Investigating Incipient Speciation in *Arabidopsis lyrata* from Patterns of Transmission Ratio Distortion

Johanna Leppälä,*,†,1 Folmer Bokma,† and Outi Savolainen*
*Department of Biology and Biocenter, University of Oulu, Oulu, 90014, Finland, †Department of Ecology and Environmental Sciences, Umeå University, Umeå S-90187, Sweden

ABSTRACT Our understanding of the development of intrinsic reproductive isolation is still largely based on theoretical models and thorough empirical studies on a small number of species. Theory suggests that reproductive isolation develops through accumulation of epistatic genic incompatibilities, also known as Bateson–Dobzhansky–Muller (BDM) incompatibilities. We can detect these from marker transmission ratio distortion (TRD) in hybrid progenies of crosses between species or populations, where TRD is expected to result from selection against heterospecific allele combinations in hybrids. TRD may also manifest itself because of intragenomic conflicts or competition between gametes or zygotes. We studied early stage speciation in *Arabidopsis lyrata* by investigating patterns of TRD across the genome in F_2 progenies of three reciprocal crosses between four natural populations. We found that the degree of TRD increases with genetic distance between crossed populations, but also that reciprocal progenies may differ substantially in their degree of TRD. Chromosomes AL6 and especially AL1 appear to be involved in many single- and two-locus distortions, but the location and source of TRD vary between crosses and between reciprocal progenies. We also found that the majority of single- and two-locus TRD appears to have a gametic, as opposed to zygotic, origin. Thus, while theory on BDM incompatibilities is typically illustrated with derived nuclear alleles proving incompatible in hybrid zygotes, our results suggest a prominent role for distortions emerging before zygote formation.

COMPLETE reproductive isolation between natural populations is usually achieved through several reproductive barriers, acting at different developmental stages. In plants prezygotic reproductive isolation acts either prepollination (*e.g.*, habitat, temporal, or pollinator isolation) or postpollination (*e.g.*, unilateral incompatibility or conspecific pollen precedence). Postzygotic isolating mechanisms either reduce viability, fertility, or fitness of hybrids or prevent development of a hybrid altogether. The effect of postzygotic reproductive barriers may be environment dependent (*e.g.*, Hatfield and Schluter 1999) or independent of the environment. The latter, intrinsic postzygotic barriers, cause reduced viability or fertility of hybrids and have been studied extensively, *e.g.*, in *Drosophila* (for a review see Coyne and Orr 2004; Presgraves 2010).

Despite the importance of development of reproductive isolation as an evolutionary process, and despite considerable study, precise molecular genetic mechanisms of postzygotic isolation have rarely been described (Presgraves 2010; Rieseberg and Blackman 2010). It is generally assumed, however, that intrinsic postzygotic isolation is due to Bateson–Dobzhansky–Muller incompatibilities (BDMI; Orr 1995; Orr and Turelli 2001). These incompatibilities emerge when in (usually) isolated populations new alleles arise, which upon secondary contact do not function together. In a few cases single genes involved in BDMIs have been identified (reviewed by Presgraves 2010) and from *Drosophila* even a pair of loci is known (Brideau *et al.* 2006), but generally the mechanism by which BDMI cause postzygotic isolation is unknown.

Genic incompatibilities that cause reduced hybrid viability or fertility may lead to non-Mendelian segregation of

doi: 10.1534/genetics.113.152561
Manuscript received August 7, 2012; accepted for publication May 3, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152561/-/DC1.

[1] Corresponding author: Division of Plant Biology, Department of Biosciences, Viikinkaari 1, 00014 University of Helsinki, Helsinki, Finland. E-mail: johanna.leppala@helsinki.fi

alleles and genotypes (transmission ratio distortion, TRD) in hybrid progenies of experimental crosses. Thus, BDMI manifest themselves by distorting transmission ratios of alleles, a phenomenon commonly observed in experimental crosses between species (Jenczewski *et al.* 1997; Rieseberg *et al.* 2000). Loci involved in BDMI have been mapped by analyzing TRD of alleles in the offspring of experimental crosses between (sub)species or populations (Vogl and Xu 2000; Harushima *et al.* 2001). In plants, locations of transmission ratio-distorted loci (TRDL) have been used as indicators of genetic incompatibilities in *Mimulus* (Fishman *et al.* 2001), rice (Harushima *et al.* 2001), *Eucalyptus* (Myburg *et al.* 2004), tomato (Moyle and Graham 2006), *Ceratodon* moss (McDaniel *et al.* 2008), *Ceratopteris* fern (Nakazato *et al.* 2007), and *Arabidopsis thaliana* (Salomé *et al.* 2011).

TRD is not always a sign of BDMI, but may also arise for various other reasons during development of gametes or after the gametes are formed but before fertilization. The former could be a selfish segregation distorter or meiotic drive system where a certain allele gets transmission advantage over the other, *e.g.*, in meiosis of germ cells (reviewed by Lyttle 1991; Werren 2011). Selfish genetic changes have been found, surprisingly often, to be involved in formation of hybrid dysfunction (Presgraves 2010). Selfish genetic elements and the genomic conflicts they cause have been recognized as important factors in promoting evolutionary change (Werren 2011). Presgraves (2010) concluded that intrinsic genomic instability underlies evolution of speciation genes (genes causing hybrid sterility or inviability) more often than classical adaptive incompatibilities developing in both diverging populations. Currently most of the examples derive from a few model species and more investigations are needed (Johnson 2010).

TRD could also be due to other prefertilization events such as competition between gametes, *e.g.*, variable pollen tube growth rate (reviewed by Howard 1999). In angiosperms, competition would be more likely to occur between male gametes, because of their greater number and longer haploid phase compared to female gametes. TRD could also be due solely to factors acting after fertilization, such as differential viability of hybrid genotypes, competition between zygotes on maternal resources (Korbecka *et al.* 2002), or as a combination of both gametic and zygotic factors.

Transmission ratios may be distorted due to epistatic interactions not only between nuclear genes, but also between nuclear genes and cytoplasmic factors, such as mitochondrial or chloroplast genes. To be able to investigate the role of cytoplasmic factors in generating TRD, one must perform reciprocal crosses, so that the two reciprocal hybrid progenies share the same nuclear genome but differ in cytoplasm. Reciprocal crosses are feasible with hermaphroditic species and have therefore been most commonly conducted in plants. The role of cytonuclear interactions in performance of reciprocal hybrids has been demonstrated in many species (reviewed by Levin 2003). Apart from cytoplasmic male sterility, little is known about the role of cytonuclear interactions in causing postzygotic reproductive isolation (Rieseberg and Blackman 2010).

Designs of experimental crosses differ, but studies of TRD are usually started with two parents representing different populations, subspecies, or species. However, these parents have often been derived from selfed lines. While this simplifies mapping algorithms and increases statistical power to detect TRDL, it provides little information about allelic variation within natural populations (except if several independent crosses were produced). Only a few studies have examined variation in incompatibility alleles within populations in nature (*e.g.*, Christie and Macnair 1987; Reed and Markow 2004; Sweigart *et al.* 2007; Koide *et al.* 2008a; Martin and Willis 2010; Gérard and Presgraves 2012), but the importance of polymorphic hybrid incompatibility is becoming more acknowledged (Cutter 2012).

In addition to the interest in understanding how, and at what developmental stage, intrinsic postzygotic reproductive isolation acts, interesting questions that remain to be answered concern the rate of development of postzygotic reproductive isolation. Earlier studies on *Drosophila*, *Lepidoptera*, frogs, and some plant genera (reviewed in Coyne and Orr 2004) imply that intrinsic postzygotic isolation develops gradually over time. If transmission ratio distortion reflects reproductive isolation, then the degree of TRD should also increase with the genetic distance between populations or species in experimental crosses. It has been predicted that the number of genic incompatibilities increases between taxa faster than linearly with time (Orr 1995; Orr and Turelli 2001). This seemingly simple prediction has been surprisingly difficult to test in practice: even if genetic distances between populations or species are known, it is difficult to compare the degree of TRD between crosses that differ in statistical power to detect and locate TRD loci because of different numbers, locations, and types of genetic markers and different numbers of hybrid individuals. The same applies to mapping of quantitative trait loci for reproductive isolation. Hence, the first empirical efforts supporting this prediction were published only recently (Matute *et al.* 2010; Moyle and Nakazato 2010), and it remains poorly known whether and how the degree of TRD increases with genetic distance between populations.

Here we investigate TRD in crosses between isolated populations of *A. lyrata* and map transmission ratio distorting loci. Earlier studies in this species indicated that TRD is common in crosses of genetically distant populations (Kuittinen *et al.* 2004) but rare in crosses within populations (Leppälä *et al.* 2008). We compare three experimental crosses, at different genetic distances as assessed with microsatellite (Muller *et al.* 2008) and sequence data (Wright *et al.* 2003; Ross-Ibarra *et al.* 2008; Pyhäjärvi *et al.* 2012). The most distant populations crossed (F_{ST} = 0.62; Pyhäjärvi *et al.* 2012) represent the subspecies *lyrata* and *petraea*. The intermediate (F_{ST} = 0.35) and closely (F_{ST} = 0.20) related crosses are between populations of ssp. *petraea*. This allows us to investigate whether TRD increases with genetic distance between the populations crossed.

We used novel algorithms that allow us to distinguish gametic and zygotic TRD, as well as to detect epistatic two-locus TRD from crosses of outbred individuals. The outbred F_2 design of our crosses (Figure 1), with two differently heterozygous F_1, allows us to evaluate the roles of within-population allelic variation, male and female function, and cytoplasm for TRD loci. Because one of the populations is involved in all of the three crosses, we can also examine whether the same or different genomic regions are associated with TRD when crossed with different populations.

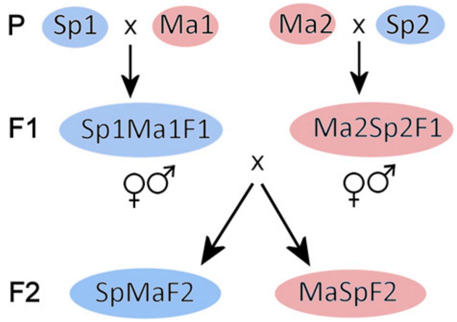

Figure 1 The crossing design for all three reciprocal crosses of *A. lyrata* examplefied by the Sp × Ma cross. The four parents (P), two F_1 individuals, and F_2 progenies are shown. The blue background indicates Spiterstulen and red Mayodan cytoplasm. Note that the F_1 individuals were crossed reciprocally (Sp1Ma1 × Ma2Sp2 and Ma2Sp2 × Sp1Ma1) so that four different alleles (Sp1, Sp2, Ma1, Ma2) could segregate in both reciprocal F_2 progenies. The image has been modified from Leppälä and Savolainen (2011).

Materials and Methods

Crosses

To study transmission ratio distortion we used genetic marker data from three crosses between different *A. lyrata* populations. For a cross between Spiterstulen, Norway (Sp), and Mayodan, North Carolina (Ma), crossing details and plant growth conditions are described in Leppälä and Savolainen (2011). In brief, two unrelated plants from both populations were crossed to produce F_1 hybrids (Figure 1). The crosses were done reciprocally so that F_1's carry different cytoplasm. From both crosses one F_1 was used to produce the F_2 generation. Again the crosses were conducted reciprocally to gain F_2 progenies with different cytoplasmic backgrounds (referred to as SpMaF_2 when cytoplasm from Sp and MaSpF_2 when cytoplasm from Ma).

The same crossing design was used when crossing Spiterstulen with Stubbsand, Sweden (Stu). To make crosses comparable, the same plants from Spiterstulen were used as the parents for the F_1. They were crossed in the same manner, so that the Sp individual that acted as a pollen recipient in the Sp × Ma cross also acted as a pollen recipient in the Sp × Stu cross. The F_2 individuals from the Sp × Stu cross were grown in the same greenhouse at the same time as the Sp × Ma F_2.

The third cross was conducted between Spiterstulen and Plech, Germany (Pl). For this cross, different Sp individuals were used as parents for the F_1 than those used in the crosses described above. Here the crossing design also differed in that the two Sp parents were used as pollen recipients and the two Pl individuals as pollen donors in both initial crosses to achieve F_1. The F_1 were crossed as described above to get two F_2 populations, both carrying Sp cytoplasm. To distinguish between the F_2 reciprocal progenies, they are referred to as SpPl2F_2 and SpPl3F_2. The growth conditions of the F_2 are described in Quilot-Turion *et al.* (2013).

Genotyping

DNA was isolated and genotyping performed for Sp × Ma and Sp × Stu F_2 as described in Leppälä and Savolainen (2011) and for Sp × Pl F_2 as in Kemi *et al.* (2013). For Sp × Ma 391 F_2 individuals were genotyped (204 for Sp cytoplasm and 187 for Ma cytoplasm) with 76 SNP, microsatellite, and Cleaved Amplified Polymorphic Sequences (CAPS) markers (for details see Leppälä and Savolainen 2011). Sp × Stu was genotyped with the same methods as Sp × Ma. For Sp × Stu, 389 F_2 plants were genotyped (195 for Sp cytoplasm and 194 for Stu cytoplasm) with 65 SNP, microsatellite, and CAPS markers. The markers used to genotype Sp × Stu F_2 and the origins of the primer sequences are in Supporting Information, Table S1. In the Sp × Pl cross, 529 F_2 individuals were genotyped (326 SpPl2F_2 and 203 SpPl3F_2). For the Sp × Pl cross, 40 microsatellite and CAPS markers were genotyped. For details see Kemi *et al.* (2013).

Analysis of transmission ratio distortion

Each of the n F_2 offspring of F_1 heterozygotes ab and cd must have one of the four possible genotypes ac, ad, bc, and bd. Because the genotype of each offspring is independent of that of the other offspring, the log-likelihood of the observed numbers of genotypes x is given by the multinomial distribution

$$L = \ln\left(\frac{n!}{x_{ac}!x_{ad}!x_{bc}!x_{bd}!}p_{ac}^{x_{ac}}p_{ad}^{x_{ad}}p_{bc}^{x_{bc}}p_{bd}^{x_{bd}}\right), \quad (1)$$

where $\Sigma x = n$ and p denotes the expected frequency of a genotype.

Under Mendelian segregation we expect the four genotypes in equal frequencies (*i.e.*, all $p = 0.25$) and substituting these values in Equation 1 we obtain the likelihood of the observed numbers of genotypes in the absence of transmission ratio distortion, L_0. Deviations of observed genotype numbers x from their expectations p may be due to chance, due to differential transmission of alleles in gametes or due to differential survival of genotypes in the zygotic phase. Let us first consider differential gametic transmission of alleles. If the male parent transfers allele a with probability q_{m}, then the probability that allele b is transferred is $1 - q_{\mathrm{m}}$. Similarly, we denote the probability that the female transmits allele c q_{f}, so that allele d is transmitted with probability $1- q_{\mathrm{f}}$. Given q_{m} and q_{f}, the expected frequencies of the genotypes are: $p_{ac}= q_{\mathrm{m}}q_{\mathrm{f}}$, $p_{ad}= q_{\mathrm{m}}(1 - q_{\mathrm{f}})$, $p_{bc}= (1 - q_{\mathrm{m}})q_{\mathrm{f}}$, and $p_{bd}= (1 - q_{\mathrm{m}})(1 - q_{\mathrm{f}})$. These expected frequencies can be substituted in Equation 1 to obtain the likelihood, and we can use standard algorithms to find the values of q_{m} and q_{f} that maximize the likelihood of the observed genotype frequencies under gametic phase transmission ratio distortion, L_{gametic}. Finally, differential zygotic survival or fertilization success may cause unequal genotype frequencies. In that case, the only constraint on p is $\Sigma p = 1$, and the likelihood of the observed genotype frequencies L_{zygotic} is maximized when expected frequencies are equal to observed frequencies, *i.e.*, $p_{ac}= x_{ac}/n$ (and similarly for the other genotypes).

Distinguishing gametic from zygotic TRD

We can evaluate which of the above models best explains the observed numbers of genotypes x using likelihood ratio tests. Let L_0 and L_{a} denote the maximized log-likelihoods of the observed genotype frequencies under the null and alternative hypothesis, respectively. The sampling distribution of the logarithmic-likelihood ratio $-2(L_0- L_{\mathrm{a}})$ follows a χ^2-distribution with degrees of freedom equal to the difference in the number of unknown parameters under the hypotheses being compared. Under Mendelian segregation all $p = 0.25$ so that there are no free parameters. With gametic phase TRD there are two unknown parameters (q_{m} and q_{f}) that determine p, and if TRD arises at the zygotic stage there is only one constraint ($\Sigma p = 1$) so that there are three unknown parameters. Thus, to test for gametic TRD, we evaluate the χ^2 distribution with $2 -0 = 2$ degrees of freedom at $-2(L_0/L_{\mathrm{gametic}})$. To test for zygotic TRD, we evaluate the χ^2 distribution with $3 - 0 = 3$ degrees of freedom at $-2(L_0/L_{\mathrm{zygotic}})$. In case of gametic phase TRD also L_{zygotic} (and not only L_{gametic}) will be substantially higher than L_0. Therefore we evaluate the χ^2 distribution with $3 - 2 = 1$ degrees of freedom at $-2(L_{\mathrm{gametic}}/L_{\mathrm{zygotic}})$. If zygotic TRD explains the data better than gametic TRD, while gametic TRD explains the data better than Mendelian segregation, this indicates zygotic TRD in addition to gametic phase TRD. Thus, comparing likelihoods of the observed genotype frequencies under different hypotheses as illustrated in Figure S1, we can identify loci with gametic, zygotic, and with both gametic and zygotic TRD. We considered genomic regions as transmission ratio distorted when $P < 0.001$.

Epistatic TRD

Transmission ratio distortion can be caused by single-locus factors, but also by interactions between two (or more) loci. To detect two-locus epistatic TRD we calculated two-locus genotype frequencies for all pairs of markers (excluding markers at the same chromosome), calculated expected frequencies by cross tabulation, and evaluated the probability that observed and expected genotype frequencies differ by chance using χ^2 tests. Because we calculated expected frequencies using cross tabulation, observed single-locus genotype frequencies are taken into account when detecting epistatic TRD.

Just like single-locus TRD, also epistatic TRD may emerge both in the gametic and in the zygotic phase. To detect gametic phase epistatic TRD we calculated expected two-locus allelic combinations by cross-tabulation from observed single-locus allelic frequencies. For example, if the pollen donor (male parent) transmitted allele a at locus 1 with observed frequency ${q_{\mathrm{m}}}^1$, and allele a at locus 2 with observed frequency ${q_{\mathrm{m}}}^2$, then the probabilities of observing two-locus allelic combinations a^1a^2, a^1b^2, a^2b^1, and a^2b^2 are ${q_{\mathrm{m}}}^1$ ${q_{\mathrm{m}}}^2$, ${q_{\mathrm{m}}}^1$ $(1 - {q_{\mathrm{m}}}^2)$, $(1 - {q_{\mathrm{m}}}^1)$ ${q_{\mathrm{m}}}^2$, and $(1 - {q_{\mathrm{m}}}^1)(1 - {q_{\mathrm{m}}}^2)$, respectively. We evaluated deviation of observed two-locus allelic combinations from their expectations using χ^2 tests. Parents were analyzed separately since at the gametic stage interaction between parents is not yet plausible.

Missing genotypes and informativeness

Above we assumed that genotypes are known with certainty for each individual, but at several loci with partially informative markers or pseudomarkers (see below) that is not the case. We first inferred missing genotypes using Wijsman's (1987) genotyping rules. Subsequently, flanking markers were used to infer genotype probabilities at partly informative genotyped loci and at pseudomarker loci. For single-locus analysis of TRD, individuals were assigned genotypes so as to maximize the sum of the log-likelihood from Equation 1, plus the log-likelihood of the assigned genotypes. For analyses of epistatic TRD this was computationally not feasible, so individuals were simply assigned the genotype that was most likely based on the flanking markers. Details on the algorithms we used for TRD mapping are in File S1.

Comparing TRD between crosses

Because numbers and positions of markers differ between crosses, we cannot compare the degree of TRD between crosses and reciprocal progenies directly by calculating the number or percentage of distorted markers. As a simple alternative, we used the percentage of the genome showing single locus TRD. We inserted pseudomarkers so as to have intermarker distances <1 cM and then approximated the percentage of the genome showing TRD as the percentage of distorted (pseudo)markers.

As we use many markers, using the conventional P-value of 0.05 as a threshold may result in a high false-positive rate. However, using a standard correction for multiple testing based on the number of markers would lead to unacceptable false-negative rates, not only because markers are statistically dependent through linkage, but also because the markers do not provide repeated tests of the same hypothesis. Therefore we made a somewhat arbitrarily trade-off between avoiding type I and type II errors and used a threshold of $P = 0.001$, but we confirmed that results are qualitatively similar when using different thresholds in the range 0.05–0.0001.

Results

All three reciprocal F_2 progenies had transmission ratio distorted markers. The reciprocal progenies of each cross were examined separately for TRD because of possible cytoplasmic effects, or differences in heterogeneous F_1 parental plants when acting as pollen donors or recipients. In the progeny of the cross between the most distant populations ($F_{ST} = 0.62$) SpMaF_2, 33 of 76 markers (43% of χ^2 tests at $\alpha = 0.05$, not corrected for multiple testing) deviated from expected Mendelian genotype frequencies, and in the reciprocal MaSpF_2 progeny 26 of 76 markers had non-Mendelian inheritance (34% at $\alpha = 0.05$). With a more conservative threshold of $\alpha = 0.001$ the percentages of distorted markers remained high (SpMaF_2 20% and MaSpF_2 23%). At intermediate genetic distance between populations ($F_{ST} = 0.35$), in the SpPl2F_2 progeny 13 of the 40 markers (33% at $\alpha = 0.05$) showed non-Mendelian segregation and in the reciprocal SpPl3F_2 9 of the 40 markers had distorted segregation ratios (23% at $\alpha = 0.05$). Finally, at the smallest genetic distance between populations ($F_{ST} = 0.20$), the progeny had a lower percentage of transmission ratio distorted markers on average. However, only 8 of 65 deviated from Mendelian segregation in the SpStuF_2 (12% at $\alpha = 0.05$), while the reciprocal StuSpF_2 progeny had 20 transmission ratio distorted markers (31% at $\alpha = 0.05$).

To more precisely identify the genomic regions showing TRD and to investigate at which level (gametic, zygotic, or both) the distortions likely developed, we applied TRD mapping methods, to compare observed genotype frequencies to expectations under the null model (Mendelian segregation), and under models allowing for gametic and zygotic phase distortion. In all crosses, most of the TRD regions were best explained by the gametic model (Table 1), but locations and types of TRD differ between progenies (Figure 2, Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, and Figure S7).

Single-locus gametic TRD

We examined more closely the transmission of the two F_1 parents' alleles in the regions showing TRD. To visualize how the F_1 parents transmitted their alleles, we plotted observed allele frequencies for each parent over the whole chromosome (Figure 3). As can be seen from Figure 3, gametic TRD can potentially arise from four different sources (male and female function in F_1 parent 1 and in F_1 parent 2) in a single cross, because both F_1 parents were used as pollen donors and recipients.

In the SpMaF_2, three regions on three chromosomes experienced strong TRD (Figure S2), of which two (on chromosomes AL1 and AL7) were best explained as gametic. On AL1 excessive transmission of the Sp allele by the pollen donor appeared to be the source of most of the distortion (Figure 3A), although transmission of the pollen recipient's alleles was biased in the same direction. On chromosome AL7 at least two loci were involved because the F_1 pollen donor transmitted more of its Sp allele on the upper arm of the chromosome, while the pollen recipient F_1 transmitted more of its Ma allele at the lower chromosome arm (Figure 3A). Two of the three TRD regions observed in the SpMaF_2, namely those on AL1 and AL6, were distorted also in the reciprocal MaSpF_2 cross. Just like in the SpMaF_2, TRD on chromosome AL1 in MaSpF_2 was largely due to excess transmission of the Sp allele by the F_1 pollen donor, but in the MaSpF_2 a similar excess transmission of the Sp allele was observed in the pollen recipient, although less severe than in the pollen donor (Figure 3A). The other TRD region that the reciprocal progenies had in common (on AL6) showed single-locus zygotic TRD. The two other gametic TRD regions in MaSpF_2, which were not shared by SpMaF_2, were on chromosomes AL3 and AL8, and in both these regions the pollen donor transmitted predominantly its Sp allele.

In the Sp × Pl cross ($F_{ST} = 0.35$) more TRD regions were observed in the SpPl2F_2 progeny than in its reciprocal. In SpPl2F_2 three regions (AL1, AL3, and AL6) were found to be involved in formation of TRD at gametic level. The region on AL1 was also found in the reciprocal SpPl3F_2 where it was the only region that showed significant TRD. Distortion in this region was most pronounced around the marker RHL1 and appears to be due in both reciprocal progenies only to a single F_1 parent (Sp1Pl1F_1) transmitting its Pl allele in excess. (Sp1Pl1F_1 acted as a pollen donor for SpPl3F_2 progeny and as pollen recipient for SpPl2F_2 progeny.) The two additional TRD regions found in SpPl2F_2, on AL3 and AL6, were also apparently of gametic origin, with the Sp allele of Sp1Pl1F_1 (AL3) and Sp alleles of both F_1 parents (AL6) overrepresented (Figure 3B).

In the Sp × Stu cross ($F_{ST} = 0.20$) one gametic TRD region was observed on AL1 in the SpStuF_2 progeny and two regions in its reciprocal progeny, on AL1 and AL6. The region at the lower arm of chromosome AL1 was common for both reciprocal progenies, with an excess of Stu alleles transmitted by both parents in both reciprocal crosses (Figure 3C). The other TRD region in StuSpF_2 was on AL6 where both parents transmitted an excess of Stu alleles.

Two-locus gametic TRD

We also examined two-locus genotypes for TRD at gametic and zygotic level. We describe here the most significant epistatic interactions where $P < 0.0001$. In the SpMa$F_{2,}$ one

Table 1 The number of single-locus TRD regions and the percentage of markers in TRD ($P < 0.001$, or $P < 0.01$ in parentheses) in three reciprocal crosses of *A. lyrata*

Measurement	Phase	Sp × Ma		Sp × Pl		Sp × Stu	
F_{st} between populations[a]		0.62		0.35		0.20	
Total lenght of the genetic map (cM)		514		511		500	
Number of markers		76		40		65	
Reciprocal progenies		$SpMaF_2$	$MaSpF_2$	$SpPl2F_2$	$SpPl3F_2$	$SpStuF_2$	$StuSpF_2$
F_2 progeny size		204	187	326	203	195	194
Single-locus TRD regions at $P < 0.001$ (0.01)	Gametic	2 (6)	3 (6)	3 (5)	1 (4)	1 (1)	2 (6)
	Zygotic	—	—	—	—	—	—
	Both	1(1)	1(1)	—	—	0 (1)	0 (1)
% of genome distorted at $P < 0.001$ (0.01)	Gametic	23 (38)	23 (34)	14 (18)	10 (19)	2 (2)	9 (23)
	Zygotic	1 (3)	1 (1)	0.7 (4)	0 (0.3)	0.6 (1)	0.6 (2)

[a] F_{ST} values from Pyhäjärvi *et al.* (2012).

epistatic gametic interaction was observed in the pollen donor and one in the pollen recipient. The interaction in the pollen recipient ($SpMaF_1$) was between AL1 and AL5 ($P = 3.1 \times 10^{-5}$; Figure S8). The region of AL1 involved in this epistatic interaction was the same as that seen at the single locus level. However, while at the single locus level in AL1 excess transmission of Sp alleles was observed, here the combinations of Sp at AL1 and Sp at AL5 and Ma at AL1 and Ma at AL5 were less frequent than expected. For the pollen donor ($MaSpF_1$) the epistatic interaction was between AL1 and AL4 ($P = 1.2 \times 10^{-9}$; Figure S8). The region in AL1 was not the same as that for single-locus TRD in this progeny and also not for the two-locus interaction in the pollen recipient ($SpMaF_1$), but the pattern was the same; combinations with alleles from the same population were less frequent than expected and gametes combining alleles from different populations were more common.

For the reciprocal progeny ($MaSpF_2$) also two significant gametic two-locus interactions were observed. The interactions were different from those in the reciprocal progeny, and both were observed in the pollen donor ($SpMaF_1$) (Figure S8). The interactions were between AL1 and AL4 and between AL3 and AL5 (Figure 4). Both interactions showed a similar pattern; gametes combining alleles from the different populations (at these interacting loci) were less frequent than expected, and gametes combining alleles from the same population were more common. This was exactly opposite to the pattern observed in the two-locus gametic interactions in the reciprocal F_2 progeny.

In the Sp × Pl cross only one, gametic, two-locus interaction ($P = 7.7 \times 10^{-7}$; Figure S9) was found, namely between AL1 and AL3 in the $SpPl2F_2$. The interacting regions were also observed to be in TRD at the gametic single-locus level. The interaction was seen in the pollen recipient ($Sp1Pl1F_1$) and was also observed but weaker ($P = 0.001$; Figure S9) in the reciprocal cross where $Sp1Pl1F_1$ was acting as a pollen donor. In both reciprocal crosses AL1–AL3 combinations of parental alleles (SpSp or PlPl) were in excess and the combinations of alleles from different populations on the same gamete were less frequent than expected.

In the Sp × Stu cross one gametic two-locus interaction was found in the pollen donor for $SpStuF_2$. The interaction was between AL2 and AL4 ($P = 3.3 \times 10^{-5}$; Figure S10), neither of which showed strong single-locus TRD. Just as in the Sp × Pl cross, within-population combinations between the loci were in excess (SpSp and StuStu) and between-population allelic combinations were less frequent than expected.

Zygotic TRD

Zygotic TRD was rarely detected in these crosses. The strongest single-locus zygotic TRD was detected in both Sp × Ma reciprocal progenies, in the top of chromosome AL6 (Figure S2 and Figure S3). Both heterozygotes were less frequent than expected while parental population homozygotes were in excess (Figure S2 and Figure S3). There might have been some gametic contribution to this case of TRD as the $SpMaF_1$ parent, when acting as a pollen donor, predominantly transmitted its Sp allele, whereas otherwise allele frequencies were as expected (Figure 3).

The other zygotic TRD was a two-locus interaction observed in both reciprocal progenies of the Sp × Stu cross. Upon closer examination of two-locus genotype ratios, it appeared that likely two or three loci from AL1 were interacting with one locus at AL6. Of the interacting regions on chromosomes AL1 and AL6, only that on the lower arm of AL1 also showed single-locus TRD. The two-locus distortion maps to the end of the lower arm of AL1 in both reciprocal crosses, but in $StuSpF_2$ also the upper arm seems to be involved in the interaction (Figure 4). Some TRD was already observed in the gametes of $SpStuF_1$ (both as pollen donor and pollen recipient; Figure S10). In the F_2 generation one two-locus genotype was entirely absent from the progeny: no F_2 individual was Sp homozygote at AL1 (marker nga280/LAS) and Stu homozygote at AL6 (AT4G04350) (for two-locus genotype counts see Table S2). Similarly no F_2 individuals in $SpStuF_2$ were Stu homozygotes at AL1 (AT1G31930) and Sp homozygotes at AL6 (AT4G04350) or in $StuSpF_2$ Stu homozygotes at AL1 (F20D22) and Sp homozygotes at AL6 (AT4G04350).

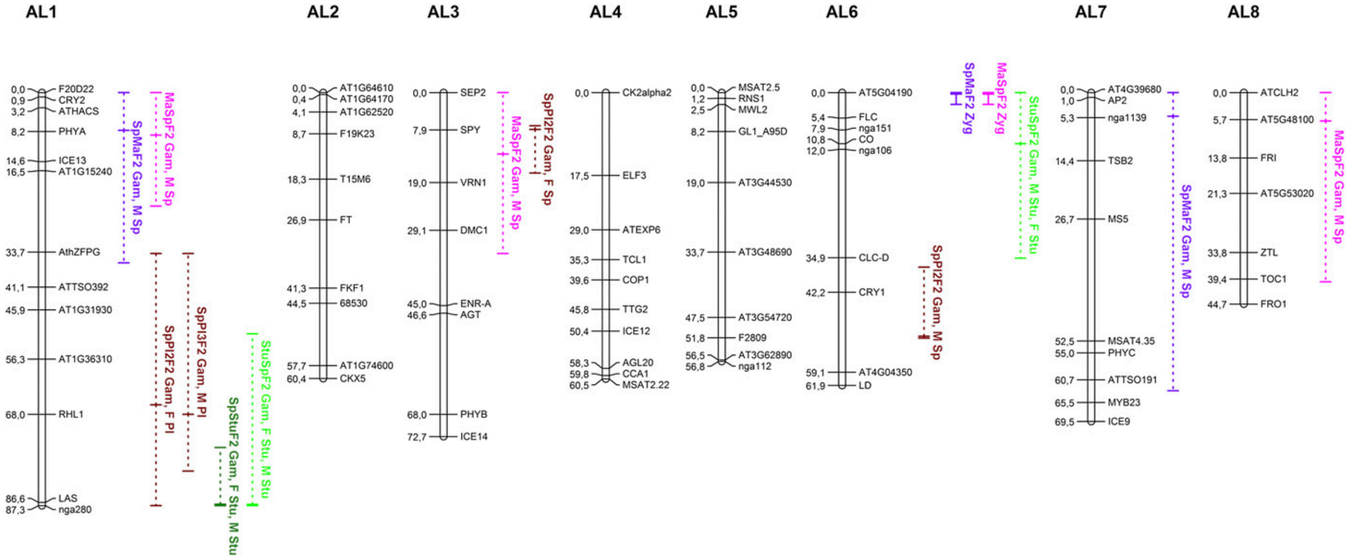

Figure 2 Locations of TRD regions in *A. lyrata* crosses on a Sp × Ma map modified from Leppälä and Savolainen (2011). Vertical lines indicate regions with significant ($P < 0.001$) gametic (dashed line) or zygotic (solid line) TRD. Horizontal ticks mark the likelihood peaks of TRD within such regions. The text indicates which reciprocal progeny was involved and whether gametic (Gam) or zygotic (Zyg) TRD was inferred. With gametic TRD the parent (F, female; M, male) with most distorted allelic ratios and the preferentially transferred allele (*e.g.*, Sp or Ma) is marked.

Comparisons between the crosses

TRD was clearly more prevalent in the distant Sp × Ma cross than in the crosses between the less diverged populations, at gametic as well as zygotic, one-locus and two-locus levels. We compared the percentage of the genome showing single-locus distortion (either gametic or zygotic, at $P < 0.001$) between different crosses (Figure 5). (Using a different significance threshold for P did not alter the results qualitatively; see Table 1.) On average, the degree of distortion increased with genetic distance between populations. However, between the Sp × Pl and Sp × Stu crosses the difference in degree of TRD was not clear. This was mainly due to a large difference in the degree of TRD between the Sp × Stu reciprocal progenies. While the Sp × Pl reciprocal progenies (both with the Sp cytoplasm) showed an approximately equal degree of TRD, the StSpF$_2$ was distinctly more distorted than its reciprocal F$_2$ progeny, where transmission showed almost no deviation from the Mendelian expectation.

It has been suggested that the number of genic incompatibilities increases between taxa faster than linearly with time (Orr 1995; Orr and Turelli 2001). To test that "snowballing" prediction, we used linear and exponential functions to describe the increase of the degree of TRD with genetic distance (Figure 5). As is evident from Figure S11, the increase is best described by linear functions, irrespective of whether we measure the degree of TRD as the proportion of the genome showing TRD or as the number of TRD regions. Formal model comparison using Akaike's information criterion confirms this visual impression.

As may be apparent from what was written above, the regions showing TRD and the inferred origin of distortion in these regions were mostly different between different crosses. Nevertheless, some patterns emerge: both single-locus and epistatic TRD was often found arising from chromosomes AL1 and AL6. All crosses showed TRD on AL1, and in the Sp × Pl and Sp × Stu crosses this TRD maps onto the same chromosomal region, at the end of the lower chromosome arm (Figure 2). In both crosses, at AL1, the Sp allele had been transmitted less frequently than expected, although in the Sp × Pl cross this happened in only one of the F$_1$ parents (Sp1Pl1) whereas in the Sp × Stu cross both parents were transmitting fewer Sp alleles—both when acting as pollen donor and recipient. The other chromosome that commonly showed TRD was AL6, but the strongest TRD did not map onto the same regions of the chromosome, suggesting that the loci causing TRD on AL6 differ between crosses.

Discussion

Gametic vs. zygotic TRD

In all three crosses between *A. lyrata* populations, most observations of non-Mendelian segregation are best explained by processes acting at the gametic stage. This could be partly due to hybrids experiencing genic incompatibilities or genomic conflicts already in the F$_1$ generation. Alternatively, it is possible that processes acting at the zygotic stage result in a pattern of distortion that mimics gametic TRD. For single-locus zygotic TRD to look like gametic TRD, selection must act against at least one of the genotypes combining parental population alleles (*e.g.*, Sp1Sp2) and against one of the genotypes combining alleles from different populations (*e.g.*, Sp1Ma2). (In this example, selection against these genotypes would mimic selection against the Sp1

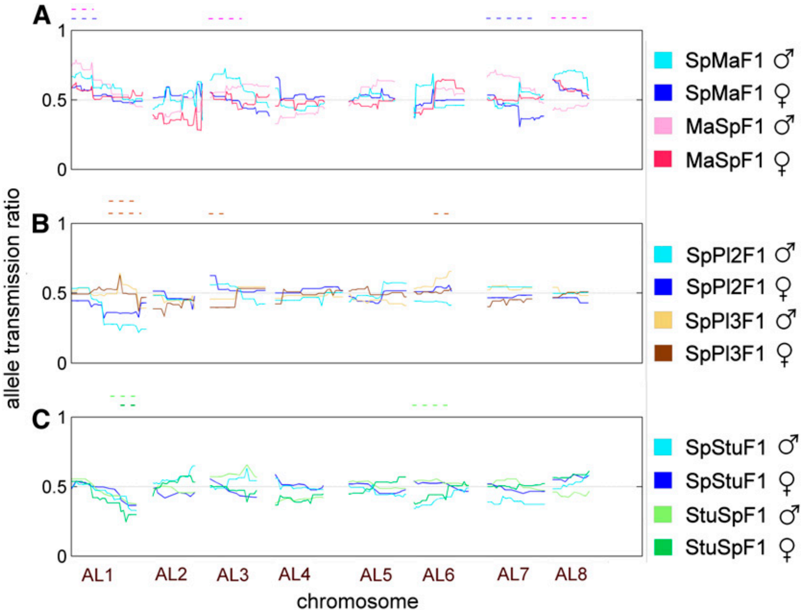

Figure 3 Allele transmission ratio of Spiterstulen alleles from the F_1 to the F_2 for each parent, separately in male (light color) and female function (dark color), over all eight chromosomes. (A) Sp × Ma, (B) Sp × Pl, (C) Sp × Stu. F_1 parents with Sp cytoplasm are in blue ($SpMaF_1$, $Sp1Pl1F_1$, and $SpStuF_1$). $MaSpF_1$ is in red (A) and $StuSpF_1$ in green (C). The $Sp2Pl2F_1$ parent had cytoplasm from Sp but is represented in brown (B). The Mendelian expectation of equal segregation of both alleles, 0.5, is marked with a gray dashed line. The significant gametic TRD regions have been marked above the chromosomes as in Figure 2.

allele.) If zygotic TRD were common in these crosses, it would need to be of the kind that mimics gametic TRD, as we detected only two clear examples of zygotic-stage TRD. It seems unlikely that most of the zygotic TRD would mimic gametic TRD. Therefore, we believe that most of the gametic TRD that we reported could really be due to events before zygote formation.

Of the two cases of zygotic TRD, one region in the Sp × Ma cross was best explained by the zygotic single-locus model. The other zygotic TRD region was one interacting pair of chromosomes in the Sp × Stu cross where certain genotypic combinations were completely absent from the F_2 progeny. The latter case seems like a classical example of a recessive BDM incompatibility, with selection against individuals that are homozygous at both loci but for alleles from different populations. We also detected two (or three) regions in AL1 interacting with the same locus in AL6, but because our analysis was limited to two-locus interactions we could not detect whether the interaction was between more than two loci. In *A. thaliana* a similar two-locus interaction was found to cause hybrid embryo lethality because functional copies of a duplicate gene were not at the same locus in different accessions (Bikard *et al.* 2009). Most TRD loci in *A. lyrata*, however, appeared to be the best explained by the gametic model.

It should be emphasized that the F_2 seeds, once formed, had high viability. This high viability of F_2 hybrids has been observed earlier in the cross between subspecies (Sp × Ma), where F_2 viability was good but male fertility was reduced (Leppälä and Savolainen 2011). Additional F_2 plants from this cross were grown in their native habitats (Norway and North Carolina). In North Carolina, hybrid fitness was intermediate to the parental populations but in Norway the F_2 hybrids surprisingly showed heterosis (Leinonen *et al.* 2011). Thus, the observed TRD in F_2 would not be expected to be due to low seed germination success or high mortality in later life stages.

The origin of gametic TRD was traced back to the F_1 parents, to see which parent (if not both) caused the distortion and which of the alleles were under- or overrepresented. Most commonly both single-locus and two-locus TRD was caused by a single F_1 parent or both F_1 parents when acting as a pollen donor (half of the gametic single locus TRD regions and four out of six gametic two-locus regions). Thus, reduced male fertility or pollen competition would be a likely source of TRD. Further support for a possible role of reduced male fertility in causing TRD comes from analysis of fertility of the Sp × Ma F_2 progeny (Leppälä and Savolainen 2011). While F_2 hybrids did not appear to suffer reduced female fertility, male fertility was strongly affected, and three out of five male fertility QTL were located in TRD regions. In this cross, male fertility was observed to be reduced also in the F_1 generation, but unfortunately the male fertility of the two parental F_1 plants was not studied. Thus both TRD and fertility reductions in *A. lyrata* appear to be due mostly to problems in male function. This may be connected to haploid (postmeiotic) gene expression which is abundant during male gametophyte development (reviewed in Borg *et al.* 2009). Whether transcription of diploid sporocytes (by masking effects of recessive deleterious incompatibilities) is more important for female than male gametophyte development has not been thoroughly studied yet (Muralla *et al.* 2011).

It is important to note that TRD in male gametes may be due to BDM incompatibilities that reduce fertility, but also to other processes such as pollen competition or action of segregation distorters or some other meiotic drivers. We cannot distinguish with certainty what process was responsible for TRD in male gametes in our study. The role

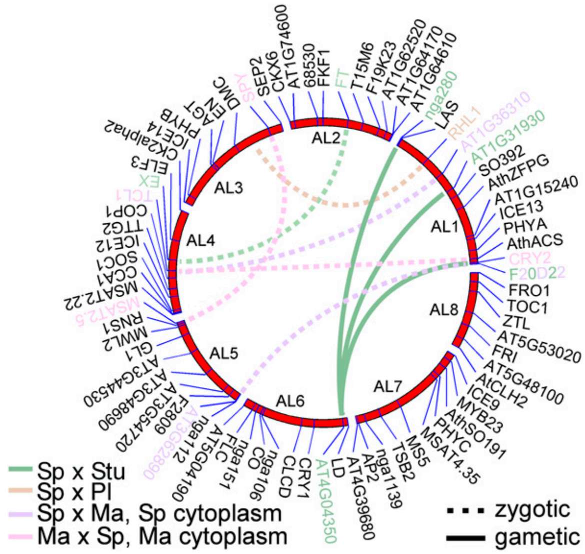

Figure 4 Loci experiencing epistatic two-locus interactions are marked with color and connected with a dashed (gametic interaction) or solid line (zygotic interaction).

of pollen competition as a source of gametic TRD could be assessed in future studies with pollen competition or pollen tube growth rate assays.

We also detected three sex-independent single-locus transmission ratio distorters where the F_1 parents in male as well as female function transmitted alleles from the same population in excess. These cases were observed on the lower arm of AL1 in the Sp × Stu and Sp × Pl cross (but only in one of the F_1 parents) and on the upper arm of AL1 in the Sp × Ma cross (although here male function was a source of higher TRD; Figure 3). These could have been caused by sex-independent segregation distorter loci that are fixed but suppressed in their population of origin. This type of segregation distorter advances its own transmission in the naïve hybrid background without suppressors. Sex-independent TRD systems have been identified in tomato (Rick 1969) and rice (Sano *et al.* 1979; Sano 1992), where one such system is mainly caused by a single locus (Koide *et al.* 2008b) with some unlinked modifiers (Koide *et al.* 2012). The classic meiotic drive systems known from mouse and *Drosophila* are restricted to males (reviewed in Lyttle 1991) but some female segregation distorters are now known, *e.g.*, the knob chromosomes in maize (Buckler *et al.* 1999), B chromosomes (reviewed in Palestis *et al.* 2004), and female meiotic drive in *Mimulus* (Fishman and Saunders 2008). Sex-independent TRD systems seem to be rarer than those affecting just male or female function. The connection between hybrid incompatibility and meiotic drive (and other genomic conflicts) has become recently better established (reviewed by Johnson 2010; McDermott and Noor 2010). To determine whether this type of segregation distorters is involved in causing TRD in F_2 hybrid progenies of *A. lyrata* would require further study.

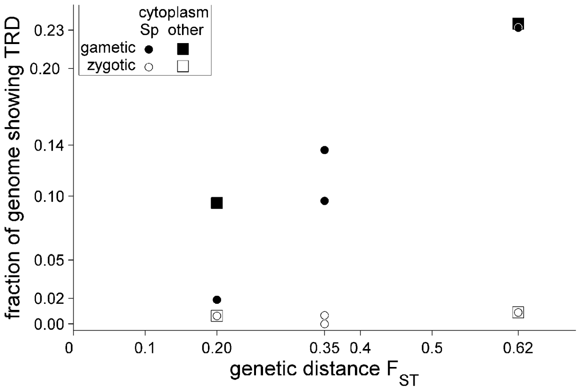

Figure 5 The fraction of the hybrid genome showing single-locus TRD ($P < 0.001$) as a function of genetic distance between crossed populations (F_{ST}, estimated from nucleotide sequences by Pyhäjärvi *et al.* (2012). The crosses from left to right: Sp × Stu, Sp × Pl, Sp × Ma.

The role of cytoplasmic factors

Genetic incompatibilities affecting hybrid fitness are often thought to develop between nuclear genes, but they may as well develop between nuclear and cytoplasmic factors. Reciprocal crosses allow us to investigate the role of cytoplasmic factors, but such crosses require hermaphrodite organisms and have therefore been restricted largely to studies in plants. In our crosses the F_1 parents had different nuclear genotypes, so we could not distinguish between cytoplasmic effects and nuclear polymorphisms within populations. However, the present results suggest that either cytoplasm or allelic polymorphisms (or a combination of these) play a major role, as the extent and especially location of TRD differ substantially between reciprocal crosses. Kuittinen *et al.* (2004) studied a cross between *A. lyrata* populations and reached the same conclusion: only one TRD region of the seven observed was shared between their reciprocal progenies.

Here, the difference between reciprocal progenies was most pronounced in the Sp × Stu cross, where 2% of the genome ($P < 0.01$, gametic model *vs.* null model) showed TRD in the SpStuF_2, against 23% in the StuSpF_2 reciprocal. In the more distant St × Ma cross the difference between reciprocals in the degree of TRD was less pronounced, but in both crosses the nature and location of TRDL differed substantially between reciprocal progenies. A similar result was obtained from analysis of male fertility of the Sp × Ma F_2: both the extent of fertility reduction and the underlying QTL differed between reciprocal F_2 progenies (Leppälä and Savolainen 2011). In addition to the role of cytoplasmic factors, alleles within a population also had different effects on male fertility of Sp × Ma cross (Leppälä and Savolainen 2011).

In one of our reciprocal crosses (Sp × Pl, at intermediate genetic distance), the role of cytoplasm *vs.* within population allelic variation could be evaluated because the cytoplasm was from the same population. In this cross the degree of TRD differed relatively little between the reciprocal F_2 (18% *vs.* 19% of genome in TRD, $P < 0.01$, gametic *vs.* null model), but within-population variation was also found: two of the TRD regions were found in only one of the progenies (SpPl2F_2) and the shared TRD region was sex independent, *i.e.*, the source of TRD could be traced back to a single F_1 parent. However, based on a single cross we cannot conclude whether one mechanism is more likely to be involved in TRD than the other.

TRD increases with genetic distance

As an indicator of genic incompatibility, the degree of TRD is expected to increase with genetic distance between populations. It is, however, difficult to compare the degree of TRD between crosses from different populations or species. As the results of the present study illustrate, TRD may be a single-locus or multiple-locus phenomenon, and TRD at any specific locus may emerge from pollen donor, pollen recipient, or both parents or in the hybrid zygotes. Furthermore, TRD at any locus may differ in severity, from a slight distortion of transmission ratios to complete absence of some genotypes. For these reasons it is difficult to compare the degree of TRD even between reciprocal crosses from the same parents. Comparisons involving distant populations or species may be further complicated by different numbers, types, and positions of marker loci, making it hard to tell whether a difference in degree of TRD is a real property of the populations studied or the result of difference in statistical power to detect TRD. Consequently, it is difficult to determine from comparison of published studies whether the degree of TRD increases with genetic distance.

Because of the difficulties associated with expressing the degree of TRD as a scalar value, we decided to use a simple measure for the sake of tractability: the percentage of the genome showing distortion. This measure does not solve the problems described above, but it is tractable, it can be calculated for almost any study, and it has one attractive property: a larger number of TRDL increases the percentage of the genome showing TRD, and further, more severe TRD will be apparent over a larger part of a chromosome because of linkage disequilibrium. Thus, measuring the degree of TRD as a percentage of the genome showing TRD hides many interesting aspects of the underlying data, but it does provide a figure that naturally integrates the number of TRDL and the severity of their effects in a simple measure.

In line with expectation, TRD has been found to increase from intraspecific to interspecific experimental crosses (Jenczewski *et al.* 1997; Rieseberg *et al.* 2000). By contrast, in natural populations of *Mimulus*, the level of TRD did not differ between intra- and interspecific crosses (Hall and Willis 2005). However, because of lack of genetic data, Hall and Willis (2005) were unable to conclude whether the *Mimulus guttatus* populations they studied were less diverged than the species pair used in an earlier study (Fishman *et al.* 2001). Good estimates of genetic distance are needed for testing directly whether crosses between more distant populations show higher TRD. A recent study by Salomé *et al.* (2011) did not find any relation between sequence divergence of parental accessions and F_2 segregation distortion in *A. thaliana*. For the *A. lyrata* populations studied here, good microsatellite and sequence-based estimates of genetic divergence were available (Muller *et al.* 2008; Ross-Ibarra *et al.* 2008; Pyhäjärvi *et al.* 2012). We found that the proportion of the genome showing TRD increased with genetic distance between populations and was clearly the highest in the most distant pair of populations crossed. This is in line with earlier studies in *A. lyrata*, which showed that the level of within-population TRD is very low (Leppälä *et al.* 2008) while several TRDL were mapped from a cross between genetically differentiated populations (Kuittinen *et al.* 2004) and from a backcross population between *A. lyrata* and its close relative *A. halleri* (Willems *et al.* 2007). Similarly, intraspecific rice crosses suggest increasing number of TRDL with increasing genetic distance between parental accessions (Matsubara *et al.* 2011). Our results do not support the snowballing hypothesis that the degree of TRD increases exponentially with time (Orr 1995; Orr and Turelli 2001): both the number of TRD regions and the percentage of the genome showing TRD increase approximately linearly with genetic distance.

An interesting feature of the overview of TRDL in the three crosses studied here was the uneven distribution of TRD over the genome, and in particular the degree of distortion of chromosomes AL1 and AL6: all three experimental crosses had TRDL on chromosomes 1 and 6. The same chromosomes showed TRD in an *A. lyrata* cross studied by Kuittinen *et al.* (2004) and in an interspecific cross between *A. lyrata* and *A. halleri* (Willems *et al.* 2007). In addition, in *A. thaliana* F_2 progenies, Salomé *et al.* (2011) found most TRD on chromosomes 1 and 5, in areas that locate to chromosomes 1 and 6 in *A. lyrata*. We do not currently have an explanation for this observation, but it does suggest that the degree of TRD varies across the genome and that in the genus *Arabidopsis* especially some regions of chromosome 1 seem to be involved. This suggests that we may learn more about the processes causing TRD, and its relation with genetic distance, by comparing nucleotide variation, gene content, and other variation between genomic regions with no TRD and regions with high numbers of TRD loci.

Acknowledgments

The authors thank M. Otsukka and A. Airikainen for their help in genotyping the Sp x Stu F_2 progeny. This study was supported by the Swedish Research Council and Umeå University (FB); work by J.L. and O.S. was supported by the Academy of Finland's Research Council for Biosciences and Environment (O.S.), by Biocenter Oulu (O.S.), and by the Finnish Population Genetics Graduate School (J.L.).

Literature Cited

Bikard, D., D. Patel, C. Le Mette, V. Giorgi, C. Camilleri *et al.*, 2009 Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science 323: 623–626.

Borg, M., L. Brownfield, and D. Twell, 2009 Male gametophyte development: a molecular perspective. J. Exp. Bot. 60: 1465–1478.

Brideau, N. J., H. A. Flores, J. Wang, S. Maheshwari, X. Wang *et al.*, 2006 Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*. Science 314: 1292–1295.

Buckler, E. S., T. L. Phelps-Durr, C. S. K. Buckler, R. K. Dawe, J. F. Doebley *et al.*, 1999 Meiotic drive of chromosomal knobs reshaped the maize genome. Genetics 153: 415–426.

Christie, P., and M. R. Macnair, 1987 The distribution of postmating reproductive isolating genes in populations of the yellow monkey flower, *Mimulus guttatus*. Evolution 41: 571–578.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Sunderland, MA.

Cutter, A. D., 2012 The polymorphic prelude to Bateson-Dobzhansky-Muller incompatibilities. Trends Ecol. Evol. 27: 209–218.

Fishman, L., and A. Saunders, 2008 Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. Science 322: 1559–1562.

Fishman, L., A. J. Kelly, E. Morgan, and J. H. Willis, 2001 A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. Genetics 159: 1701–1716.

Gérard, P. R., and D. C. Presgraves, 2012 Abundant genetic variability in *Drosophila simulans* for hybrid female lethality in interspecific crosses to *Drosophila melanogaster.* Genet. Res. 94: 1–7.

Hall, M. C., and J. H. Willis, 2005 Transmission ratio distortion in intraspecific hybrids of *Mimulus guttatus*: implications for genomic divergence. Genetics 170: 375–386.

Harushima, Y., M. Nakagahra, M. Yano, T. Sasaki, and N. Kurata, 2001 A genome-wide survey of reproductive barriers in an intraspecific hybrid. Genetics 159: 883–892.

Hatfield, T., and D. Schluter, 1999 Ecological speciation in sticklebacks: environment-dependent hybrid fitness. Evolution 53: 866–873.

Howard, D. J., 1999 Conspecific sperm and pollen precedence and speciation. Annu. Rev. Ecol. Syst. 30: 109–132.

Jenczewski, E., M. Gherardi, I. Bonnin, J. M. Prosperi, I. Olivieri *et al.*, 1997 Insight on segregation distortions in two intraspecific crosses between annual species of *Medicago* (Leguminosae). Theor. Appl. Genet. 94: 682–691.

Johnson, N. A., 2010 Hybrid incompatibility genes: remnants of a genomic battlefield? Trends Genet. 26: 317–325.

Kemi, U., A. Niittyvuopio, T. Toivainen, A. Pasanen, B. Quilot-Turion *et al.*, 2013 Role of vernalization and of duplicated FLOWERING LOCUS C in the perennial *Arabidopsis lyrata*. New Phytol. 197: 323–335.

Koide, Y., K. Onishi, D. Nishimoto, A. R. Baruah, A. Kanazawa *et al.*, 2008a Sex-independent transmission ratio distortion system responsible for reproductive barriers between Asian and African rice species. New Phytol. 179: 888–900.

Koide, Y., M. Ikenaga, N. Sawamura, D. Nishimoto, K. Matsubara *et al.*, 2008b The evolution of sex-independent transmission ratio distortion involving multiple allelic interactions at a single locus in rice. Genetics 180: 409–420.

Koide, Y., Y. Shinya, M. Ikenaga, N. Sawamura, K. Matsubara *et al.*, 2012 Complex genetic nature of sex-independent transmission ratio distortion in Asian rice species: the involvement of unlinked modifiers and sex-specific mechanisms. Heredity 108: 242–247.

Korbecka, G., P. G. L. Klinkhamer, and K. Vrieling, 2002 Selective embryo abortion hypothesis revisited: a molecular approach. Plant Biol. 4: 298–310.

Kuittinen, H., A. A. de Haan, C. Vogl, S. Oikarinen, J. Leppälä *et al.*, 2004 Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. Genetics 168: 1575–1584.

Leinonen, P. H., D. L. Remington, and O. Savolainen, 2011 Local adaptation, phenotypic differentiation and hybrid fitness in diverged natural populations of *Arabidopsis lyrata*. Evolution 65: 90–107.

Leppälä, J., and O. Savolainen, 2011 Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. Evolution 65: 2959–2972.

Leppälä, J., J. S. Bechsgaard, M. H. Schierup, and O. Savolainen, 2008 Transmission ratio distortion in *Arabidopsis lyrata*: effects of population divergence and the S-locus. Heredity 100: 71–78.

Levin, D. A., 2003 The cytoplasmic factor in plant speciation. Syst. Bot. 28: 5–11.

Lyttle, T. W., 1991 Segregation distorters. Annu. Rev. Genet. 25: 511–557.

Martin, N. H., and J. H. Willis, 2010 Geographical variation in postzygotic isolation and its genetic basis within and between two *Mimulus* species. Phil. Trans. R. Soc. B. 365: 2469–2478.

Matsubara, K., K. Ebana, T. Mizubayashi, S. Itoh, T. Ando *et al.*, 2011 Relationship between transmission ratio distortion and genetic divergence in intraspecific rice crosses. Mol. Genet. Genomics 286: 307–319.

Matute, D. R., I. A. Butler, D. A. Turissini, and J. A. Coyne, 2010 A test of the snowball theory for the rate of evolution of hybrid incompatibilities. Science 329: 1518–1521.

McDaniel, S. F., J. H. Willis, and A. J. Shaw, 2008 The genetic basis of developmental abnormalities in interpopulation hybrids of the moss *Ceratodon purpureus*. Genetics 179: 1425–1435.

McDermott, S. R., and M. A. F. Noor, 2010 The role of meiotic drive in hybrid male sterility. Phil. Trans. R. Soc. B. 365: 1265–1272.

Moyle, L. C., and E. B. Graham, 2006 Genome-wide associations between hybrid sterility QTL and marker transmission ratio distortion. Mol. Biol. Evol. 23: 973–980.

Moyle, L. C., and T. Nakazato, 2010 Hybrid incompatibility "snowballs" between *Solanum* species. Science 329: 1521–1523.

Muller, M. H., J. Leppälä, and O. Savolainen, 2008 Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. Heredity 100: 47–58.

Muralla, R., J. Lloyd, and D. Meinke, 2011 Molecular foundations of reproductive lethality in *Arabidopsis thaliana*. PLoS ONE 6: e28398.

Myburg, A. A., C. Vogl, A. R. Griffin, R. R. Sederoff, and R. W. Whetten, 2004 Genetics of postzygotic isolation in eucalyptus: whole-genome analysis of barriers to introgression in a wide interspecific cross of *Eucalyptus grandis* and *E. globulus*. Genetics 166: 1405–1418.

Nakazato, T., M.-K. Jung, E. A. Housworth, L. H. Rieseberg, and G. J. Gastony, 2007 A genomewide study of reproductive barriers between allopatric populations of a homosporous fern, *Ceratopteris richardii*. Genetics 177: 1141–1150.

Orr, H. A., 1995 The population-genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139: 1805–1813.

Orr, A. H., and M. Turelli, 2001 The evolution of postzygotic isolation: accumulating Dobzhansky–Muller incompatibilities. Evolution 55: 1085–1094.

Palestis, B. G., R. Trivers, A. Burt, and R. N. Jones, 2004 The distribution of B chromosomes across species. Cytogenet. Genome Res. 106: 151–158.

Presgraves, D. C., 2010 The molecular evolutionary basis of species formation. Nat. Rev. Genet. 11: 175–180.

Pyhäjärvi, T., E. Aalto, and O. Savolainen, 2012 Time scales of divergence and speciation among natural populations and subspecies of *Arabidopsis lyrata* (Brassicaceae). Am. J. Bot. 99: 1–9.

Quilot-Turion, B., J. Leppälä, P. H. Leinonen, P. Waldmann, O. Savolainen *et al.*, 2013 Genetic changes in flowering and morphology in response to adaptation to a high-latitude environment in *Arabidopsis lyrata*. Ann. Bot. 111: 957–958.

Reed, L. K., and T. A. Markow, 2004 Early events in speciation: polymorphism for hybrid male sterility in *Drosophila*. Proc. Natl. Acad. Sci. USA 101: 9009–9012.

Rick, C. M., 1969 Controlled introgression of chromosomes of *Solanum pennellii* into *Lycopersicon esculentum*: segregation and recombination. Genetics 62: 753–768.

Rieseberg, L. H., and B. K. Blackman, 2010 Speciation genes in plants. Ann. Bot. 106: 439–455.

Rieseberg, L. H., S. J. E. Baird, and K. A. Gardner, 2000 Hybridization, introgression, and linkage evolution. Plant Mol. Biol. 42: 205–224.

Ross-Ibarra, J., S. I. Wright, J. P. Foxe, A. Kawabe, L. DeRose-Wilson *et al.*, 2008 Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS ONE 3: e2411.

Salomé, P. A., K. Bomblies, J. Fitz, R. A. E. Laitinen, N. Warthmann *et al.*, 2011 The recombination landscape in *Arabidopsis thaliana* F_2 populations. Heredity 108: 447–455.

Sano, Y., 1992 Genetic comparisons of chromosome 6 between wild and cultivated rice. Jpn. J. Breed 42: 561–572.

Sano, Y., Y.E. Chu, and H. I. Oka, 1979 Genetic studies of speciation in cultivated rice, 1. Genic analysis for the F1 sterility between O. sativa L. and O. glaberrima Steud. Jpn. J. Genet 54: 121–132.

Sweigart, A. L., A. R. Mason, and J. H. Willis, 2007 Natural variation for a hybrid incompatibility between two species of *Mimulus*. Evolution 61: 141–151.

Vogl, C., and S. Z. Xu, 2000 Multipoint mapping of viability and segregation distorting loci using molecular markers. Genetics 155: 1439–1447.

Werren, J. H., 2011 Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc. Natl. Acad. Sci. USA 108: 10863–10870.

Wijsman, E. M., 1987 A deductive method of haplotype analysis in pedigrees. Am. J. Hum. Genet. 41: 356–373.

Willems, G., D. B. Drager, M. Courbot, C. Gode, N. Verbruggen *et al.*, 2007 The genetic basis of zinc tolerance in the metallophyte *Arabidopsis halleri* ssp *halleri* (brassicaceae): an analysis of quantitative trait loci. Genetics 176: 659–674.

Wright, S., B. Lauga, and D. Charlesworth, 2003 Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol. Ecol. 12: 1247–1263.

Communicating editor: S. I. Wright

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152561/-/DC1

Investigating Incipient Speciation in *Arabidopsis lyrata* from Patterns of Transmission Ratio Distortion

Johanna Leppälä, Folmer Bokma, and Outi Savolainen

DOI: 10.1534/genetics.113.152561

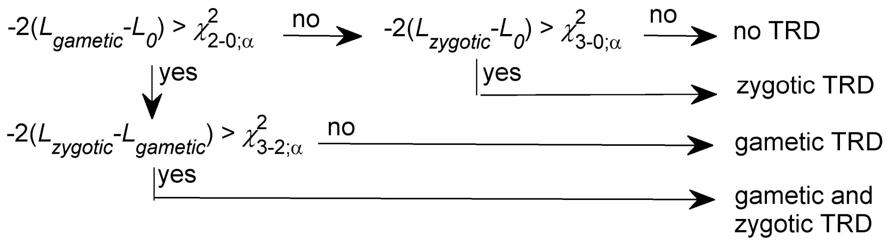
$-2(L_{gametic}-L_0) > \chi^2_{2-0;\alpha}$
no
$-2(L_{zygotic}-L_0) > \chi^2_{3-0;\alpha}$
no
no TRD
yes
yes
zygotic TRD
$-2(L_{zygotic}-L_{gametic}) > \chi^2_{3-2;\alpha}$
no
gametic TRD
yes
gametic and zygotic TRD

Figure S1 Schematic view of the distinction between gametic and zygotic TRD. First, the maximum log-likelihoods of the genotype frequencies under the null (i.e. Mendelian), gametic, and zygotic model are calculated. Then, likelihood ratios are compared to cumulative chi-square distributions with appropriate degrees of freedom (see main text) evaluated at significance level alpha. If this comparison suggests that the data is better explained by the alternative hypothesis, the arrow labeled "yes" is followed, otherwise the arrow labeled "no" is followed.

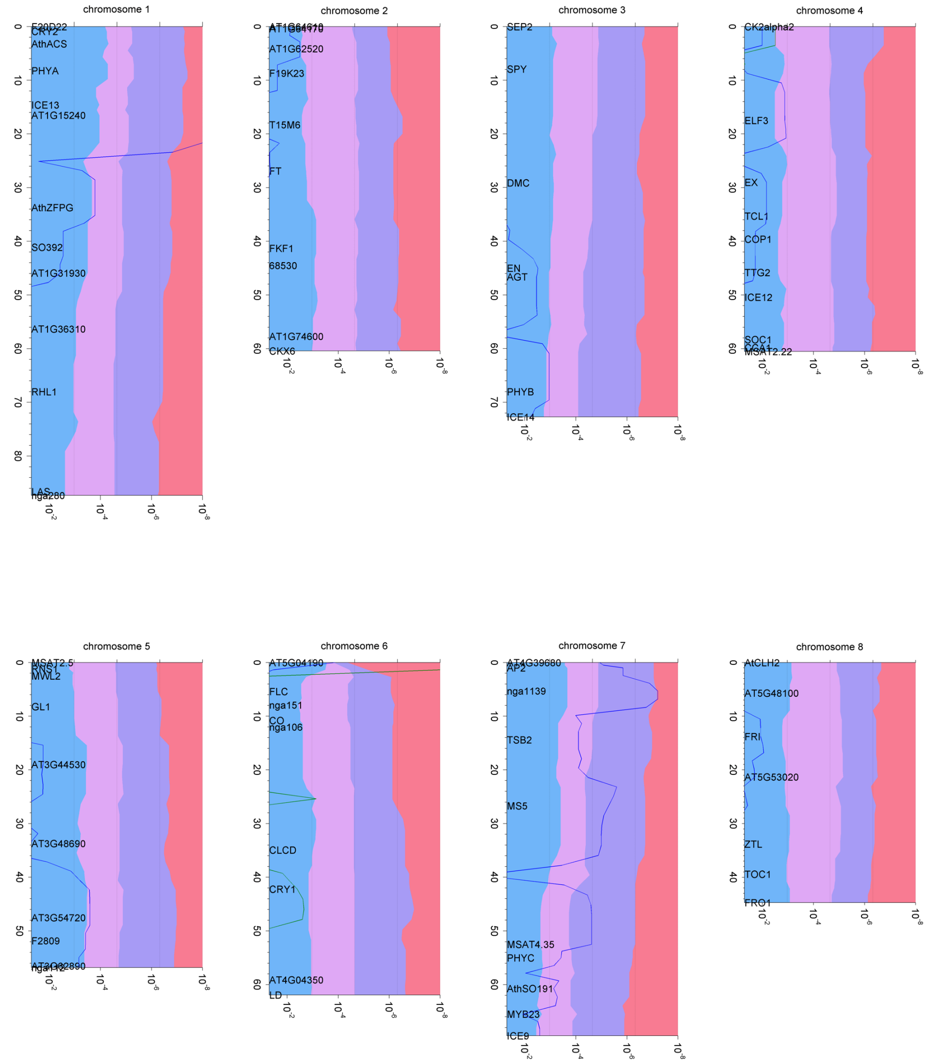

Figure S2 Single-locus TRD analyses for SpMaF_2. The *p*-value of the likelihood ratio tests are on the horizontal axes and genetic distances in cM on the y-axis. The blue line indicates gametic vs. null, and the green line indicates zygotic vs. null. The frequencies of the four F_2 genotypes are shown in different colours; from left to right: Sp homozygote (Sp1Sp2, blue), Sp1Ma2, Ma1Sp2, and Ma homozygote (Ma1Ma2, red).

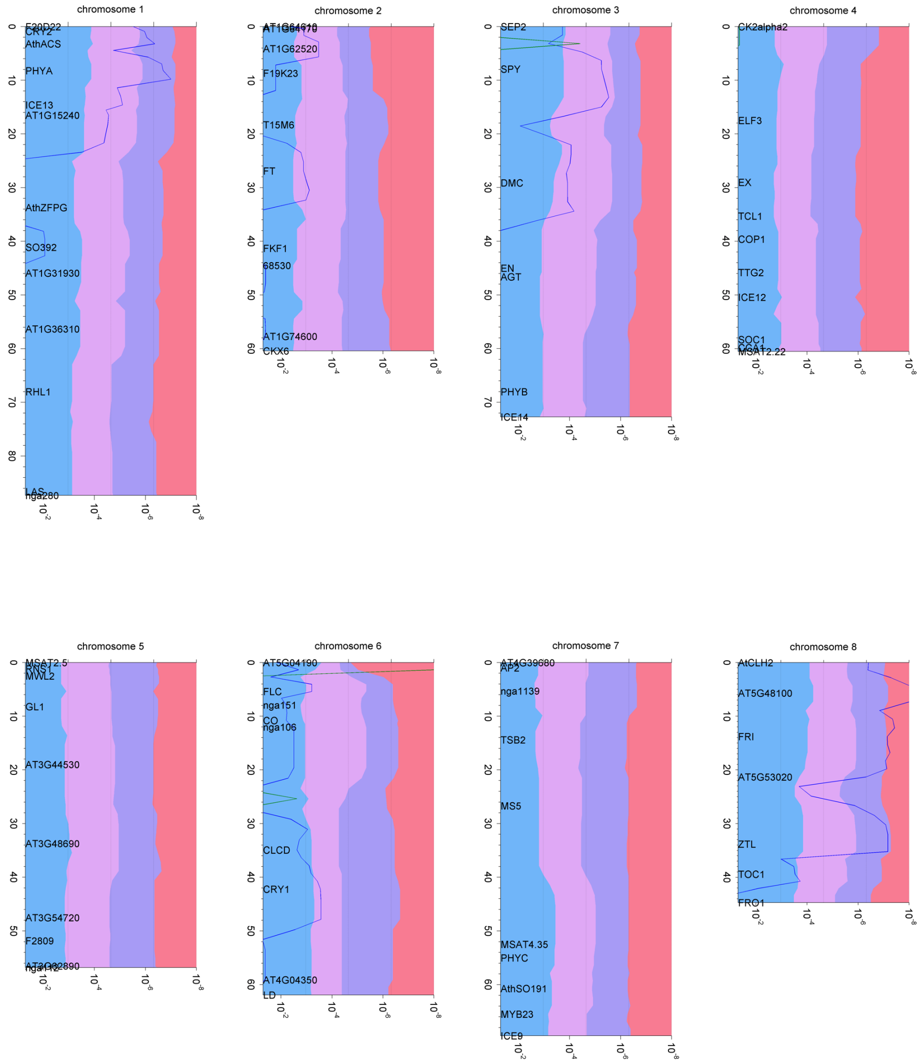

Figure S3 Single-locus TRD analyses for MaSpF_2. The *p*-value of the likelihood ratio tests are on the horizontal axes and genetic distances in cM on the y-axis. The blue line indicates gametic vs. null, and the green line indicates zygotic vs. null. The frequencies of the four F_2 genotypes are shown in different colours; from left to right: Sp homozygote (Sp1Sp2, blue), Sp1Ma2, Ma1Sp2 and Ma homozygote (Ma1Ma2, red).

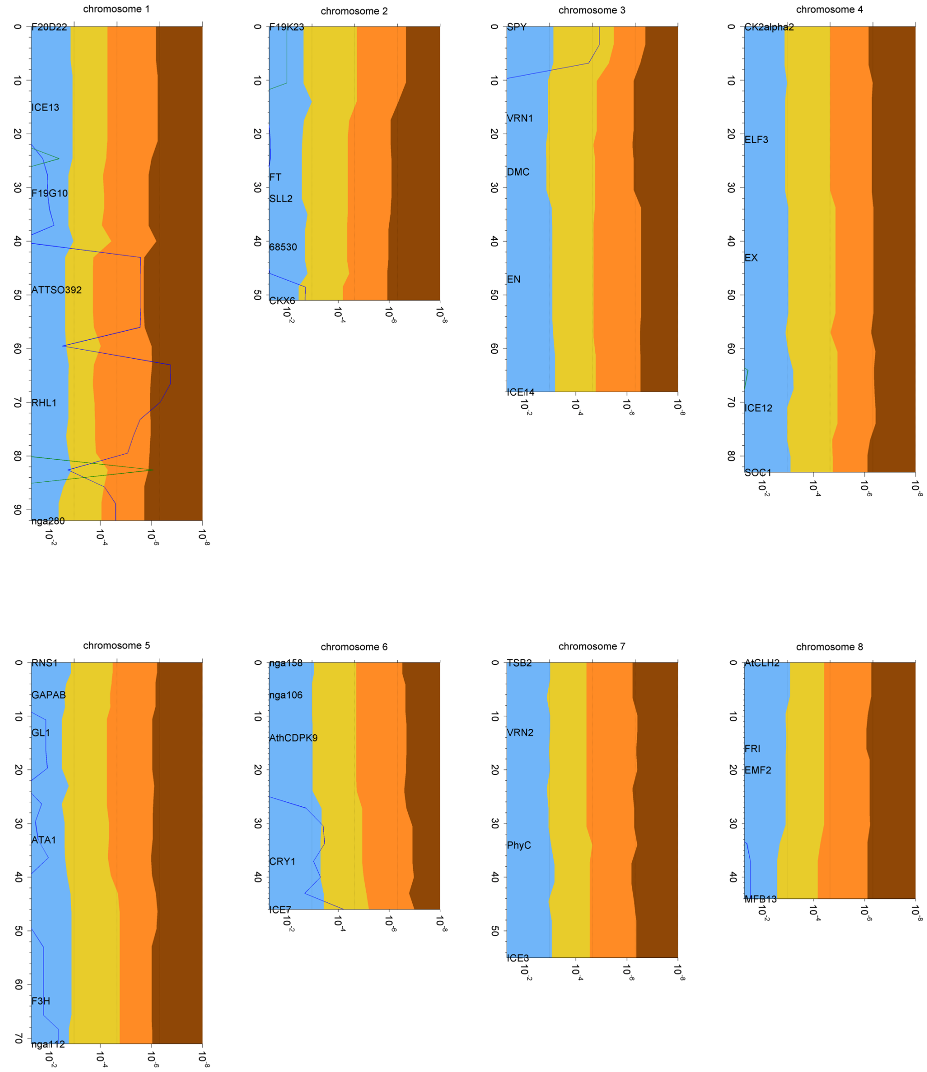

Figure S4 Single-locus TRD analyses for SpPl2F$_2$. The *p*-value of the likelihood ratio tests are on the horizontal axes and genetic distances in cM on the y-axis. The blue line indicates gametic vs. null, and the green line indicates zygotic vs. null. The frequencies of the four F$_2$ genotypes are shown in different colours; from left to right: Sp homozygote (Sp1Sp2, blue), Sp1Pl2, Pl1Sp2 and Pl homozygote (Pl1Pl2, brown).

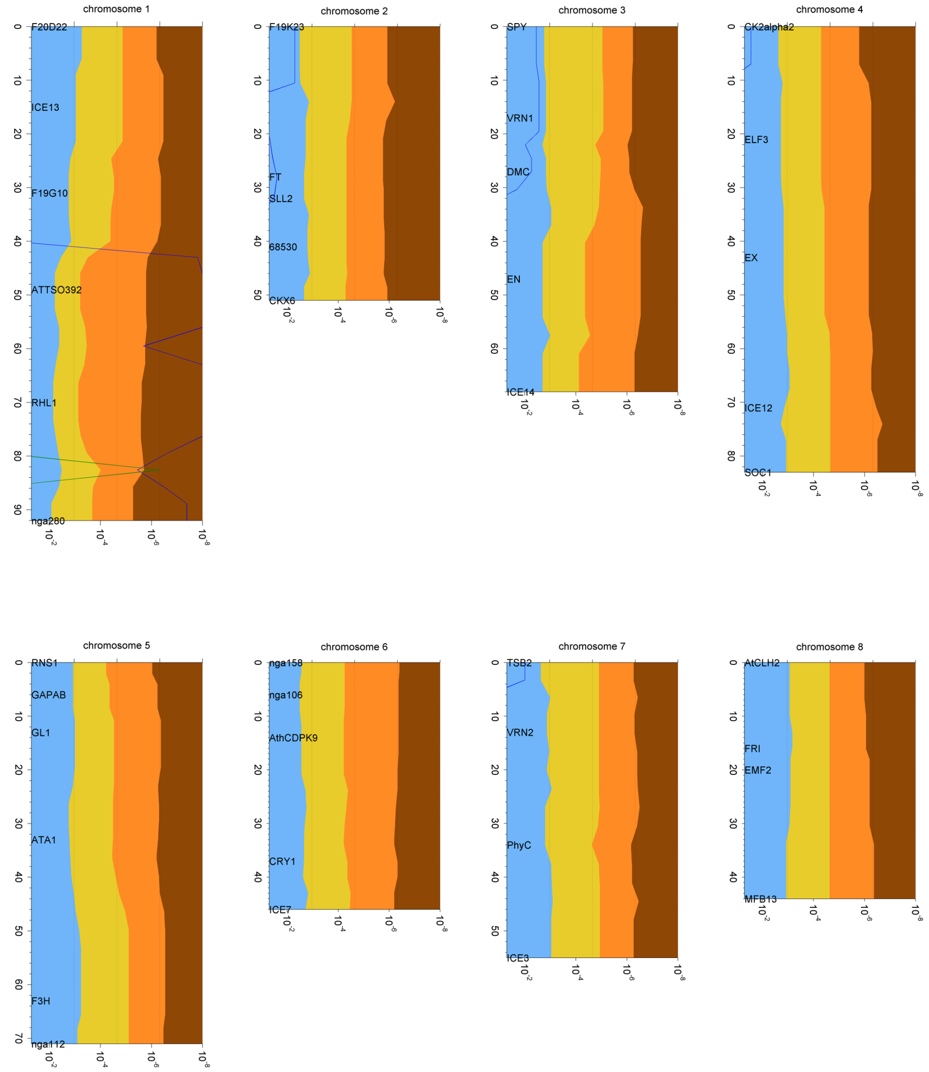

Figure S5 Single-locus TRD analyses for SpPl3F_2. The *p*-value of the likelihood ratio tests are on the horizontal axes and genetic distances in cM on the y-axis. The blue line indicates gametic vs. null, and the green line indicates zygotic vs. null. The frequencies of the four F_2 genotypes are shown in different colours; from left to right: Sp homozygote (Sp1Sp2, blue), Sp1Pl2, Pl1Sp2 and Pl homozygote (Pl1Pl2, brown).

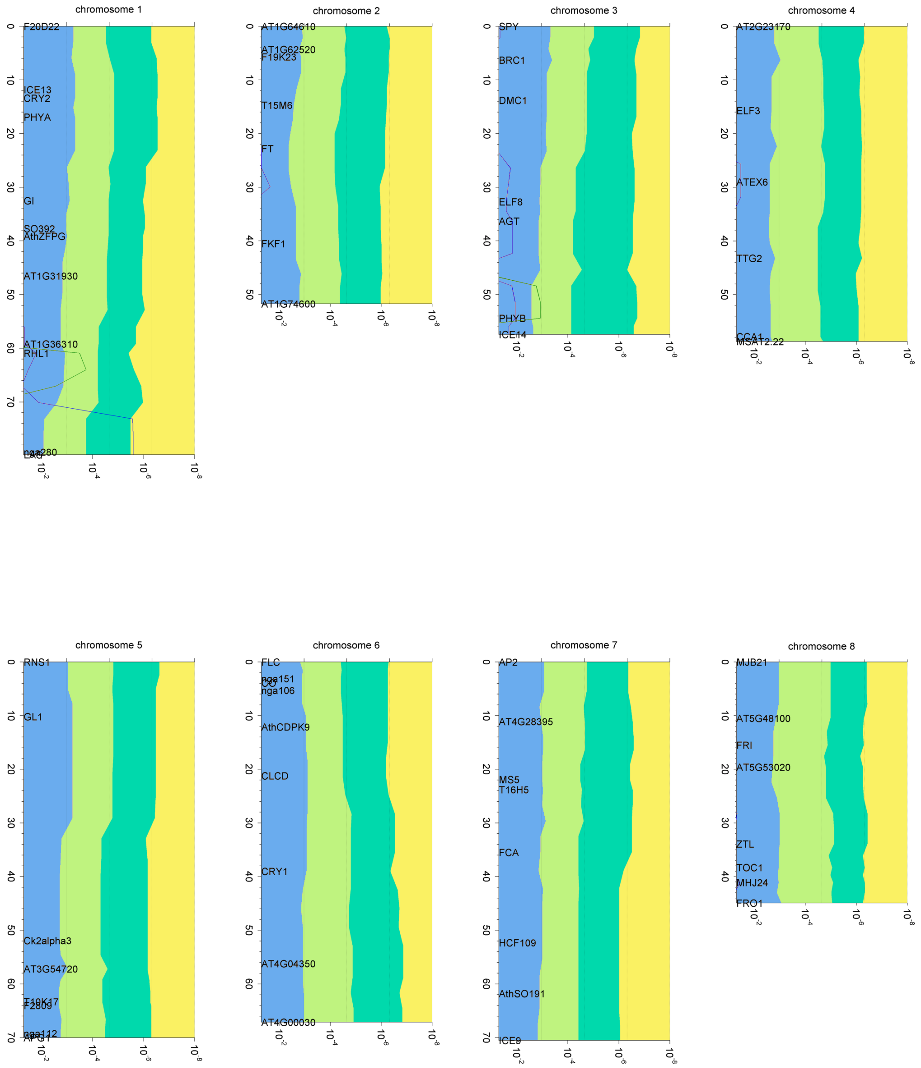

Figure S6 Single-locus TRD analyses for SpStuF_2. The *p*-value of the likelihood ratio tests are on the horizontal axes and genetic distances in cM on the y-axis. The blue line indicates gametic vs. null, and the green line indicates zygotic vs. null. The frequencies of the four F_2 genotypes are shown in different colours; from left to right: Sp homozygote (Sp1Sp2, blue), Sp1Stu2, Stu1Sp2 and Stu homozygote (Stu1Stu2, yellow).

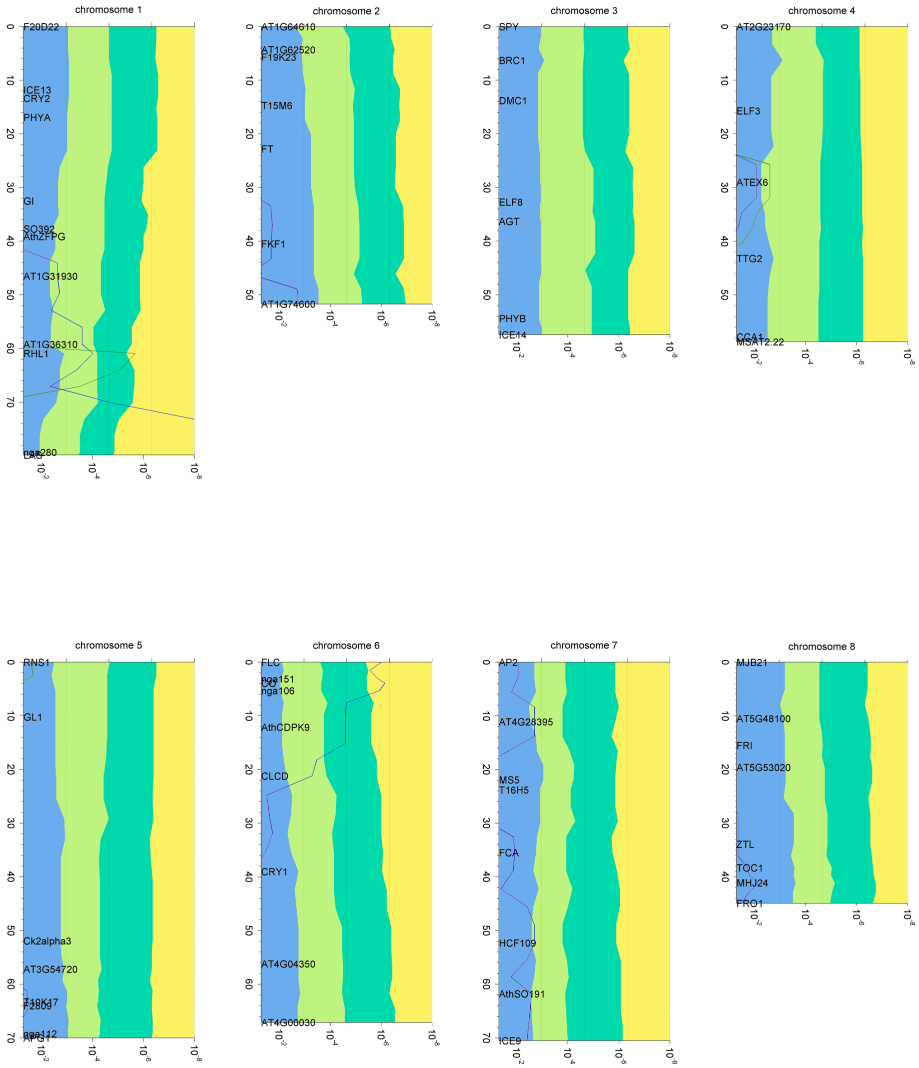

Figure S7 Single-locus TRD analyses for StuSpF$_2$. The *p*-value of the likelihood ratio tests are on the horizontal axes and genetic distances in cM on the y-axis. The blue line indicates gametic vs. null, and the green line indicates zygotic vs. null. The frequencies of the four F$_2$ genotypes are shown in different colours; from left to right: Sp homozygote (Sp1Sp2, blue), Sp1Stu2, Stu1Sp2 and Stu homozygote (Stu1Stu2, yellow).

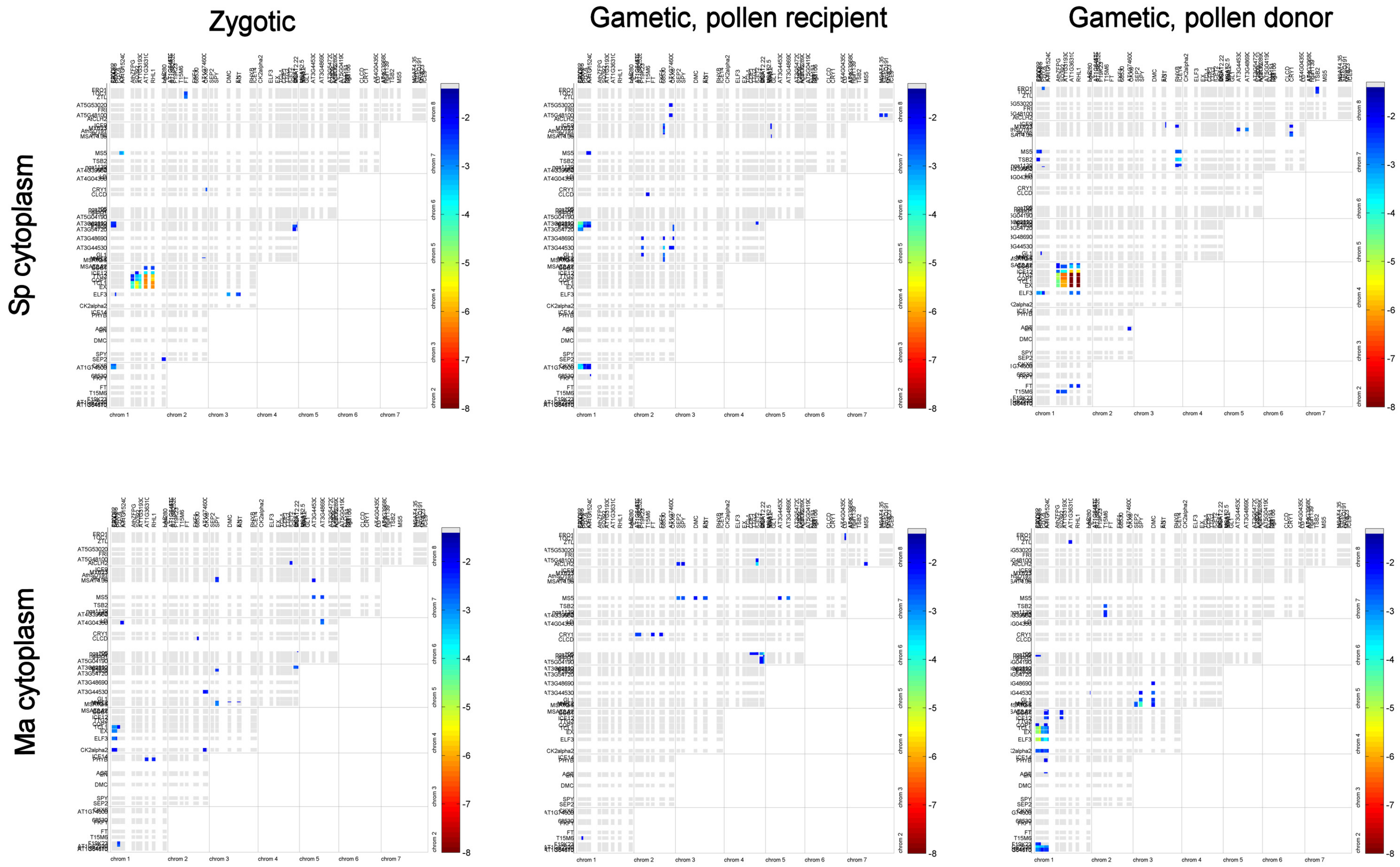

Figure S8 Two-locus TRD analyses for Sp x Ma reciprocal crosses. In each plot the molecular markers are indicated on horizontal axis starting from AL1 to AL8 from left to right. Similarly in vertical axis the marker order runs from AL1 to AL8 from down to up. The colours indicate *p*-values (between 0.01 - 1x10^{-8}) from two-locus χ^2 tests. Any *p*>0.01 is shown in grey.

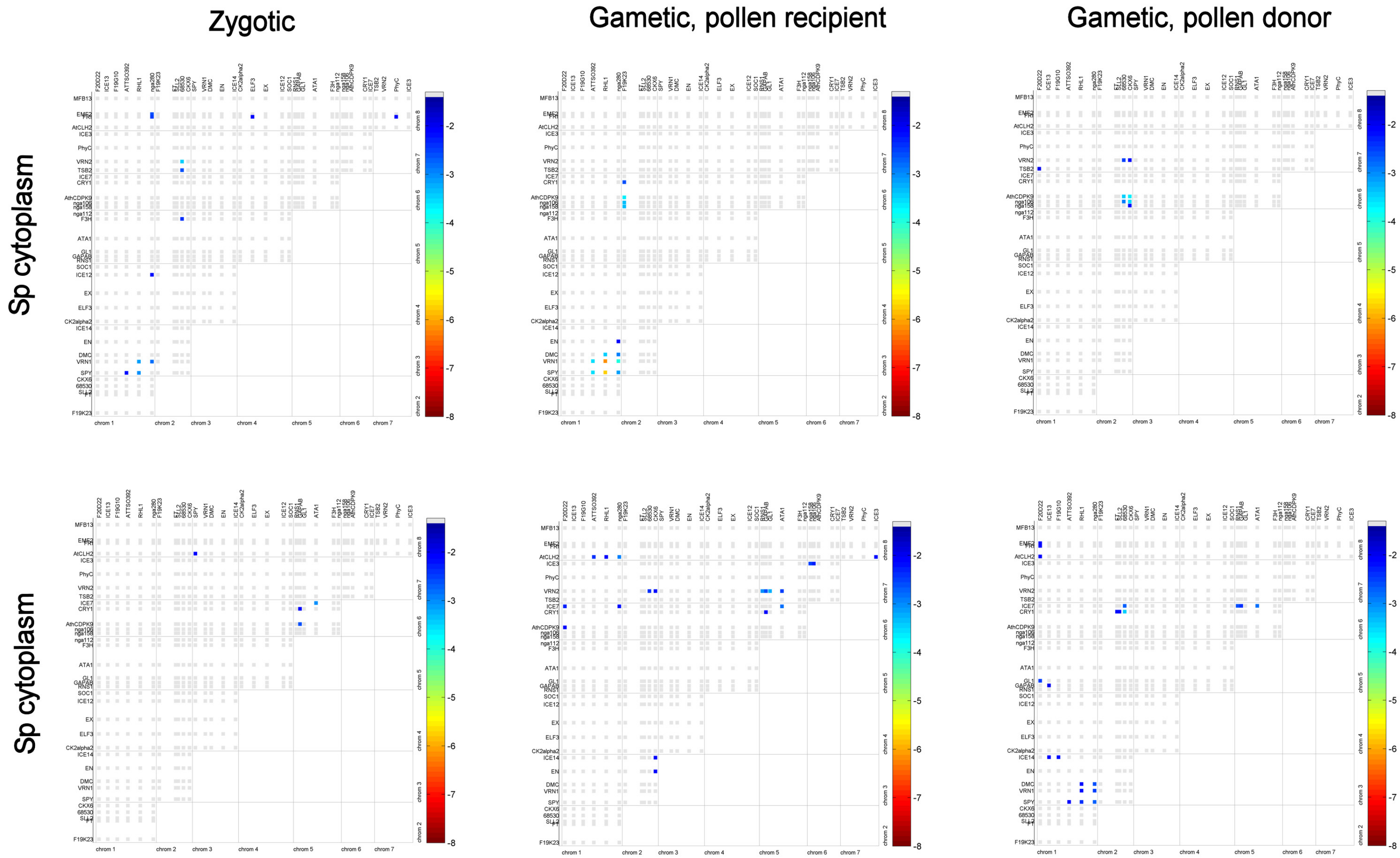

Figure S9 Two-locus TRD analyses for Sp x Pl reciprocal crosses. In each plot the molecular markers are indicated on horizontal axis starting from AL1 to AL8 from left to right. Similarly in vertical axis the marker order runs from AL1 to AL8 from down to up. The colours indicate *p*-values (between 0.01 - $1x10^{-8}$) from two-locus χ^2 tests. Any *p*>0.01 is shown in grey.

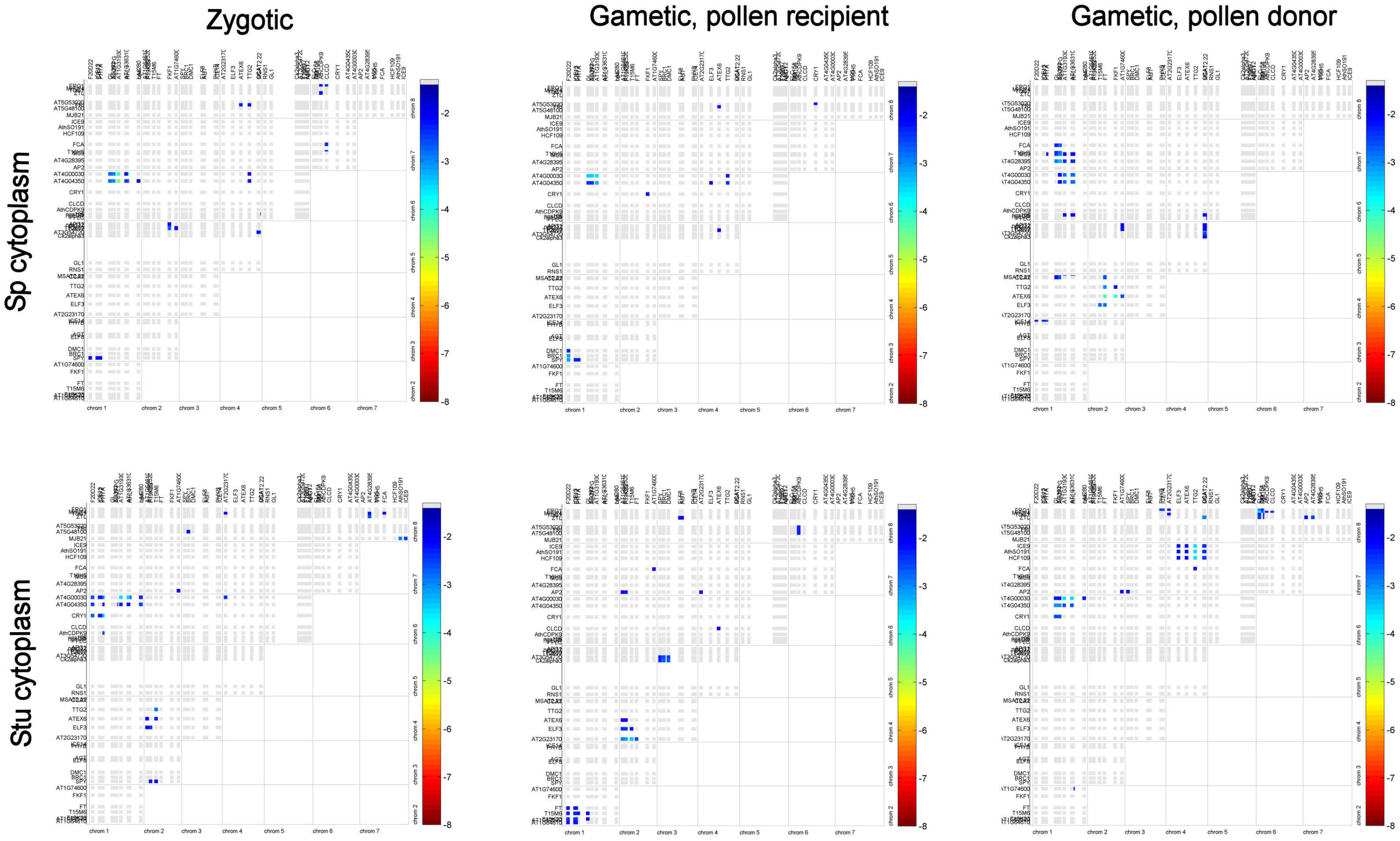

Figure S10 Two-locus TRD analyses for Sp x Stu reciprocal crosses. In each plot the molecular markers are indicated on horizontal axis starting from AL1 to AL8 from left to right. Similarly in vertical axis the marker order runs from AL1 to AL8 from down to up. The colours indicate *p*-values (between 0.01 - 1x10^{-8}) from two-locus χ^2 tests. Any *p*>0.01 is shown in grey.

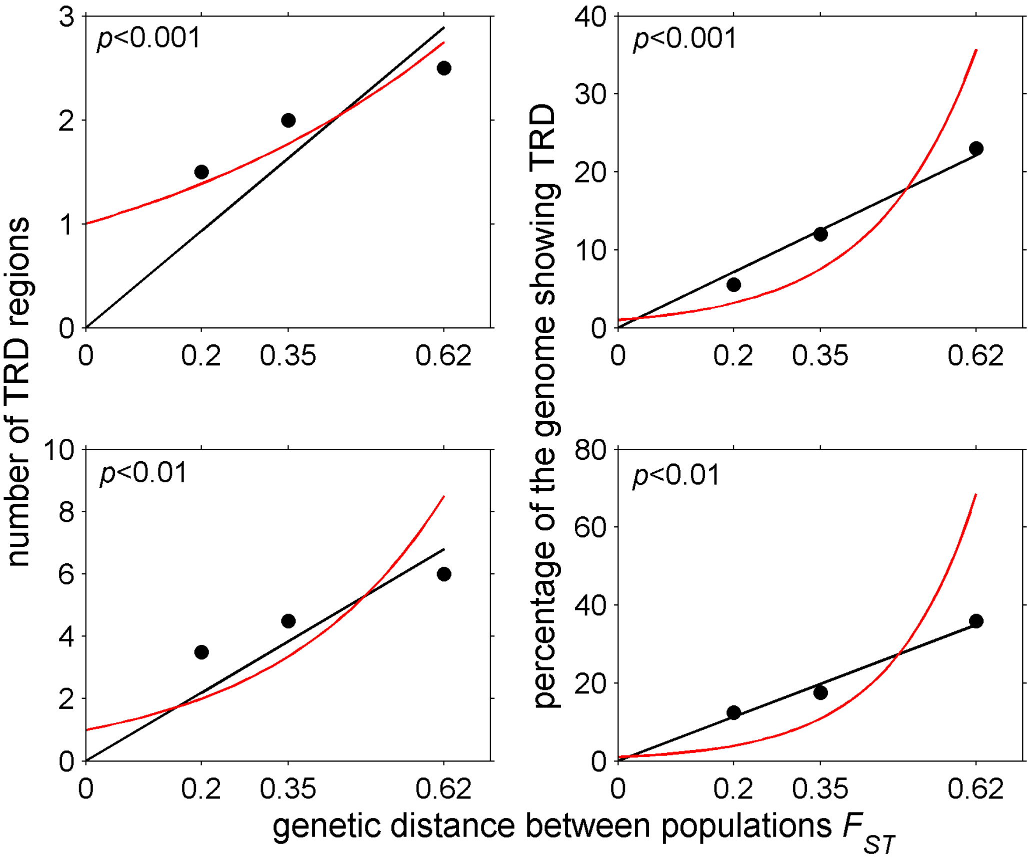

Figure S11 Analysis of "snowballing" of TRD. Fit of linear (black lines) and exponential (red lines) functions to the increase of the number of TRD regions (left panels) and percentage of the genome showing TRD (right panels) with genetic distance between hybridized populations (F_{ST}). Upper panels show results when a significance threshold of *p*<0.001 is used, lower panels show results for a threshold of *p*<0.01

Table S1 Origins of primer sequences for molecular markers genotyped for Sp x Stu cross. Marker types and locus in *A. thaliana* are indicated to all markers.

Chromosome	Locus name	Marker type	Locus or BAC in *A. thaliana*	Origin of primer sequences	Forward primer	Reverse primer
AL1	F20D22	Microsat	AT1G04120	Clauss *et al*. 2002		
AL1	CRY2	SNP	AT1G04400	Kuittinen et al. 2004		
AL1	PHYA	SNP	AT1G09570	Kuittinen et al. 2004		
AL1	ICE13	Microsat	AT1G13220	Clauss et al. 2002		
AL1	GI	SNP	AT1G22770	Kuittinen et al. 2004		
AL1	AthZFPG	Microsat	AT1G24625	Clauss et al. 2002		
AL1	ATTSO392	Microsat	AT1G30630	Clauss et al. 2002		
AL1	AT1G31930	SNP	AT1G31930	Ross-Ibarra et al. 2008		
AL1	AT1G36310	SNP	AT1G36310	Hansson et al. 2006		
AL1	RHL1	CAPS/*Sma*I	AT1G48380	Kuittinen et al. 2004		
AL1	LAS	SNP	AT1G55580	Leppälä & Savolainen 2011		
AL1	nga280	Microsat	AT1G55840	Clauss et al. 2002		
AL2	AT1G64610	SNP	AT1G64610	Hansson et al. 2006		
AL2	AT1G62520	SNP	AT1G62520	Hansson et al. 2006		
AL2	F19K23	Microsat	AT1G62050	Clauss et al. 2002		
AL2	T15M6	Microsat	AT1G58180	Leppälä & Savolainen 2011		
AL2	FT	SNP	AT1G65480	Kuittinen et al. 2004		
AL2	FKF1	SNP	AT1G68050	A. Niittyvuopio		
AL2	AT1G74600	SNP	AT1G74600	Ross-Ibarra et al. 2008		
AL3	SPY	SNP	AT3G11540	Kuittinen et al. 2004		
AL3	BRC1	SNP	AT3G18550		ATTGCTCCCTTTTAGCCCTTC	TCTCTCGTCCTTGGACAACTTC
AL3	DMC1	SNP	AT3G22880	Kuittinen et al. 2004		
AL3	ELF8	SNP	AT2G06210		GCTGCTAATGATGCGACTGAT	ACTTCCACTTGCAGCTTCTTG
AL3	AGT	SNP	AT2G16870	Leppälä & Savolainen 2011		
AL3	PHYB	SNP	AT2G18790	Kuittinen et al. 2004		
AL3	ICE14	Microsat	AT2G20310	Clauss et al. 2002		
AL4	AT2G23170	SNP	AT2G23170	Ross-Ibarra et al. 2008		
AL4	ELF3	SNP	AT2G25930	Kuittinen et al. 2004		

Chromosome	Locus name	Marker type	Locus or BAC in *A. thaliana*	Origin of primer sequences	Forward primer	Reverse primer
AL4	ATEX6	SNP	AT2G28950	Kuittinen et al. 2004		
AL4	TTG2	SNP	AT2G37260	Leppälä & Savolainen 2011		
AL4	CCA1	SNP	AT2G46830	A. Niittyvuopio		
AL4	MSAT2.22	Microsat	AT2G47960	Loudet et al. 2002		
AL5	RNS1	SNP	AT2G02990	Kuittinen et al. 2004		
AL5	GL1_A95D	SNP	AT3G27920	Kivimäki et al. 2007		
AL5	CK2alpha3	dCAPS/*Vsp*I	AT3G50000		GGAAGCCTTGGTCCAAATTCAT**T**AA	CACATGTTCGAGTTATGTTACGTG
AL5	AT3G54720	SNP	AT3G54720	Ross-Ibarra et al. 2008		
AL5	T10K17	Microsat	T10K17		CAAAAGTTGGTGGTAGTGG	CACGCAAAATTACAATCTCTG
AL5	F2809	Microsat	AT3G57320	Leppälä & Savolainen 2011		
AL5	nga112	Microsat	AT3G62650	Clauss et al. 2002		
AL5	APG1	SNP	AT3G63410		TTACCTTCCCCAAGGGTTTAG	AGCTGCTAGAGTTCCCAGGAG
AL6	FLC	SNP	AT5G10140	A. Niittyvuopio		
AL6	nga151	Microsat	AT5G14480	Bell & Ecker 1994		
AL6	CO	SNP	AT5G15840	Kuittinen et al. 2004		
AL6	nga106	Microsat	AT5G16520	Bell & Ecker 1994		
AL6	AthCDPK9	Microsat	MQM1	Clauss et al. 2002		
AL6	CLC-D	SNP	AT5G26240	Kuittinen et al. 2004		
AL6	CRY1	CAPS/*Bam*HI	AT4G08920	Kuittinen et al. 2004		
AL6	AT4G04350	SNP	AT4G04350	Ross-Ibarra et al. 2008		
AL6	AT4G00030	SNP	AT4G00030	Ross-Ibarra et al. 2008		
AL7	AP2	SNP	AT4G36920	Kuittinen et al. 2004		
AL7	AT4G28395	SNP	AT4G28395	Ponce et al. 1999		
AL7	MS5	SNP	AT4G20900	Leppälä & Savolainen 2011		
AL7	T16H5	Microsat	T16H5		TGGCAGTACCTATCTATCGTA	CGGAATTAGGGATTTCAGA
AL7	FCA	SNP	AT4G16280	Kuittinen et al. 2004		
AL7	HCF109	SNP	AT5G36170		AGAGCTTCTGCTGGTTTGGAG	TCGCCAGTTGACTTCTCTCCT
AL7	ATTSO191	Microsat	AT5G37780	Clauss et al. 2002		
AL7	ICE9	Microsat	AT5G40340	Clauss et al. 2002		
AL8	MJB21	Microsat	MJB21		AAAGTAAGCCAAGCGTCAT	AACTAACAAAAAGCGGAGAAG

Chromosome	Locus name	Marker type	Locus or BAC in *A. thaliana*	Origin of primer sequences	Forward primer	Reverse primer
AL8	AT5G48100	SNP	AT5G48100	Ross-Ibarra et al. 2008		
AL8	FRI	SNP & indel	AT4G00650	Kuittinen et al. 2004		
AL8	AT5G53020	SNP	AT5G53020	Ross-Ibarra et al. 2008		
AL8	ZTL	SNP	AT5G57360	A. Niittyvuopio		
AL8	TOC1	SNP	AT5G61380	A. Niittyvuopio		
AL8	MHJ24	Microsat	MHJ24	Clauss et al. 2002		
AL8	FRO1	SNP	AT5G67590	Leppälä & Savolainen 2011		

Table S2 Zygotic two-locus interaction between AL1 and AL6 in Sp x Stu cross. Observed and expected (in parentheses) two locus genotype counts.

		Sp cytoplasm				Stu cytoplasm				
		AL6, AT4G04350 (56 cM)					**AL6**, AT4G04350 (56 cM)			
		Sp1Sp2	Sp1Stu2	Sp2Stu1	Stu1Stu2		Sp1Sp2	Sp1Stu2	Sp2Stu1	Stu1Stu2
AL1, F20D22 (0 cM)	Sp1Sp2	14 (13)	13 (16)	17 (16)	11 (9)	Sp1Sp2	16 (12)	13 (14)	10 (16)	10 (13)
	Sp1Stu2	13 (9)	10 (10)	9 (11)	4 (6)	Sp1Stu2	12 (8)	7 (9)	17 (10)	9 (8)
	Sp2Stu1	15 (12)	14 (15)	13 (15)	8 (8)	Sp2Stu1	13 (11)	13 (13)	17 (14)	9 (12)
	Stu1Stu2	4 (12)	18 (14)	17 (14)	9 (8)	Stu1Stu2	0 (10)	16 (12)	10 (14)	16 (11)
		Sp1Sp2	Sp1Stu2	Sp2Stu1	Stu1Stu2		Sp1Sp2	Sp1Stu2	Sp2Stu1	Stu1Stu2
AL1, AT1G31930 (47 cM)	Sp1Sp2	18 (10)	9 (13)	10 (13)	6 (7)	Sp1Sp2	13 (9)	6 (11)	6 (12)	4 (10)
	Sp1Stu2	18 (12)	16 (14)	7 (15)	8 (8)	Sp1Stu2	17 (11)	15 (13)	14 (14)	10 (11)
	Sp2Stu1	10 (9)	8 (11)	17 (12)	4 (7)	Sp2Stu1	8 (8)	9 (10)	14 (11)	12 (9)
	Stu1Stu2	0 (14)	22 (17)	22 (17)	14 (10)	Stu1Stu2	3 (13)	19 (15)	20 (17)	18 (14)
		Sp1Sp2	Sp1Stu2	Sp2Stu1	Stu1Stu2		Sp1Sp2	Sp1Stu2	Sp2Stu1	Stu1Stu2
AL1, LAS (80 cM)	Sp1Sp2	7 (5)	7 (6)	8 (7)	0 (4)	Sp1Sp2	10 (5)	5 (6)	3 (6)	0 (5)
	Sp1Stu2	17 (11)	6 (14)	19 (14)	5 (8)	Sp1Stu2	11 (10)	10 (12)	16 (13)	6 (11)
	Sp2Stu1	10 (12)	19 (14)	10 (15)	10 (8)	Sp2Stu1	8 (11)	11 (13)	10 (14)	9 (11)
	Stu1Stu2	12 (17)	23 (21)	19 (21)	17 (12)	Stu1Stu2	12 (15)	23 (18)	25 (20)	29 (17)

File S1
TRD mapping algorithms.

Genotype probabilities: The crossing design (Fig.1) allows us to infer the population of origin of alleles at fully informative marker loci, and the origin of some alleles at partially informative loci. Hence, we can make inferences about the population of origin of the alleles at the remaining marker loci, i.e. we can infer haplotypes. Let us consider a marker with alleles abc and d where alleles a and c are from population 0 and alleles b and d from population 1. Instead of the name of the allele (a or b) we can use the population of origin as the "phase" of the maternal F_1 allele, i.e. 0 if allele a was inherited from the female F_1 parent, and 1 if allele b was inherited from the female F_1 parent. Similarly, we can write 0 if allele c was inherited from the male F_1 parent, and 1 if allele d was inherited from the male F_1 parent. Thus, we can re-write the four possible genotypes ac, ad, bc, and bd as "phases" 00, 01, 10, and 11, respectively. These haplotype phases correspond to genotypes at fully genotyped markers, but their advantage as compared to genotypes is that they are comparable between loci, so that they can be used to infer genotypes at pseudomarkers. For example, at another locus alleles b and c may originate from population 0, so that genotype bc is assigned haplotype phase 00.

Haplotype phases of flanking markers are generally used to infer haplotype phases of pseudomarkers in QTL mapping. Considering the phase of just the maternal allele, if both flanking markers are in phase 0, the pseudomarker is more likely to be in phase 0 than in phase 1. More precisely, if φ is the (unknown) phase of the pseudomarker, the probability that the pseudomarker is in phase 0 is:

Eq. S1
$$p(\varphi = 0) = \frac{p(\varphi = 0|\varphi_l)p(\varphi_r|\varphi = 0)}{p(\varphi_l|\varphi_r)}$$

where the phases of the left and right flanking markers are indicated with subscripts *l* and *r*. The alternative (φ=1) has probability 1-p(φ =0). The conditional probabilities in Eq. S1 depend on the distances between the pseudomarker and its flanking markers. Let us express the distance between the pseudomarker and the left flanking marker as a recombination fraction, d_l. Then, the probability of the pseudomarker phases is given by Haldane's map function:

Eq. S2
$$p(\varphi|\varphi_l) = \begin{cases} 0.5\exp(-2d_l) & \varphi \neq \varphi_l \\ 1 - 0.5\exp(-2d_l) & \varphi = \varphi_l \end{cases}$$

In the case a flanking marker is lacking (for example if we consider the first or last marker on a chromosome) the distance *d* on that side is infinite so that $p(\varphi|\varphi_l) = 0.5$ and the probability of the phase of the pseudomarker is influenced only by the remaining flanking marker. Thus, we can infer haplotype phases (and hence transmission ratios) using flanking markers.

The above method to infer pseudomarker phases is widely used, but requires modification for the present purpose, where we consider an experimental cross between natural, outcrossing populations. Consider for example a locus where allele a originates from population 0 and allele b from population 1, and where both F_1 parents are have genotype ab, so that the possible F_2 genotypes are aa, ab, and bb. Genotypes aa and ab represent phases 00 and 11, respectively, while genotype ab is either 01 or 10. Hence, an individual with genotype ab at this locus provides no phase information (both alleles are equally likely to stem from both populations) even though it clearly provides information about transmission ratios (as it is neither 00 nor 11). In order to employ the information about transmission ratios provided by partly informative markers such as the example above above, we must extend Haldane's mapping function to incorporate both maternal and paternal alleles simultaneously, and to more than two flanking markers.

The extension of the mapping function to incorporate both alleles is rather straightforward. Let r denote the recombination rate on a very short distance, for example r = 0.01 per centimorgan (cM). Then, the probability that two flanking markers one cM apart are in phases 00 and 10 would equal r. These markers are in phases 00 and 11 only if recombination occurs twice, that is with probability r^2. We can conveniently write all the possible transitions between the 4 phases in the 4x4 transition matrix Q:

$$Q = \begin{matrix} -2r-r^2 & r & r & r^2 \\ r & -2r-r^2 & r^2 & r \\ r & r^2 & -2r-r^2 & r \\ r^2 & r & r & -2r-r^2 \end{matrix}$$

where rows and columns refer to 00, 01, 10, and 11, respectively. Entries on the off-diagonal are chosen so that rows sum to 0. We can also write the probability of the phases as a matrix, corresponding to the row and columns order of Q (i.e. 00, 01, 10, and 11). For example, at a genotyped marker the phase may be 00, which can be written:

$$P_A = \begin{matrix} 1 \\ 0 \\ 0 \\ 0 \end{matrix}$$

The probabilities of the phases at a flanking marker B at distance d_{AB} are then given by the Chapman-Kolmogorov equation:

Eq. S3 $$P_{B|A} = P_A \exp(d_{AB} Q)$$

where the exponent is a matrix exponent. For example, for P_A above and d_{AB}=5cM, $P_{B|A}$=[0.9066; 0.0453; 0.0453; 0.0027]. Like Haldane's mapping function, equation S3 accounts for multiple recombination events, assuming that these are independent, random events. In other words, equation S3 is Haldane's mapping function applied to both maternal and paternal alleles simultaneously, and written in matrix representation for mathematical convenience.

To employ the information provided by partly informative markers we must also extend phase inference to multiple flanking loci. Consider for example a pseudomarker flanked by a partly informative marker, which in turn is flanked at close distance by a fully informative marker. The partly informative marker itself provides little information to infer the phase of a pseudomarker next to it, while (due to close linkage) we could be rather certain that it is in the same phase as the fully informative marker. To make use of all the information provided by the genotypes, we must therefore use all fully and partially informative markers to infer phase probabilities at any (pseudo)marker locus.

Just as expressed in equation S1 for a single flanking marker, the phase probabilities of a (pseudo)marker are determined by two components: all the markers to the left, and all the markers to the right. Let us denote the phase probabilities of the *i*-th marker given the markers to the left as $P_{i|l}$, and the phase probabilities given the markers to the right as $P_{i|r}$. We calculate the phase probabilities as:

Eq.S4 $$P_i = \frac{P_{i|l}P_{i|r}}{\sum P_{i|l}P_{i|r}}$$

where the product is element-wise, and the summation over phases. The divisor, just like in equation S1, assures that P sums to unity.

In order to obtain $P_{i|l}$ we calculate sequentially, starting from the leftmost marker on the chromosome and proceeding to the right (using equation A3) $P_{i|l} = P_{i-1} \exp(d_{i(i-1)}Q)$. (For the first marker on the chromosome P_{i-1} =[0.25 0.25 0.25 0.25], reflecting that all phases are equally likely a priori.) If the *i*-th marker is (partly) informative, we can set some elements of $P_{i|l}$ to zero, and re-scale the remaining probabilities so that $P_{i|l}$ sums to unity. Thus, $P_{i|l}$ can be regarded as the phase probabilities of the *i*-th marker if these were determined only by the markers to the left of it. Starting from the rightmost marker, we can similarly calculate $P_{i|r}$ as the phase probabilities of the *i*-th marker if these were only determined by the markers to the right. Finally, we use equation S5 to calculate the phase probabilities P for every marker.

Likelihood maximization As explained above, (pseudo)marker phases can be analyzed for TRD as genotypes. At fully informative markers, phases are known with certainty, but at partly informative markers and pseudomarkers phases can only be assigned probabilities. This has implications for calculating the likelihood of genotype frequencies (Eq. S1): At partially informative loci the likelihoods L_f under different hypotheses depend on the assignment of phases. Thus, we should still maximize the likelihood, but the likelihood will now consist of two components: L_f, and a component representing the likelihoods of the phase assignments. Let $L_{\varphi,j}$ denote the log-likelihood of the phases of the *j*-th F_2 individual. If the phase is known with certainty (e.g. at a fully informative marker) this likelihood will be $L_{\varphi,j}$ =log(1)=0. If the phase is not certain, but for example P=[0.9066; 0.0453; 0.0453; 0.0027] then $L_{\varphi,j}$ = log 0.0453 if the individual is assigned phase 01. Maximizing the likelihood now involves choosing the phases for the *n* F_2 individuals in such a way that it maximizes the likelihood:

Eq. S5 $$L = L_f + \sum_{j=1}^{n} L_\varphi$$

It should be noted that the number of unknown phases is a property of the data that is independent of the hypothesis that is being evaluated. Therefore unknown phases do not affect the difference in the numbers of estimated parameters between hypotheses, i.e. the number of degrees of freedom of the χ^2-distribution used to compare likelihoods.

Maximizing the likelihood is not an easy task (except at fully informative markers). If all individuals are assigned the phase that is most likely, L_φ is maximized, but this may lead to genotype frequencies that render L_f sub-optimal. In an F_2 of n individuals with k possible phases, there are k^n possible phase assignments across individuals. It is clear that for realistic n, the number of assignments is truly large, and exhaustive search for the assignment that maximizes L is prohibitive. Therefore, we use an iterative algorithm to attempt to find the phase assignment that maximizes the likelihood.

1. The algorithm used to maximize the likelihood starts by assigning every individual a plausible phase. For every individual, the initial probability of the i-th phase was calculated as p_i*P, where p_i is the expected frequency of the i-th phase (Eq. 1) and P the probability of this phase according to the flanking markers (Eq. S4). The individual was then assigned the most probable phase. (That is the phase suggested by the flanking markers (i.e. suggested by P), except when that genotype is not expected to be observed (i.e. $p_i = 0$) based on the TRD hypothesis being evaluated.) The likelihoods (L_f and L_φ) of the initial assignment are then calculated using equation S5.

2. For every individual, and for every possible alternative phase, it is calculated how the likelihood (both L_f and L_φ) would change. For example, if an individual is currently assigned phase 10, there are three alternatives, 00, 01, and 11, each of which may have a different effect on L_f as well as L_φ.

3. The single individual and alternative phase is selected that results in the greatest increase in likelihood L.

Steps 2 and 3 are repeated until no further improvement of the likelihood can be achieved. It should perhaps be emphasized that in every iteration, only one individual is assigned a different haplotype in step 3.