFK-GEC-GENGL-- 48
AhS12/1-81 1 MKSASLLIVSY--IFISLVLC-HVQEVEA-----QKKK-C--TDSQVFG-GVC-AFNRN-- 46
SCR_NT1/1-85 1 MKSASLLIVSY--IFISLVLC-HVQEVEA-----QKTK-C--TDSQVFG-GVC-AFNRN-- 46
Conservation
*95667796 51--189477 85-*6697*8-----3345-*--4245381-*2*-42543--
AlS37/1-86 50 LRCKEDIAKNFR-PSRPFEC-DCQTFD-----QGRICYCKK----CLV---------- 86
AhS32/1-93 47 KVCIDEI-KSRNKFNTPNHC-GCTNLG-----SSIFCQCKL---PCKRQASTRNLVD 93
AlS18/1-78 49 SFCRAVF-NYQV-----KNC-RCLPHW-----DRQVCYCRI----C------------ 78
AlS1/1-80 48 RECAADF-YRIK--VIVTRC-SCRDFL-----KSRICDCKI----C------------ 80
AhS15/1-81 49 KVCMEDF-K------KKPLKCIKCLDYKPL---EIHHCTCRF----C------------ 81
AhS12/1-81 47 QGCIADFARDFK--KKAWDC-ECINTVH---NAHLCRCKV----C------------ 81
SCR_NT1/1-85 47 QGCIADFARDLK--KKAWEC-ESINTVH---NAHLCRCKV----CSVLL-------- 85
Conservation
31*3448-5100--00134*-374541-----4344*4*94----*-----------

Figure 2. *S*-locus structure of the Siberian NT1 selfing *A. lyrata* population. **a**) Phylogenetic tree of SCR proteins reveals clustering of NT1 SCR (green) and AhS12. **b**) Comparison of the *S*-locus region of the *A. lyrata* NT1 genome assembly with the *A. halleri* S12 haplotype (Durand et al. 2014). Links between *S*-loci are colored according to the BLAST scores from highest (blue) to lowest (gray). *SCR*, *SRK* and flanking U-box and *ARK3* genes have green, orange and purple borders, respectively. *SRK* genes appear to be completely absent from the *S*-locus of the NT1 *A. lyrata* selfing accession. The only BLAST hit to SRK is a spurious hit to *ARK3* as they both encode receptor-like serine/threonine kinases. **c**) Protein sequence alignment of *S*-locus *SCR* genes from *A. halleri* and *A. lyrata*, including NT1. One of the eighth conserved cysteines important for structural integrity has been lost from the NT1 SCR protein.

Population level re-sequencing shows that selfing Siberian *A. lyrata* is a progenitor of *A. kamchatica*

The remaining plants (nine in total) which we collected during the same expedition (Supplementary Data 1) were sequenced using Illumina NovaSeq (150bpPE). Additionally, we sequenced 10 herbarium samples of *A. lyrata* from Taymyr, Yakutia, Kamchatka and Chukotka dating from 1958–2014 using the same platform (See Methods and Supplementary Data 1). The herbarium samples were obtained from the Moscow University Herbarium (Seregin 2020). Our dataset also included previously published whole genome resequencing data from the diploid *A. lyrata* samples collected in the same region (Novikova et al. 2016) (Supplementary Data 1, Figure 3a) and European *A. lyrata* samples (Takou et al. 2021) as outgroups, for a total of 18 samples. After mapping all samples to the NT1 reference genome, we obtained on average 23x coverage for each individual (Supplementary Data 1).

To determine whether multiple selfing populations might exist in the examined geographic region, we first calculated the percent of heterozygous sites for each individual(Supplementary Data 1, Figure 3a). Two modes on the heterozygosity levels were apparent in our *A. lyrata* dataset, which we assign as selfing (~0.4%, indicated with yellow markers) and outcrossing (~0.9% within *A. lyrata* samples, indicated by green markers). Such assignment is also supported by our observations of the live individuals growing in the greenhouse: NT1 populations produced seeds without crosses, whereas NT8 and NT12 populations did not. Allotetraploid *A. kamchatica* co-occuring in the same geographical region is also self-compatible. To ensure that none of our *A. lyrata* samples were misclassified, we included allotetraploid *A. kamchatica* samples in the dataset and mapped them in the same way to the NT1 *A. lyrata* reference without separating subgenomes. The majority of the SNPs in *A. kamchatica* represent divergent sites between the two subgenomes, which explains its high heterozygosity levels, clearly distinct from selfing *A. lyrata* samples (Figure 3).

We genotyped *S*-loci of all the short-read sequenced accessions in our dataset by using a genotyping pipeline for *de novo* discovery of divergent alleles (Genete et al. 2020). Using both *SCR* and *SRK* sequences as the reference allele databases (Guo et al. 2011; Goubet et al. 2012; Takou et al. 2020) for the discovery pipeline, we found that all the low-heterozygosity samples matched the same *S*-haplogroup – AhS12 (Fig. 3b). This suggests that all the selfing accessions originated from the same breakdown of self-incompatibility event. For each outcrossing individual we find two different SRK alleles, but only either one or no SCR alleles. This lower number of SCR alleles is likely due to an incomplete SCR database rather than these genes being absent in outcrossing individuals. Interestingly, six out of seven selfing accessions (including the reference NT1) still had *SCR* and no *SRK*, whereas accession MW0079456 had *SRK* and no *SCR*. Protein alignment of this *SRK* gene with functional *SRK* of the same haplogroup (AkSRK-D) (Tsuchimatsu et al. 2012) showed that all the structurally important cysteines are conserved in MW0079456, suggesting that *SRK* is putatively functional in MW0079456 (Supplementary Fig. 11).

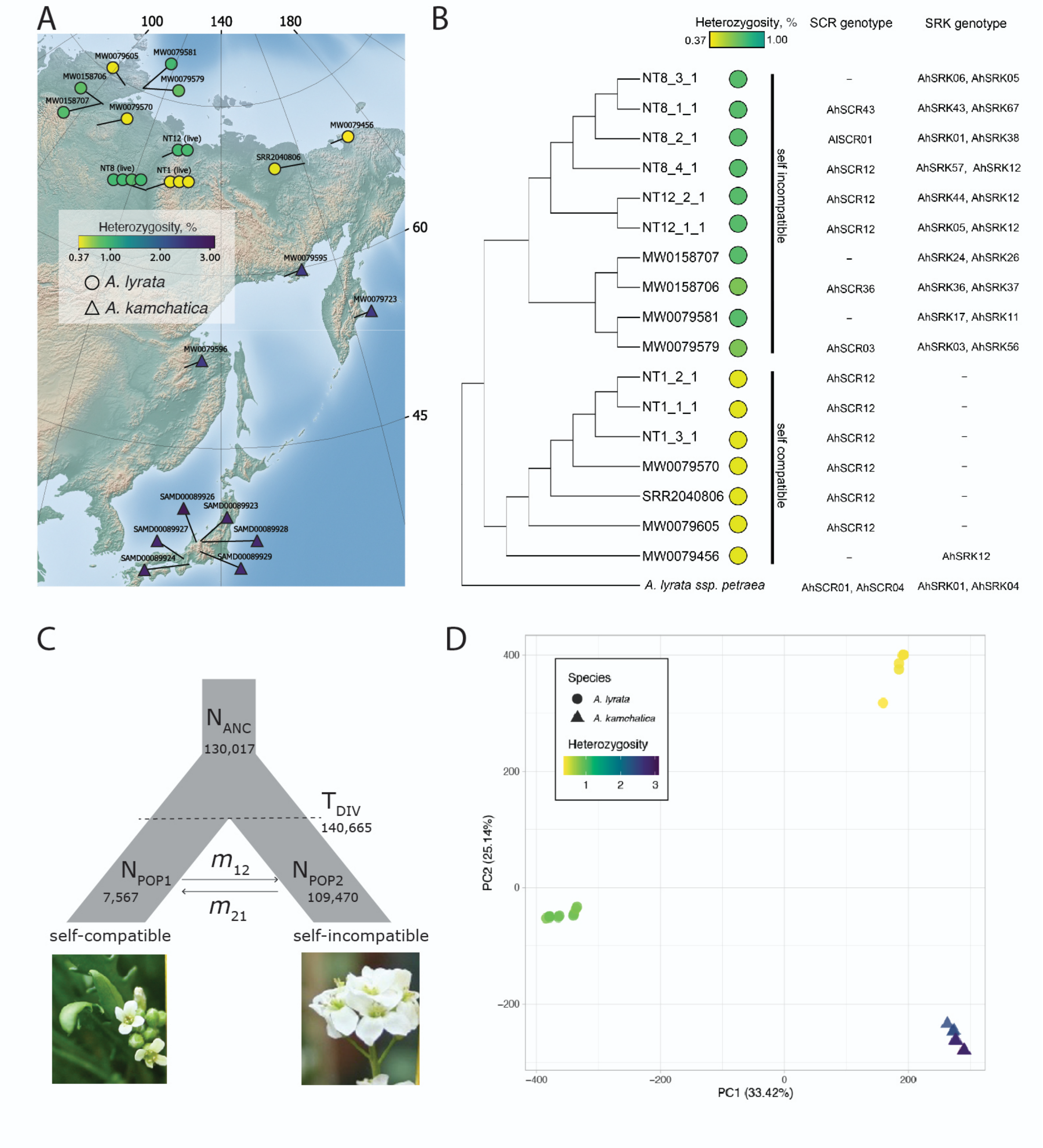


Figure 3. (A) Map of short-read sequenced Siberian *A. lyrata* (circles) and *A. kamchatica* (triangles). Live *A. lyrata* accessions names start with NT, herbarium sample names start with MW, a previously published sample of *A. lyrata* has been assigned the SRR prefix, and *A. kamchatica* samples start with SAMD. Colors indicate heterozygosity per sample, calculated by the percent of heterozygous sites. (B) Neighbor-joining tree of Siberian *A. lyrata* accessions with heterozygosity and genotyped *SCR* snd *SRK* alleles. (C) Best-fit demographic model of divergence and asymmetric migration between selfing and outcrossing lineages, with parameter estimates (D) Principal component analysis on the total genomic SNPs shows that selfing *A. lyrata* is genetically closer to *A. kamchatica* than the outcrossing populations on PC1.

The observation that all the Siberian selfing *A. lyrata* accessions have the same S haplotype suggests that they originated from a single breakdown of self-incompatibility event. The calculated total nucleotide diversity in 10kb windows for selfing *A. lyrata* has a mean value of 0.11% (95% CI [0.105 – 0.118]), which is about 7.5 times lower compared to 0.84% (95% CI [0.818 – 0.87]) in the outcrossing Siberian *A. lyrata* population. Though the selfing lineage in Siberia likely originated from a single founder, the joint allele frequency spectrum between selfing and outcrossing Siberian *A. lyrata* shows considerable amount of shared polymorphism (1,059,027 SNPs shared, vs 394,906 SNPs private to the selfer lineage; Supplemental Figure 13). This may be because the founder was a heterozygous outcrosser and a certain amount of gene flow does occur between these lineages, as self-compatibility does not prevent plants from mating with outcrossers.

To further investigate the relationships between selfing and outcrossing populations and to date the origin of the self-incompatibility breakdown, we implemented a series of demographic models in fastsimcoal26 (Excoffier et al. 2013). The best-fit model is shown (Fig. 3C), which includes divergence between selfers and outcrossers with subsequent asymmetric introgression between populations. The estimate of divergence time (TDIV) in this model is ~140 ka (140,665, 95% CI 132,499 - 133,841). All tested models can be viewed in Supplementary Figure 13, with corresponding parameters in Supplementary Table 2 and input files in Supplementary Data 3.

Allotetraploid *A. kamchatica* originated multiple times from different founding pairs of *A. lyrata* and *A. halleri* in East Asia (Dart et al. 2004; Shimizu et al. 2005; Shimizu-Inatsugi et al. 2009; Schmickl et al. 2010; Novikova et al. 2018). The most frequent S haplotype in *A. kamchatica* inherited from *A. lyrata* is Ak-D, which matches AhS12 (Tsuchimatsu et al. 2012), the same haplotype we found in the selfing Siberian *A. lyrata* populations. The phylogeny of the first exon of *SRK* from *A. kamchatica*, selfing Siberian *A. lyrata*, outcrossing *A. lyrata* and *A. halleri* shows that selfing *A. lyrata* is most closely related to *A. kamchatica* (Supplementary Fig. 12A). The principal component analysis based on 10,423,108 biallelic SNPs suggests an overall closer genome relatedness between Siberian selfing *A. lyrata* and *A. kamchatica* compared to Siberian outcrossing *A. lyrata* and *A. kamchatica* (Figure 3D). This is also seen in the hierarchical clustering analysis, where selfing Siberian *A. lyrata* populations are the closest to *A. kamchatica* (Supplementary Fig. 12B).

Discussion

We have found previously undescribed selfing populations of *A. lyrata* in Siberia, which originated independently from the North American selfing *A. lyrata*. Siberian selfing populations match only one S haplotype AhS12 (AlS42), whereas transition to selfing in the North American populations at the Great Lakes is associated with haplogroups AhS1 (AlS1), AhS31 (AlS19) and AhS29 (AlS13) (Hu et al. 2011; Mable et al. 2017). The difference in haplotypes associated with selfing lineages in Siberia and North America supports their independent origin. A transition to selfing is often associated with changes in flower morphology (Sicard and Lenhard 2011; Tsuchimatsu and Fujii 2022) which we observed in the Siberian but not in the North American selfing accessions (Supplementary Fig. 1B-D). The lack of so-called “selfing syndrome” in the latter was described previously (Carleial et al. 2017). Transition to selfing in the North American *A. lyrata* probably happened after or during colonization of the area, around ~10 Kya (Carleial et al. 2017), which is much more recent compared to our estimates of the Siberian selfer ~140 Kya. Similarly, in the outcrossing species *Leavenworthia alabamica*, two independent selfing lineages have been described, with the older one (~150 Kya) showing an obvious selfing syndrome whereas the most recent selfing lineage (~48 Kya) did not (Busch et al. 2011).

Selfing accessions can be considered natural inbred lines which are especially useful in genomics, as assembly of their genomes is not complicated by long heterozygous stretches. So far, only one selfing accession (MN47) of *A. lyrata* from North America has been fully assembled and serves as a reference for these species (Hu et al. 2011). However, while a single reference genome provides a useful resource for short-read re-sequencing-based population genetic studies (Novikova et al. 2016; The 1001 Genomes Consortium 2016), reference bias is an increasingly recognized problem. Using long and proximity-ligation reads we assembled high-quality genomes of the Siberian selfing *A. lyrata* accession NT1 and re-assembled North American *A. lyrata* MN47 accession. We found five inversions ranging from 0.3 to 2.4 Mb in length in between these independently evolved selfing accessions (Fig. 1, Supplementary Table 1). Large genomic structural rearrangements, especially inversions, can prevent chromosomal pairing and drive reproductive isolation and speciation (Rieseberg 2001; Stevison et al. 2011; McGaugh and Noor 2012; Ayala et al. 2013; Jeffares et al. 2017). In these circumstances selfing probably increases tolerance to such rearrangements and can even promote their fixation. For example, karyotypic changes from 8 to 5 chromosomes in *A. thaliana* are linked to a transition to self-compatibility at about 500 Kya (Durvasula et al. 2017). *A. lyrata* transitions to selfing are more recent but are consistent with this observation. Interestingly, the inversions found within *A. thaliana* (Jiao and Schneeberger 2020; Goel and Schneeberger 2022) and within *A. lyrata* (this study) are comparable in size: up to 2.5 Mb and 2.4 Mb respectively. However, to fully corroborate that selfing genomes are more tolerant to large structural rearrangements one must compare the results to the outcrossing genomes which are not yet available at comparable quality because heterozygosity renders them harder to assemble.

Control of selfing by the *S*-locus in *Arabidopsis* is largely determined by two genes, the male *SCR* and female *SRK*, recombination between which leads to loss of function. Multiple transposable and repeated elements populate *S*-loci, and their low homology across S-haplotypes helps to prevent recombination between different haplotypes in heterozygous individuals (Nasrallah 2005). The length of the *S*-locus can reach up to 100 kb, which makes it extremely hard to assemble with short reads. Our long-read-based genome assembly of *A. lyrata* NT1 contains a fully-assembled *S*-locus (Figure 2), in which we manually annotated *SCR* by BLAST analysis of all known *SCR* sequences in *Arabidopsis*. The *SRK* gene was absent from our assembly. Mapping of the short reads from the *A. lyrata* NT1 accession to *A. halleri* AhS12 sequence of the same haplotype also did not yield any coverage of the *SRK* gene (Supplementary Fig. 4), so we conclude that *SRK* was lost from the NT1 genome. However, this does not prove that the loss of *SRK* is the causal mutation leading to selfing, as the SCR protein of NT1 *A. lyrata* appears to have lost a functionally important cysteine residue (Kusaba et al. 2001; Mishima et al. 2003; Tsuchimatsu et al. 2010), and may have lost expression in flower buds. Genotyping of the *S*-locus in other selfing *A. lyrata* accessions reveals that all of them share the same *S*-haplotype 12, which suggests their shared origin. Moreover, one of the selfing

A. lyrata accessions has *SRK*, which appears structurally intact based on the first exon (Supplementary Figure 11), but seems to lack *SCR* (accession number MW0079456, Figure 3). This excludes the possibility that the initial cause of the transition to selfing was female-driven and suggests that the loss of one cysteine from the SCR protein sequence, or loss of expression, triggered selfing in Siberian *A. lyrata*.

Moreover, our results show that Siberian selfing *A. lyrata* contributed to the origin of the allotetraploid *A. kamchatica* or that they shared a common selfing ancestor. We explored the relationships among all Siberian *A. lyrata* accessions with *A. kamchatica* using PCA and hierarchical clustering. Principal components analysis (Fig 3D) shows that *A. kamchatica* clusters closely to self-compatible Siberian *A. lyrata* on PC1, which is consistent with the sister relationship between *A. kamchatica* and self-compatible *A. lyrata* in the neighbor joining tree (Supplementary Figure 6A). A tree of *A. lyrata* and *A. kamchatica* accessions which share the AhS12 haplotype (based on exon 1 of the *SRK* gene; Figure 6B), shows self-compatible *A. lyrata* accession is nested within a clade of *A. kamchatica* accessions, providing further support for their shared origin. Furthermore, our demographic modeling suggests the Siberian selfing lineage originated approximately 140kya, prior to the establishment of *A. kamchatica*, which was estimated by (Paape et al. 2018) at approximately 137kya. Thus it is plausible that at least one of the multiple polyploid origins of *A. kamchatica* included this selfing Siberian *A. lyrata* lineage as a parental genome donor.

Because *SRK* in the *A. lyrata* subgenome of *A. kamchatica* (Ak-D, same haplotype as AhS12 and AlS42) is functional (Tsuchimatsu et al. 2012), we conclude that the common ancestor of *A. kamchatica* and self-compatible Siberian *A. lyrata* had a non-functioning *SCR*. Our results show that indeed, SCR from the NT1 accession is not recognised by a functional SRK of the same haplogroup from NT8.4-24 accession (Supplementary Fig. 10). Whether the initial loss-of-function in the SCR protein was due to a loss of a structurally important cysteine residue (Figure 2C) or a loss of expression (Supplementary Fig. 9A-B) is unclear. Transitioning to selfing through degradation of male specificity gene is consistent with the recurrent pattern in the evolution of self-compatibility (reviewed in (Shimizu and Tsuchimatsu 2015). An *S*-haplotype with non-functional SCR and functional SRK will produce pollen of higher fitness, as it will be compatible with all other *S*-haplotypes including itself. In contrast, a *S*-haplotype with a functional SCR and a non-functional SRK will produce pollen that will be self-compatible but incompatible with the fraction of the population carrying the same, albeit fully functional, *S*-haplotype. Pistils with a non-functional SRK do not have a higher fitness unless pollen availability is very limited, making fixation of the male-driven selfing more likely (Tsuchimatsu and Shimizu 2013).

A previous study showed that self-compatibility in *A. kamchatica* was male (SCR)-driven in the more dominant *S*-haplotype inherited from *A. lyrata* (Ah12/Al42/Ak-D) (Tsuchimatsu et al. 2012). Our results add nuance to this story by showing that *A. lyrata* was already selfing when it contributed to the allotetraploid *A. kamchatica* in a cross with *A. halleri*, rather than selfing evolving within allopolyploid *A. kamchatica*. This means that the transition to selfing in *A. kamchatica* was most likely immediate. Similar examples where a dominant loss-of-function mutation in one of the progenitors facilitated transition to selfing in allotetraploids exist in *A. suecica* (Novikova et al. 2017), *Capsella bursa-pastoris* (Bachmann et al. 2021) and *Brassica napus* (Okamoto et al. 2007; Kitashiba and Nasrallah 2014). It should be noted that both progenitors of *A. kamchatica* also co-exist in Europe without forming other allotetraploids (Clauss and Koch 2006; Schmickl et al. 2010), although it is possible to create such an interspecific cross (Sarret et al. 2009). Both *A. lyrata* and *A. halleri* in Europe are strictly outcrossing. Based on these observations, we speculate that allotetraploid establishment in Brassicaceae requires one self-compatible parent with a dominant loss-of-function haplotype.

Materials and Methods

Plant collection and growth

We collected seeds from individual plants from three populations in Siberia (NT1, NT8 and NT12): three individuals from NT1 (NT1_1, NT1_2, NT1_3), four from NT8 (NT8_1, NT8_2, NT8_3, NT8_4) and two from NT12 (NT12_1, NT12_2). Collected seeds were grown in the greenhouse at 21°C, under 16 hours of light per day until a full rosette was formed, after which plants were moved to open frames outside on the grounds of the Max Planck Institute for Plant Breeding Research in Cologne, Germany. We grew several seeds per collected bag of seeds from individual plants, each was given an additional number extension (e.g., NT1_1**_1,** NT1_1**_2**, etc.). In this work we only used the plants with last extension 1.

Pollen tube staining

Almost mature flower buds were opened and, after removing the anthers, manually pollinated. Pistils were collected 2–3 hours after pollination, fixed for 1.5 hours in 10% acetic acid in ethanol and softened in 1 M NaOH overnight. Before staining, the tissue was washed three times in KPO_4 buffer (pH 7.5). For staining we submerged the tissue in 0.01% aniline blue for 10–20 minutes. After that, pistils were transferred to slides into mounting media and observed under UV light (Lu 2011).

Long-read sequencing

DNA extraction, library preparation and long-read sequencing of the NT1 and MN47 *A. lyrata* accessions were performed by the Max Planck-Genome-centre Cologne, Germany (https://mpgc.mpipz.mpg.de/home/). High molecular weight DNA was isolated from 1.5 g material with a NucleoBond HMW DNA kit (Macherey Nagel). Quality was assessed with a FEMTOpulse device (Agilent) and quantity measured by a Quantus fluorometer (Promega). HiFi libraries were then prepared according to the manual "Procedure & Checklist - Preparing HiFi SMRTbell® Libraries using SMRTbell Express Template Prep Kit 2.0" with an initial DNA fragmentation by g-Tubes (Covaris) and final library size selection on BluePippin (Sage Science). Size distribution was again controlled by FEMTOpulse (Agilent). Size-selected libraries were then sequenced on a Sequel II device with Binding Kit 2.0 and Sequel II Sequencing Kit 2.0 for 30 h (Pacific Biosciences).

Short-read sequencing

Plant material was processed in two different ways, indicated by type I and II in Supplementary table 1.

Type I: herbarium material was extracted in a dedicated clean-room facility (Ancient DNA Laboratory, Department of Archaeology, University of Cambridge). The lab has strict entry and surface decontamination protocols, and no nucleic acids are amplified in the lab. For each accession, leaf and/or stem tissue was placed in a 2 mL tube with 2 tungsten carbide beads and ground to a fine powder using a Qiagen Tissue Lyser. Each batch of extractions included a negative extraction control (identical but without tissue). DNA was extracted using the DNeasy Plant Mini Kit (Qiagen). Libraries preparation and sequencing were performed by Novogene LTD (UK). Sequencing libraries were generated using NEBNext® DNA Library Prep Kit following manufacturer's recommendations and indices were added to each sample. The genomic DNA is randomly fragmented to a size of 350bp by shearing, then DNA fragments were end polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. The PCR products were purified (AMPure XP system) and resulting libraries were analyzed for size distribution by Agilent 2100 Bioanalyzer and quantified using real-time PCR.

Type II:
Genomic DNA was isolated with the "NucleoMag© Plant " kit from Macherey and Nagel (Düren, Germany) on the KingFisher 96Plex device (Thermo) with programs provided by Macherey and Nagel. Random samples were selected for a quality control to ensure intact DNA as a starting point for library preparation. TPase-based libraries were prepared as outlined by (Rowan et al. 2019) on a Sciclone (PerkinElmer) robotic device. Short-read (PE 150bp) sequencing was performed by Novogene LTD (UK), using a NovaSeq 6000 S4 flow cell Illumina system.

Transcriptome sequencing

We used three flash-frozen open flowers of the *A. lyrata* NT1 accession as input material for RNA and sRNA sequencing. RNA was extracted by the RNeasy Plant Kit (Qiagen) including an on-column DNase I treatment. Quality was assessed by Agilent Bioanalyser and the amount was calculated by an RNA-specific kit for Quantus (Promega). An Illumina-compatible library was prepared with the NEBNext® Ultra™ II RNA Library Prep Kit for Illumina ® and finally sequenced on a HiSeq 3000 at the Max Planck-Genome-centre Cologne.

PacBio *de novo* assembly and annotation of NT1 and MN47 *A. lyrata* accessions

Raw PacBio reads of NT1 were assembled using Hifiasm assembler (Cheng et al. 2021) in the default mode, choosing the primary contig graph as our resulting assembly. The completeness of our assembly was assessed using BUSCO (Seppey et al. 2019) with Brassicales_odb10 set. Repeated sequences were masked using RepeatMasker (Smit et al. 2013-2015) with the merged libraries of RepBase *A. thaliana* repeats and NT1 *A. lyrata* repeats, which we modeled with RepeatModeler (Smit and Hubley 2008-2015). Then, annotation from the reference MN47 genome (Rawat et al. 2015) was transferred to our NT1 repeat-masked assembly by using Liftoff (Shumate and Salzberg 2020). Contigs were reordered according to their alignment to the reference chromosomes and updated gene and repeat annotations using RagTag (Alonge et al. 2019) in the scaffolding mode without correction. Assembly of MN47 PacBio reads was done using the Hifiasm assembler with the same parameters.

HiC sequencing of NT1 *A. lyrata* accession

A chromatin-capture library of the NT1 *A. lyrata* accession was prepared by the Max Planck-Genome-centre Cologne, Germany. We followed the Dovetail® Omni-C® Kit starting with 0.5 g of fresh weight as input. Libraries were quantified and quality assessed by capillary electrophoresis (Agilent Tapestation) and then sequenced at the Novogene Ltd (UK), using a NovaSeq Illumina system.

Mapping of Hi-C reads for the *A. lyrata* accessions NT1 and MN47

To validate the assembled scaffolds of *A. lyrata*, we used proximity-ligation short read Hi-C data. For NT1, Hi-C reads were mapped to the repeat-masked NT1 genome assembly, using the mapping pipeline proposed by the manufacturer (https://omni-c.readthedocs.io/en/latest/index.html). The Dovetail Omni-C processing pipeline is based on BWA (Li and Durbin 2009), pairtools (https://github.com/mirnylab/pairtools) and Juicertools (Durand et al. 2016). We mapped the Hi-C reads for MN47 (released previously (Zhu et al. 2017)) to a repeat masked MN47 genome (Hu et al. 2011) and to a repeat masked version of the newly assembled MN47 genome (in this paper) using HiCUP (version 0.6.1) (Wingett et al. 2015). The assemblies were manually examined using Juicebox (Robinson et al. 2018). Plots of the HiC contact matrix were made using the function hicPlotMatrix from HiCExplorer (Wolff et al. 2020) (version 3.7.2).

Synteny analysis of *A. lyrata*, *A. suecica* and *C. rubella* genomes

Synteny analysis was done by performing an all-against-all BLASTp search using the CDS sequences of both genomes. We used SynMap (Haug-Baltzell et al. 2017), a tool from the online platform CoGe, with the default parameters for DAGChainer. The Quota Align algorithm was used to decide on the syntenic depth, employing the default parameters. Syntenic blocks were not merged. The results were visualized using the R (version 4.1.2) library 'circlize' (version 0.4.13), as well as using plotsr (version 0.5.3) (Goel and Schneeberger 2022) for the supplementary figures.

Validation of structural variants between NT1 and MN47 *A. lyrata* accessions

To validate the inversions (Supplementary Table 1) we used PacBio, Hi-C data and synteny analysis results. Guided by synteny analyses, we first identified inversion breakpoints. Then, we investigated the long read map at these regions and either confirmed their contiguity or manually flipped the genomic region, followed by another round or long read map investigation (Supplementary Fig x). To map the PacBio HiFi reads we used Winnowmap (Jain et al. 2020). As the last step, we analyzed the Hi-C contact maps in the same regions to show that there is no evidence for alternative genome assembly configurations (Supplementary Figure 2-6).

A. lyrata NT1 *S*-locus genotyping and manual annotation

We manually annotated the *S*-locus in our initial assembly before the reference-guided reordering and scaffolding. In the transferred annotation resulting from Liftoff (Shumate and Salzberg 2020) we found both of the flanking genes (U-box and ARK3) in the same contig. The final coordinates of the *S*-locus in the NT1 assembly on scaffold 7 are 9,291,658bp to 9,336,246bp. The length of the assembled NT1 *A. lyrata S*-locus including both flanking genes is about 44.5Kbp. We mapped PacBio long reads back to the assembled NT1 genome using minimap2 (Li 2018) with default parameters in order to make sure that there are no obvious gaps in coverage or break points (Supplementary Fig. 8). Similarly to Zhang et al. (Zhang et al. 2019), we blasted the *SRK* and *SCR* sequences from all the known *S*-haplotypes across *Arabidopsis* and *Capsella* to the *A. lyrata* NT1 *S*-locus, finding a single hit at the *SCR* gene from the AhS12 haplogroup. We constructed a comparative structure plot of *A. lyrata* NT1 and *A. halleri* S12 (Genbank accession KJ772374) *S*-loci (Fig. 2b) using the R library genoPlotR (Guy et al. 2010). We aligned SCR protein sequences using MAFFT with default parameters and estimated a phylogenetic tree with RaxML (Stamatakis 2014) using the BLOSUM62 substitution model and visualized the alignment (Fig. 2c) using Jalview2 (Waterhouse et al. 2009). The phylogenetic tree was visualized using R package “ape” (Paradis et al. 2004). To check for presence of sRNA precursor sequences in the *A. lyrata* NT1 *S*-locus, we extracted sRNA precursor regions from *A. halleri S*-locus assemblies published by Durand et al. (2014) and performed BLASTn search of their sequences against our assembly. To genotype S haplotypes in the samples resequenced with the short reads, we used the *S*-locus genotyping pipeline NGSgenotyp from (Genete et al. 2020). In order to check for expression of *SCR* in the anther of NT1 we also used the NGSgenotyp pipeline on RNA-seq data from flower buds, using a reference database with published sequences of *SCR*, and the *SCR* paralog AL3G38610 expressed in anthers as a positive control.

Short read mapping, variant calling, and variant annotation

We first filtered the short paired-end reads (150bp) for adapter contamination using bbduk.sh script from BBMap (38.20) (Bushnell 2014) with the following parameters settings: ktrim=r k=23 mink=11 hdist=1 tbo tpe qtrim=rl trimq=15 minlen=70. Then we mapped the reads to the MN47 and NT1 *A. lyrata* genome with bwa mem (0.7.17) (Li and Durbin 2009), marking shorter split reads as secondary (-M parameter). We marked potentially PCR duplicated reads with picard MarkDuplicates (http://broadinstitute.github.io/picard/), sorted and indexed the bam file with samtools (Li et al. 2009). To call variants, we used the HaplotypeCaller algorithm from GATK (McKenna et al. 2010) (3.8). We then ran GenotypeGVCF from GATK including non-variant sites on the entire sample set to generate a vcf. To estimate heterozygosity levels for samples mapped to MN47, we calculated the proportion of heterozygous sites within all the confidently called sites.

Tree estimation

Genome-wide SNP tree

We filtered the vcf generated above to include only biallelic SNPs, which resulted in 10,423,108 SNPs. We further filtered this dataset to include only Siberian *A. lyrata* and an outgroup (excluding *A. kamchatica* from this portion). These data were read into R (version) and from them we estimated a neighbor-joining tree using the nj function. We then visualized the neighbor-joining tree as a cladogram using ggtree (cite) and annotated the tips with associated data.

SRK tree

We assembled partial SRK sequences from Siberian *A. lyrata* and *A. kamchatika* accessions based on short-read sequencing data using the assembly step of the S-locus genotyping pipeline NGSgenotyp (Genete et al. 2020), aligned them with MAFFT (Katoh and Standley 2013), estimated a maximum likelihood phylogeny using RaXML (Stamatakis 2014) and visualized the tree using R library "ape" (Paradis et al. 2004).

PCR identification of AhS12 haplotype

For DNA extraction, 1cm of leaf material was frozen in liquid nitrogen and ground to a powder. We added 400 µl UltraFastPrep Buffer to the powdered tissue, then mixed, vortexed, and finally spun for five minutes at 5000 rpm. We then took 300 µl of the supernatant, added 300 µl isopropanol, and mixed by inversion. We again spun for five minutes at 5000 rpm, then discarded the supernatant and dried 10-30 minutes at 37°C. The pellet was resuspended in 200 µl 1xTE and stored at 4°C. We amplified the AhSRK12 allele by PCR using 1.5 µl of DNA solution and previously published primers (forward ATCATGGCAGTGGAACACAG, reverse CAAATCAGACAACCCGACCC) (Ruggiero et al. 2008). We ran 35 cycles consisting of 30 s at 94°C, 30 s annealing at 56.8°C, and a 40 s extension at 72°C. We visualized PCR products via gel electrophoresis using 1.5% agarose gel with GelGreen® nucleic acid stain (Supplementary figure). Accessions identified with SRK 12 (NT8.4-24) and without SRK 12 (NT8.4-20) were used in crosses (Supplementary Figure 8b-d).

Principal components analysis

For PCA, we used the set of 10,423,108 biallelic SNPs shared by Siberian *A. lyrata* and *A. kamchatica*, removing any sites with missing data. We then performed a principal components analysis in R using the *prcomp* function of the 'stats' package and plotted the first and second principal components using ggplot2 (Wickham 2016).

Demographic modeling

We calculated nucleotide diversity using all biallelic and non-variant sites in 10kb windows with custom script uploaded to github (https://github.com/novikovalab/selfing_Alyrata) Confidence intervals for the median of the distribution were calculated using the basic bootstrap method in the R package 'boot'.

To prepare an unfolded joint allele frequency spectrum of the seven self-compatible accessions and the ten self-incompatible accessions, we first filtered out missing data from the SNP-only vcf and then first polarized it using a European *A. lyrata* accession as an outgroup (to determine ancestral vs derived SNPs). Following Nordborg & Donnelly ((Nordborg and Donnelly 1997)) we excluded sites heterozygous in the selfing population and treated selfers as haploid. We then generated the joint allele frequency spectrum using easySFS (https://github.com/isaacovercast/easySFS). EasySFS produces output ready for use in fastsimcoal2 (fsc26) (Excoffier et al. 2013; Excoffier et al. 2021) , which we then used for demographic modeling. We tested three models for the origin of self-compatibility in Siberian *A.*

lyrata, (A) simple divergence, (B) divergence with symmetrical introgression (migration) (C) divergence with asymmetrical introgression.

For each model, we initiated 50 fastsimcoal2 runs. We then chose the best run for each model (the run with the best likelihood scores) and from that best run calculated the Aikake Information Criterion for the model. After selecting the model with the best AIC score, we used the maximum likelihood parameter file to generate 200 pseudo observations of joint SFS for bootstrapping. For each of the 200 pseudo observations, we initiated 50 fastsimcoal2 runs, then selected the best run for each model based on likelihood scores as above. The resulting parameter estimates from the 200 replicates were used to calculate the 95% confidence intervals in R. Site frequency spectra and other fastsimcoal2 input files (.tpl and .est) can be found in Supplementary Data 3.

Acknowledgements

We thank the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/) for performing short- and long- read libraries and PacBio sequencing in this study. We thank the curator of the Moscow University Herbarium (MW) - Alexey Segerin for providing *A. lyrata* samples for sequencing. We thank Vincent Castric and Mathieu Genete for sharing the up-to-date SRK database for the *S*-locus genotyping pipeline. We thank Jo Osborn, University of Cambridge, for assisting with the herbarium extractions. Field research was supported by the Austrian Science Fund (FWF grant P30208-B-29) and held within the state assignment of the Papanin Institute for Biology of Inland Waters Russian Academy of Sciences (theme 121051100099-5).

Data availability

The raw Illumina short reads for the 12 re-sequenced samples used in this study have been submitted to the ENA database under the project number PRJEB50329 (ERP134897). Individual accession names are listed in the Supplementary Data1. Raw PacBio HiFi reads of NT1 and MN47, Hi-C reads of NT1, RNAseq reads of NT1, and the genome assembly and annotation of *A. lyrata* NT1 and MN47 have been submitted to ENA database under the same project number PRJEB50329 (ERP134897).

Supplementary Materials

Supplementary Data 1. List of sequenced and analyzed *A. lyrata* accessions

Supplementary Data 2. List of all possible inversions between *A. lyrata* NT1, MN47v1, MN47v2, *A. suecica* and *C. rubella*.

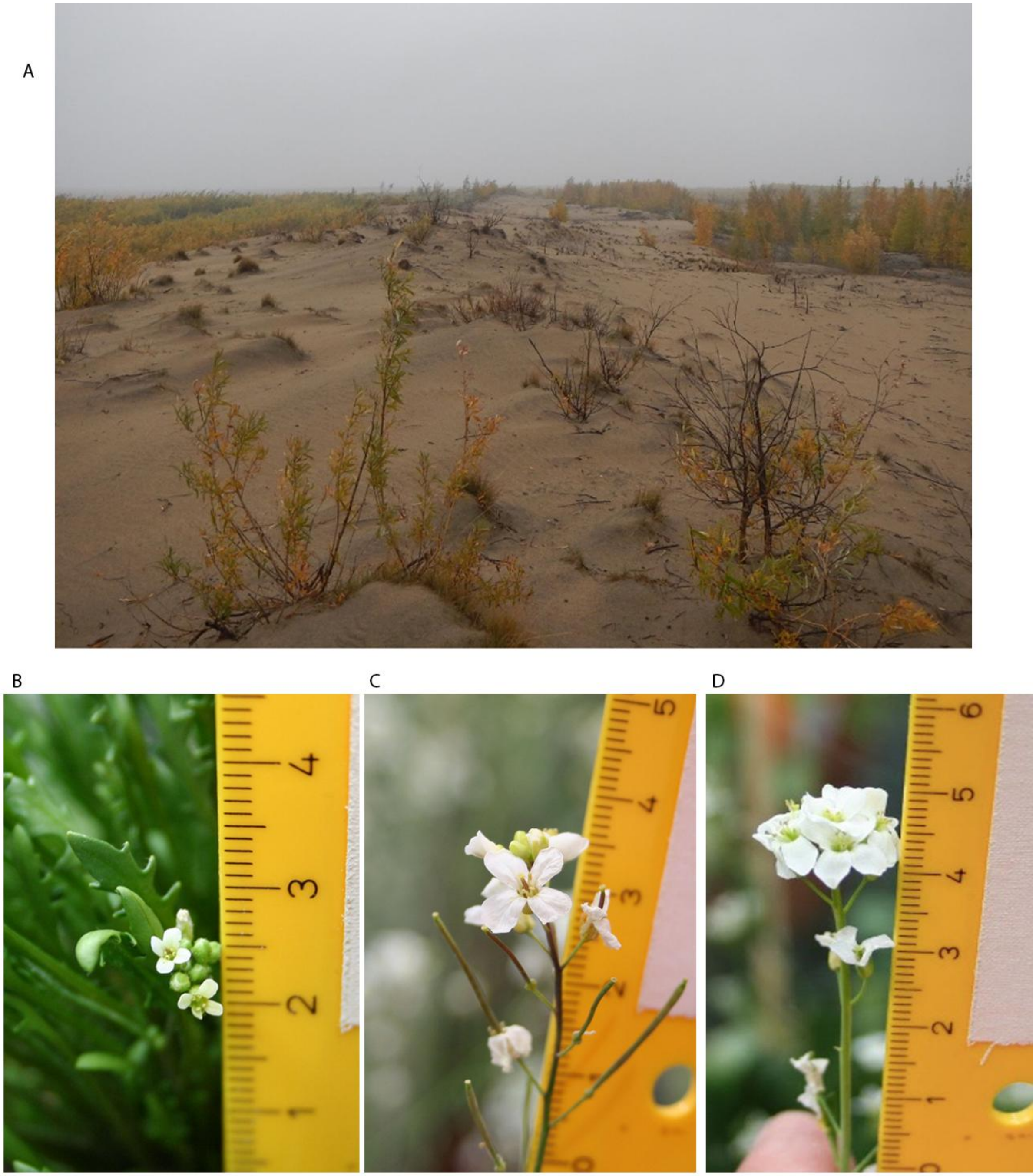
Supplementary Data 3: Input files used for demographic modeling

Supplementary Table 1. Inversions between *A. lyrata* NT1 and MN47v2, as reported by CoGe Synmap after confirmation using Hi-C and examination of long reads mapping (see Supplementary Fig. 3-7).

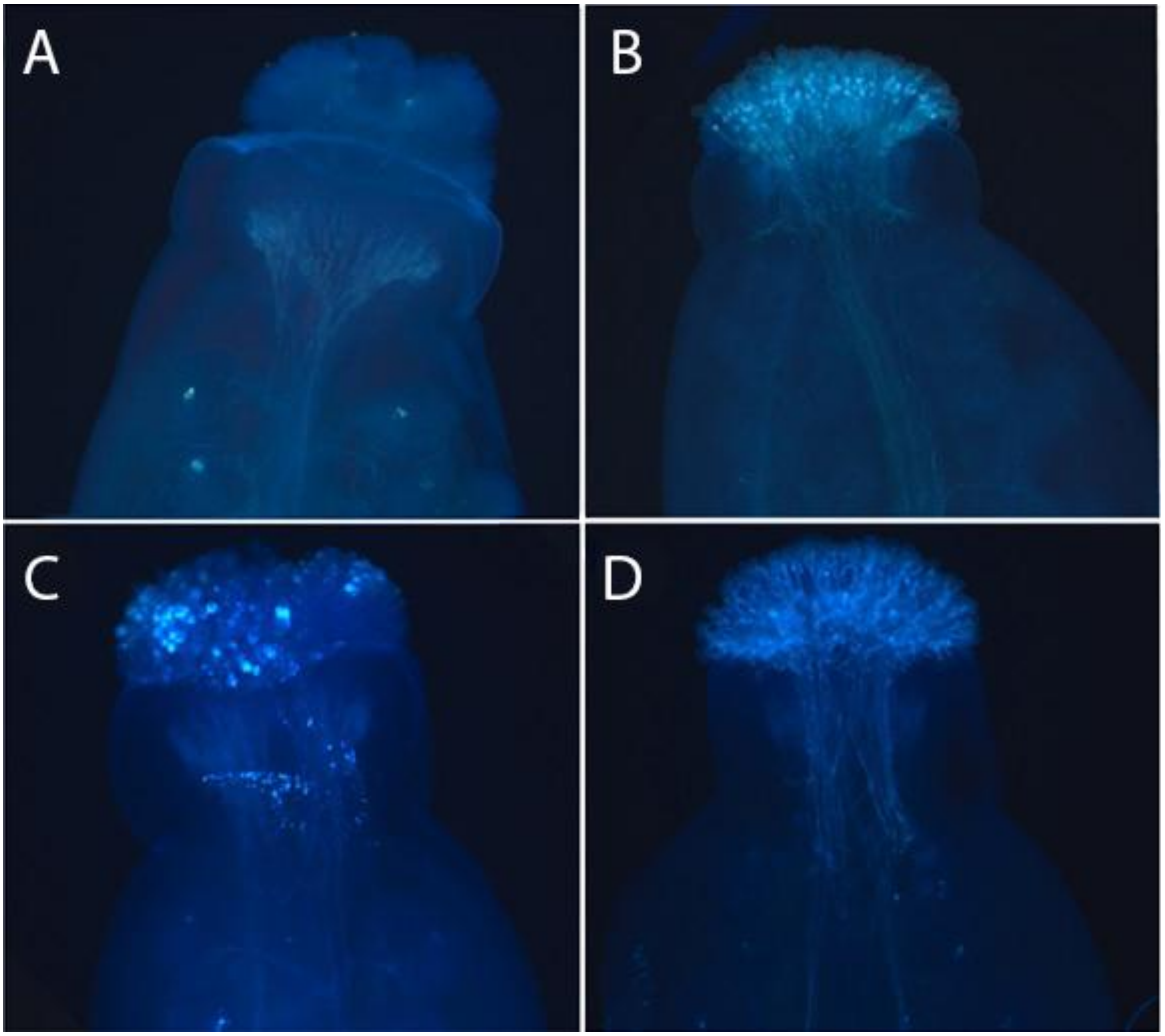
***A. lyrata* - MN47 v2**	**start**	**end**	**length**	***A. lyrata* - NT1**	**start**	**end**	**length**
chr 1	14625309	17069605	2444296	chr 1	17471724	14410341	3061383
chr 3	22131279	24228593	2097314	chr 3	26447300	24296296	2151004
chr 4	18459828	18815810	355982	chr 4	19135397	18871811	263586
chr 6	17833645	19172561	1338916	chr 6	19669769	18509483	1160286
chr 2	4890607	5375291	484684	chr 2	4823531	4397789	425742

Supplementary Table 2. Demographic parameter estimates.
Models correspond to Supplementary Figure 13. Parameters estimated by Fastsimcoal26. Model with lowest AIC score highlighted in gray, with 95% confidence interval shown. Additional details (.tpl and .est files) in Data Supplement 3.

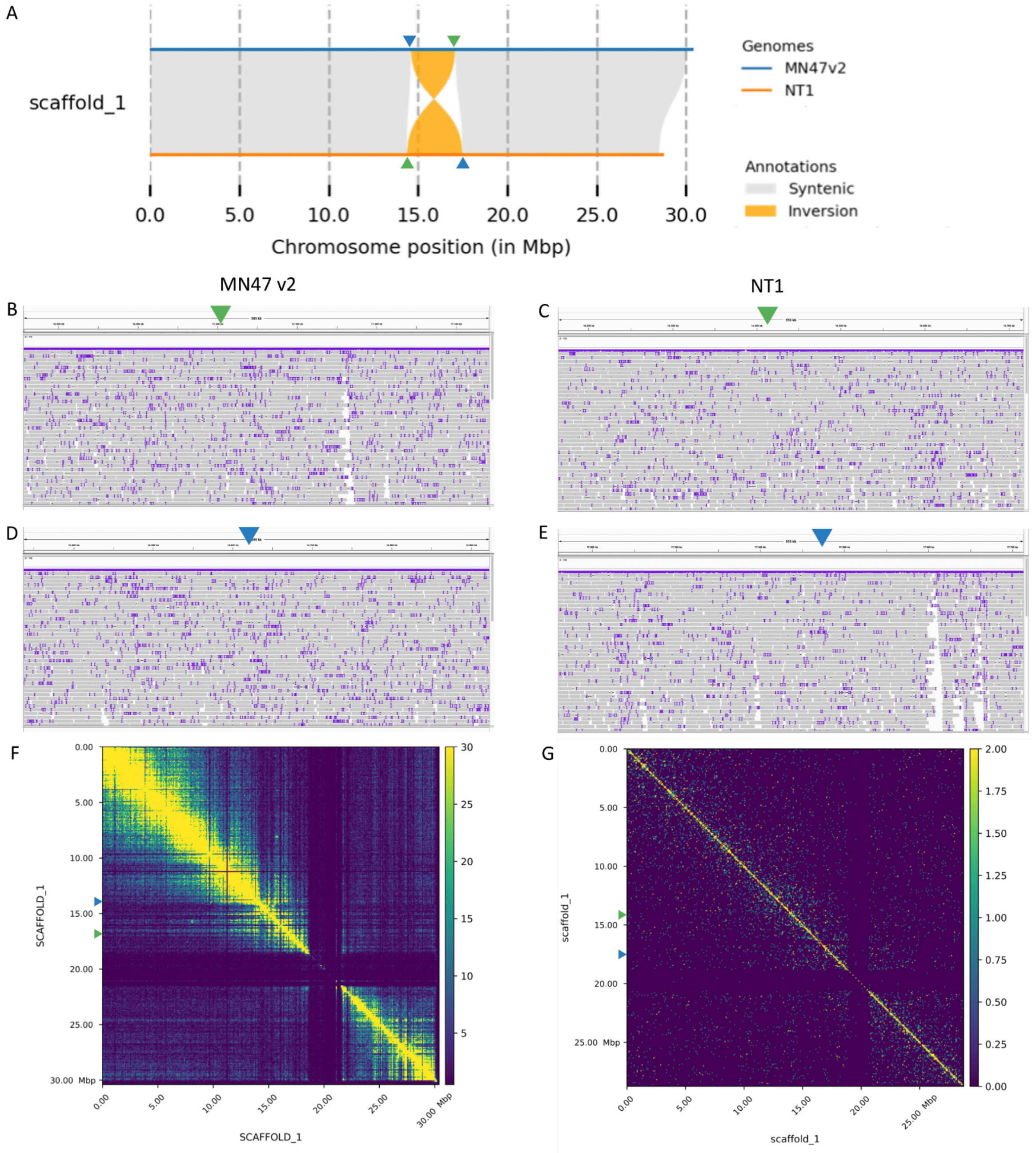
Corresponding model in Supplemental figure 13	Parameter estimates								
	NPOP1	**NPOP2**	**NANC**	**TDIV**	**migr12**	**migr21**	**MaxEstLhood**	**deltaL**	**AIC**
A	20,894	189,622	286,394	62,323	-	-	-15413165.6	3051174.4	30826339.2
B	21,092	204,004	265,858	139,479	8.60E-07	-	-15380333.8	3018342.6	30760677.6
C	15,133	218,939	260,034	140,665	5.67E-07	1.90E-06	-15373338.2	3011347.0	30746688.5
2.5%	15,099	214,977	266,628	132,499	5.34E-07	1.86E-06	-	-	-
97.5%	15,275	216,125	268,515	133,841	5.48E-07	1.92E-06			



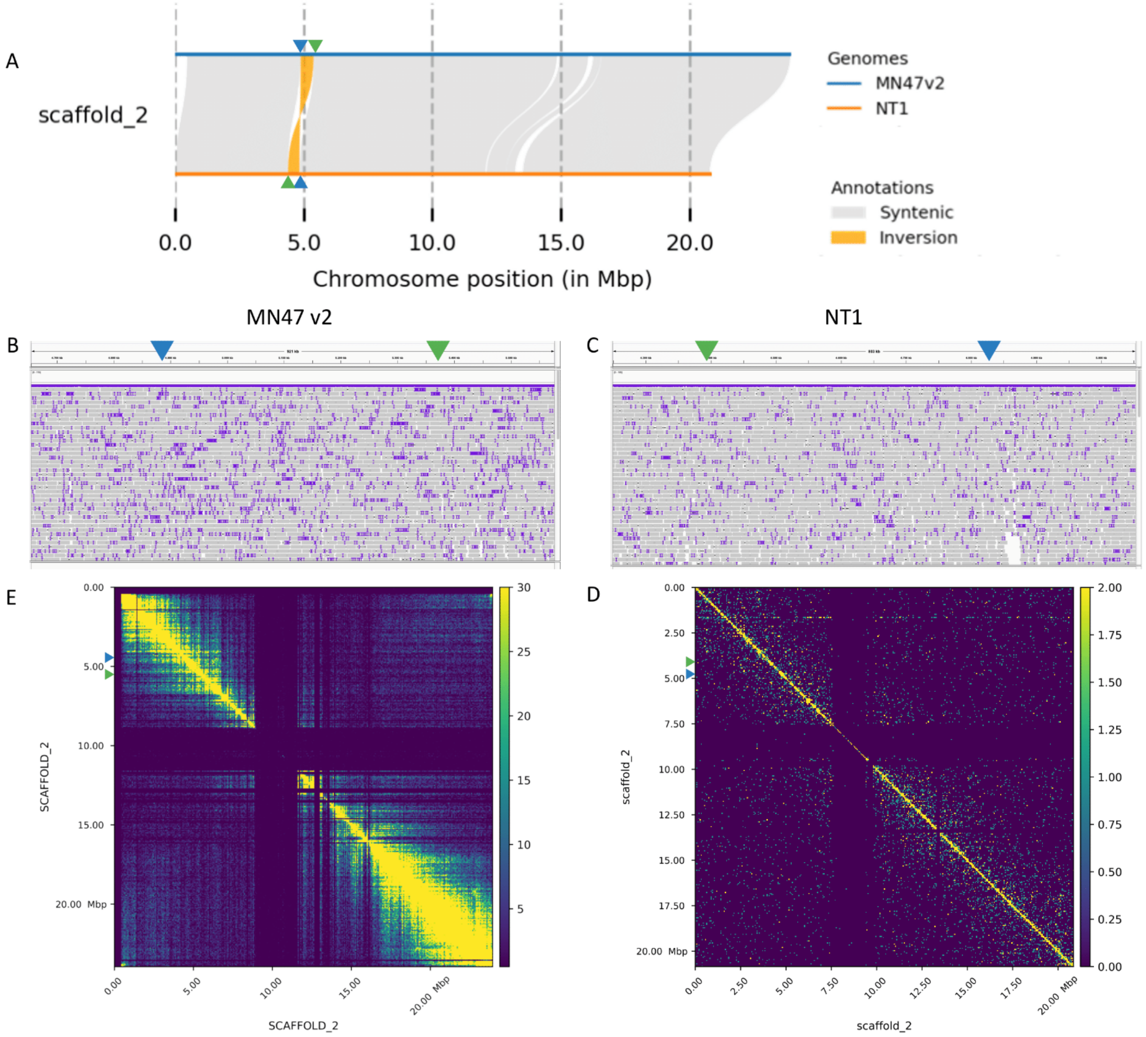
Supplementary Figure 1. (A) Photo of the collection site of the NT1 accession, showing the natural habitat of *A. lyrata* on an island in the course of the Lena River. Flowers of Siberian selfing NT1 (B) are clearly smaller in size compared to North American selfing MN47 (C) and outcrossing Siberian (D) *A. lyrata* accessions growing in the greenhouse.



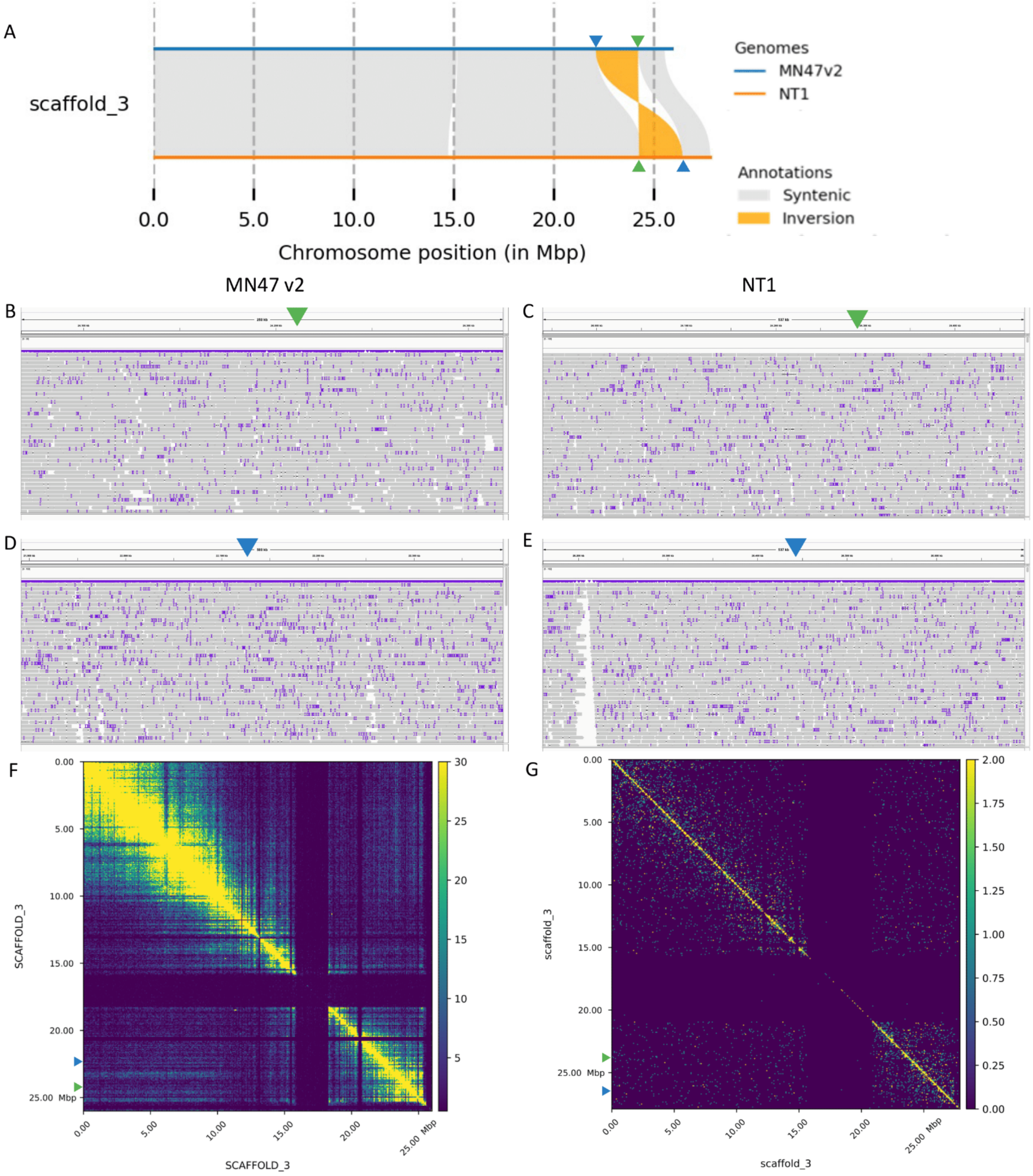
Supplementary Figure 2. Pollen tube staining confirming self-compatibility of the NT1. (A) Unpollinated stigma. (B) Self-compatible reaction: pollen tube growth on stigma of *A. lyrata* NT1 flower after self-pollination. (C) Self-incompatible reaction in outcrossing accession after self-pollination: pollen on the top of the stigma. (D) Compatible reaction: pollen tube growth in outcrossing *A. lyrata* accession after cross-pollination.



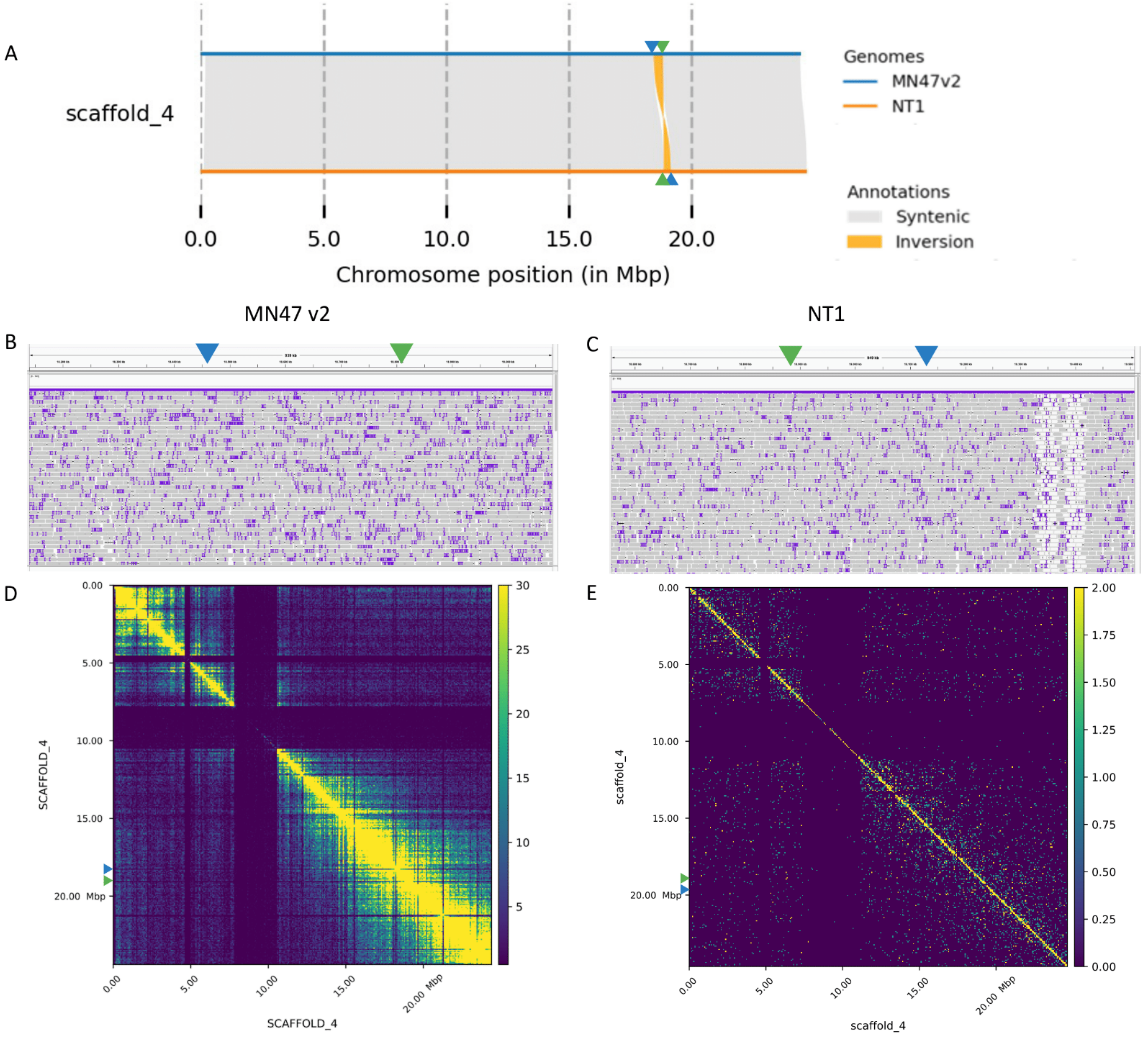
Supplementary Figure 3. Validation of an inversion on chromosome 1 between *A. lyrata* NT1 and MN47 accessions. a) Synteny and rearrangement plot of chromosome 1. (b,d) IGV snapshot showing PacBio HiFi reads of MN47 mapped to MN47 v2 assembly around inversion breakpoints (Supplementary Table 1, first row). (c,e) IGV snapshot showing PacBio HiFi reads of NT1 mapped to the NT1 assembly around inversion breakpoints (Supplementary Table 1, first row). The continuity of the long reads mapped in (b-e) suggests the inferred inversion is real. The results of the Hi-C contact maps of MN47 (f) and NT1 (g) show that a majority of the paired reads are mapped in cis, again validating the inversion. The colors on the Hi-C heatmaps (f,g) correspond to the density of the paired reads from low (dark blue) to high (yellow). The triangle markers in (a-g) indicate the start (blue triangle) and the stop (green triangle) of inversion.



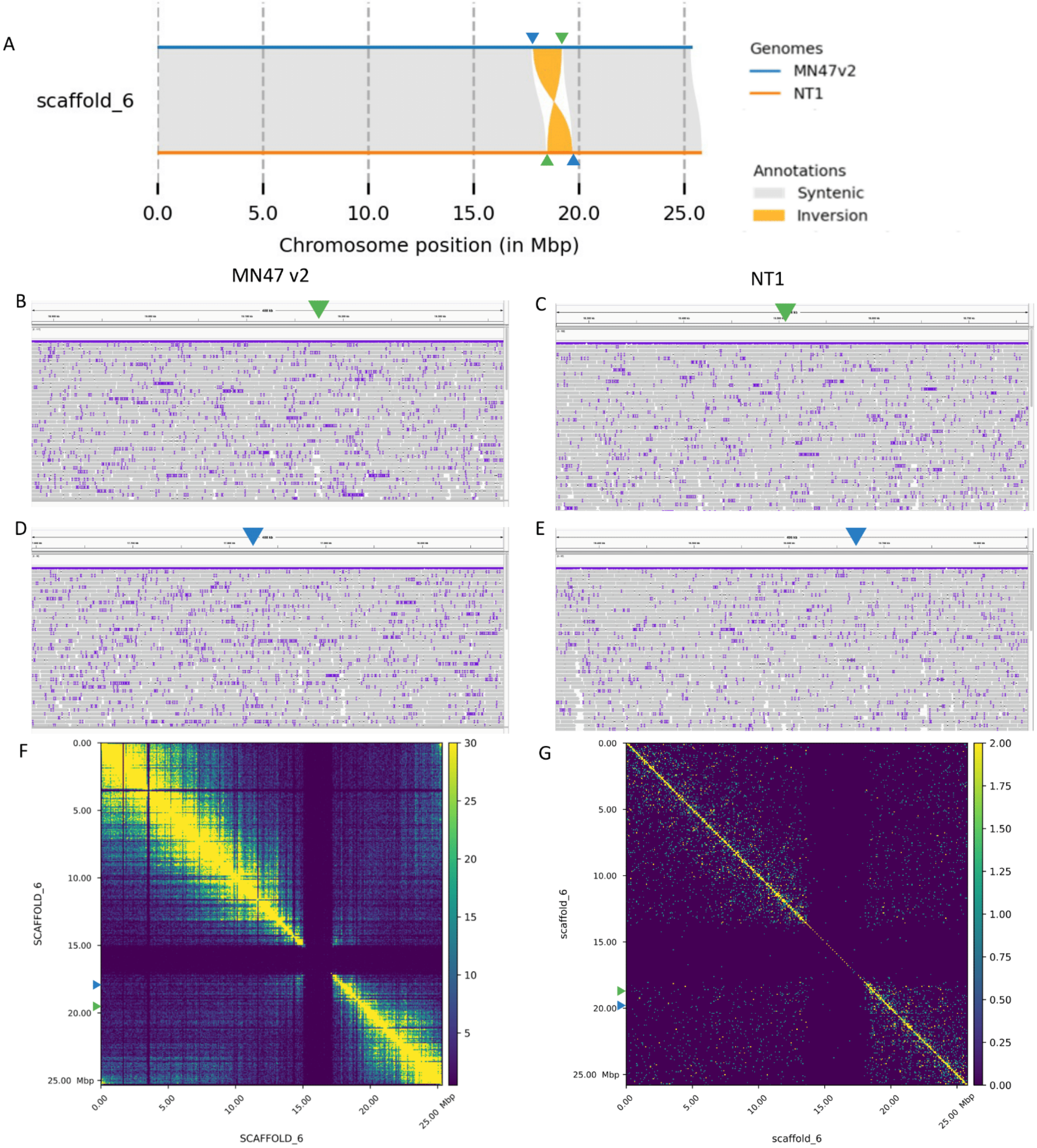
Supplementary Figure 4. Validation of an inversion on chromosome 2 between *A. lyrata* NT1 and MN47 accessions. a) Synteny and rearrangement plot of chromosome 2. b) IGV snapshot showing PacBio HiFi reads of MN47 mapped to MN47 v2 assembly around inversion breakpoints (Supplementary Table 1, second row). c) IGV snapshot showing PacBio HiFi reads of NT1 mapped to the NT1 assembly around inversion breakpoints (Supplementary Table 1, second row). The continuity of the long reads mapped in (b-c) suggests the inferred inversion is real. The results of the Hi-C contact maps of MN47 (d) and NT1 (e) show that a majority of the paired reads are mapped in cis, again validating the inversion. The colors on the Hi-C heatmaps (d,e) correspond to the density of the paired reads from low (dark blue) to high (yellow). The triangle markers in (a-e) indicate the start (blue triangle) and the stop (green triangle) of inversion.



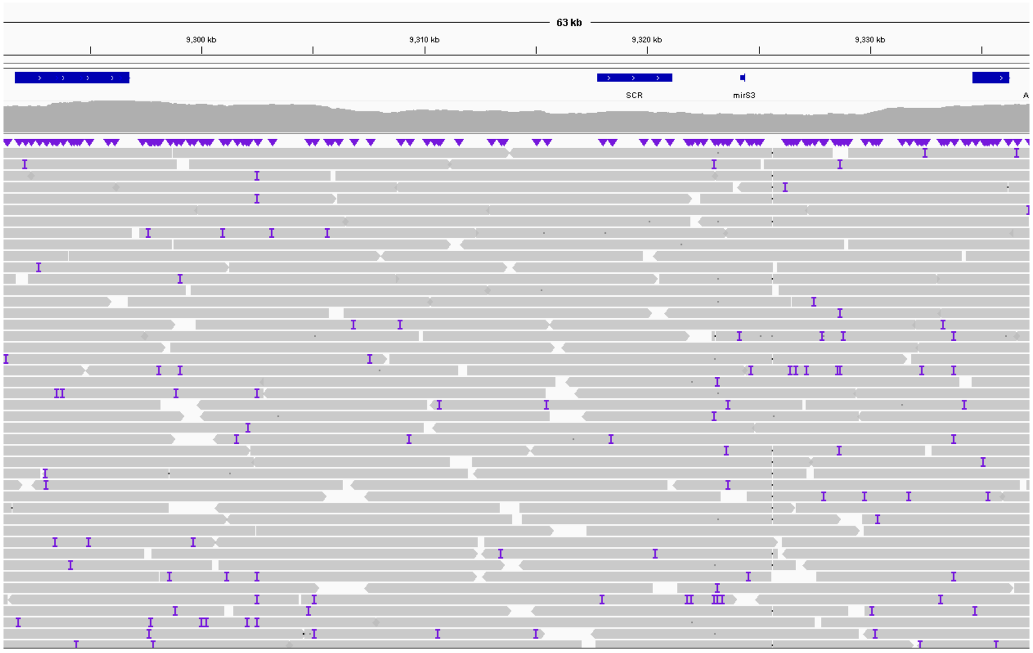
Supplementary Figure 5. Validation of an inversion on chromosome 3 between *A. lyrata* NT1 and MN47 accessions. a) Synteny and rearrangement plot of chromosome 3. (b,d) IGV snapshot showing PacBio HiFi reads of MN47 mapped to MN47 v2 assembly around inversion breakpoints (Supplementary Table 1, third row). (c,e) IGV snapshot showing PacBio HiFi reads of NT1 mapped to the NT1 assembly around inversion breakpoints (Supplementary Table 1, third row). The continuity of the long reads mapped in (b-e) suggests the inferred inversion is real. The results of the Hi-C contact maps of MN47 (f) and NT1 (g) show that a majority of the paired reads are mapped in cis, again validating the inversion. The colors on the Hi-C heatmaps (f,g) correspond to the density of the paired reads from low (dark blue) to high (yellow). The triangle markers in (a-g) indicate the start (blue triangle) and the stop (green triangle) of inversion.



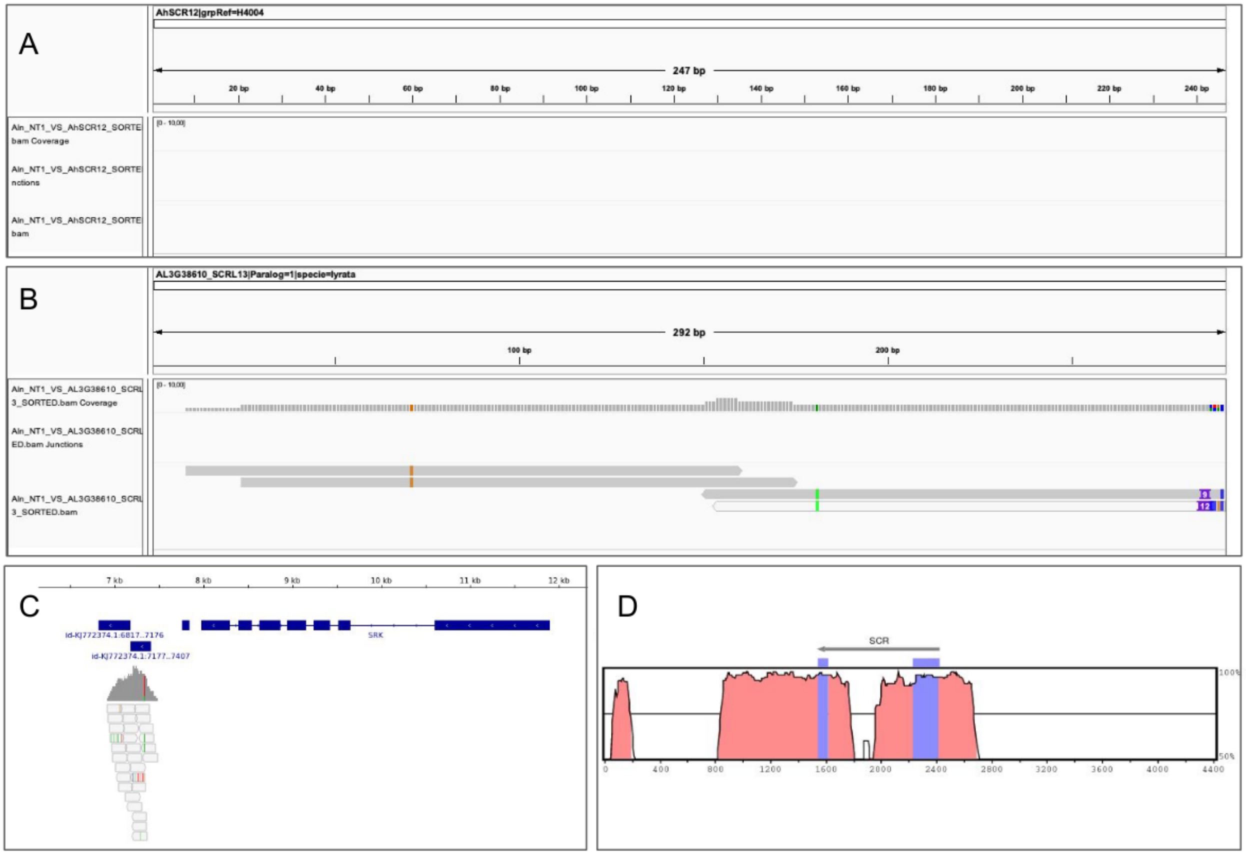
Supplementary Figure 6. Validation of an inversion on chromosome 4 between *A. lyrata* NT1 and MN47 accessions. a) Synteny and rearrangement plot of chromosome 4. b) IGV snapshot showing PacBio HiFi reads of MN47 mapped to MN47 v2 assembly around inversion breakpoints (Supplementary Table 1, fourth row). c) IGV snapshot showing PacBio HiFi reads of NT1 mapped to the NT1 assembly around inversion breakpoints (Supplementary Table 1, fourth row). The continuity of the long reads mapped in (b-c) suggests the inferred inversion is real. The results of the Hi-C contact maps of MN47 (d) and NT1 (e) show that a majority of the paired reads are mapped in cis, again validating the inversion. The colors on the Hi-C heatmaps (d,e) correspond to the density of the paired reads from low (dark blue) to high (yellow). The triangle markers in (a-e) indicate the start (blue triangle) and the stop (green triangle) of inversion.



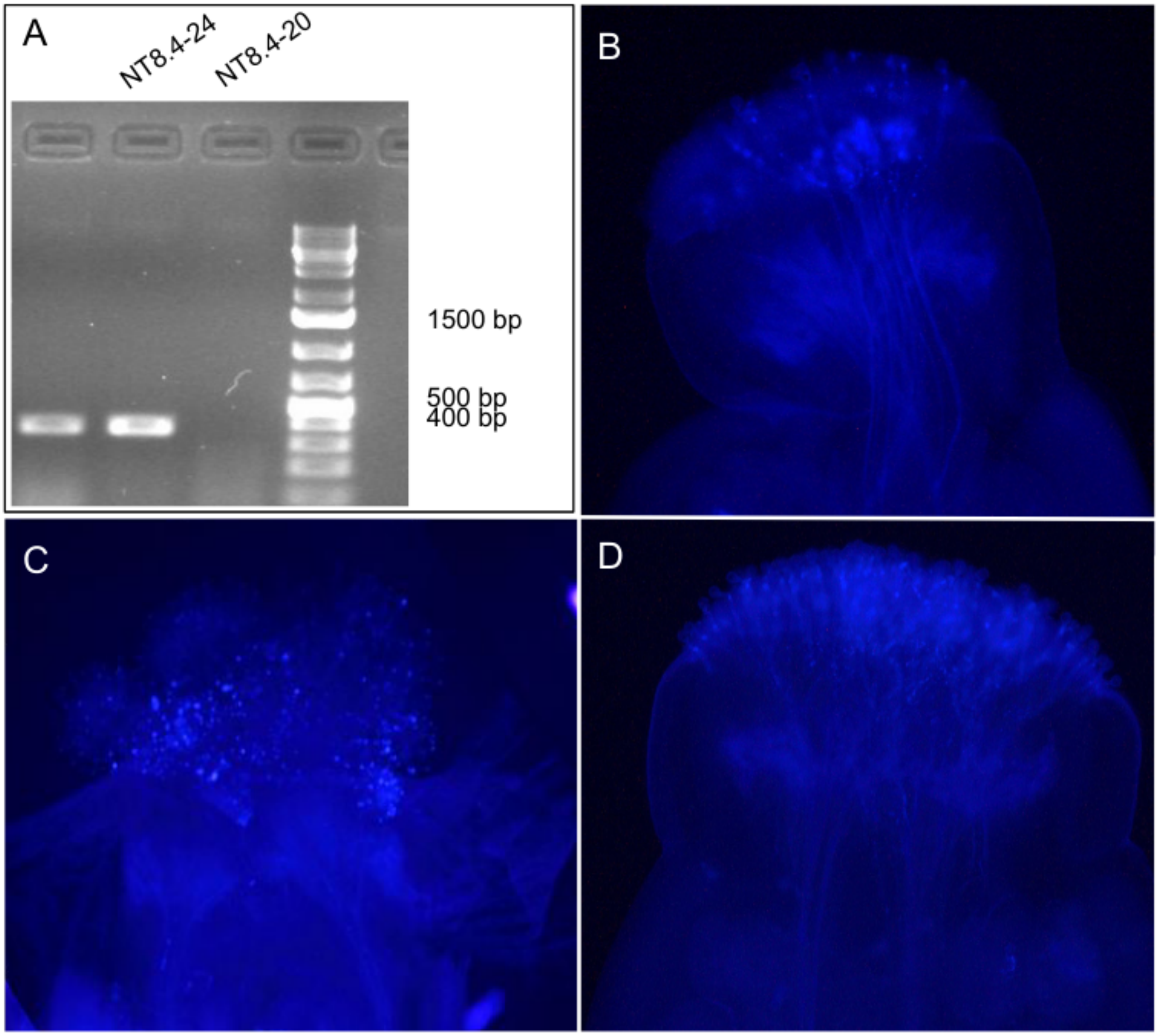
Supplementary Figure 7. Validation of an inversion on chromosome 6 between *A. lyrata* NT1 and MN47 accessions. a) Synteny and rearrangement plot of chromosome 6. (b,d) IGV snapshot showing PacBio HiFi reads of MN47 mapped to MN47 v2 assembly around inversion breakpoints (Supplementary Table 1, fifth row). (c,e) IGV snapshot showing PacBio HiFi reads of NT1 mapped to the NT1 assembly around inversion breakpoints (Supplementary Table 1, fifth row). The continuity of the long reads mapped in (b-e) suggests the inferred inversion is real. The results of the Hi-C contact maps of MN47 (f) and NT1 (g) show that a majority of the paired reads are mapped in cis, again validating the inversion. The colors on the Hi-C heatmaps (f,g) correspond to the density of the paired reads from low (dark blue) to high (yellow). The triangle markers in (a-g) indicate the start (blue triangle) and the stop (green triangle) of inversion.



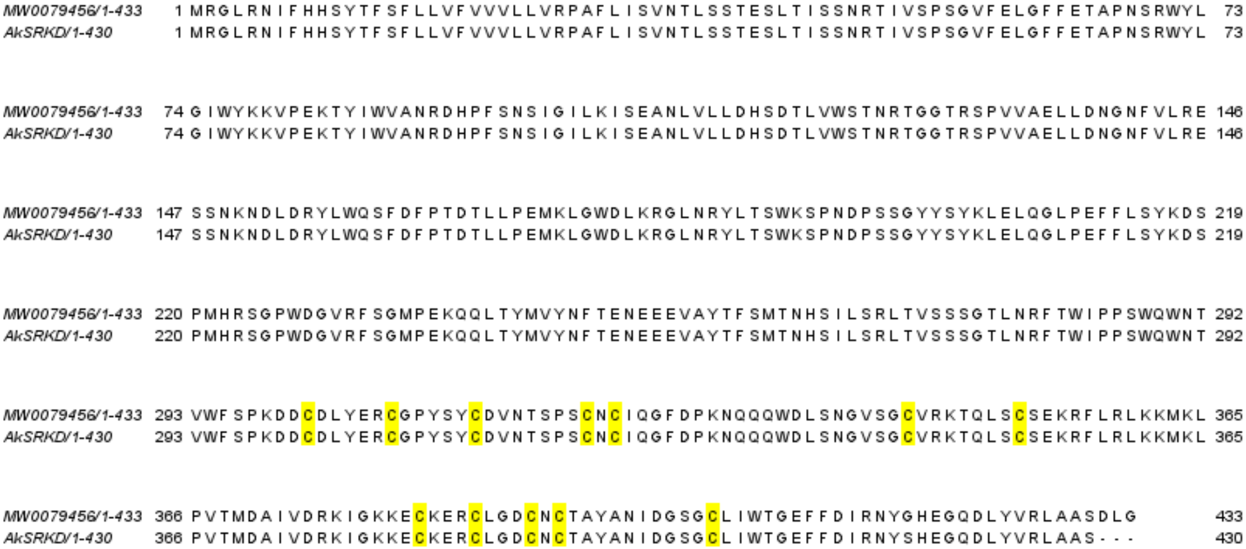
Supplementary Figure 8. IGV (Robinson et al. 2011) snapshot showing raw PacBio HiFi long-read coverage along the assembled *S*-locus in NT1 *A. lyrata* accession, on Scaffold 7. ARK3 and U-box flanking genes are marking the two ends of the *S*-locus, SCR gene is present, SRK is missing, mirS3 precursor is annotated manually. No gaps in the long read mapping suggests the completeness of the *S*-locus assembly.



Supplementary Figure 9. IGV snapshot of the *A. lyrata* NT1 short reads from RNA-seq on flower buds mapping to (A) the AhS12 SCR sequence, (B) the AL3G38610 SCR paralog, serving as positive control. The results suggest absence of expression of the SCR gene in NT1, although insufficient sequencing depth cannot be ruled out. (C) IGV snapshot of the *A. lyrata* NT1 short reads mapping to the combined reference of NT1 genome (with absent SRK) and AhS12 sequence with AhSRK12. The absence of reads mapping to AhSRK12 confirms the loss of SRK from the NT1 *A. lyrata* genome. (D) Vista plot comparing the SCR region within the S-locus of Ah12 from *Arabidopsis halleri* (used as a reference) with that of NT1. Strong similarity is observed in the promoter region suggesting the absence of major rearrangements that could have led to loss of expression.

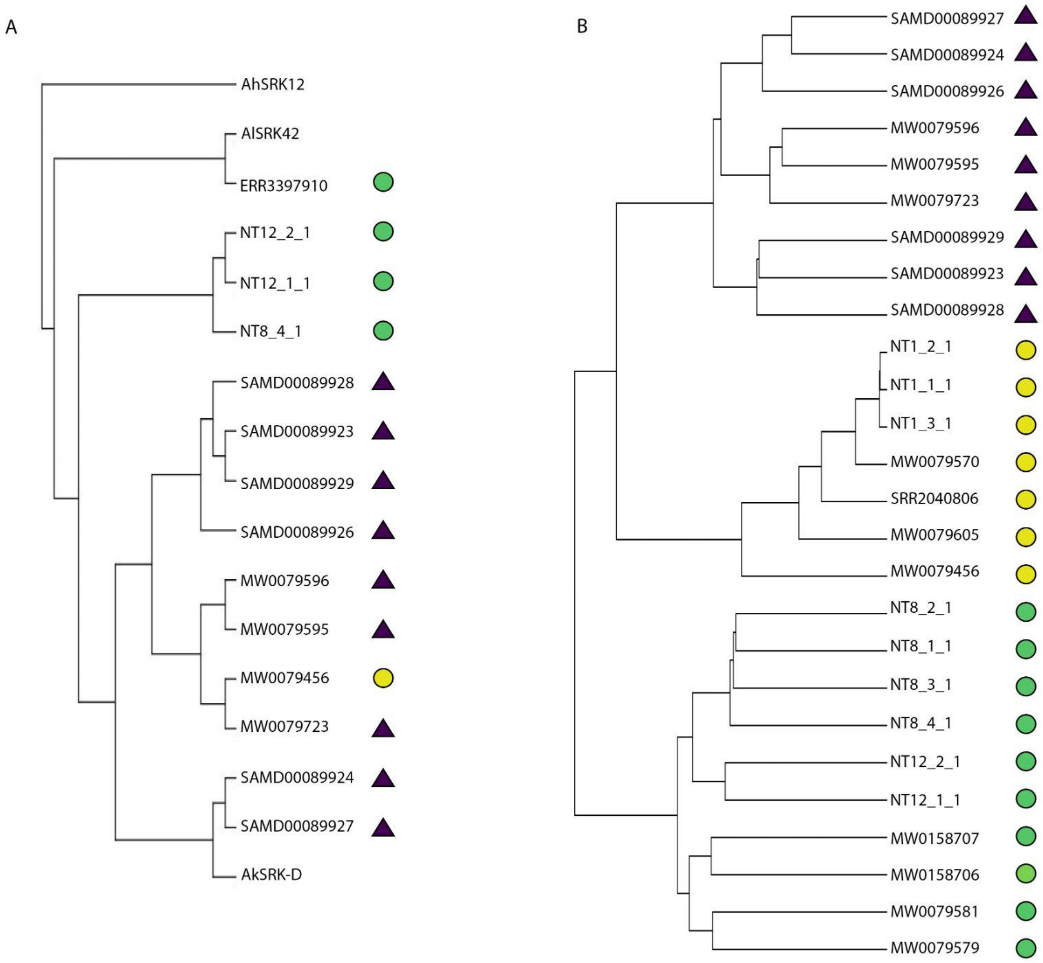


Supplementary figure 10. (A) Gel picture with PCR products and 1kb+ ladder identifying accessions with AhSRK12 (two positive bands with expected size of 360 bp). The sample in the middle (NT8.4-24) was used for crosses with NT1 (B) and another outcrossing accession without AhSRK12 (on the right, NT8.4-20) was used in (D). (B) NT8.4-24 (SRK12) ♀ x NT1 (SCR12) ♂ = compatible. This suggests that SCR12 in NT1 is not recognised by SRK12 at NT8-4-24 and therefore SCR12 in NT1 is not functional. (C) NT8.4-24 (SRK12) ♀ x NT8.4-24 (SCR12) ♂ = incompatible. Serves as a negative control and confirms that NT8.4-24 accession is an obligate outcrosser and SRK is functional in NT8.4-24. (D) NT8.4-20 (not SRK12) ♀ x NT8.4-24 (SCR12) ♂ = compatible. Serves as a positive control.

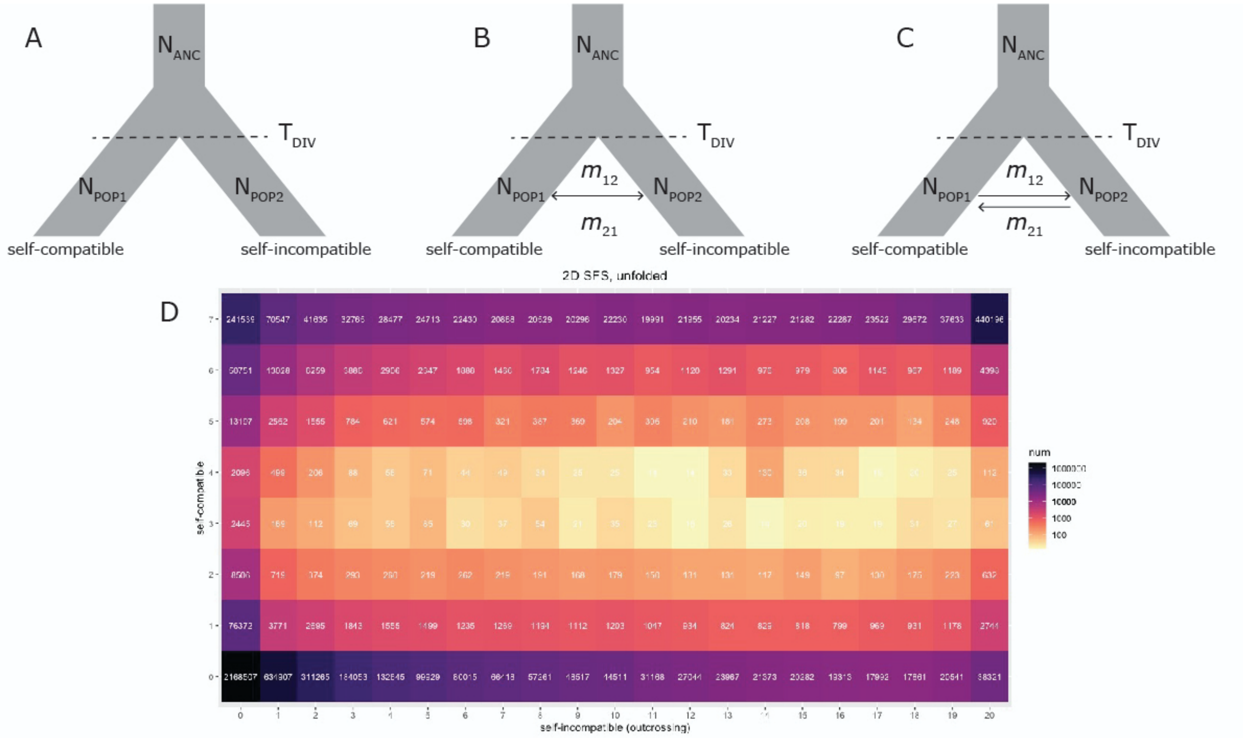


```
MW0079456/1-433    1 MRGLRNIFHHSYTFSFLLVFVVVLLVRPAFLISVNTLSSTESLTISSNRTIVSPSGVFELGFFETAPNSRWYL  73
AkSRKD/1-430       1 MRGLRNIFHHSYTFSFLLVFVVVLLVRPAFLISVNTLSSTESLTISSNRTIVSPSGVFELGFFETAPNSRWYL  73

MW0079456/1-433   74 GIWYKKVPEKTYIWVANRDHPFSNSIGILKISEANLVLLDHSDTLVWSTNRTGGTRSPVVAELLDNGNFVLRE 146
AkSRKD/1-430      74 GIWYKKVPEKTYIWVANRDHPFSNSIGILKISEANLVLLDHSDTLVWSTNRTGGTRSPVVAELLDNGNFVLRE 146

MW0079456/1-433  147 SSNKNDLDRYLWQSFDFPTDTLLPEMKLGWDLKRGLNRYLTSWKSPNDPSSGYYSYKLELQGLPEFFLSYKDS 219
AkSRKD/1-430     147 SSNKNDLDRYLWQSFDFPTDTLLPEMKLGWDLKRGLNRYLTSWKSPNDPSSGYYSYKLELQGLPEFFLSYKDS 219

MW0079456/1-433  220 PMHRSGPWDGVRFSGMPEKQQLTYMVYNFTENEEEVAYTFSMTNHSILSRLTVSSSGTLNRFTWIPPSWQWNT 292
AkSRKD/1-430     220 PMHRSGPWDGVRFSGMPEKQQLTYMVYNFTENEEEVAYTFSMTNHSILSRLTVSSSGTLNRFTWIPPSWQWNT 292

MW0079456/1-433  293 VWFSPKDDCDLYERCGPYSYCDVNTSPSCNCIQGFDPKNQQQWDLSNGVSGCVRKTQLSCSEKRFLRLKKMKL 365
AkSRKD/1-430     293 VWFSPKDDCDLYERCGPYSYCDVNTSPSCNCIQGFDPKNQQQWDLSNGVSGCVRKTQLSCSEKRFLRLKKMKL 365

MW0079456/1-433  366 PVTMDAIVDRKIGKKECKERCLGDCNCTAYANIDGSGCLIWTGEFFDIRNYGHEGQDLYVRLAASDLG      433
AkSRKD/1-430     366 PVTMDAIVDRKIGKKECKERCLGDCNCTAYANIDGSGCLIWTGEFFDIRNYSHEGQDLYVRLAAS---      430
```

Supplementary Figure 11. Protein sequence alignment of the first exon of SRK gene from the MW0079456 accession with AkSRK-D (Tsuchimatsu et al. 2012). All the structurally important cysteines are retained (highlighted yellow) suggesting functional integrity of the protein.



Supplementary Figure 12. (A) Phylogenetic tree with first exons of *SRK* sequences from *A. lyrata* AlSRK42 and *A. kamchatica* (AkSRK-D) rooted on *A. halleri* AhSRK12. Note that AlSRK42, AkSRK-D and AhSRK12 are members of the same haplogroup, which is older than Arabidopsis species. *S*-loci phylogeny between haplogroups does not follow species phylogeny, but it follows the species phylogeny within haplogroups. SRKs of Siberian selfing and outcrossing *A. lyrata* are more similar to SRK sequences from *A. kamchatica* than to European *A. lyrata* with the same haplogroup*.* The sequences of AhSRK12 (Goubet et al. 2012), AlSRK42 (Castric et al. 2008) and AkSRK-D (Tsuchimatsu et al. 2012) were published previously. (B) Hierarchical clustering of Siberian *A. lyrata* accessions with *A. kamchatica* based on whole-genome sequences. Symbols indicating species (circles for *A. lyrata* and triangles for *A. kamchatica*) and heterozygosity levels (yellow - low, green - medium and purple - high) correspond to Figure 3.



Supplementary Figure 13. Models applied for demographic modeling with fastsimcoal2. (A) Divergence of self-compatible and self-incompatible populations, without gene flow between them. (B) Model A plus equal bidirectional gene flow (migration) between self-compatible and self-incompatible lineages. (C) Model B but with asymmetric gene flow (migration) between self-compatible and self-incompatible lineages. (D) Unfolded 2D SFS of selfers and outcrossers.

References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20:224.

Ayala D, Guerrero RF, Kirkpatrick M. 2013. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution* 67:946–958.

Bachmann JA, Tedder A, Fracassetti M, Steige KA, Lafon-Placette C, Köhler C, Slotte T. 2021. On the origin of the widespread self-compatible allotetraploid Capsella bursa-pastoris (Brassicaceae). *Heredity* [Internet]. Available from: http://dx.doi.org/10.1038/s41437-021-00434-9

Boggs NA, Dwyer KG, Shah P, McCulloch AA, Bechsgaard J, Schierup MH, Nasrallah ME, Nasrallah JB. 2009. Expression of distinct self-incompatibility specificities in Arabidopsis thaliana. *Genetics* 182:1313–1321.

Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, Novikova PY, Nordborg M. 2021. Gradual evolution of allopolyploidy in Arabidopsis suecica. *Nat Ecol Evol* [Internet]. Available from: http://dx.doi.org/10.1038/s41559-021-01525-w

Busch JW, Joly S, Schoen DJ. 2011. Demographic signatures accompanying the evolution of selfing in Leavenworthia alabamica. *Mol. Biol. Evol.* 28:1717–1729.

Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States) Available from: https://www.osti.gov/servlets/purl/1241166

Carleial S, van Kleunen M, Stift M. 2017. Small reductions in corolla size and pollen: ovule ratio, but no changes in flower shape in selfing populations of the North American Arabidopsis lyrata. *Oecologia* 183:401–413.

Castric V, Bechsgaard J, Schierup MH, Vekemans X. 2008. Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection.Bergelson J, editor. *PLoS Genet.* 4:e1000168.

Castric V, Vekemans X. 2004. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. *Mol. Ecol.* 13:2873–2889.

Charlesworth D, Vekemans X, Castric V, Glémin S. 2005. Plant self-incompatibility systems: a molecular evolutionary perspective. *New Phytol.* 168:61–69.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18:170–175.

Clauss MJ, Koch MA. 2006. Poorly known relatives of Arabidopsis thaliana. *Trends Plant Sci.* 11:449–459.

Dart S, Kron P, Mable BK. 2004. Characterizing polyploidy inArabidopsis lyratausing chromosome counts and flow cytometry. *Can. J. Bot.* 82:185–197.

Dukić M, Bomblies K. 2022. Male and female recombination landscapes of diploid Arabidopsis arenosa. *Genetics* [Internet]. Available from: http://dx.doi.org/10.1093/genetics/iyab236

Durand E, Méheust R, Soucaze M, Goubet PM, Gallina S, Poux C, Fobis-Loisy I, Guillon E, Gaude T, Sarazin A, et al. 2014. Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science* 346:1200–1205.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* 3:95–98.

Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Picó FX, Alonso-Blanco C, et al. 2017. African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 114:5213–5218.

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

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. 2021. *fastsimcoal2*: demographic inference under complex evolutionary scenarios. *Bioinformatics* [Internet] 37:4882–4885. Available from: http://dx.doi.org/10.1093/bioinformatics/btab468

Foxe JP, Stift M, Tedder A, Haudry A, Wright SI, Mable BK. 2010. Reconstructing origins of loss of self-incompatibility and selfing in North American Arabidopsis lyrata: a population genetic context. *Evolution* 64:3495–3510.

Fujii S, Takayama S. 2018. Multilayered dominance hierarchy in plant self-incompatibility. *Plant Reprod.* 31:15–19.

Genete M, Castric V, Vekemans X. 2020. Genotyping and De Novo Discovery of Allelic Variants at the Brassicaceae Self-Incompatibility Locus from Short-Read Sequencing Data. *Mol. Biol. Evol.* 37:1193–1201.

Goel M, Schneeberger K. 2022. plotsr: Visualising structural similarities and rearrangements between multiple genomes. *bioRxiv* [Internet]:2022.01.24.477489. Available from: https://www.biorxiv.org/content/10.1101/2022.01.24.477489v1

Goubet PM, Berges H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl AC, Fobis-Loisy I, Vekemans X, et al. 2012. Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS Genet.* 8:e1002495.

Griffin PC, Willi Y. 2014. Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American Arabidopsis lyrata. *Ecol. Lett.* 17:484–490.

Guo YL, Zhao X, Lanz C, Weigel D. 2011. Evolution of the S-locus region in Arabidopsis relatives. *Plant Physiol.* 157:937–946.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26:2334–2335.

Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. 2017. SynMap2 and

SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* 33:2197–2198.

Henry IM, Dilkes BP, Tyagi A, Gao J, Christensen B, Comai L. 2014. The BOY NAMED SUE quantitative trait locus confers increased meiotic stability to an adapted natural allopolyploid of Arabidopsis. *Plant Cell* 26:181–194.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43:476–481.

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. 2020. Weighted minimizer sampling improves long read mapping. *Bioinformatics* 36:i111–i118.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8:14061.

Jiao W-B, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* 11:989.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

Kitashiba H, Nasrallah JB. 2014. Self-incompatibility in Brassicaceae crops: lessons for interspecific incompatibility. *Breed. Sci.* 64:23–37.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME. 2001. Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing A. lyrata and its autogamous relative A. thaliana. *Plant Cell* 13:627–643.

Le Veve A, Burghgraeve N, Genete M, Lepers-Blassiau C, Takou M, De Meaux J, Mable BK, Durand E, Vekemans X, Castric V. 2022. Long-term balancing selection and the genetic load linked to the self-incompatibility locus in Arabidopsis halleri and A. lyrata. *bioRxiv* [Internet]:2022.04.12.487987. Available from: https://www.biorxiv.org/content/10.1101/2022.04.12.487987v1

Levin DA. 1975. Minority Cytotype Exclusion in Local Plant Populations. *Taxon* 24:35–43.

Levin DA. 2012. Mating system shifts on the trailing edge. *Ann. Bot.* 109:613–620.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Llaurens V, Billiard S, Leducq J-B, Castric V, Klein EK, Vekemans X. 2008. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in Arabidopsis halleri? *Evolution* 62:2545–2557.

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V, et al. 2013. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* 45:884–890.

Lu Y. 2011. Arabidopsis Pollen Tube Aniline Blue Staining. *Bio Protoc.* [Internet] 1. Available from: https://bio-protocol.org/e88

Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I. 2006. Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. *Proc. Natl. Acad. Sci. U. S. A.* 103:5224–5229.

Mable BK, Beland J, Di Berardo C. 2004. Inheritance and dominance of self-incompatibility alleles in polyploid Arabidopsis lyrata. *Heredity* 93:476–486.

Mable BK, Hagmann J, Kim S-T, Adam A, Kilbride E, Weigel D, Stift M. 2017. What causes mating system shifts in plants? Arabidopsis lyrata as a case study. *Heredity* 118:52–63.

Mable BK, Robertson AV, Dart S, Di Berardo C, Witham L. 2005. Breakdown of self-incompatibility in the perennial Arabidopsis lyrata (Brassicaceae) and its genetic consequences. *Evolution* 59:1437–1448.

Mable BK, Schierup MH, Charlesworth D. 2003. Estimating the number, frequency, and dominance of S-alleles in a natural population of Arabidopsis lyrata(Brassicaceae) with sporophytic control of self-incompatibility. *Heredity* 90:422–431.

McGaugh SE, Noor MAF. 2012. Genomic impacts of chromosomal inversions in parapatric Drosophila species. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367:422–429.

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

Mishima M, Takayama S, Sasaki K, Jee JG, Kojima C, Isogai A, Shirakawa M. 2003. Structure of the male determinant factor for Brassica self-incompatibility. *J. Biol. Chem.* 278:36389–36395.

Nasrallah JB. 2005. Recognition and rejection of self in plant self-incompatibility: comparisons to animal histocompatibility. *Trends Immunol.* 26:412–418.

Nasrallah JB. 2019. Self-incompatibility in the Brassicaceae: Regulation and mechanism of self-recognition. *Curr. Top. Dev. Biol.* 131:435–452.

Nordborg M, Donnelly P. 1997. The coalescent process with selfing. *Genetics* 146:1185–1195.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, et al. 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* 48:1077–1082.

Novikova PY, Hohmann N, Van de Peer Y. 2018. Polyploid Arabidopsis species originated around recent glaciation maxima. *Curr. Opin. Plant Biol.* 42:8–15.

Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R, Fedorenko OM, Holm S, Sall T, Prat E, et al. 2017. Genome Sequencing Reveals the Origin of the Allotetraploid Arabidopsis suecica. *Mol. Biol. Evol.* 34:957–968.

Okamoto S, Odashima M, Fujimoto R, Sato Y, Kitashiba H, Nishio T. 2007. Self-compatibility in Brassica napus is caused by independent mutations in S-locus genes. *Plant J.* 50:391–400.

Paape T, Briskine RV, Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, Tanaka K, Nishiyama T, Sabirov R, Sese J, et al. 2018. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. *Nat. Commun.* 9:3909.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.

Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D, Pecinka A, Schneeberger K. 2015. Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLoS One* 10:e0137391.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16:351–358.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29:24–26.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* 6:256–258.e1.

Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. 2019. An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics* 213:771–787.

Ruggiero MV, Jacquemin B, Castric V, Vekemans X. 2008. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in Arabidopsis halleri. *Genet. Res.* 90:37–46.

Sarret G, Willems G, Isaure M-P, Marcus MA, Fakra SC, Frérot H, Pairis S, Geoffroy N, Manceau A, Saumitou-Laprade P. 2009. Zinc distribution and speciation in Arabidopsis halleri x Arabidopsis lyrata progenies presenting various zinc accumulation capacities. *New Phytol.* 184:581–595.

Schmickl R, Jorgensen MH, Brysting AK, Koch MA. 2010. The evolutionary history of the Arabidopsis lyrata complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evol. Biol.* 10:98.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York. p. 227–245.

Seregin A. 2020. Moscow University Herbarium (MW). Available from: http://dx.doi.org/10.15468/CPNHCC

Shimizu-Inatsugi R, Lihová J, Iwanaga H, Kudoh H, Marhold K, Savolainen O, Watanabe K, Yakubov VV, Shimizu KK. 2009. The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. *Mol. Ecol.* 18:4024–4048.

Shimizu KK, Fujii S, Marhold K, Watanabe K, Kudoh H. 2005. Arabidopsis kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh and A. kamchatica subsp. kawasakiana (Makino) K. Shimizu & Kudoh, New Combinations. *Acta phytotaxonomica et geobotanica* 56:163–172.

Shimizu KK, Tsuchimatsu T. 2015. Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. *Annu. Rev. Ecol. Evol. Syst.* 46:593–622.

Shumate A, Salzberg SL. 2020. Liftoff: accurate mapping of gene annotations. *Bioinformatics* [Internet]. Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1016

Sicard A, Lenhard M. 2011. The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Ann. Bot.* 107:1433–1443.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* 45:831–835.

Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0 http://www.repeatmasker.org.

Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. .

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stebbins GL. 1957. Self Fertilization and Population Variability in the Higher Plants. *Am. Nat.* 91:337–354.

Stevison LS, Hoehn KB, Noor MAF. 2011. Effects of Inversions on Within- and Between-Species Recombination and Divergence. *Genome Biol. Evol.* 3:830–841.

Takayama S, Isogai A. 2005. Self-incompatibility in plants. *Annu. Rev. Plant Biol.* 56:467–489.

Takou M, Hämälä T, Koch EM, Steige KA, Dittberner H, Yant L, Genete M, Sunyaev S, Castric V, Vekemans X, et al. 2021. Maintenance of Adaptive Dynamics and No Detectable Load in a Range-Edge Outcrossing Plant Population. *Mol. Biol. Evol.* 38:1820–1836.

Takou M, Hämälä T, Koch E, Steige KA, Dittberner H, Yant L, Genete M, Sunyaev S, Castric V, Vekemans X, et al. 2020. Maintenance of adaptive dynamics and no detectable load in a range-edge out-crossing plant population. *bioRxiv* [Internet]:709873. Available from: https://www.biorxiv.org/content/10.1101/709873v3

Tarutani Y, Shiba H, Iwano M, Kakizaki T, Suzuki G, Watanabe M, Isogai A, Takayama S. 2010. Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. *Nature* 466:983–986.

The 1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166:481–491.

Tsuchimatsu T, Fujii S. 2022. The selfing syndrome and beyond: diverse evolutionary consequences of mating system transitions in plants. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 377:20200510.

Tsuchimatsu T, Kaiser P, Yew CL, Bachelier JB, Shimizu KK. 2012. Recent loss of self-incompatibility by degradation of the male component in allotetraploid Arabidopsis kamchatica. *PLoS Genet.* 8:e1002838.

Tsuchimatsu T, Shimizu KK. 2013. Effects of pollen availability and the mutation bias on the fixation of mutations disabling the male specificity of self-incompatibility. *J. Evol. Biol.* 26:2221–2232.

Tsuchimatsu T, Suwabe K, Shimizu-Inatsugi R, Isokawa S, Pavlidis P, Städler T, Suzuki G, Takayama S, Watanabe M, Shimizu KK. 2010. Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene. *Nature* 464:1342–1346.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Available from: http://ggplot2.org

Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S. 2015. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* 4:1310.

Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, Grüning BA. 2020. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* 48:W177–W184.

Wyatt R. 1988. Phylogenetic aspects of the evolution of self-pollination. In: Gottlieb LD, Jain SK, editors. Plant Evolutionary Biology. Dordrecht: Springer Netherlands. p. 109–131.

Zhang T, Qiao Q, Novikova PY, Wang Q, Yue J, Guan Y, Ming S, Liu T, De J, Liu Y, et al. 2019. Genome of Crucihimalaya himalaica, a close relative of Arabidopsis, shows ecological adaptation to high altitude. *Proc. Natl. Acad. Sci. U. S. A.* [Internet]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30894495

Zhao H, Zhang Y, Zhang H, Song Y, Zhao F, Zhang Y ’e, Zhu S, Zhang H, Zhou Z, Guo H, et al. 2022. Origin, loss, and regain of self-incompatibility in angiosperms. *Plant Cell* 34:579–596.

Zhu W, Hu B, Becker C, Doğan ES, Berendzen KW, Weigel D, Liu C. 2017. Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biol.* 18:157.