The genetic architecture of recurrent segregation distortion in *Arabidopsis thaliana*

Danelle K. Seymour[1], Eunyoung Chae, Burak I. Ariöz, Daniel Koenig[2], Detlef Weigel

Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[1]Current address: Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA

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

21 **Short title**

22 Segregation distortion in *A. thaliana*

23

24 **Key words**

25 Segregation distortion, *Arabidopsis thaliana*, genetic incompatibility, allele frequency

26 distortion

27

28 **Corresponding author**

29 Detlef Weigel

30 Max Planck Institute for Developmental Biology

31 Spemannstrasse 37-39

32 D-72076 Tübingen

33 Germany

34 +49-(0)7071-601 1411

35 email: weigel@weigelworld.org

36

Abstract

The equal probability of transmission of alleles from either parent during sexual reproduction is a central tenet of genetics and evolutionary biology. Yet, there are many cases where this rule is violated. Such violations limit intraspecific gene flow and can facilitate the formation of genetic barriers, a first step in speciation. Biased transmission of alleles, or segregation distortion, can result from a number of biological processes including epistatic interactions between incompatible loci, gametic selection, and meiotic drive. Examples of these phenomena have been identified in many species, implying that they are universal, but comprehensive species-wide studies of segregation distortion are lacking. We have performed a species-wide screen for distorted allele frequencies in over 500 segregating populations of *Arabidopsis thaliana* using reduced-representation genome sequencing. Biased transmission of alleles was evident in up to a quarter of surveyed populations. Most populations exhibited distortion at only one genomic region, with some regions being repeatedly affected in multiple populations. Our results begin to elucidate the species-level architecture of biased transmission of genetic material in *A. thaliana*, and serve as a springboard for future studies into the basis of intraspecific genetic barriers.

Introduction

At the genetic level, evolution is the change in the frequency of allelic variants over time. While in many cases the strength of selection is too low for these changes to be detected within a few generations, a unique opportunity to directly study such changes

is offered in cases where selection coefficients are high. In such a situation, competition between alleles can be seen already in the distribution of heterozygous progeny (a/A). It is manifested as a deviation from the 1:2:1 Mendelian ratio of diploid genotypes (a/a, a/A, A/A), termed allelic or segregation distortion. Deviation from this ratio has important implications for population dynamics. Favoring inheritance of an allele from one grandparent over that from the other grandparent implies that certain genotypic combinations may be unfit (FISHMAN AND SAUNDERS 2008; PHADNIS AND ORR 2009; MCDERMOTT AND NOOR 2010). Depending on the underlying mechanism, complete obstruction to the free flow of genetic information can be an irreversible step on the path towards speciation (reviewed in (PRESGRAVES 2010)).

Segregation distortion, which is quite commonly observed in nature, can be the result of deleterious epistatic interactions between incompatible loci, gametic selection, or meiotic drive (reviewed in (LYTTLE 1991)). Perhaps epistatic interactions of the Bateson-Dobzhansky-Muller type are the best-studied examples (ORR 1996). Alone, the mutations causal for incompatibilities in progeny are innocuous in their native genetic environment. But when combined, their reduced fitness or lethality removes incompatible genotypic combinations from the population. Examples of two-locus incompatibilities have been identified in and between several eukaryotic species (reviewed in (ORR AND PRESGRAVES 2000; BOMBLIES AND WEIGEL 2007; RIESEBERG AND WILLIS 2007)) and causal loci are frequently associated with fast molecular evolution (reviewed in (ORR AND PRESGRAVES 2000; BOMBLIES AND WEIGEL 2007)). Likely, strong epistatic incompatibilities are a common topic in the literature not only due to their role in

speciation, but also because they are easy to detect.

Far fewer examples of meiotic drive and gametic selection have been characterized. Meiotic drive refers to the preferential inheritance of one chromosome during meiosis and is most easily discovered during female gametogenesis (SANDLER *et al.* 1959), as only one of the four meiotic products will become the egg nucleus. This creates the opportunity for "selfish" loci to position themselves favorably so that they their transmission is favored in the next generation. Some known examples of female drive involve changes in either centromeric or other heterochromatic regions (MALIK AND HENIKOFF 2002; FISHMAN AND SAUNDERS 2008), possibly favoring transmission of the drive chromosome by increasing its affinity for the meiotic machinery (STURTEVANT AND DOBZHANSKY 1936; RHOADES 1942; SANDLER *et al.* 1959; HARTL *et al.* 1967; RHOADES *et al.* 1967; DUNN AND BENNETT 1968; ZIMMERING *et al.* 1970; FISHMAN AND SAUNDERS 2008). Many known drive loci are located on sex chromosomes (especially in various *Drosophila* species) and are associated with inversions or other cytological changes (STURTEVANT AND DOBZHANSKY 1936; ZIMMERING *et al.* 1970; FISHMAN AND SAUNDERS 2008). Drive loci on sex chromosomes are more readily identified because they alter the sex ratio, which is easily noticed without molecular biology assays.

Transmission biases arising after formation of the haploid gametes are classified as instances of gametic selection. Due to the differences of male and female gametogenesis, gametic selection can be more easily detected in males. Sperm is produced from all four meiotic products, and each of these haploid sperm cells can compete for the ability to fertilize the ovule. A classic example of gametic selection

involves growth of the pollen tube that delivers the male gametes of plants (SNOW *et al.* 2000). For example, differential pollen tube growth can improve the reproductive success of the genotype that elongates more quickly.

A few instances of segregation distortion are well understood, but knowledge of the species-wide prevalence of the phenomenon is mostly missing. Despite the apparent ubiquity of segregation distortion, it is unclear how often epistatic incompatibilities, gametic selection, or meiotic drive are the cause. In *A. thaliana*, segregation distortion due to partially or fully recessively acting alleles has been observed repeatedly in different experimental population designs (LISTER AND DEAN 1993; MITCHELL-OLDS 1995; ALONSO-BLANCO *et al.* 1998; LOUDET *et al.* 2002; WERNER *et al.* 2005; SIMON *et al.* 2008; TÖRJÉK *et al.* 2008; BALASUBRAMANIAN *et al.* 2009; SALOMÉ *et al.* 2012). The largest published study to date in *A. thaliana* examined segregation distortion in 17 F_2 populations, over half of which exhibited evidence of distortion (SALOMÉ *et al.* 2012). Although *A. thaliana* is typically a self-fertilizing species, outcrossing in nature can be quite common, implying that opportunities for unequal transmission shaping genetic diversity exist (BOMBLIES *et al.* 2010). On the other hand, the preference for inbreeding creates a system sensitized for detection of intraspecific distortion, since accessions collected from nature are typically homozygous throughout the genome. Cross-fertilization between accessions removes an allele from its native, homozygous context, thus creating an opportunity for biased transmission, which in turn makes *A. thaliana* an ideal system for the identification of preferentially inherited loci.

We have surveyed over 500 segregating F_2 populations for segregation distortion

in order to characterize the contribution of biased transmission to the generation of intraspecific genetic barriers. Segregating F_2 populations were derived from intercrossing 80 distinct, resequenced *A. thaliana* accessions spanning the Eurasian range of the species (CAO *et al.* 2011). For this large survey, populations were genotyped in pools using reduced-representation high-throughput sequencing to estimate allelic ratios. In addition to documenting the prevalence of segregation distortion in *A. thaliana*, we have also begun to dissect the population-wide genetic architecture of segregation distortion. The crosses and genomic regions we have characterized provide a platform with which to dissect the relative contribution of deleterious epistatic interactions, male gametic selection, and female drive meiotic to biased inheritance.

Materials and Methods

Germplasm. The F_2 populations were generated by intercrossing 80 natural *Arabidopsis thaliana* accessions with whole-genome resequencing information (CAO *et al.* 2011). Intercrossing was facilitated by induced male sterility which was achieved by artificial miRNA (amiR) mediated knock-down of the floral homeotic gene APETALA3 (AP3) (CHAE *et al.* 2014). One half of F_1 plants were transgene-free and able to produce F_2 progeny through self-fertilization, as each original female grandparent was hemizygous for the amiR transgene. In total, 583 F_2 populations were generated using 67 of the 80 natural accessions as the female grandparent. All 80 accessions were used as the male grandparent and, on average, each grandparent contributed to 14.7 F_2

145 populations. Germplasm information can be found in Table 1 and grandparental seed
146 availability is listed in Table S1.

147 **Growth conditions.** At least 300 individuals from each F_2 population were sown
148 onto 0.5x MS medium (0.7% agar; pH 5.6). Prior to plating, seeds were gas sterilized for
149 16 hours using 40 ml of household bleach (1-4%) and 1.5 ml of concentrated HCl.
150 Seeds were stratified at 4°C in the dark for 8 days and then plates were shifted to 23°C
151 long day conditions (16 h light:8 h dark). After 5 days, seedlings were harvested in bulk
152 and flash frozen in liquid nitrogen.

153 **DNA extraction and GBS library preparation.** DNA was extracted from each
154 pool of F_2 individuals using a CTAB procedure (2% CTAB, 1.4 M NaCl, 100 mM Tris (pH
155 8), 20 mM EDTA (pH 8)) (SPRINGER 2010). DNA integrity was confirmed by gel
156 electrophoresis, and DNA quantification was performed using the Qubit fluorimeter
157 (Qubit BR assay) (Thermo Fisher Scientific, Waltham, MA). For library preparation, 300
158 ng of each DNA sample were diluted in 27 μl. Restriction enzyme-mediated reduced-
159 representation libraries were generated using KpnI, which is predicted to cleave the *A.*
160 *thaliana* reference genome into 8,366 fragments. The library preparation protocol is
161 detailed in (ROWAN *et al.* 2017). Briefly, DNA was digested and then ligated to barcoded
162 adapter sequences with sticky ends complementary to the KpnI cleavage site. After
163 ligation, 96 barcoded samples were pooled and then sheared using the Covaris S220
164 instrument (Covaris, Woburn, MA). Next, end-repair, dA-tailing, a second universal
165 adapter ligation, and PCR enrichment were performed using the Illumina compatible
166 NEBNext DNA Library Prep Master Mix Set (NEB, Ipswich, MA). Library quality was

167 determined using the Agilent 2100 Bioanalyzer (DNA 1000 kit) (Agilent, Santa Clara,
168 CA) and libraries were normalized (10 nM) based on library quantification (ng/μl) and
169 mean fragment length. Sequencing was performed on the Illumina HiSeq 2000 (Illumina,
170 San Diego, CA). Adapter sequences can be found in (ROWAN *et al.* 2017).

171 **SNP identification and allele frequency estimation.** SHORE software (v0.9.0)
172 (OSSOWSKI *et al.* 2008) was used for all analyses described in this section. Sequencing
173 reads were barcode sorted and quality filtered. During quality filtering the restriction
174 enzyme overhang was also trimmed using SHORE import. Reads for each bulked
175 population were then aligned to the TAIR10 reference genome allowing for two
176 mismatches using SHORE mapflowcell. After alignment, SNPs were called with SHORE
177 qVar using default parameters. Read counts for both the reference and non-reference
178 base were extracted for each polymorphic position. SNPs were filtered further using the
179 grandparental whole-genome information and read counts for the female grandparental
180 allele were output only for positions expected to be segregating between the two initial
181 grandparents based on the resequencing data (CAO *et al.* 2011). The allele frequency of
182 the female grandparental allele was calculated for each polymorphic position as the
183 number of reads containing the female grandparental allele divided by the total number
184 of reads covering that position.

185 **Modeling of allele frequency and significance testing for allelic distortion.**
186 High read coverage was sought for each library to enable accurate allele frequency
187 estimation. The realized median coverage of the population bulks was 78x. The
188 distribution of read coverage per library is shown in Fig S1A.

Even with high read coverage, allele frequency estimates were still noisy. To generate accurate allele frequency estimates, the allele frequency was modeled in 5 Mb sliding windows (0.5 Mb steps). We used a beta-binomial model to account for variation in the true allele frequency as well as stochastic variation that arises from read sampling. From the optimized model we extracted the alpha and beta parameters from each genomic window. These parameters describe the shape of the probability distribution in each window, and from these parameters the mean allele frequency as well as the 95% confidence intervals were estimated. Using these estimates, a non-parametric statistical test was performed to assess whether the allele frequency estimates were significantly different from 50%, the expected frequency for non-distorted genomic regions. A false discovery correction (FDR) was performed to account for the number of genomic windows tested per population (n = 240). After allele frequency estimation, quality control measures culled low quality bulks. Populations were excluded from subsequent analysis for the following reasons: 1) having a genome-wide average allele frequency greater than 0.75, 2) exhibiting either confidence intervals (CI) larger than 0.40 or noisy confidence intervals across the genome (standard deviation of CI width greater than 0.15), or 3) displaying three or more chromosomes with windows that did not attain model convergence. After quality control, 492 populations remained for subsequent analyses.

Identification of distorted regions. Two thresholds were used to identify significantly distorted genomic windows. The first approach utilized p-value estimates from the non-parametric statistical test performed on each window. False discovery rate

(FDR) corrections were applied to account for the number of tested genomic windows (n = 240, $p < 0.05$). Distorted populations were required to have at least five adjacent genomic windows on the biased chromosome with significant FDR corrected p-values. Populations with statistically significant segregation distortion are listed in Table 1.

The second, less conservative approach identified outliers by calculating Z-scores for each genomic window relative to the mean allele frequency of all surveyed F_2 populations (0.5029). Allele frequencies for each window were derived from the beta-binomial model predictions. Genomic windows with allele frequency estimates greater than 2.5 times the population-wide standard deviation (0.0382) were considered to be distorted. A distorted F_2 population was required to contain five genomic windows with significant Z-scores on the chromosomes containing the locus of interest. Distorted populations identified using extreme Z-scores are listed in Table 1.

Interval identification using whole-genome resequencing. Six F_2 populations displayed severe distortion at one of six distinct genomic regions (Fig S4). 1,500 individuals were sown from each of these six populations onto 0.5x MS medium (0.7% agar; pH 5.6) as described for the initial screen. DNA was extracted from each population bulk using a standard CTAB preparation (2% CTAB, 1.4 M NaCl, 100 mM Tris (pH 8), 20 mM EDTA (pH 8)). Illumina TruSeq libraries were prepared according to manufacturer's guidelines using 1 μg of starting material per population. Libraries were sequenced on an Illumina HiSeq 3000 instrument (Illumina, San Diego, CA). Twenty-one nucleotide long k-mers were identified directly from the short reads using jellyfish (v2.2.3) (MARCAIS AND KINGSFORD 2011) with the following arguments: -m 21 -s 300M -t

10 -C. Not only does jellyfish identify all unique k-mers, but it also calculates the occurrence, or coverage, of each k-mer. The distribution of 21-mer coverages is shown in Figure S3 for each population. 21-mers found in only one of the two grandparental genomes (coverage < 25X) were aligned to the TAIR10 genome using bwa aln (LI AND DURBIN 2009). Only perfect matches were allowed. A 1 Mb sliding window (50 kb steps) was used to plot the 21-mer coverage across the distorted chromosome in each population. Regions of the genome with reduced coverage of 21-mers are located within the candidate interval (Fig 6B, S4). Interval boundaries were delineated by merging all windows with values within 1x coverage of the minimal window in the candidate region.

Interval identification for distortion bulked segregant analysis. Bulked segregant analysis (MICHELMORE *et al.* 1991) was used to narrow the candidate intervals for Star-8, ICE49, and ICE63. Sequencing reads from the original screen were combined for all distorted populations sharing the grandparent of interest, resulting in a distorted bulk. Those that shared the grandparent, but did not exhibit distortion, were combined separately, resulting in a normal bulk. Positions segregating between the grandparent of interest and all other members of the bulk were identified. The positions segregating in the distorted bulk are not shared with those segregating in the normal bulk. By combining reads from multiple populations, a median of 806 to 1135x coverage was achieved at each segregating position. Candidate intervals were calculated from the maximally distorted position to any flanking segregating site that was within 5% of the peak allele frequency (Table 2).

Material and data availability: Seeds for grandparental lines are available from the Arabidopsis Biological Resource Center (ABRC) or the European Arabidopsis Stock Center (NASC); stock identifiers are listed in Table S1. The source code to generate allele frequency estimates and the raw allele frequencies for each F_2 population are located in the following github repository: https://github.com/dkseym/F2_Segregation_Distortion.

Results

Frequent segregation distortion in intraspecific *A. thaliana* F_2 populations. The incidence of segregation distortion, a molecular signature of genetic conflict, was surveyed in 583 F_2 populations generated from naturally inbred accessions that represent much of the Eurasian genetic diversity in *A. thaliana* (CAO *et al.* 2011). The studied F_2 populations were derived from crosses between 67 accessions used as female and male grandparents, and a further 13 that were used only as male grandparents (CAO *et al.* 2011). The number of crosses performed per accession ranged from 3 to 34, with a median of 14 F_2 populations generated from each grandparent.

F_2 seeds were sown on plates, stratified at 4°C to break dormancy, and then grown for five days in 23°C long days. At least 300 individuals per F_2 population were harvested in bulk for genome-wide genotyping-by-sequencing (GBS), implemented as restriction enzyme-mediated reduced-representation sequencing. Based on previous reduced-representation approaches (BAIRD *et al.* 2008; MONSON-MILLER *et al.* 2012), a

custom protocol was developed to adapt this method to the specific requirements of our system. Accurate allele frequency estimate in bulks requires high sequencing coverage at each segregating site. The selected restriction enzyme, KpnI, cuts infrequently in the *A. thaliana* genome, allowing high coverage to be achieved for a portion of the genome, about 1%, with moderate sequencing effort. Whole-genome SNP information was available for the inbred grandparents (CAO *et al.* 2011), facilitating identification of informative sites. We attained an average of 78x coverage per F_2 population (Fig S1A), and an average of 2,509 sites were segregating in any given population (Fig S1B).

Regions displaying significant segregation distortion, as indicated by deviation from the expected 1:1 ratio of grandparental alleles, were identified by modeling the allele frequency in 5 Mb sliding windows, with 0.5 Mb steps. Using the beta-binomial model estimates of allele frequencies together with the confidence intervals of the estimates, a non-parametric statistical test was performed in each window. In total, 62 populations exhibited regions of significant segregation distortion after false discovery rate (FDR) correction for the number of tested windows ($n = 240$, $p < 0.05$). When considering only the 492 populations passing quality control measures, 62 (12.6%) of these were found to harbor genomic regions with significant distortion (Fig S2). This is a rather conservative estimate of the incidence of segregation distortion in our crosses, because the ability to detect significant distortion is highly dependent on the size of the confidence interval estimates (i.e., the coverage of each population).

To generate a less conservative estimate of the number of distorted regions, we also used a Z-score outlier approach. Any region with allele frequencies greater than 2.5

297 standard deviations from the combined population mean was considered to be distorted.
298 This less conservative approach identified 122 (24.8%) of the 492 populations with at
299 least a single distorted region (Fig 1). All regions identified via the FDR method were
300 also detected using the Z-score outlier approach.

301 An example of a chromosome with a distorted region that was identified using
302 both methods is shown in Figure 2. Although we did not screen the complete diallel of
303 possible F_2 combinations, we did survey populations that sampled a large fraction of the
304 genetic space covered by the 80 founders (Fig 1, Fig S2). That segregation distortion is
305 evident in up to 24% of surveyed F_2 populations suggests that intraspecific genetic
306 barriers are much more common than previously anticipated.

307 **The dynamics of segregation distortion in *A. thaliana*.** The genetics of
308 segregation distortion is dictated by the biological process driving the observed non-
309 Mendelian inheritance. To understand the relative contribution of different processes
310 such as genetic incompatibility, meiotic drive, and gametic selection, we determined
311 how many genomic regions showed segregation distortion in our data set.

312 Regardless of identification method – FDR or Z-score outlier –, the majority of
313 populations exhibited distortion at only a single locus (Fig 3A). If classical Bateson-
314 Dobzhansky-Muller genetic incompatibilities were driving segregation distortion in our
315 populations, we would expect two distorted regions per population, unless the
316 responsible loci were linked. We also found that distortion occurs on all five
317 chromosomes, although distorted regions are most frequently located on chromosome 1
318 (Fig 3B).

The alleles in distorted regions that are favored to be inherited are derived from many grandparental accessions. Of the 80 accessions used as founders, over 50 gave rise to F_2 populations exhibiting significant segregation distortion. Some grandparents were especially notable, such as Star-8. Regions with alleles contributed by Star-8 were distorted in 60% of F_2 populations (40% for the FDR threshold) (Fig 4A,B).

If genetic barriers are primarily caused by genetic drift as individuals diverge from a common ancestor, we would expect more distantly related accessions to give rise to distortion more frequently (LEPPALA *et al.* 2013). As the grandparental accessions had been sampled from eight geographic regions representative of Eurasian genetic diversity (CAO *et al.* 2011), we were able to test if genetic diversity between the two F_2 grandparents was correlated with the probability of segregation distortion. We found no significant difference between the genetic distances of grandparents of distorted populations compared to grandparents of non-distorted populations (Wilcoxon rank-sum test, 1% significance threshold; p=0.03 [Z-score outlier distortion list], p = 0.11 [FDR list]) (Fig 5A,B). That genetic diversity is not a strong predictor of segregation distortion suggests that genetic drift, which becomes more notable after longer periods of separation, is not necessarily the most important driver of intraspecific genetic barriers in *A. thaliana*.

Refining candidate intervals surrounding distorted loci. To begin to understand which processes are responsible for the observed segregation distortion, we sought to define the minimal size of distorted genomic intervals. Genotyping F_2 individuals in bulk enabled screening of a large number of test populations, but without

341 genotype information from individual segregants to estimate recombination breakpoints,
342 most candidate regions are not much smaller than entire chromosome arms.

343 Since we did not know a priori which populations would be the most informative
344 to study in detail, we designed two strategies to narrow the candidate regions to
345 facilitate subsequent fine-mapping. First, we increased the density of informative
346 markers about 200 fold by whole-genome resequencing of six populations with severe
347 segregation distortion. We also increased the number of recombination events in these
348 populations by analysis of 1,500 F_2 individuals from each of the six populations. We
349 sequenced these bulks to approximately 40x coverage. Although this coverage was
350 lower than the average 78x coverage we had used in our GBS analyses, by integrating
351 over multiple markers, together with the larger number of F_2 individuals and thus
352 recombination events, we expected this to substantially improve our power to delineate
353 distorted regions.

354 Unfortunately, exploratory analyses indicated that the lower coverage at
355 individual markers is accompanied by increased stochasticity in allele frequency
356 estimates. We therefore took advantage of local linkage disequilibrium to diminish that
357 noise. Short stretches of unique 21 nucleotide (nt) sequences (known as k-mers or 21-
358 mers) were identified in the raw sequencing reads of each F_2 population. Any 21-mer
359 sequence shared between grandparents should occur at the average genome-wide
360 coverage, and when we plotted 21-mer frequencies, we found a major found peak of 21-
361 mer coverage around 40x, the average per-population whole-genome coverage, in all
362 six populations, as expected (Fig 6A, S3). In contrast, 21-mers present in only one of

363 the two parents should have approximately half as much coverage, and a second peak,
364 resulting from a much smaller number of 21-mers, was apparent in all populations as
365 well (Fig 6A, S3).

366 To narrow down candidate intervals, we extracted 21-mers that were predicted to
367 be present in only one of the two grandparents. Regions of the genome that are
368 distorted should display a decrease in coverage of such grandparent-specific 21-mers
369 near the causal locus. We used a sliding window approach (1 Mb windows, 50 kb steps)
370 to calculate the average coverage of such 21-mers. Using this strategy, we were able to
371 narrow the intervals surrounding four of the six candidate loci to less than 5 Mb, and in
372 one case to 1.5 Mb (Table 2, Fig 6B, S4).

373 In a complementary approach, we sought to refine candidate regions by
374 obtaining a more precise estimate of local allele frequency. To this end, we greatly
375 increased sequencing coverage by combining information from cases with shared
376 grandparents and the same distorted regions. As mentioned earlier, some
377 grandparental accessions contributed alleles that were favored in multiple F_2
378 populations. Star-8, ICE63, and ICE49 contributed alleles that were favored in at least
379 40% of crosses of these to other accessions (based on the Z-score outlier method), with
380 the same regions being favored in all distorted populations sharing a particular
381 grandparent. Using a bulked segregant analysis approach (MICHELMORE *et al.* 1991), we
382 generated two pools of reads for each grandparent. One comprised the sequencing
383 reads from all distorted populations and the other contained the sequencing reads from
384 all non-distorted populations. The allele frequency of SNPs was calculated for sites

385 segregating between the focal grandparent and all other accessions in either the
386 distorted pool or the non-distorted pool.

387 A median coverage of at least 806x was achieved at each segregating site, vastly
388 improving the accuracy of our estimates. For one grandparent, Star-8, we narrowed the
389 interval to 2.0 Mb, in the middle of the top arm of chromosome 1, where recombination
390 is high (Table 2, Fig 6C). This strategy was less successful for the other two
391 grandparents, ICE63 and ICE49, likely because of the distortion being less strong in
392 these cases as well as the location of the distorted regions near the centromere or on
393 the distal chromosome arm, both parts of the chromosome where recombination is
394 reduced (Table 2, Fig S5).

395 ## Discussion

396 Despite the ubiquity of non-Mendelian segregation of alleles in natural populations, the
397 genetic and molecular characterization of the responsible loci has been lagging
398 (reviewed in (ZIMMERING *et al.* 1970; LYTTLE 1991; LYON 2003; FISHMAN AND SAUNDERS
399 2008; PHADNIS AND ORR 2009; HAMMOND *et al.* 2012; LARRACUENTE AND PRESGRAVES
400 2012). Such systems are most easily studied, when distortion is severe and differences
401 in phenotypically distinct progeny classes are obvious (reviewed in (ZIMMERING *et al.*
402 1970)). Because sexual dimorphism is common, many of the earliest known cases were
403 discovered because sex-ratio deviated greatly from 1:1 (reviewed in (ZIMMERING *et al.*
404 1970)). The effects of an allele that is preferentially inherited can be neutralized in a
405 population by fixation of the allele or by the evolution of secondary modifiers. Many

406 cases of segregation distortion were discovered in interspecific crosses (CAMERON AND
407 MOAV 1957; MAGUIRE 1963; SIRACUSA *et al.* 1991; TAO *et al.* 2001; FISHMAN AND
408 SAUNDERS 2008; ZANDERS *et al.* 2014), not because the phenomenon is more common
409 in interspecific hybrids, but because the severity of distortion is extreme in the absence
410 of species-specific modifiers, sometimes reaching fixation in only a generation or two
411 (FISHMAN AND SAUNDERS 2008). The same loci responsible for segregation distortion in
412 interspecific crosses may also underlie unexpected intraspecific segregation patterns.
413 However, in intraspecific crosses, allele frequencies are often only perturbed by a few
414 percent (LYTTLE 1991; FISHMAN AND SAUNDERS 2008), and without molecular genotyping
415 techniques, such subtle allelic distortion will go mostly undetected.

416 Exploiting advances in sequencing and genotyping technology, we have been
417 able to characterize segregation distortion in hundreds of intraspecific crosses. The
418 identification of distorted regions greatly depends on sequencing coverage; in our
419 system, a 10% deviation in absolute allele frequency becomes significant with
420 approximately 100x sequence coverage, and more subtly distorted regions could be
421 detected with even higher coverage. Similar pooled genotyping approaches have been
422 used to identify distorted loci in other systems (CUI *et al.* 2015; BELANGER *et al.* 2016a;
423 BELANGER *et al.* 2016b; WEI *et al.* 2017), illustrating the general power of this approach.

424 Although *A. thaliana* is self-compatible, outcrossing is reasonably common, and
425 descendants of recent outcrossing events are easily found in wild stands of this species
426 (BOMBLIES *et al.* 2010). By surveying a broad collection of germplasm for non-Mendelian
427 inheritance, we could confirm that allelic distortion is a common feature of F_2

428 populations, implying that allelic distortion has a major impact on shaping local genetic
429 diversity. Not only do distorted loci segregate in up to a quarter of all F_2 populations, but
430 multiple genomic regions contribute to this phenomenon, with the degree of distortion
431 varying both by population and by locus. Intraspecific distortion loci that have been
432 identified in other systems typically occur at low population frequencies (HICKEY AND
433 CRAIG 1966; PERKINS AND BARRY 1977; HIRAIZUMI AND THOMAS 1984; HAMMER *et al.*
434 1989; MCMULLEN *et al.* 2009; HOU *et al.* 2015; FRAGOSO *et al.* 2017), although there are
435 exceptions, such as the tightly linked *zeel-1* and *peel-1* genes in *C. elegans* (SEIDEL *et*
436 *al.* 2008; BEN-DAVID *et al.* 2017). The low frequency of the causal alleles has been
437 hypothesized to result from antagonistic modifier loci having evolved in response to the
438 fitness costs that are often linked to distortion loci (reviewed in (ZIMMERING *et al.* 1970;
439 LYTTLE 1991)). In an interspecific *Drosophila* cross, the causal locus itself is responsible
440 for both the distortion phenotype and for reduced gamete success (PHADNIS AND ORR
441 2009). We have found multiple cases of genomic regions that are distorted in one or
442 very few population(s), suggesting that frequency of distortion alleles is often low in *A.*
443 *thaliana* as well. This could be because these alleles are older, giving sufficient time for
444 modifiers to evolve and rise to high frequency. If these are linked, we would not have
445 detected them as separate genomic loci, as our mapping resolution was mostly
446 chromosome arm scale.

447 Of particular interest are regions that are repeatedly distorted across many
448 populations at extreme frequencies. For example, the Star-8 region on chromosome 1 is
449 significantly favored in ~50% of crosses, with this region being inherited by up to 70 or

even 80% of the progeny. This could be an example of a young allele for which suppressors have not yet evolved, or it could be that the balance between fitness costs (if any) and the degree of distortion is stable at this frequency. The *D* locus in *Mimulus guttatus* is perhaps the best example of a stable distortion polymorphism, in this case caused by meiotic drive (FISHMAN AND SAUNDERS 2008). The measured degree of distortion at this locus (58:42) is predicted by the associated decrease in pollen viability (FISHMAN AND SAUNDERS 2008). This allele is segregating in about half of all individuals from a natural population (FISHMAN AND SAUNDERS 2008). Other instances of distortion loci segregating at intermediate frequencies are known, but the evolutionary dynamics of these cases are not as well characterized (reviewed in (ZIMMERING *et al.* 1970; LYTTLE 1991))

A peculiarity of allelic distortion in our panel of *A. thaliana* crosses is that in most cases, only a single genomic region is inherited in a non-Mendelian fashion. Classic meiotic drive systems consist of a distorter locus and a responder locus, with the two being almost always linked through an inversion or genetic rearrangement that reduces recombination between them (STALKER 1961; WU AND BECKENBACH 1983; SILVER 1985; LYTTLE 1991). As a result, classic drive loci are inherited as a single distorted genomic region. Our results are reminiscent of such cases, suggesting that several such loci are segregating in *A. thaliana*, although we cannot currently infer the number of genes in the mapping intervals responsible for segregation distortion.

Apart from meiotic drive, more conventional two-locus deleterious interactions conforming to the Bateson-Dobzhansky-Muller model of genetic incompatibilities can

also perturb expected allelic (and genotypic) segregation ratios. A survey in *D. melanogaster* showed intraspecific genetic incompatibilities due to epistatic interaction between two (often unlinked) loci are not uncommon, with natural strains carrying an average of 1.15 incompatible loci (CORBETT-DETIG *et al.* 2013). Hybrid incompatibility is a common feature in both plants in animals, with many known cases of deleterious epistatic interactions between two nuclear loci segregating in *A. thaliana* (BOMBLIES *et al.* 2007; ALCÁZAR *et al.* 2009; BIKARD *et al.* 2009; VLAD *et al.* 2010; DURAND *et al.* 2012; CHAE *et al.* 2014; AGORIO *et al.* 2017; PLÖTNER *et al.* 2017). In our set of crosses, simultaneous distortion at two independent genomic regions was the exception. In our design, incompatible interactions would only be detectable if the F_1 was fertile and dominance relationship between alleles was such that over 10% of the progeny did not give rise to seedlings. In other words, if both genes acted completely recessively and the doubly homozygous progeny failed to grow, they still would not be noticed in our segregation distortion scans. We note that even in cases where two independent genomic regions are significantly distorted in a single population, the absence of genotype data for individuals does not allow us to explicitly examine if these regions genetically interact. Although the nature of our experimental design has not yet revealed the species-wide architecture of partially or fully recessive epistatic interactions segregating in *A. thaliana*, this can be addressed in future studies by genotyping individuals instead of pools.

While a handful of classical segregation distortion loci has been molecularly characterized in detail (reviewed in (ZIMMERING *et al.* 1970; LYTTLE 1991; LYON 2003;

494 LARRACUENTE AND PRESGRAVES 2012)), the molecular nature of most loci is still
495 unknown. As a result, there is still much to be learned about the biological processes
496 and evolutionary forces leading to uneven segregation, including whether such alleles
497 are more likely to be evolutionarily old or young. For example, numerous cases of hybrid
498 incompatibilities in *A. thaliana* are due to interactions between disease resistance
499 genes, which have very divergent alleles, both because of rapid evolution and long-term
500 balancing selection (BOMBLIES *et al.* 2007; ALCÁZAR *et al.* 2009; DURAND *et al.* 2012;
501 CHAE *et al.* 2014). The fast evolution of centromeres and other satellite sequence
502 repeats, a result of intragenomic conflict, has also been shown to cause or to be closely
503 linked to allelic distortion (WU *et al.* 1988; FISHMAN AND SAUNDERS 2008; CHMATAL *et al.*
504 2014; MAHESHWARI *et al.* 2015). In our crosses, distorted regions often localized near
505 centromeres.

506 Whether the conflict arises in interspecific or intraspecific crosses, it appears that
507 natural selection, not genetic drift, is often responsible for the evolution of non-
508 Mendelian inheritance. In support of this, we found little correlation between the degree
509 of genetic differentiation between the grandparental accessions and the probability of
510 observing allelic distortion in their progeny, in line with what has been seen in a much
511 smaller panel of F_2 populations (SALOMÉ *et al.* 2012).

512 To conclude, by surveying a large number of F_2 populations descending from 80
513 genetically diverse grandparents, we were able to identify numerous genomic regions in
514 *A. thaliana* that are not transmitted in a Mendelian fashion. Considering that our
515 statistical power would not have allowed us to discover complete absence of genotypes

516 resulting from higher-order epistatic interactions, it is likely that the regions we identified
517 are only the tip of the iceberg. Notably, the majority of accessions tested contributed
518 such distorted alleles, emphasizing the ubiquity of alleles that are unevenly transmitted.
519 Together, these findings confirm the findings from other systems that genetic barriers
520 segregating within wild species are more common that previously thought (SEIDEL *et al.*
521 2008; CORBETT-DETIG *et al.* 2013; HOU *et al.* 2015).

522 Author contributions

523 D.K.S., D.K., E.C. and D.W. conceived the project. D.K.S., E.C. and B.I.A. generated
524 the material and data. D.K.S. and D.K. analyzed the data. D.K.S. and D.W. wrote the
525 manuscript with contributions from all authors.

526 Acknowledgments

527 This work was supported by ERC AdG IMMUNEMESIS and the Max Planck Society.

528 References

529 Agorio, A., S. Durand, E. Fiume, C. Brousse, I. Gy *et al.*, 2017 An *Arabidopsis* natural
530 epiallele maintained by a feed-forward silencing loop between histone and DNA.
531 PLoS Genet 13**:** e1006551.
532 Alcázar, R., A. V. Garcia, J. E. Parker and M. Reymond, 2009 Incremental steps toward
533 incompatibility revealed by *Arabidopsis* epistatic interactions modulating salicylic
534 acid pathway activation. Proc Natl Acad Sci USA 106**:** 334-339.

535 Alonso-Blanco, C., A. J. Peeters, M. Koornneef, C. Lister, C. Dean *et al.*, 1998
536 Development of an AFLP based linkage map of Ler, Col and Cvi *Arabidopsis*
537 *thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line
538 population. Plant J 14**:** 259-271.
539 Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver *et al.*, 2008 Rapid SNP
540 discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3**:**
541 e3376.
542 Balasubramanian, S., C. Schwartz, A. Singh, N. Warthmann, M. C. Kim *et al.*, 2009 QTL
543 mapping in new *Arabidopsis thaliana* advanced intercross-recombinant inbred
544 lines. PLoS ONE 4**:** e4318.
545 Belanger, S., I. Clermont, P. Esteves and F. Belzile, 2016a Extent and overlap of
546 segregation distortion regions in 12 barley crosses determined via a Pool-GBS
547 approach. Theor Appl Genet 129**:** 1393-1404.
548 Belanger, S., P. Esteves, I. Clermont, M. Jean and F. Belzile, 2016b Genotyping-by-
549 sequencing on pooled samples and its use in measuring segregation bias during
550 the course of androgenesis in barley. Plant Genome 9.
551 Ben-David, E., A. Burga and L. Kruglyak, 2017 A maternal-effect selfish genetic element
552 in *Caenorhabditis elegans*. Science 356**:** 1051-1055.
553 Bikard, D., D. Patel, C. Le Mette, V. Giorgi, C. Camilleri *et al.*, 2009 Divergent evolution
554 of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science
555 323**:** 623-626.

556 Bomblies, K., J. Lempe, P. Epple, N. Warthmann, C. Lanz *et al.*, 2007 Autoimmune
557 response as a mechanism for a Dobzhansky-Muller-type incompatibility
558 syndrome in plants. PLoS Biol 5**:** e236.
559 Bomblies, K., and D. Weigel, 2007 Hybrid necrosis: autoimmunity as a potential gene-
560 flow barrier in plant species. Nat Rev Genet. 8**:** 382-393.
561 Bomblies, K., L. Yant, R. Laitinen, S.-T. Kim, J. D. Hollister *et al.*, 2010 Local-scale
562 patterns of genetic variability, outcrossing and spatial structure in natural stands
563 of *Arabidopsis thaliana*. PLoS Genet 6**:** e1000890.
564 Cameron, D. R., and R. M. Moav, 1957 Inheritance in *Nicotiana tabacum*. XXVII. Pollen
565 killer, an alien genetic locus inducing abortion of microspores not carrying it.
566 Genetics 42**:** 326-335.
567 Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender *et al.*, 2011 Whole-
568 genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43**:**
569 956-963.
570 Chae, E., K. Bomblies, S. T. Kim, D. Karelina, M. Zaidem *et al.*, 2014 Species-wide
571 genetic incompatibility analysis identifies immune genes as hot spots of
572 deleterious epistasis. Cell 159**:** 1341-1351.
573 Chmatal, L., S. I. Gabriel, G. P. Mitsainas, J. Martinez-Vargas, J. Ventura *et al.*, 2014
574 Centromere strength provides the cell biological basis for meiotic drive and
575 karyotype evolution in mice. Curr Biol 24**:** 2295-2300.
576 Corbett-Detig, R. B., J. Zhou, A. G. Clark, D. L. Hartl and J. F. Ayroles, 2013 Genetic
577 incompatibilities are widespread within species. Nature 504**:** 135-137.

578 Cui, Y., F. Zhang, J. Xu, Z. Li and S. Xu, 2015 Mapping quantitative trait loci in selected
579 breeding populations: A segregation distortion approach. Heredity 115**:** 538-546.
580 Dunn, L. C., and D. Bennett, 1968 A new case of transmission ratio distortion in house
581 mouse. Proc Narl Acad Sci U S A 61**:** 570-573.
582 Durand, S., N. Bouché, E. Perez Strand, O. Loudet and C. Camilleri, 2012 Rapid
583 establishment of genetic incompatibility through natural epigenetic variation. Curr
584 Biol 22**:** 326-331.
585 Fishman, L., and A. Saunders, 2008 Centromere-associated female meiotic drive entails
586 male fitness costs in monkeyflowers. Science 322**:** 1559-1562.
587 Fragoso, C. A., M. Moreno, Z. Wang, C. Heffelfinger, L. J. Arbelaez *et al.*, 2017 Genetic
588 architecture of a rice nested association mapping population. G3 7**:** 1913-1926.
589 Hammer, M. F., J. Schimenti and L. M. Silver, 1989 Evolution of mouse chromosome 17
590 and the origin of inversions associated with t haplotypes. Proc Natl Acad Sci U S
591 A 86**:** 3261-3265.
592 Hammond, T. M., D. G. Rehard, H. Xiao and P. K. Shiu, 2012 Molecular dissection of
593 *Neurospora* Spore killer meiotic drive elements. Proc Natl Acad Sci U S A 109**:**
594 12093-12098.
595 Hartl, D. L., Hiraizum.Y and J. F. Crow, 1967 Evidence for sperm dysfunction as
596 mechanism of segregation distortion in *Drosophila melanogaster*. Proc Natl Acad
597 Sci U S A 58**:** 2240-2245.
598 Hickey, W. A., and G. B. Craig, Jr., 1966 Distortion of sex ratio in populations of *Aedes*
599 *aegypti*. Can J Genet Cytol 8**:** 260-278.

Hiraizumi, Y., and A. M. Thomas, 1984 Suppressor systems of Segregation Distorter (SD) chromosomes in natural populations of *Drosophila melanogaster*. Genetics 106**:** 279-292.

Hou, J., A. Friedrich, J. S. Gounot and J. Schacherer, 2015 Comprehensive survey of condition-specific reproductive isolation reveals genetic incompatibility in yeast. Nat Commun 6**:** 7214.

Larracuente, A. M., and D. C. Presgraves, 2012 The selfish *Segregation Distorter* gene complex of *Drosophila melanogaster*. Genetics 192**:** 33-53.

Leppala, J., F. Bokma and O. Savolainen, 2013 Investigating incipient speciation in *Arabidopsis lyrata* from patterns of transmission ratio distortion. Genetics 194**:** 697-708.

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

Lister, C., and C. Dean, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J 4**:** 745-750.

Loudet, O., S. Chaillou, C. Camilleri, D. Bouchez and F. Daniel-Vedele, 2002 Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in *Arabidopsis*. Theor. Appl. Genet. 104**:** 1173-1184.

Lyon, M. F., 2003 Transmission ratio distortion in mice. Annu Rev Genet 37**:** 393-408.

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

Maguire, M. P., 1963 High transmission frequency of a *Tripsacum* chromosome in corn. Genetics 48**:** 1185-1194.

622 Maheshwari, S., E. H. Tan, A. West, F. C. Franklin, L. Comai *et al.*, 2015 Naturally
623 occurring differences in CENH3 affect chromosome segregation in zygotic
624 mitosis of hybrids. PLoS Genet 11**:** e1004970.
625 Malik, H. S., and S. Henikoff, 2002 Conflict begets complexity: the evolution of
626 centromeres. Curr Opin Genet Dev 12**:** 711-718.
627 Marcais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel
628 counting of occurrences of k-mers. Bioinformatics 27**:** 764-770.
629 McDermott, S. R., and M. A. Noor, 2010 The role of meiotic drive in hybrid male sterility.
630 Philos Trans R Soc Lond B Biol Sci 365**:** 1265-1272.
631 McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al.*, 2009 Genetic
632 properties of the maize nested association mapping population. Science 325**:**
633 737-740.
634 Michelmore, R. W., I. Paran and R. V. Kesseli, 1991 Identification of markers linked to
635 disease-resistance genes by bulked segregant analysis: a rapid method to detect
636 markers in specific genomic regions by using segregating populations. Proc Natl
637 Acad Sci U S A 88**:** 9828-9832.
638 Mitchell-Olds, T., 1995 Interval mapping of viability loci causing heterosis in
639 *Arabidopsis*. Genetics 140**:** 1105-1109.
640 Monson-Miller, J., D. C. Sanchez-Mendez, J. Fass, I. M. Henry, T. H. Tai *et al.*, 2012
641 Reference genome-independent assessment of mutation density using restriction
642 enzyme-phased sequencing. BMC Genomics 13**:** 72.

643 Orr, H. A., 1996 Dobzhansky, Bateson, and the genetics of speciation. Genetics 144:
644 1331-1335.
645 Orr, H. A., and D. C. Presgraves, 2000 Speciation by postzygotic isolation: forces,
646 genes and molecules. Bioessays 22: 1085-1094.
647 Ossowski, S., K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann *et al.*, 2008
648 Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome
649 Res. 18: 2024-2033.
650 Perkins, D. D., and E. G. Barry, 1977 The cytogenetics of *Neurospora*. Adv Genet 19:
651 133-285.
652 Phadnis, N., and H. A. Orr, 2009 A single gene causes both male sterility and
653 segregation distortion in *Drosophila* hybrids. Science 323: 376-379.
654 Plötner, B., M. Nurmi, A. Fischer, M. Watanabe, K. Schneeberger *et al.*, 2017 Chlorosis
655 caused by two recessively interacting genes reveals a role of RNA helicase in
656 hybrid breakdown in *Arabidopsis thaliana*. Plant J, doi doi: 10.1111/tpj.13560.
657 Presgraves, D. C., 2010 The molecular evolutionary basis of species formation. Nat Rev
658 Genet 11: 175-180.
659 Rhoades, M. M., 1942 Preferential segregation in maize. Genetics 27: 0395-0407.
660 Rhoades, M. M., E. Dempsey and A. Ghidoni, 1967 Chromosome Elimination in Maize
661 Induced by Supernumerary B Chromosomes. Proc Natl Acad Sci U S A 57:
662 1626-1632.
663 Rieseberg, L. H., and J. H. Willis, 2007 Plant speciation. Science 317: 910-914.

664 Rowan, B. A., D. K. Seymour, E. Chae, D. S. Lundberg and D. Weigel, 2017 Methods
665 for genotyping-by-sequencing. Methods Mol Biol 1492**:** 221-242.
666 Salomé, P. A., K. Bomblies, J. Fitz, R. A. Laitinen, N. Warthmann *et al.*, 2012 The
667 recombination landscape in Arabidopsis thaliana F_2 populations. Heredity 108**:**
668 447-455.
669 Sandler, L., Y. Hiraizumi and I. Sandler, 1959 Meiotic drive in natural populations of
670 *Drosophila melanogaster*. I. the Cytogenetic Basis of Segregation-Distortion.
671 Genetics 44**:** 233-250.
672 Seidel, H. S., M. V. Rockman and L. Kruglyak, 2008 Widespread genetic incompatibility
673 in *C. elegans* maintained by balancing selection. Science 319**:** 589-594.
674 Silver, L. M., 1985 Mouse t haplotypes. Annu Rev Genet 19**:** 179-208.
675 Simon, M., O. Loudet, S. Durand, A. Bérard, D. Brunel *et al.*, 2008 Quantitative trait loci
676 mapping in five new large recombinant inbred line populations of *Arabidopsis*
677 *thaliana* genotyped with consensus single-nucleotide polymorphism markers.
678 Genetics 178**:** 2253-2264.
679 Siracusa, L. D., W. G. Alvord, W. A. Bickmore, N. A. Jenkins and N. G. Copeland, 1991
680 Interspecific backcross mice show sex-specific differences in allelic inheritance.
681 Genetics 128**:** 813-821.
682 Snow, A. A., T. P. Spira and H. Liu, 2000 Effects of sequential pollination on the
683 success of "fast" and "slow" pollen donors in *Hibiscus moscheutos* (Malvaceae).
684 Am J Bot 87**:** 1656-1659.

685 Springer, N. M., 2010 Isolation of plant DNA for PCR and genotyping using organic
686 extraction and CTAB. Cold Spring Harb Protoc 2010**:** pdb prot5515.
687 Stalker, H. D., 1961 The genetic systems modifying meiotic drive in *Drosophila*
688 *paramelanica*. Genetics 46**:** 177-202.
689 Sturtevant, A. H., and T. Dobzhansky, 1936 Geographical distribution and cytology of
690 "sex ratio" in *Drosophila pseudoobscura* and related species. Genetics 21**:** 473-
691 490.
692 Tao, Y., D. L. Hartl and C. C. Laurie, 2001 Sex-ratio segregation distortion associated
693 with reproductive isolation in *Drosophila*. Proc Natl Acad Sci U S A 98**:** 13183-
694 13188.
695 Törjék, O., R. C. Meyer, M. Zehnsdorf, M. Teltow, G. Strompen *et al.*, 2008 Construction
696 and analysis of two reciprocal *Arabidopsis* introgression line populations. J Hered
697 99**:** 396-406.
698 Vlad, D., F. Rappaport, M. Simon and O. Loudet, 2010 Gene transposition causing
699 natural variation for growth in *Arabidopsis thaliana*. PLoS Genet 6**:** e1000945.
700 Wei, K. H., H. M. Reddy, C. Rathnam, J. Lee, D. Lin *et al.*, 2017 A pooled sequencing
701 approach identifies a candidate meiotic driver in *Drosophila*. Genetics 206: 451-
702 465.
703 Werner, J. D., J. O. Borevitz, N. Warthmann, G. T. Trainer, J. R. Ecker *et al.*, 2005
704 Quantitative trait locus mapping and DNA array hybridization identify an *FLM*
705 deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci U S A
706 102**:** 2460-2465.

Wu, C. I., and A. T. Beckenbach, 1983 Evidence for extensive genetic differentiation between the sex-ratio and the standard arrangement of *Drosophila pseudoobscura* and *D. persimilis* and identification of hybrid sterility factors. Genetics 105**:** 71-86.

Wu, C. I., T. W. Lyttle, M. L. Wu and G. F. Lin, 1988 Association between a satellite DNA sequence and the *Responder of Segregation Distorter* in *D. melanogaster*. Cell 54**:** 179-189.

Zanders, S. E., M. T. Eickbush, J. S. Yu, J. W. Kang, K. R. Fowler *et al.*, 2014 Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. Elife 3**:** e02630.

Zimmering, S., L. Sandler and B. Nicolett, 1970 Mechanisms of meiotic drive. Annu Rev Genet 4**:** 409-436.

Tables

Table 1. Germplasm information for surveyed F_2 populations. All crosses are listed, with those passing quality control (QC) indicated with a "1". Similarly, "1" and "0" indicates whether distortion was detected using FDR significance testing of beta-binomial modeling of allele frequencies or Z-score deviation.

Table 2. Candidate intervals for distorted loci. ND, not determined.

Figure legends

Figure 1. Z-score estimated segregation distortion is evident in a wide range of crosses. Genotypic combinations surveyed in this F_2 screen are shown in blue, and populations with significant segregation distortion based on Z-score metrics in green. Grandparental accessions are ordered by the geographic location of their collection (CAO *et al.* 2011). Female grandparents are located on the y-axis and male grandparents on the x-axis.

Figure 2. A representative F_2 population, POP035 (ICE63 x Vash-1), with significant segregation distortion. Distortion in this population was detected with both thresholds (FDR and Z-score outlier). (A) The beta-binomial modeled allele frequency (blue) across each chromosome is plotted in the upper panel. 95% confidence intervals are indicated by the shaded grey area and the expected frequency of 0.5 is marked by the dashed black line. (B) The $-\log_{10}$ of the p-value derived from the non-parametric statistical test. The dashed black line in this panel represents the FDR corrected (n = 240) significance threshold ($p < 0.05$).

Figure 3. Genomic properties of distorted loci. (A) The fraction of surveyed F_2 populations that exhibited segregation distortion at either one or two genomic loci. (B)

750 The number of populations containing distorted loci that reside on each of the five *A.*
751 *thaliana* chromosomes.
752
753 **Figure 4. Many grandparental accessions contributed biased alleles.** Each
754 grandparent contributed its genetic material to a median of 14 distinct F_2 populations.
755 Plotted is the fraction of F_2 populations with one shared grandparent that are
756 significantly distorted as measured either by (A) because of FDR corrected deviation
757 from beta-binomial modeled allele frequencies, or (B) 2.5x Z-score deviation.
758
759 **Figure 5. Genetic distance between grandparental accessions is not predictive of**
760 **biased allelic transmission.** A box plot of genetic distances between the
761 grandparental accessions of normal (grey) and distorted (colored) F_2 populations. At a
762 significance threshold of $p < 0.01$, the genetic distances between grandparents of
763 distorted populations determined from FDR corrected deviation (A) or 2.5x Z-score
764 deviation (B) is not significantly different from that of normal populations (Wilcoxon rank
765 sum test). Genetic distance was calculated as the number of segregating sites over the
766 number of interrogated sites. All positions were required to have complete coverage
767 across all 80 grandparental accessions.
768
769 **Figure 6. Mapping intervals refined using k-mer coverage and bulked segregant**
770 **analysis. (**A) The coverage of unique 21 nt k-mers is plotted for POP035 (ICE63 x
771 Vash-1) after whole-genome resequencing. The first peak in coverage represents 21-

772 mers found in only one of the two grandparents (red arrow), while the second, larger
773 peak represents those sequences found in both (black arrow). (B) The upper panel
774 displays the beta-binomial modeled allele frequency estimates (blue) and their 95%
775 confidence intervals (grey) for POP035 as described in the legend for Figure 2. In the
776 lower panel, the coverage of 21-mers unique to only one of the two grandparents
777 (coverage < 25x) is plotted in 1 Mb sliding windows (50 kb steps). Coverage decreases
778 in the candidate regions. Intervals (grey box) are defined by merging windows with
779 values within 1x coverage of the minimal window in each population. (C) Bulked
780 segregant analysis was performed for Star-8, an accession that repeatedly contributed
781 distorted loci. Sequencing reads were combined for populations exhibiting distortion
782 when crossed with Star-8, and for populations not exhibiting distortion when crossed to
783 Star-8 (normal pool). A candidate interval (grey box) was obtained by merging all
784 segregating positions within 5% of the maximal allele frequency.
785
786
787

Supplemental tables

Table S1. Germplasm identifiers.

Supplemental figure legends

Figure S1. Reduced-representation sequencing reliably enriches for 1% of the *A. thaliana* genome. (A) Mean sequencing coverage at sites segregating in each F_2 population. (B) Number of sites segregating in each F_2 population. The mean observed number of segregating sites (2,500) is comparable to the expected number of segregating sites derived from previously published resequencing data (CAO *et al.* 2011).

Figure S2. Statistically significant segregation distortion is evident in a wide range of crosses. Genotypic combinations surveyed in this F_2 screen are shown in blue, and populations with significant segregation distortion based on non-parametric statistical tests of beta-binomial modeled allele frequencies in green. Grandparental accessions are ordered by the geographic region of their collection (CAO *et al.* 2011). Female grandparents are located on the y-axis and male grandparents on the x-axis.

Figure S3. Distribution of unique 21-mers in whole-genome resequencing data. The coverage of unique 21 nt k-mers is plotted for each of the six populations that underwent whole-genome resequencing. The first peak in coverage represents 21-mers found in only one of the two grandparents, while the second, more prominent peak represents those found in both.

812

813 **Figure S4. 21-mer coverage from whole-genome resequencing can be used to**
814 **refine mapping intervals.** For each population, the upper panel displays the beta-
815 binomial modeled allele frequency estimates (blue) and their 95% confidence intervals
816 (grey) as described in the legend for Figure 2. In the lower panel, the coverage of 21-
817 mers unique to only one of the two grandparents (coverage < 25x) is plotted in 1 Mb
818 sliding windows (50 kb steps). Coverage decreases in the candidate regions. Intervals
819 (grey box) are defined by merging windows with values within 1x coverage of the
820 minimal window in each population. No candidate region was defined for POP064 as
821 coverage decrease coincides with the centromere, not the distorted region.

822

823 **Figure S5. Increasing the number of analyzed segregants can be used to refine**
824 **mapping intervals.** Bulked segregant analysis was performed for grandparental
825 accessions that repeatedly contributed distorted loci (Star-8 [Figure 6C], ICE63 [shown
826 here], and ICE49). Sequencing reads were combined for populations exhibiting
827 distortion or not exhibiting distortion when crossed to the focal grandparent. An average
828 of over 800x coverage was achieved at sites segregating between the focal accessions
829 and all other members in the bulk. A candidate interval (grey box) was obtained by
830 merging all segregating positions within 5% of the maximal allele frequency. Data for
831 ICE49 not shown, as there were too few segregating sites.

832

Figure 1

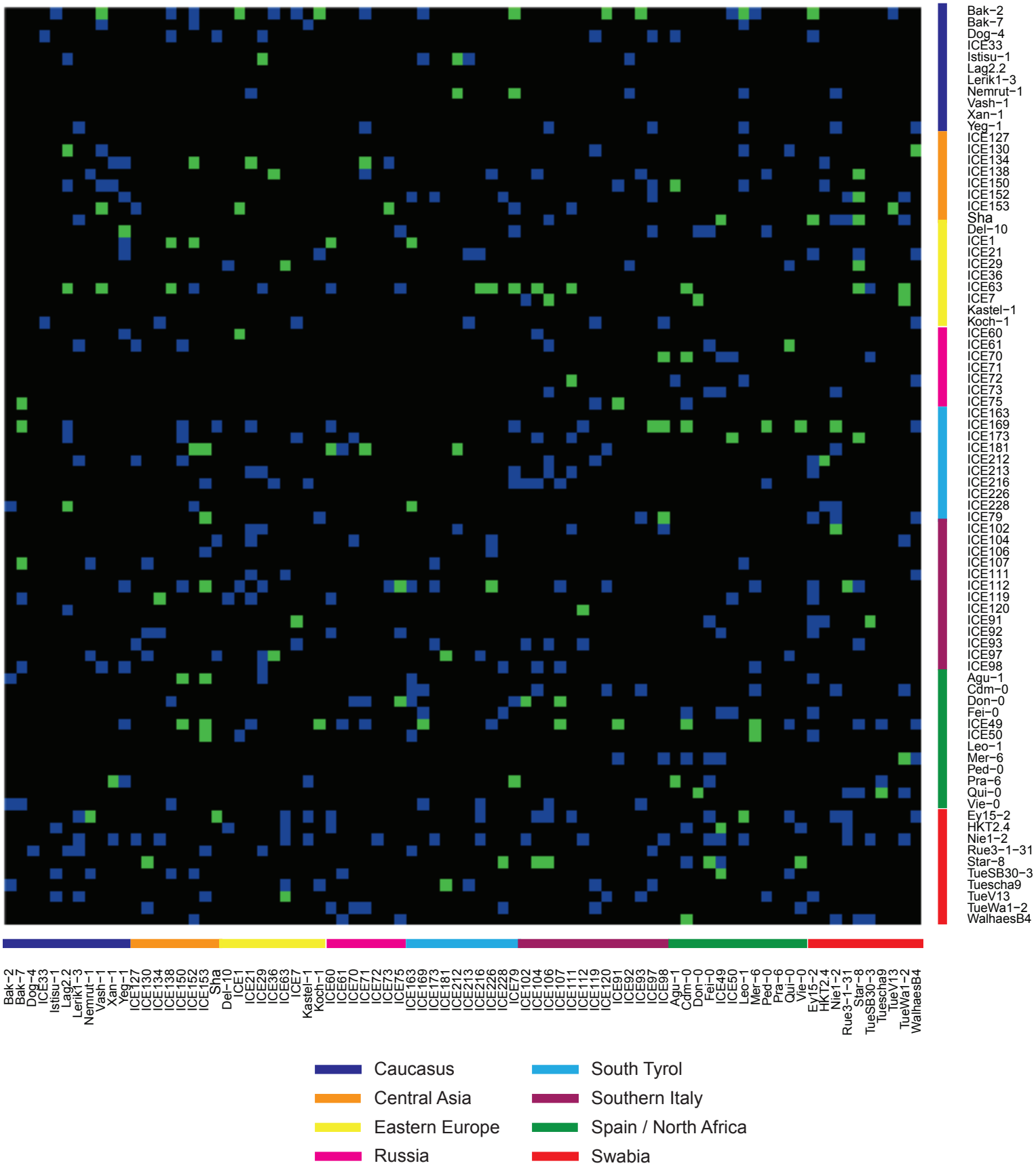

Bak−2
Bak−7
Dog−4
ICE33
Istisu−1
Lag2.2
Lerik1−3
Nemrut−1
Vash−1
Xan−1
Yeg−1
ICE127
ICE130
ICE134
ICE138
ICE150
ICE152
ICE153
Sha
Del−10
ICE1
ICE21
ICE29
ICE36
ICE63
ICE7
Kastel−1
Koch−1
ICE60
ICE61
ICE70
ICE71
ICE72
ICE73
ICE75
ICE163
ICE169
ICE173
ICE181
ICE212
ICE213
ICE216
ICE226
ICE228
ICE79
ICE102
ICE104
ICE106
ICE107
ICE111
ICE112
ICE119
ICE120
ICE91
ICE92
ICE93
ICE97
ICE98
Agu−1
Cdm−0
Don−0
Fei−0
ICE49
ICE50
Leo−1
Mer−6
Ped−0
Pra−6
Qui−0
Vie−0
Ey15−2
HKT2.4
Nie1−2
Rue3−1−31
Star−8
TueSB30−3
Tuescha9
TueV13
TueWa1−2
WalhaesB4
Caucasus
Central Asia
Eastern Europe
Russia
South Tyrol
Southern Italy
Spain / North Africa
Swabia

Figure 2

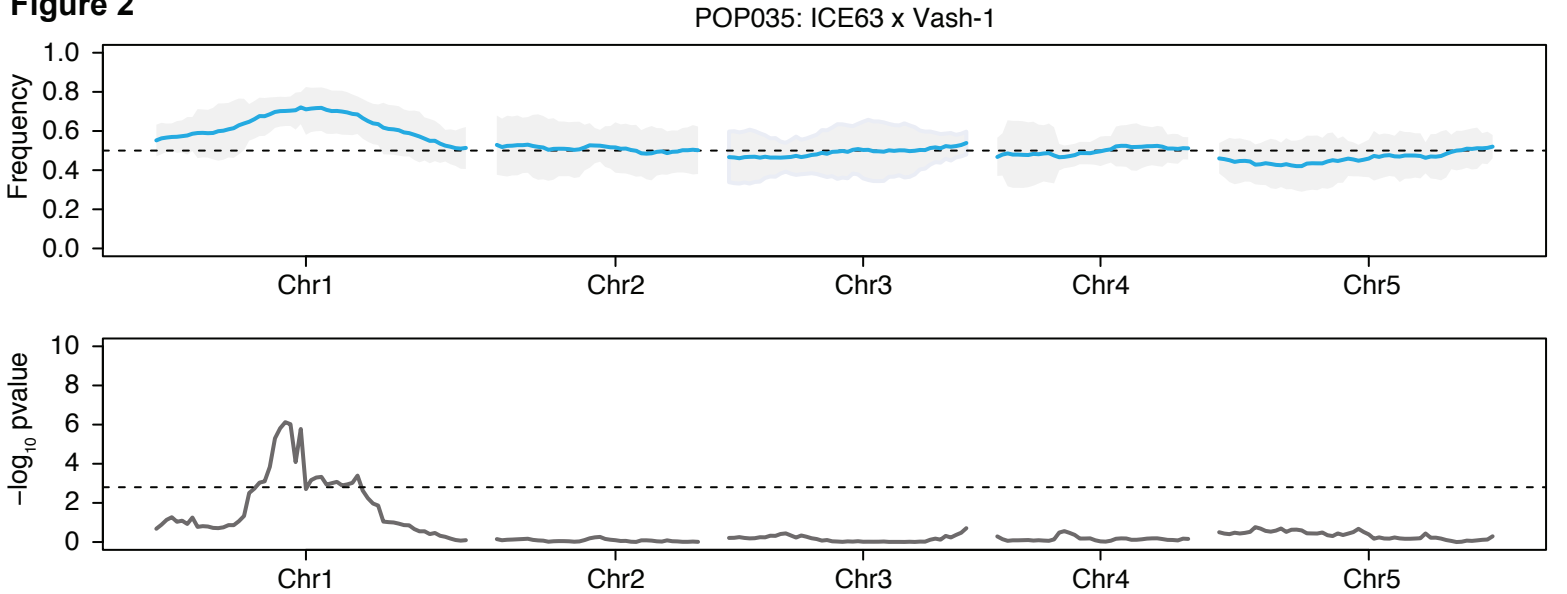

POP035: ICE63 x Vash-1
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
Chr1
Chr2
Chr3
Chr4
Chr5
−log10 pvalue
10
8
6
4
2
0
Chr1
Chr2
Chr3
Chr4
Chr5

Figure 3

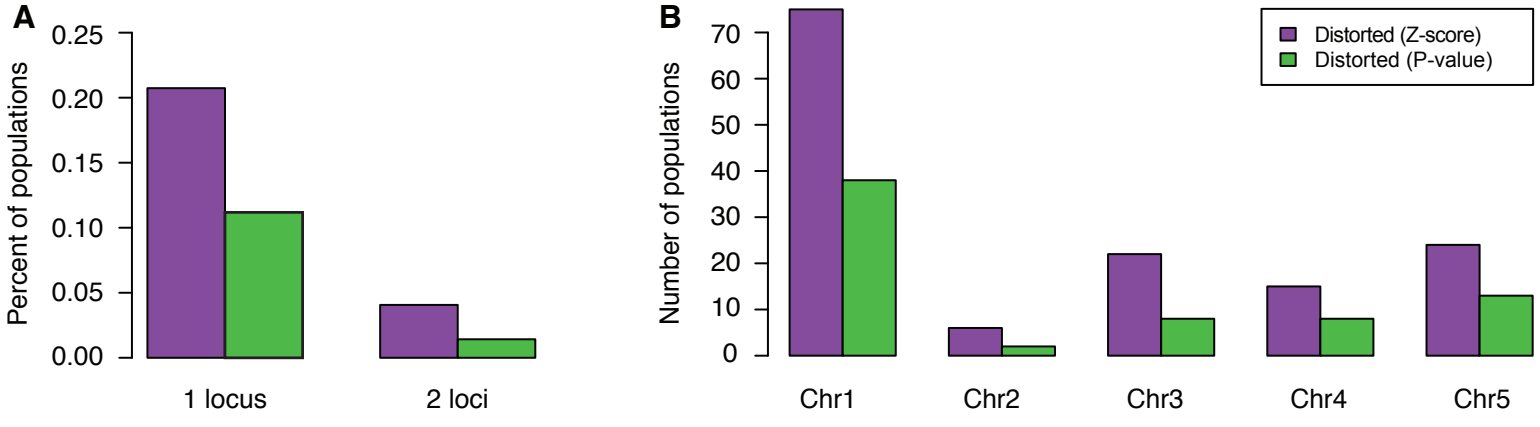

A
Percent of populations
0.25
0.20
0.15
0.10
0.05
0.00
1 locus
2 loci
B
Number of populations
70
60
50
40
30
20
10
0
Distorted (Z-score)
Distorted (P-value)
Chr1
Chr2
Chr3
Chr4
Chr5

Figure 4

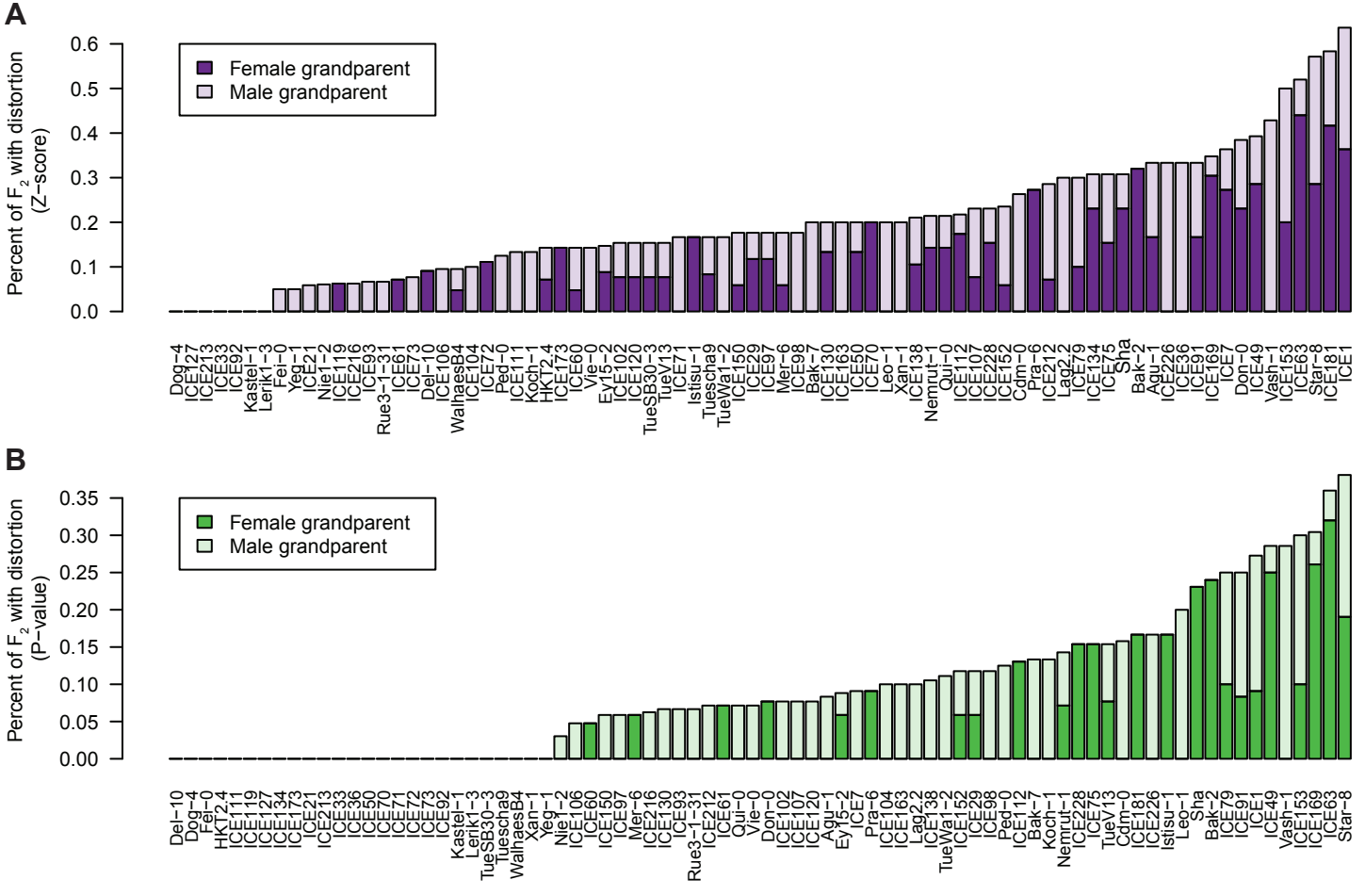

A
Percent of F2 with distortion (Z−score)
Female grandparent
Male grandparent
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Dog−4
CE127
CE213
ICE33
ICE92
Kastel−1
Lerik1−3
Fei−0
Yeg−1
ICE21
Nie1−2
ICE119
ICE216
ICE93
Rue3−1−31
ICE61
ICE73
Del−10
ICE106
WalhaesB4
ICE104
ICE72
Ped−0
ICE111
Koch−1
HKT2.4
ICE173
ICE60
Vie−0
Ey15−2
ICE102
ICE120
TueSB30−3
TueV13
ICE71
Istisu−1
Tuescha9
TueWa1−2
ICE150
ICE29
ICE97
Mer−6
ICE98
Bak−7
ICE130
ICE163
ICE50
ICE70
Leo−1
Xan−1
ICE138
Nemrut−1
Qui−0
ICE112
ICE107
ICE228
ICE152
Cdm−0
Pra−6
ICE212
Lag2.2
ICE79
ICE134
ICE75
Sha
Bak−2
Agu−1
ICE226
ICE36
ICE91
ICE169
ICE7
Don−0
ICE49
Vash−1
ICE153
ICE63
Star−8
ICE181
ICE1
B
Percent of F2 with distortion (P−value)
Female grandparent
Male grandparent
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Del−10
Dog−4
Fei−0
HKT2.4
ICE111
ICE119
ICE127
ICE134
ICE173
ICE21
ICE213
ICE33
ICE36
ICE50
ICE70
ICE71
ICE72
ICE73
ICE92
Kastel−1
Lerik1−3
TueSB30−3
Tuescha9
WalhaesB4
Xan−1
Yeg−1
Nie1−2
ICE106
ICE60
ICE150
ICE97
Mer−6
ICE216
ICE130
ICE93
Rue3−1−31
ICE212
ICE61
Qui−0
Vie−0
Don−0
ICE102
ICE107
ICE120
Agu−1
Ey15−2
ICE7
Pra−6
ICE104
ICE163
Lag2.2
ICE138
TueWa1−2
ICE152
ICE29
ICE98
Ped−0
ICE112
Bak−7
Koch−1
Nemrut−1
ICE228
ICE75
TueV13
Cdm−0
ICE181
ICE226
Istisu−1
Leo−1
Sha
Bak−2
ICE79
ICE91
ICE1
ICE49
Vash−1
ICE153
ICE169
ICE63
Star−8

Figure 5

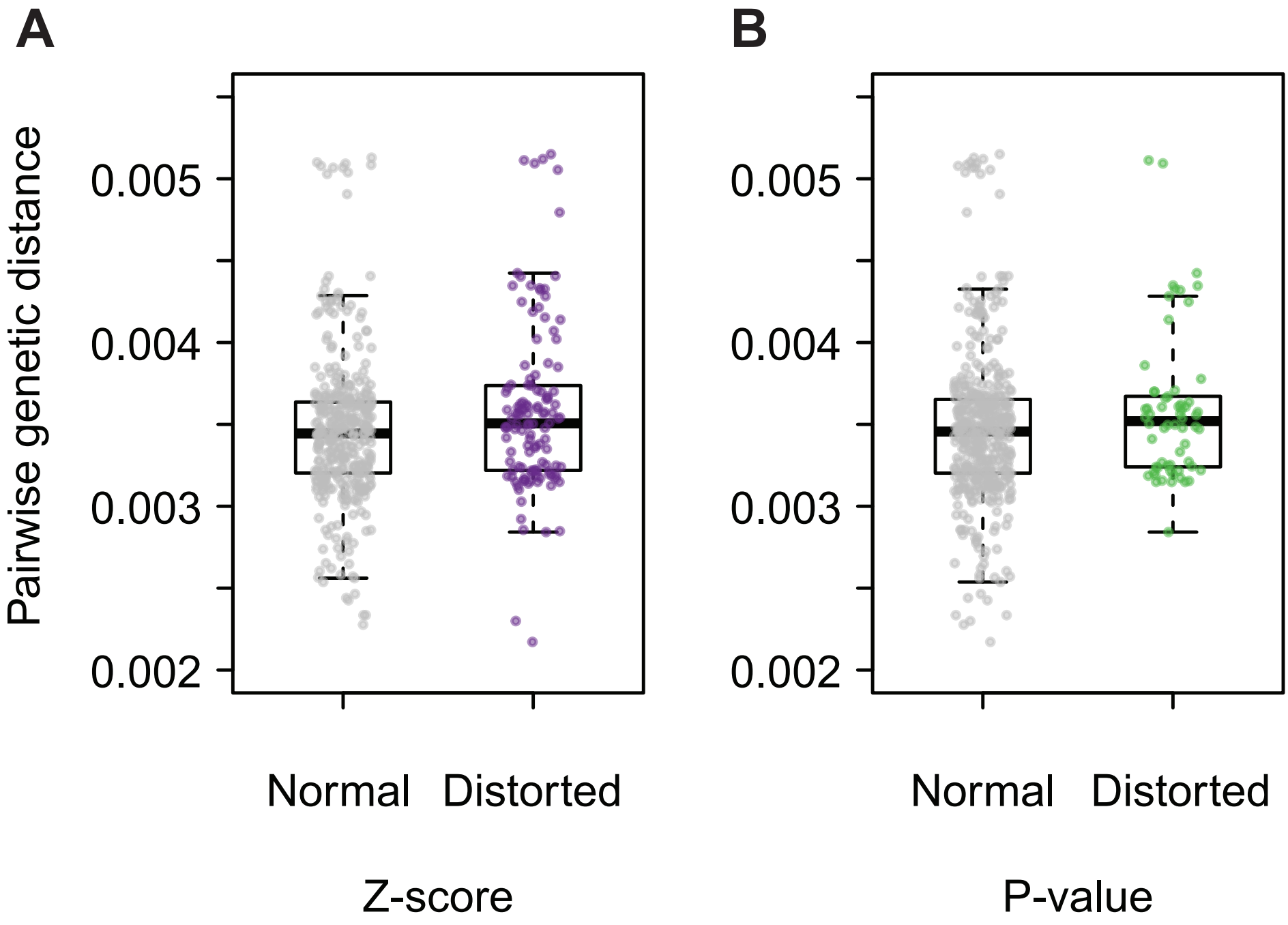

A
B
Pairwise genetic distance
0.005
0.004
0.003
0.002
Normal
Distorted
Z-score
P-value

Figure 6

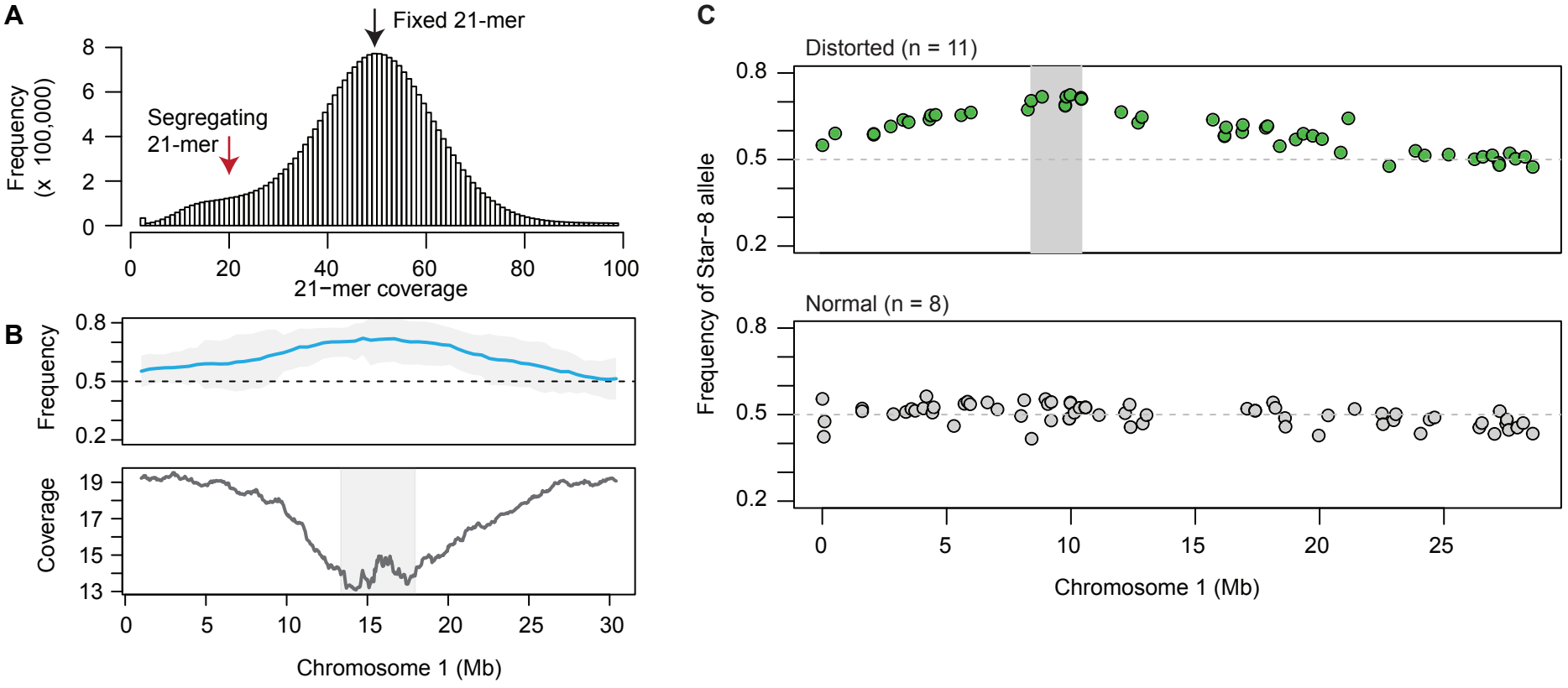

A
Fixed 21-mer
Segregating
21-mer
Frequency
(x 100,000)
21−mer coverage
B
Frequency
Coverage
Chromosome 1 (Mb)
C
Distorted (n = 11)
Normal (n = 8)
Frequency of Star−8 allele
Chromosome 1 (Mb)

Figure S1

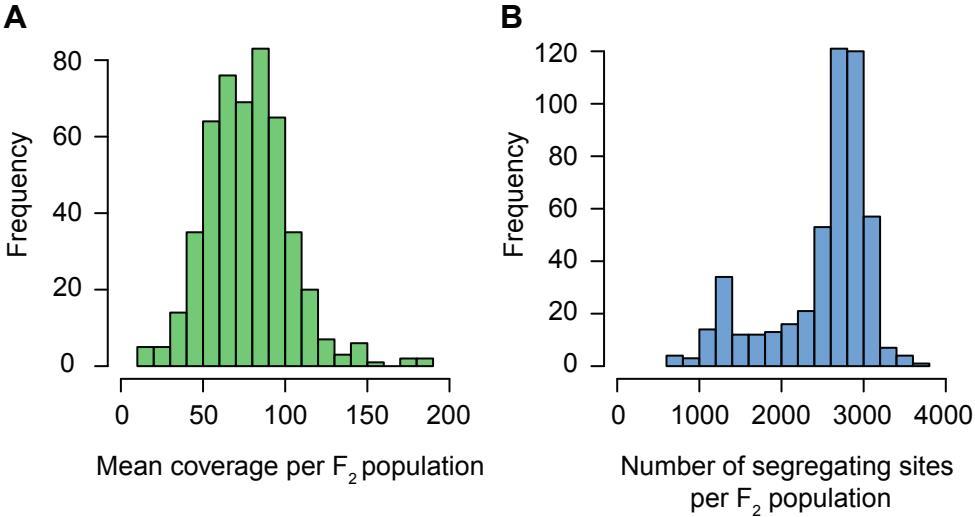

A
Frequency
0
20
40
60
80
0
50
100
150
200
Mean coverage per F2 population
B
Frequency
0
20
40
60
80
100
120
0
1000
2000
3000
4000
Number of segregating sites
per F2 population

Figure S2

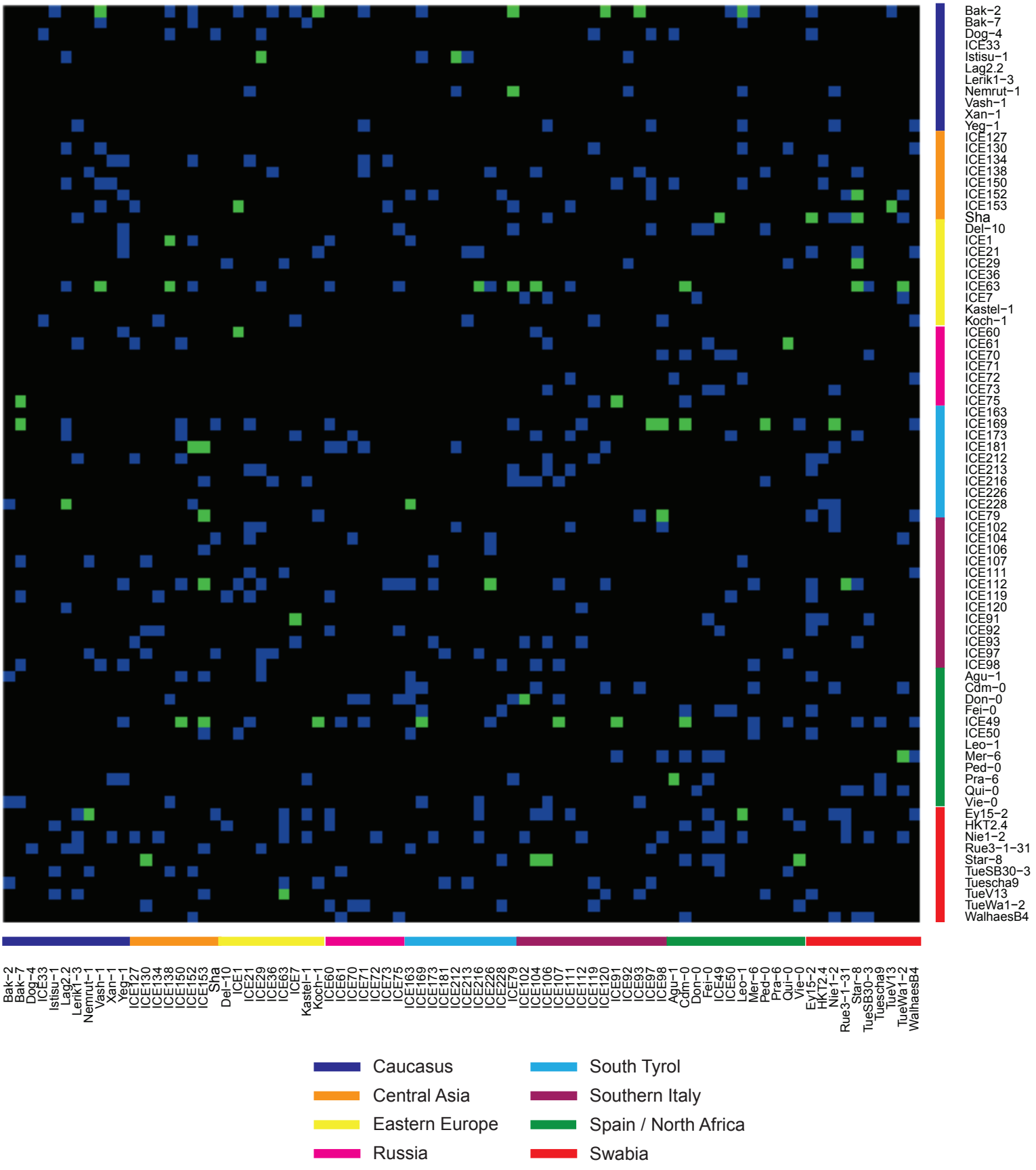

Bak−2
Bak−7
Dog−4
ICE33
Istisu−1
Lag2.2
Lerik1−3
Nemrut−1
Vash−1
Xan−1
Yeg−1
ICE127
ICE130
ICE134
ICE138
ICE150
ICE152
ICE153
Sha
Del−10
ICE1
ICE21
ICE29
ICE36
ICE63
ICE7
Kastel−1
Koch−1
ICE60
ICE61
ICE70
ICE71
ICE72
ICE73
ICE75
ICE163
ICE169
ICE173
ICE181
ICE212
ICE213
ICE216
ICE226
ICE228
ICE79
ICE102
ICE104
ICE106
ICE107
ICE111
ICE112
ICE119
ICE120
ICE91
ICE92
ICE93
ICE97
ICE98
Agu−1
Cdm−0
Don−0
Fei−0
ICE49
ICE50
Leo−1
Mer−6
Ped−0
Pra−6
Qui−0
Vie−0
Ey15−2
HKT2.4
Nie1−2
Rue3−1−31
Star−8
TueSB30−3
Tuescha9
TueV13
TueWa1−2
WalhaesB4
Caucasus
Central Asia
Eastern Europe
Russia
South Tyrol
Southern Italy
Spain / North Africa
Swabia

Figure S3

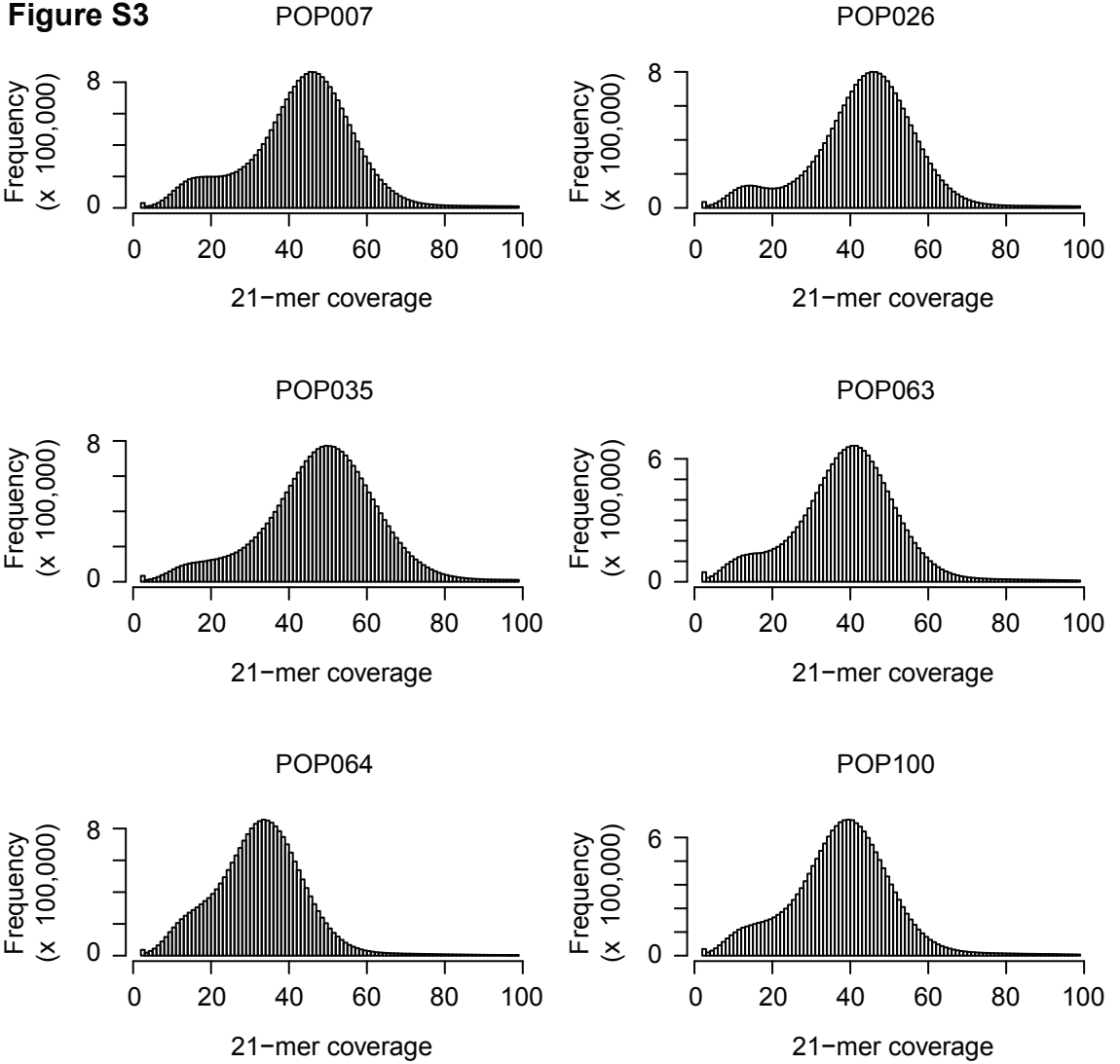

POP007
POP026
POP035
POP063
POP064
POP100
Frequency
(x 100,000)
21−mer coverage
0
20
40
60
80
100
8
6

Figure S4

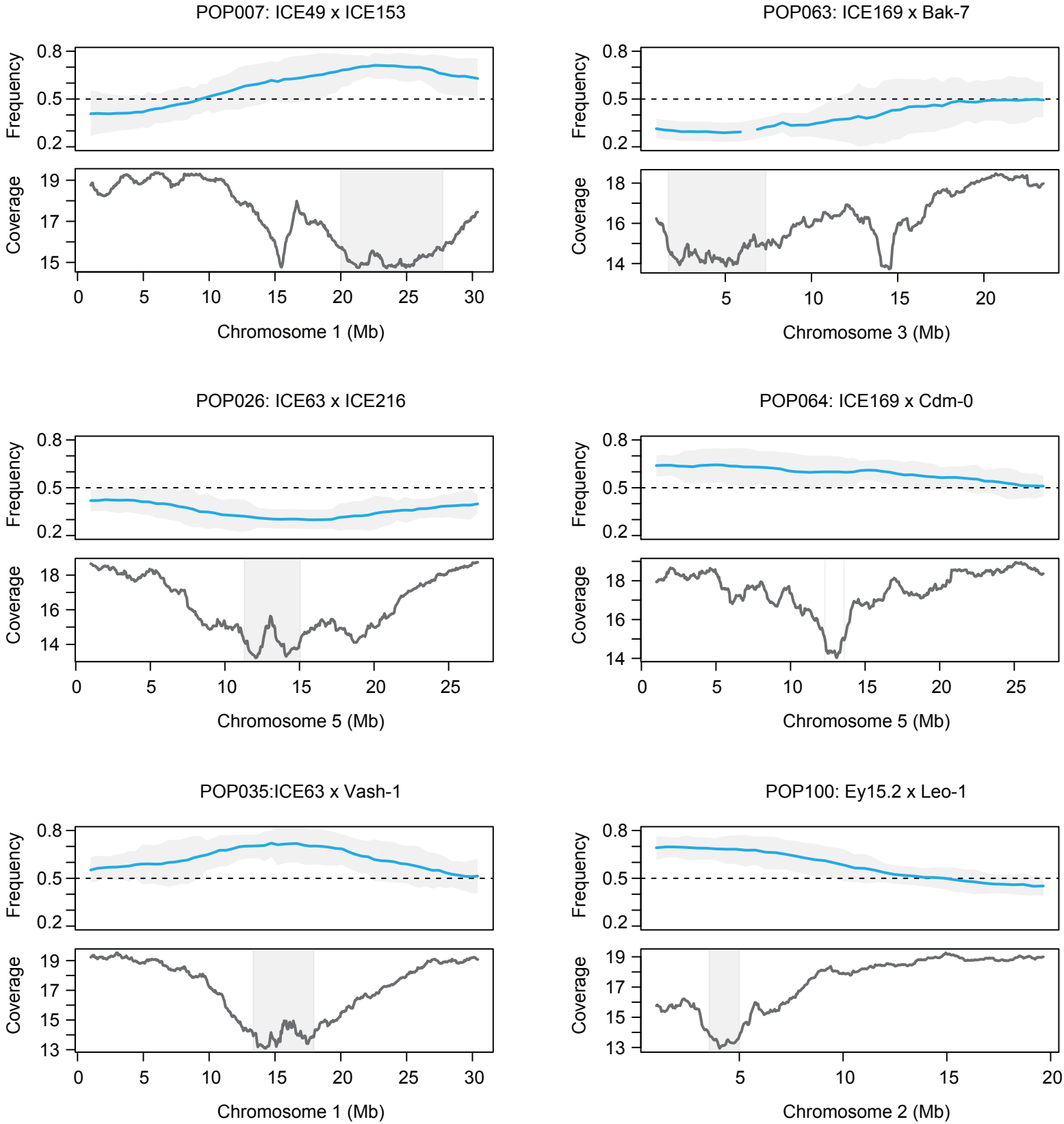
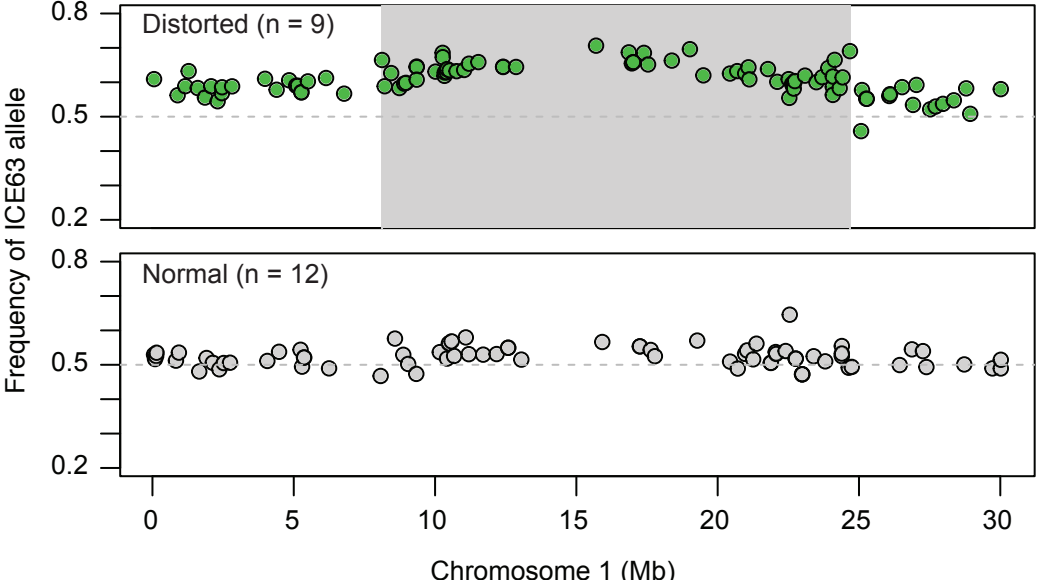

POP007: ICE49 x ICE153
Frequency
Coverage
Chromosome 1 (Mb)
POP063: ICE169 x Bak-7
Chromosome 3 (Mb)
POP026: ICE63 x ICE216
Chromosome 5 (Mb)
POP064: ICE169 x Cdm-0
Chromosome 5 (Mb)
POP035:ICE63 x Vash-1
Chromosome 1 (Mb)
POP100: Ey15.2 x Leo-1
Chromosome 2 (Mb)

Figure S5


Distorted (n = 9)
Normal (n = 12)
Frequency of ICE63 allele
0.8
0.5
0.2
0
5
10
15
20
25
30
Chromosome 1 (Mb)