Open access • Posted Content • DOI:10.1101/829622

Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype — Source link

Kaichi Huang, Rose L. Andrew, Rose L. Andrew, Gregory L. Owens ...+4 more authors

Institutions: University of British Columbia, University of New England (United States), University of California, Berkeley, Duke University

Published on: 04 Nov 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Helianthus petiolaris, Incipient speciation, Population and Ecotype

Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype

Kaichi Huang[1†], Rose L. Andrew[1,2], Gregory L. Owens[1,3], Kate L. Ostevik[1,4], Loren H. Rieseberg[1]

[1]Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

[2]School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia

[3]Department of Integrative Biology, University of California, Berkeley, CA, USA

[4]Department of Biology, Duke University, Durham, NC, USA

[†]Author for correspondence: kaichi.huang@botany.ubc.ca

ABSTRACT

Both models and case studies suggest that chromosomal inversions can facilitate adaptation and speciation in the presence of gene flow by suppressing recombination between locally adapted alleles. Until recently, however, it has been laborious and time-consuming to identify and genotype inversions in natural populations. Here we apply RAD sequencing data and newly developed population genomic approaches to identify putative inversions that differentiate a sand dune ecotype of the prairie sunflower (*Helianthus petiolaris*) from populations found on the adjacent sand sheet. We detected seven large genomic regions that exhibit a different population structure than the rest of the genome and that vary in frequency between dune and non-dune populations. These regions also show high linkage disequilibrium and high heterozygosity between, but not within haplotypes, consistent with the behavior of large inversions, an inference subsequently validated in part by comparative genetic mapping. Genome-environment association analyses show that key environmental variables, including vegetation cover and soil nitrogen, are significantly associated with inversions. The inversions co-locate with previously described "islands of differentiation," and appear to play an important role in adaptive divergence and incipient speciation within *H. petiolaris*.

KEYWORDS

Inversions, divergence with gene flow, local adaptation, genomic islands, genome-environment association, *Helianthus petiolaris*

INTRODUCTION

Genetic differentiation between differently adapted populations can be highly variable across the genome. During the process of adaptive divergence, genomic regions under selection will display strong differentiation, while ongoing gene flow between populations will homogenize other regions, generating heterogeneous patterns of genomic divergence (Wu, 2001; Nosil, Funk, & Ortiz-Barrientos, 2009). Large islands of differentiation, namely "genomic islands of divergence", are commonly seen in recently diverging populations, ecotypes, and species, including well-known examples in *Rhagoletis* (Feder, Chilcote, & Bush, 1988), *Anopheles* (Turner, Hahn, & Nuzhdin, 2005), *Heliconius* (Nadeau et al., 2012) and *Helianthus* (Andrew & Rieseberg, 2013). The causes of these large islands are not fully understood (although see McGaugh & Noor, 2012; Berg et al., 2017). It has been proposed that divergence hitchhiking, in which gene exchange is reduced adjacent to a locus under strong divergent selection, could generate large regions of differentiation, but the conditions under which it occurs are limited (Via, 2012; Feder & Nosil, 2010). Chromosomal inversions represent another possible explanation for such islands because they can suppress recombination and impede gene flow across large genomic regions (Butlin, 2005; Hoffman & Rieseberg, 2008).

Inversions have long been viewed as important in local adaptation and speciation (Wellenreuther & Bernatchez, 2018; Dobzhansky & Sturtevant, 1938). One primary reason is that, by suppressing recombination, inversions can establish and maintain favorable

combinations of locally adapted alleles, despite gene flow with non-adapted populations (Rieseberg, 2001; Kirkpatrick & Barton, 2006). The critical importance of inversions in local adaptation has been revealed by emerging studies that document the association of inversions with adaptive traits within species (Feder, Roethele, Filchak, Niedbalski, & Romero-Severson, 2003; Lowry & Willis, 2010; Kirubakaran et al., 2016; Wellenreuther & Bernatchez, 2018 for review). Beyond their role in adaptation, inversions can preserve alleles that cause intrinsic genetic incompatibilities in hybrids, and facilitate the accumulation of new incompatibilities, thereby aiding species' persistence in the face of gene flow (Noor, Grams, Bertucci, & Reiland, 2001; Navarro & Barton, 2003). Finally, inversions can establish linkage between locally adapted alleles and those causing assortative mating, which is typically required in models of speciation with gene flow (Felsenstein, 1981; Trickett & Butlin, 1994; Servedio, 2009).

Much of what we know about inversions (at least until very recently), comes from studies of Dipteran flies, whose very large larval salivary gland chromosomes permit detection of inversions from chromosome banding patterns (Krimbas & Powell, 1992). However, in most other organisms, more time-consuming and/or expensive methods have been required, such as analyses of meiotic configurations (Heslop-Harrison, 2013), comparative genetic mapping (Kirubakaran et al., 2016), Hi-C sequencing (Dixon et al., 2018), optical mapping (Tang, Lyons, & Town, 2015), paired-end mapping (Lamichhaney et al., 2016), or long-read sequencing. The laboriousness and/or expense of these methods have hindered our understanding of the frequency and importance of inversions in natural populations. Recently, population genomic

81 approaches have been applied to detect potential inverted regions, including methods based on
82 linkage disequilibrium (LD) (Faria et al., 2019; Arostegui, Quinn, Seeb, Seeb, & McKinney, 2019)
83 and local population structure (Li & Ralph, 2019). The LD approach takes advantage of the
84 expectation that inversions will create high LD between (but not within) inversion haplotypes.
85 The local population structure approach assumes that the lack of gene flow between inversion
86 haplotypes will lead to systematic differences in patterns of genetic relatedness between
87 inverted and collinear regions. Such differences can be detected by conducting windowed
88 analyses of population structure across the genome (Li & Ralph, 2019). Both methods offer an
89 efficient means for identifying putative inversions and estimating their frequency in natural
90 populations.
91
92 In this study, we focus on the genetic architecture of adaptation in a dune-adapted ecotype of
93 the prairie sunflower *Helianthus petiolaris* Nutt. This widespread annual sunflower inhabits
94 sandy soils in the Central and Southwestern USA. However, in the Great Sand Dunes National
95 Park and Preserve (GSD), Colorado, an ecotype of this species occurs in active sand dunes. This
96 dune ecotype differs from conspecific populations, which are abundant on the sand sheet below
97 the dunes, for a number of ecologically relevant phenotypic traits, including seed size,
98 branching, and root architecture (Andrew, Ostevik, Ebert, & Rieseberg, 2012). Despite its origin
99 less than 10,000 years ago (Andrew, Kane, Baute, Grassa, & Rieseberg, 2013), multiple
100 reproductive barriers isolate the two ecotypes, including strong extrinsic selection against
101 immigrants and hybrids, conspecific pollen precedence, as well as a weak crossability barrier

(Ostevik, Andrew, Otto, & Rieseberg, 2016). Nonetheless, substantial and asymmetric gene flow have been reported between dune and non-dune populations (Andrew et al., 2012), as predicted by models of isolation with gene flow. Moreover, genetic differentiation between the ecotypes is largely restricted to several large genomic regions while background divergence is extremely low (Andrew & Rieseberg, 2013), making it a good system to study the evolution of genomic islands of divergence. The underlying mechanism for these large regions of high divergence was not previously determined, but chromosomal inversions represent a leading hypothesis given their ability to impede introgression, as well as the high rates of chromosomal evolution reported for *Helianthus* (Burke et al., 2004; Ostevik, Samuk, & Rieseberg, 2019).

Our analyses complement a recently submitted study from our group on the genetic architecture of local adaptation across three sunflower species (Todesco et al., 2019). In that study, we used whole genome shotgun sequence (WGS) data to sample genetic variation across the ranges of three sunflower species, including *H. petiolaris*. The study detected numerous large haplotypes in all three species that co-varied with ecologically relevant phenotypic, climate, and soil variation. Further analyses show that many of the haplotypes (but not all), were associated with structural variation, including inversions. One population from GSD (ten individuals) was included in this study, and it appeared to be enriched for structural variants. Thus, we also wished to validate this observation with more extensive sampling from GSD and surrounding regions, as well as to exploit a comprehensive data set on local variation in soil

fertility and plant cover on the dune and surrounding sand sheet to better assess the role of inversions in divergent adaptation with gene flow.

Specifically, we employ RAD sequence data previously generated for this system (Andrew et al., 2013) and apply a local population structure approach to detect and genotype putative inversions in this system. We also conduct additional population genomic analyses (including LD analyses) and develop two genetic maps (one for each ecotype) to further validate these inferences. Lastly, we search for associations between the genotypic data and key environmental factors, including soil nutrient availability and vegetation coverage. We address four main questions: 1) Can structural variants such as inversions be detected with RAD sequencing data? 2) If so, are they enriched in the dune habitat at GSD as previously suggested? 3) Likewise, do they correspond closely to the genomic islands of differentiation (i.e., high F_{ST} regions) previously reported between dune and non-dune sunflowers? and 4) Lastly, is there evidence that inversions contribute importantly to adaptive divergence in this system?

MATERIALS AND METHODS

Plant materials and RAD sequencing

Our study employs the plant materials and RAD sequencing (Baird et al., 2008) data set previously reported by Andrew & Rieseberg (2013) and Andrew et al. (2013). Twenty populations from dune, non-dune and intermediate habitats in the GSD were sampled (Andrew et al., 2013, Supporting Information Table S1), and five unrelated individuals from each of the 20 populations were subjected to RAD sequencing by Floragenex (Portland, OR) using the restriction enzyme *Pst*I. All samples were barcoded and sequenced with at least 60 bp reads, with a subset sequenced with 80 bp reads. The first 5 bp covering the restriction site and relatively low-quality 20 bp at 3' end of the 80 bp reads were trimmed with PRINSEQ v0.20.4 (Schmieder & Edwards, 2011), yielding reads with equal length of 55 bp, to avoid biases in alignment due to sequences of different lengths.

SNP calling

We re-called SNPs from the RAD sequencing data since much better reference genomes are now available for cultivated sunflower (*Helianthus annuus*), a close relative of *H. petiolaris*. Briefly, RAD sequences were aligned to reference genome Ha412HOv2.0 with bwa mem v0.7.17 (Li, 2013) using the default settings. Variant calling was performed with the Genome Analysis Tool Kit v4.0.8.1 (GATK; DePristo et al., 2011). Sample alignments were processed with the GATK HaplotypeCaller and samples were jointly genotyped using GATK's GenotypeGVCFs chromosome by chromosome. Variants of all chromosomes were later merged with MergeVcfs in Picard tools (http://broadinstitute.github.io/picard/). Only bi-allelic SNPs were selected for

downstream analyses. SNPs were filtered with GATK VariantFiltration with filter expression "QD < 4.0 || FS > 20.0 || MQ < 40.0 || MQRankSum < -5.0" and individual genotypes with depth less than 30 were set as missing. Loci that were non-variant or varied only due to singletons after filtering, as well as those with > 40% missing data, were excluded from the data set. Finally, SNPs with excess heterozygosity were filtered with GATK's 'VariantFiltration' filter expression "ExcessHet < 20.0" to avoid misalignment on paralogous regions.

Because the new reference genome provides physical locations of the SNPs and has much more complete chromosome coverage compared to the one used by Andrew and Rieseberg (2013), we re-calculated Weir and Cockerham's F_{ST} (Weir, 1996) between dune and non-dune ecotypes with VCFtools (Danecek et al., 2011) to examine genetic divergence across the new reference genome and re-localize regions of divergence.

Local population structure analysis

We analyzed patterns of population structure across the genome using the R package "lostruct" (Li & Ralph, 2019), in order to detect regions of abnormal population structure that might be generated by chromosomal inversions. The genome was divided into non-overlapping windows with size of 50 SNPs and principle component analysis (PCA) was calculated for each window to reflect local population structure. To measure the similarity of patterns of relatedness between windows, Euclidean distances between matrices were calculated for the first two principle

components (PCs) and then mapped using multidimensional scaling (MDS) into 40-dimensional space. Different window sizes were tested to reach the best balance between signal and noise. The SNP data set was converted to BCF format with BCFtools v1.9 (Li, 2011) before input to lostruct.

To identify localized genomic regions with extreme MDS values, we first defined outlier windows as those with absolute values greater than 4 standard deviations from the mean across all windows for each of the 40 MDS coordinates. We then tested whether outlier windows were chromosomally clustered with 1,000 permutations of windows over chromosomes to evaluate differences from random expectation where outliers are randomly distributed among chromosomes. For each MDS coordinate with more than 4 outlier windows, we selected the first chromosome with a significant excess of outliers ($p < 0.01$) for further examination. For each coordinate, outlier windows that deviated in different directions were examined separately. Adjacent outliers with less than 4 windows between them were kept as a cluster. In cases where the same chromosome had outlier clusters across multiple MDS coordinates, we calculated Pearson's product moment correlation coefficient between the MDS coordinates using sample genotype matrices and collapsed the ones with correlation > 0.8 by selecting the coordinate with the larger number of outliers. The coordinates of the putative inversions were defined by the start position of the first outlier window to the end position of the last outlier window.

While inversions are a major driver of MDS outliers detected by lostruct (Li & Ralph, 2019), MDS outliers can be generated by other processes as well, such as linked selection. Therefore, we performed a series of additional analyses to look for additional population genomic signatures of inversions. Due to suppressed recombination, haplotype blocks with different orientations should evolve largely independently, resulting in distinct nucleotide differences between them. Therefore, for an inversion segregating in a population, a PCA of population structure should divide the samples into three distinct groups representing the two inversion haplotypes, with heterozygotes between the haplotypes forming an intermediate cluster. To test this, we calculated PCAs with SNPrelate (Zheng et al., 2012) using all SNPs from each putative inversion. To identify the composition of groups of genotypes, we used the R function “kmeans” with the method developed by Hartigan and Wong (1979) to perform clustering on the first PC, using the maximum, minimum and middle of the range of PC scores as the initial cluster centers. The discreteness of the clustering was evaluated by the proportion of the between-cluster sum of squares over the total. The K-means cluster assignment was used as the genotype of the sample.

If the groups detected in the PCA represent homozygotes and heterozygote for the orientations, we expect the central group to have high heterozygosity relative to the other two groups. For each region identified, we extracted all variable sites across the outlier windows and calculated the proportion of heterozygous sites over the total as heterozygosity for each individual in each group identified by k-means clustering.

To examine the effect of recombination suppression of the putative inversions, intrachromosomal linkage disequilibrium (LD) was calculated among all SNPs with minor allele frequencies > 5%. Pairwise LD (R^2) values were calculated using PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007) for each chromosome with all samples. Values of SNPs were grouped into 1 Mb windows and the second largest R^2 value was plotted using ggplot2 (Wickham, 2016). For chromosomes with MDS outlier regions, R^2 was also calculated with individuals homozygous for the more common orientation only.

Only the regions displaying clustering of three distinct groups in the PCA with higher heterozygosity in the middle group and high LD were kept as putative inversions in downstream analyses. For each region, allele frequency differences between ecotypes were estimated using "prop.test" in R and the genotype frequency for each population was plotted onto a map of land cover classification downloaded from Multi-Resolution Land Characteristics Consortium (https://www.mrlc.gov/) at 30-m resolution.

Genetic map construction

Genetic maps of dune and non-dune ecotypes were generated using F1 testcross mapping to validate our inversion detection approach. Pollen from a single dune plant (seed collected from population 1300) and a single non-dune plant (seed collected from a new population at Latitude

247 37.724, Longitude -105.718) from GSD, was used to fertilize individuals of the male sterile *H.*
248 *annuus* HA89cms cultivar, which is highly homozygous. For each cross, the HA89cms individuals
249 that bore the most seeds (100-150 seeds) were selected to produce the F1 mapping
250 populations. Loci that are heterozygous in a wild parent are expected to segregate 1:1 in the
251 corresponding F1 population, permitting the generation of a genetic map. DNA was extracted
252 from germinated F1 seeds or, when germination failed, directly from seeds. Barcoded
253 genotyping-by-sequencing (Poland, Brown, Sorrells, & Jannink, 2012) libraries were prepared
254 using the restriction enzymes *Pst*I and *Msp*I. A depletion step with Duplex-Specific Nuclease
255 (DSN; Evrogen, Moscow, Russia) was conducted on the libraries to reduce the proportion of
256 repetitive sequences, including plastid DNA (Todesco et al., in preparation). The libraries were
257 sequenced on an Illumina Hiseq 4000 instrument to produce paired end, 100 bp reads (Illumina,
258 San Diego, CA, USA). Samples were demultiplexed using a custom Perl script that also removed
259 barcode sequences. FASTQ files were examined for quality but not trimmed. Raw reads were
260 aligned to the Ha412HOv2.0 reference genome using NextGenMap v0.5.2 (Sedlazeck,
261 Rescheneder, & Von Haeseler, 2013) and variants were called using GATK v4.0.8.1 as described
262 above for the RAD sequences. Only SNPs were kept and filtered with the expression "QD < 15.0
263 || FS > 20.0 || MQ < 40.0 || MQRankSum < -5.0", and individual genotypes with depth less than
264 30 were set as missing. Loci that were invariant after filtering and had a genotype missing rate >
265 50% were excluded.
266

267 Genetic maps were built using R/qtl (Broman, Wu, Sen, & Churchill, 2003) and R/ASMap (Taylor
268 & Butler, 2017). Individuals with fewer than 50% markers genotyped were excluded, as were
269 duplicate markers, markers with less than 50% of individuals scored, and markers with extreme
270 segregation patterns (genotype frequency <0.3 or >0.7). The "mstmap.cross" function was used
271 to construct linkage groups (LGs) with the remaining markers using a *p*-value of 10^{-15}, which was
272 chosen to minimize false linkages. Because marker phase was unknown prior to mapping, mirror
273 image LGs were generated initially, and the function "switchAlleles" was used to reverse
274 genotype scores for such LGs. Markers with segregation distortion *P*-value < 0.05 and missing
275 rate < 0.1 were pulled aside from the map, and those with more than three double crossovers
276 and markers with extreme (> 2 standard deviation) segregation distortion within a 21-marker
277 window were removed using custom functions. LGs with less than two markers were discarded.
278 Some less extreme markers that were originally placed aside were then pushed back into the
279 map and the markers were filtered again with the same criteria. This step was done twice to
280 reintroduce markers with segregation distortion *P*-value < 0.01, missing rate < 0.3 and the ones
281 with segregation distortion *P*-value < 0.001 and missing rate < 0.5. The function "calc.errorlod"
282 was also used to filter genotyping errors. Finally, very small (1-5 markers) LGs were discarded,
283 leaving 17 LGs for each ecotype.
284
285 Due to sparse marker density on the LG that corresponded to chromosome 5 after filtering,
286 markers that mapped to chromosome 5 on the *H. annuus* reference genome were extracted
287 and genetic mapping was repeated using less stringent parameters. Markers that were located

288 at the far end of LG 5, and those that disturbed synteny, were removed because they might
289 represent misaligned markers from other chromosomes. This remapping was conducted for
290 both dune and non-dune mapping populations and the new LGs were included in downstream
291 genetic map comparisons.
292
293 To compare marker orders, we took advantage of the fact that SNP markers were called against
294 the Ha412HOv2.0 reference genome. Homologous reference chromosomes for each linkage
295 groups were identified based on physical positions of markers and prior knowledge of the
296 location of translocations between *H. petiolaris* and *H. annuus* (Ostevik et al., 2019). For each
297 putative inversion, we asked whether markers from that region differed in order or genetic
298 distance with respect to the reference genome and/or between ecotypes.
299
300 **Genome-environment association analysis**
301
302 To further assess the role of putative inversions in dune adaptation in the ecotype, as well as to
303 identify the environmental variables that might be driving divergent selection pressures, we
304 used data on soil nutrient availability and vegetation coverage for each population to conduct
305 genome-environment association (GEA) analysis.
306
307 The collection and estimation of these measurements have been described in detail in a
308 previous study (Andrew et al., 2012). Additional composite variables of soil or cover data were

generated by PCA and the first three PCs (soil PC1-3 and cover PC1-3) were used in the analyses. The vegetation coverage data were arcsine-square-root-transformed and all measurements were standardized prior to PCA.

The GEA analysis was performed using BayPass v2.1, which explicitly accounts for the covariance structure among the population allele frequencies resulting from population demography (Gautier, 2015). We further filtered the SNPs by missing rate < 10% and minimum allele frequency > 10% and generated a data set of SNP frequencies for all populations. Population structure was estimated by running BayPass under the core model mode with all filtered SNPs. The covariance matrix from this analysis was then used as a control for population structure to evaluate associations of SNPs with each environmental variable. For each SNP, a Bayes factor (BF) was computed under the standard covariate model using the default importance sampling estimator approach. Scaling was performed for each environmental variable using the "-scalecov" option. Due to missing soil data in population 970, the analysis was run separately for soil variables and coverage variables.

To further examine the associations between the putative inversions and environmental variables, we also performed a GEA analysis in which putative inversions were treated as single bi-allelic loci. A SNP data set excluding SNPs from within the putative inversions was used to estimate the covariance matrix to control for the effects of the MDS outlier regions on

329 population structure. Bayes factors were calculated using the same core model mode in BayPass
330 as described above.
331
332 To calculate a significance threshold, we simulated pseudo-observed data (POD) with 1,000
333 SNPs using the "simulate.baypass" function implemented in BayPass with the covariance matrix
334 generated under the core model, and analyzed the newly created POD for each environmental
335 variable as described above. The top 1% quantile of the POD BFs was computed as the threshold
336 for significance.
337
338 **RESULTS**
339
340 **SNP Calling**
341
342 Using a high quality reference genome for cultivated *H. annuus*, 87.0% of RAD sequences were
343 aligned on average, and after variant calling with GATK, a total of 260,478 variable sites were
344 scored. Filtering produced a data set of 37,930 high-quality bi-allelic SNPs across 17
345 chromosomes of the reference, which corresponds to approximately 12 sites per Mbp. This
346 compares favorably to the 11,727 SNPs that could be positioned on chromosomes in our
347 previous analyses (Andrew & Rieseberg, 2013).
348

Analysis of patterns of genetic divergence between the dune and non-dune ecotypes yielded similar results to the previous study (Andrew & Rieseberg, 2013): low overall F_{ST} and high heterogeneity among sites with the largest clusters of outliers found on chromosomes 5,9 and 11 (Figure 1). However, highly divergent regions are more distinct and contiguous in the present study due to the larger number of SNPs and better genome assembly. In addition, a distinctive island can now be seen on the end of chromosome 7, which was not detected in the previous analysis

Detection of putative chromosomal inversions

Using a window-based local population structure analysis implemented in "lostruct," and our outlier discovery approach, we identified a total of 9 clusters of MDS outliers with our RAD SNPs (Table1, Figure 2).

In PCAs of most outlier regions, individuals were aggregated into three discrete groups on the first PC, which explained much more variation than the second PC (Table 1, Figure 2b, Supporting Information Figure S1). The discreteness was supported by the high (>0.9) proportion of the between-cluster sum of squares over the total in k-means clustering (Table 1). Moreover, in most regions, heterozygosity of the middle group was significantly higher than within the other two groups (Figure 2c, Supporting Information Figure S1). These patterns are consistent with the presence of two clusters of individuals that are homozygous for alternative

370 inversion haplotypes and an intermediate cluster of individuals that are heterozygous for the
371 inversion haplotypes with no or very little recombination between them. Two exceptions were
372 found, including one on chromosome 13 for MDS12, where samples formed only two groups in
373 the PCA and the expected pattern of heterozygosity was not observed. Likewise, samples did
374 not form distinct clusters for outlier region MDS21 on chromosome 9 (Table 1, Supporting
375 Information Figure S1). Note that outlier region for MDS21 encompasses that of MDS02, which
376 does act like a legitimate inversion, as well as an upstream region of the chromosome that
377 generally does not.
378
379 Almost all outlier clusters were also characterized by high LD, and almost all large regions of
380 high LD across the genome were identified as outlier regions in our analyses. For the two outlier
381 clusters that did not form three distinct groups in the PCA, the MDS12 outlier region on
382 chromosome 13 was characterized by high LD. Thus, we cannot rule out the possibility that this
383 is an inversion, but that heterozygotes are rare and genotypes mis-classified. MDS21 includes a
384 large high LD region, which represents the MDS02 outlier region, as well as a smaller high LD
385 region at the start. Possibly the latter represents a small inversion that is in partial LD with the
386 MDS02 outlier region. There also were a handful of very small high LD regions (e.g., on
387 Chromosome 15 from 119-123 Mbp) that might represent inversions, but they did not pass our
388 stringent criteria for MDS outliers. Lastly, while high LD was detected for the outliers when
389 compared across all samples, recombination was not restricted within the homozygous group
390 (Figure 2d, Supporting Information Figure S1, S2), except for MDS21. These results are

391 consistent with the role of inversions in altering recombination in heterozygotes while
392 recombination in homozygotes remains unaffected.
393
394 Overall, seven of the outlier clusters showed clustering of three distinct groups in PCA, higher
395 heterozygosity in the middle group and high LD across the outlier region, and were kept as
396 putative inversions for downstream analyses (Table 1). All the putative inversions, except one
397 on chromosome 9 (pet09.02), overlapped substantially with large haplotypes identified in *H.*
398 *petiolaris* using WGS data over its entire geographic distribution (Todesco et al. 2019; Table 1).
399 These 7 putative inversions occurred on 6 chromosomes. A majority of them were located near
400 the end of chromosomes, while the putative inversion on chromosome 7 (pet07.01) and the
401 larger one on chromosome 9 (pet09.01) resided in the middle sections of the chromosomes
402 (Figure 1). Each of the putative inversions contained at least 5 MDS outlier windows (i.e. 250
403 SNPs) and their sizes varied between 11 and 57 Mbp (Table 1).
404
405 All of the putative inversions displayed significant allele frequency differences between dune
406 and non-dune ecotypes (*P* ranges from 0.024 for pet09.02 to 2.92×10^{-22} for pet05.01, Table 1),
407 but the distributions of the genotypes for each inversion were variable. For several putative
408 inversions, the sand dunes are enriched with samples homozygous for one of the orientations
409 (cluster 0 or cluster 2 identified by k-means clustering) (e.g., pet11.01 and pet05.01), while
410 others showed more heterozygotes in the dunes (e.g., pet09.01) (Figure 3, Supporting
411 Information Figure S3). For pet14.01, the “dune” orientation was not found in the non-dune

habitat, although this orientation has a low frequency among samples, with only one individual identified as homozygous (Figure 3).

Most of the putative inversions were associated with regions of high F_{ST} between dune and non-dune ecotypes (Figure 1), especially in pet05.01, pet07.01, pet09.01 and pet11.01, where the largest divergence between ecotypes was found. Two exceptions were pet14.01 and pet09.02, for which the frequency of the "dune" orientation was relatively low.

Genetic maps

After SNP filtering, a total of 117 individuals and 9,926 markers from non-dune mapping population, and 128 individuals and 11,748 markers from dune mapping population, entered the map construction process. The final map for non-dune ecotype is made up of 2,559 markers at 801 unique positions with 98.5% of the map having a marker at least every 10 cM and 89.7% having a marker every 5 cM. Similarly, the map for the dune ecotype is made up of 3,077 markers at 571 unique positions with 96.8% of the map having a marker every 10 cM and 87.4% of it having a marker every 5 cM. Both of the final genetic maps correspond well with the expected 17 chromosomes and translocations found previously between *H. petiolaris* and *H. annuus* (Burke et al., 2004; Ostevik et al., 2019). The LGs are longer than the map reported by Burke et al. (2004), which is probably due to greater coverage of the genome. However, we cannot rule out the possible that a low level of genotyping error from our GBS mapping

433 approach may have contributed as well, although note that our maps are comparable in length
434 with maps for the two subspecies *of H. petiolaris* recently reported by Ostevik et al. (2019). Two
435 LGs in the dune map were unexpectedly short (D_LG2 and D_LG5; Supporting Information
436 Figure S4, S5) due to few markers from the middle of the corresponding reference
437 chromosomes, which caused the LGs to split after stringent filtering. After reconstruction with
438 less stringent parameters, LG5s in both maps were of similar size and had enough coverage for
439 map comparisons.
440
441 In map comparisons of the putative inversions, pet05.01 exhibited the expected pattern of
442 reverse marker orders between the two maps. In the map for non-dune ecotype, markers were
443 largely syntenic with the reference genome, while in the map for dune ecotype, there was a
444 continuous block of markers with inverted order relative to the reference (Figure 4a). However,
445 for pet07.01, pet09.01, pet09.02 and pet14.01, marker orders did not differ between the maps.
446 But, for pet09.01, the many markers that mapped to this region formed tight clusters in both
447 maps, indicating very low recombination in the wild non-dune and dune plants used to make
448 these maps (Figure 4, Supporting Information Figures S6). This implies that both plants are
449 heterozygous for the pet09.01 inversion, which would account for the recombination
450 suppression observed. A similar pattern of reduced recombination was seen for pet11.01 and
451 pet17.01 in the non-dune maps, but not in the map made from dune plant, in which markers
452 from the region were in reverse order compared to the reference. Interestingly, markers with

453 reverse order only covered part of the region for pet11.01, which implies the presence of an
454 adjacent low recombination region or sequential inversions (Supporting Information Figures S6).
455
456 Genotyping of the inversions in the parental plants using GBS confirms our interpretations. The
457 dune and non-dune parental plants were homozygous for different haplotypes of pet05.01 and
458 heterozygous for both haplotypes at pet09.01. For pet11.01 and pet17.01, the dune plant was
459 homozygous while the non-dune was heterozygous for the inversion, which explains the
460 clustering of markers in the non-dune maps.
461
462 **Genome-environment association analysis**
463
464 After stringent filtration, 8,383 SNPs were retained for GEA analysis. In GEA, we found several
465 large genomic regions with consistently high BF values, most of which overlapped nearly
466 perfectly with the putative inversions. When treated as single loci, the putative inversions
467 typically exhibited associations that were similar in strength to the peaks seen for the genome-
468 wide SNPs (Figure 5, Supporting Information Figures S7, S8).
469
470 The BF thresholds computed with POD ranged from 1.42 to 5.46 decibans (dB) depending on
471 environmental variables. Several putative inversions displayed significant associations with
472 environmental variables. The strongest signal of association was found for variables describing
473 vegetation cover (e.g., % forbs, % grasses, and % debris), with the most striking one being

pet05.01 with PC1 of coverage variables (Table 2). pet17.01 was also found to be associated with coverage variables, especially total cover. For soil characteristics, the strongest association was found for pet11.01 with NO_3 nitrogen. pet11.01 also displayed a significant association with PC2 of the soil variables but it was not as strong. Pet07.01 displayed significant associations with a number of soil variables but not with any of the three soil PCs. In contrast, pet05.01 was marginally associated with soil PC2, but not with any of the individual soil variables. Interestingly, % grasses is strongly associated with both pet05.01 and pet11.01, whereas % forbs is only associated with the former. This pattern might be related to nitrogen availability, since nitrogen (also associated with pet11.01) is often limiting for grasses, but not for legumes, which are the most frequent forbs on the dunes.

DISCUSSION

Genomic islands of differentiation often arise between diverging populations connected by gene flow (Feder & Nosil, 2009). While regions with higher than average differentiation can be created by divergence hitchhiking (Via, 2012), such regions are unlikely to be large or to have the sharp boundaries often reported for islands of divergence. Inversions represent a more likely explanation for large and discrete islands since recombination is reduced across the entire inverted region. Also, unlike other recombination modifiers, inversions reduce recombination between haplotypes, but not within them, which facilitates adaptive divergence. Theory

indicates that inversions will be favored if they prevent recombination between locally adapted alleles when challenged by migration of non-adapted alleles (Kirkpatrick and Barton 2006). Inversions can also facilitate speciation by preventing recombination between locally adapted alleles and those contributing to assortative mating (Ortiz-Barrientos, Engelstädter, & Rieseberg, 2016).

Despite the clear importance of inversions in adaptation and speciation, it remains difficult to identify and genotype them, especially in non-model systems. Using a population genomic approach with RAD sequencing data, we detected seven putative chromosomal inversions that separate dune and non-dune *H. petiolaris* in GSD, which we validated by a combination of population genetic and comparative genetic mapping approaches. Also, we demonstrated that inversions account for the genomic islands of high divergence between the ecotypes and contribute to ecological divergence in this system.

Identification of inversions

Employing the methods implemented in lostruct, which makes use of the effect that inversions have on population structure, we found clusters of windows with outlier MDS values, i.e. genomic regions with extreme population structure compared to the rest of the genome, and we provided multiple lines of evidence showing that the majority of these signals are left by inversions.

There are other processes that can generate a pattern of contiguous outlier MDS, such as selection coupled with gene flow, low recombination, or introgression. Linked selection can generate heterogeneous population structure across the genome (Li & Ralph, 2019), especially when selection is strong and acts in the face of gene flow, and may also generate long LD blocks. However, the regions that we identified are typically > 10Mb. It is unlikely that the effect of selection would span a region of several to tens of Mbp on the genome in the absence of structural variation. Moreover, such regions under selection are expected to generate a continuous pattern of population structure in a PCA as opposed to the three discrete clusters with higher heterozygosity in the middle cluster reported here. Lastly, the finding of high LD across putative inversions when tested across all samples, but not within putative homozygous groups, distinguishes inverted regions from other regions of reduced recombination (e.g. centromeres), because other mechanisms of recombination suppression are expected to restrict recombination in all groups of individuals. Other small, blurred-edged regions of low recombination were also found in our LD analysis (e.g., on chromosome 8 from 85-100 Mbp and chromosome 17 from 185-205 Mbp; Supporting Information Figure S2), but they displayed symmetric patterns of LD in different sample sets and were often associated with low sequence coverage, suggestive of centromeres or other heterochromatic regions. Introgression from another species can also form two distinct haplotype blocks and generate patterns similar to those of an inversion. However, gene flow and recombination will erode such patterns unless the introgression is recent.

Using genetic maps, we were able to validate one of the inversions (pet05.01) identified with population genetic data and provide additional support for three more based on suppressed recombination in putative inversion heterozygotes (pet09.01, pet11.01 and pet17.01). However, because the wild parents might have the same haplotype for pet07.01, pet09.02 and pet14.01, we were unable to corroborate them. This demonstrates one of the weaknesses of the genetic mapping approach – mapping will only detect a subset of segregating inversions. In contrast, approaches based on population genetic data provide a fine-grained and comprehensive way to search for potential inversions, and our methods appear to be robust.

Using RAD sequence data, we detected six structural variants identified from WGS data (Todesco et al., 2019) and one additional new putative inversion (pet09.02). We demonstrated that reduced representation sequencing data have the same power to detect inversions with SNP densities as low as 12 per Mbp. Moreover, with more extensive sampling across the habitat transition than that used by Todesco et al., we were able to better estimate population allele frequencies, as well as genetic divergence between ecotypes. We further demonstrated that these inversions are enriched in dune environment and that they correspond closely to genomic islands of differentiation at GSD (see below).

However, there are limitations to our approach for detecting inversions. First, while a population genomic approach such as that employed here can provide initial clues regarding the

existence of chromosomal inversions, additional independent evidence, such as comparative genetic mapping in this study or Hi-C sequencing analysis by Todesco et al. (2019), is needed to confirm the inversions for further investigation. Second, pinpointing the positions of breakpoints is not feasible given the low density of RAD markers. This can be challenging even with high-depth whole genome sequencing because of the abundance (typically) of repetitive sequences near breakpoints (Tang et al., 2015). Third, the limited genomic coverage of RAD sequence data, together with the dependence on deviations in population structure, biases detection towards large inversions with high sequence divergence. Therefore, it is not suitable for estimating the rate of origin and size distribution of chromosomal variants. However, it offers a convenient way to explore the evolutionary role of inversions because large and highly divergent inversions are also the ones that are most likely to play an important role in local adaptation and speciation. Lastly, we expect that the approach we described here could be further improved by better tuning of window size and outlier thresholds to match population sizes and SNP densities. Despite these limitations, our workflow provides a feasible and economical way of examining inversion frequencies and their evolutionary role in natural populations.

Inversions contribute to adaptive divergence

Previous work identified several large regions of differentiation that displayed signatures of divergent adaptation between dune and non-dune ecotypes in this system (Andrew &

579 Rieseberg, 2013). Our analyses showed that recombination is suppressed in these highly
580 divergent genomic regions due to chromosomal inversions. Increasing evidence suggests that
581 such islands of differentiation may be prevalent in incipient species (Turner et al., 2005; Michel
582 et al., 2010), and inversions have been shown to play an important role in maintaining
583 ecological and genetic divergence in the face of gene flow (Rieseberg, 2001; Noor et al., 2001;
584 Feder et al., 2003; Lowry & Willis, 2010). Our findings add to the growing body of case studies
585 on how structural chromosomal changes interact with local adaptation and gene flow to shape
586 the genomic landscape of divergence in early stages of speciation.
587
588 Analyses of inversion haplotype frequencies based on genotypes inferred from k-means showed
589 that all of the inversions are significantly enriched on the dunes (Table 1), suggesting that they
590 may be under selection, although for some inversions “non-dune” alleles are often found as
591 heterozygotes on the dunes. This could be due to differences in the kinds and strength of
592 selection on the inversions, but could also result from our sampling scheme. The individuals
593 used in the study were collected as seeds from mature plants, and thus reflected post-mating
594 population frequencies rather than that of living plants. If the inversions contribute to seedling
595 survival in dunes, then we likely are under-estimating frequency differences between ecotypes.
596 This is not implausible given that selection against immigrants is known to contribute strongly to
597 reproductive isolation in this system (Ostevik et al., 2016).
598

599 Additional evidence that the inversions contribute to local adaptation comes from the
600 observation that four of the inversions (pet05.01, pet09.01, pet11.01 and pet14.01) co-localize
601 with seed size QTLs identified in other work (Ostevik, 2016; Todesco et al., 2019). Large seeds
602 help plants survive burial in actively moving sand dunes (Donovan, Rosenthal, Sanchez-Velenosi,
603 Rieseberg, & Ludwig, 2010; Ostevik et al., 2016), and seed size is the most divergent phenotypic
604 trait between the ecotypes. These observations are further reinforced by the strong association
605 of pet05.01 with vegetation cover, which is negatively correlated with dune stability. Among the
606 inversion haplotypes associated with increased seed size, pet14.01 was in relatively low
607 frequency. However, this inversion underlies ecotype differentiation in another dune ecotype of
608 *H. petiolaris* (Todesco et al., 2019). Possibly, pet14.01 was only recently introduced to GSD, so it
609 will be interesting to monitor its frequency over the next 1-2 decades. Several inversions were
610 also found to be associated with soil variables in our GEA analyses. Sand dunes are
611 characterized by low nutrient availability, and a QTL for leaf N content maps to inversion
612 pet11.01 (Todesco et al., 2019), which we have shown to be associated with soil N in this study,
613 suggestive of a role in tolerance to low nutrients. Future mapping studies of related
614 physiological traits would help reveal the mechanistic basis by which inversions, especially
615 pet11.01, aid adaptation to low nutrient soils.
616
617 In the study by Todesco et al. (2019), multiple traits and soil characteristics were constantly
618 found associated with the same inversions in *H. petiolaris*. These signals could be caused by the
619 low number of samples in the dunes and the resulting selection-driven linkage of the inversions

among those samples. With denser sampling across the landscape, we were able to break the linkage of dune inversions and disentangle the effects in GEA. We show that various sets of inversions are responsible for different aspects of dune adaptation in this system.

The observation that inversions are associated with different traits and environmental factors in the dune habitat implies that the inversions are likely favored because they maintain combinations of locally advantageous alleles despite ongoing gene flow with non-adapted populations (Kirkpatrick & Barton, 2006). Models of parapatric and sympatric speciation have emphasized the importance of linkage between genes underlying local adaptation and those involved in reproductive isolation (Ortiz-Barrientos et al., 2016; Servedio, 2009; Noor et al., 2001). A key assortative mating barrier between the ecotypes is conspecific pollen precedence (Ostevik et al. 2016). Thus, a hypothesis going forward is that loci causing conspecific pollen precedence will also be located within one or more of these inversions

CONCLUSION

Using RAD sequencing data and a population genomic approach, we were able to detect multiple inversions *de novo* at low cost, determine their frequencies in natural populations, and assess their role in adaptation through GEA analyses. Localized heterogeneity of population structure caused by inversions has been detected in other systems using whole genome sequencing data (Li & Ralph, 2019). We show that inversions can also be detected with reduced

641 representation sequencing data with low SNP densities. Given the ever-expanding population
642 sequencing data available for non-model systems, we anticipate an explosion of inversion
643 reports across the plant and animal kingdoms, especially in systems where divergence appears
644 to have occurred in the face of gene flow.

ACKNOWLEDGEMENTS

We thank Qin Li for help in map plotting with R, and Shaghayegh Soudi for suggestions on BayPass. This work was supported by a China Scholarship Council scholarship (no. 201506380099) to K.H., a Killam Postdoctoral Fellowship to R.L.A., and NSERC grant (327475) to L.H.R.

REFERENCES

Andrew, R. L., Kane, N. C., Baute, G. J., Grassa, C. J., & Rieseberg, L. H. (2013). Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology*, *22*(3), 799-813.

Andrew, R. L., Ostevik, K. L., Ebert, D. P., & Rieseberg, L. H. (2012). Adaptation with gene flow across the landscape in a dune sunflower. *Molecular Ecology*, *21*(9), 2078-2091.

Andrew, R. L., & Rieseberg, L. H. (2013). Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution*, *67*(9), 2468-2482.

Arostegui, M. C., Quinn, T. P., Seeb, L. W., Seeb, J. E., & McKinney, G. J. (2019). Retention of a chromosomal inversion from an anadromous ancestor provides the genetic basis for alternative freshwater ecotypes in rainbow trout. *Molecular Ecology*, *28*(6), 1412-1427.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*(10), e3376.

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., ... & Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, *119*(6), 418.

Broman, K. W., Wu, H., Sen, Ś., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, *19*(7), 889-890.

Butlin, R. K. (2005). Recombination and speciation. *Molecular Ecology*, *14*(9), 2621-2635.

Burke, J. M., Lai, Z., Salmaso, M., Nakazato, T., Tang, S., Heesacker, A., ... & Rieseberg, L. H. (2004). Comparative mapping and rapid karyotypic evolution in the genus *Helianthus*. *Genetics*, *167*(1), 449-457.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), 7.

Coluzzi, M., Sabatini, A., Della Torre, A., Di Deco, M. A., & Petrarca, V. (2002). A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science*, *298*(5597), 1415-1418.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & McKenna, A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), 491.

Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., ... & Clark, R. (2018). Integrative detection and analysis of structural variation in cancer genomes. *Nature genetics*, *50*(10), 1388.

Dobzhansky, T., & Sturtevant, A. H. (1938). Inversions in the chromosomes of Drosophila pseudoobscura. *Genetics*, *23*(1), 28.

Donovan, L. A., Rosenthal, D. R., Sanchez-Velenosi, M., Rieseberg, L. H., & Ludwig, F. (2010). Are hybrid species more fit than ancestral parent species in the current hybrid species habitats?. *Journal of Evolutionary Biology*, *23*(4), 805-816.

Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., ... & Westram, A. M. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular ecology*, *28*(6), 1375-1393.

Feder, J. L., Chilcote, C. A., & Bush, G. L. (1988). Genetic differentiation between sympatric host races of the apple maggot fly Rhagoletis pomonella. *Nature*, *336*(6194), 61.

Feder, J. L., & Nosil, P. (2009). Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions?. *Evolution: International Journal of Organic Evolution*, *63*(12), 3061-3075.

Feder, J. L., & Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution: International Journal of Organic Evolution*, *64*(6), 1729-1747.

Feder, J. L., Roethele, J. B., Filchak, K., Niedbalski, J., & Romero-Severson, J. (2003). Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics*, *163*(3), 939-953.

Felsenstein, J. (1981). Skepticism towards Santa Rosalia, or why are there so few kinds of animals?. *Evolution*, *35*(1), 124-138.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*(4), 1555-1579.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Heslop-Harrison, J. (2013). *Pollen: development and physiology*. Londone, UK: Butterworth-Heinemann.

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation?. *Annual review of ecology, evolution, and systematics*, *39*, 21-42.

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*(1), 419-434.

Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., ... & Sonesson, A. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, *25*(10), 2130-2143.

Krimbas, C. B., & Powell, J. R. (1992). *Drosophila inversion polymorphism*. Boca Raton, FL: CRC press.

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., ... & Chen, W. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, *48*(1), 84.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987-2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289-304.

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, *8*(9), e1000500.

McGaugh, S. E., & Noor, M. A. (2012). Genomic impacts of chromosomal inversions in parapatric *Drosophila* species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 422-429.

Michel, A. P., Sim, S., Powell, T. H., Taylor, M. S., Nosil, P., & Feder, J. L. (2010). Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences*, *107*(21), 9724-9729.

Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., ... & Mallet, J. (2012). Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 343-353.

Navarro, A., & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, *57*(3), 447-459.

Noor, M. A., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, *98*(21), 12084-12088.

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375-402.

Ortiz-Barrientos, D., Engelstädter, J., & Rieseberg, L. H. (2016). Recombination rate evolution and the origin of species. *Trends in ecology & evolution*, *31*(3), 226-236.

Ostevik, K. L. (2016). *The ecology and genetics of adaptation and speciation in dune sunflowers* (Doctoral dissertation). University of British Columbia.

Ostevik, K. L., Andrew, R. L., Otto, S. P., & Rieseberg, L. H. (2016). Multiple reproductive barriers separate recently diverged sunflower ecotypes. *Evolution*, *70*(10), 2322-2335.

Ostevik, K. L., Samuk, K., & Rieseberg, L. H. (2019). Ancestral reconstruction of sunflower karyotypes reveals dramatic chromosomal evolution. *bioRxiv*, 737155.

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, *7*(2), e32253.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in ecology & evolution*, *16*(7), 351-358.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863-864.

Sedlazeck, F. J., Rescheneder, P., & Von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790-2791.

Servedio, M. R. (2009). The role of linkage disequilibrium in the evolution of premating isolation. *Heredity*, *102*(1), 51.

Tang, H., Lyons, E., & Town, C. D. (2015). Optical mapping in plant comparative genomics. *GigaScience*, *4*(1), s13742-015.

Taylor, J., & Butler, D. (2017). R package ASMap: efficient genetic linkage map construction and diagnosis. *arXiv preprint arXiv:1705.06916*.

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J. S., Soudi, S., Burge, D. O., ... & Rieseberg, L. H. (2019). Massive haplotypes underlie ecotypic differentiation in sunflowers. *bioRxiv*, 790279.

Trickett, A. J., & Butlin, R. K. (1994). Recombination suppressors and the evolution of new species. *Heredity*, *73*(4), 339.

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, *3*(9), e285.

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 451-460.

Weir, B. S. (1996). *Genetic data analysis II*. Sunderland, MA: Sinauer.

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in ecology & evolution*, *33*(6), 427-440.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wu, C. I. (2001). The genic view of the process of speciation. *Journal of evolutionary biology*, *14*(6), 851-865.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326-3328.

DATA ACCESSIBILITY

RAD sequencing data published previously: Dryad doi:10.5061/dryad.j2448.
Environmental data published previously: Dryad doi: 10.5061/dryad.158pb518.
Scripts and SNP data for genetic map construction are available upon request and will be set to GitHub and Dryad before publication, respectively.

AUTHOR CONTRIBUTIONS

K.H. and L.H.R. conceived the study; R.L.A. contributed genetic and environmental data; K.H. performed all the analyses; G.L.O. helped with the local structure analysis; K.L.O contributed to genetic map construction and synteny analysis; K.H. and L.H.R. wrote the paper; and all authors approved the final manuscript.

TABLES

TABLE 1 Clusters of MDS outliers obtained with "lostruct". MDS coordinates for which the outlier regions were identified, reference chromosomes with start and end positions of MDS outlier clusters, numbers of MDS outlier windows, variance explained by PC1 and PC2 in PCA of outlier regions, proportions of between-cluster sum of squares in k-means clustering, codes used in main text for putative inversions, as well as *P*-values of the "prop.test" for haplotype frequency differences between ecotypes are shown

MDS	Chromosome	Start (bp)	End (bp)	Number of outlier windows	PC1 variance (%)	PC2 variance (%)	Proportion of between-cluster sum of squares	Region code	*P*
MDS01	Ha412HOChr11	3587653	60627948	13	10.1	1.04	0.9851	pet11.01§	1.669×10^{-15}
MDS02	Ha412HOChr09	102388477	140632318	10	9.81	0.63	0.9914	pet09.01§	2.62×10^{-13}
MDS03	Ha412HOChr05	156436125	186198645	9	10.13	0.71	0.9898	pet05.01§	2.924×10^{-22}
MDS04	Ha412HOChr07	109423942	129416998	5	12.02	1.8	0.9746	pet07.01§	1.185×10^{-11}
MDS05	Ha412HOChr14	126811094	166275087	11	17.99	4.84	0.9521	pet14.01§	NA‡
MDS06	Ha412HOChr17	12368066	23002709	8	19.48	3.81	0.9709	pet17.01§	1.249×10^{-06}
MDS07	Ha412HOChr09	171481816	182472659	7	17.1	5.02	0.9282	pet09.02	0.02353
MDS12	Ha412HOChr13	116213599	135318474	5	15.96	6.91	0.9286	-†	-
MDS21	Ha412HOChr09	73794134	144545686	4	11.1	8.96	0.8486	-†	-

†Not included in downstream analyses because do not appear to represent inversions. ‡Only one genotype found in non-dune ecotype for this region. §Previously described by Todesco et al. (2019).

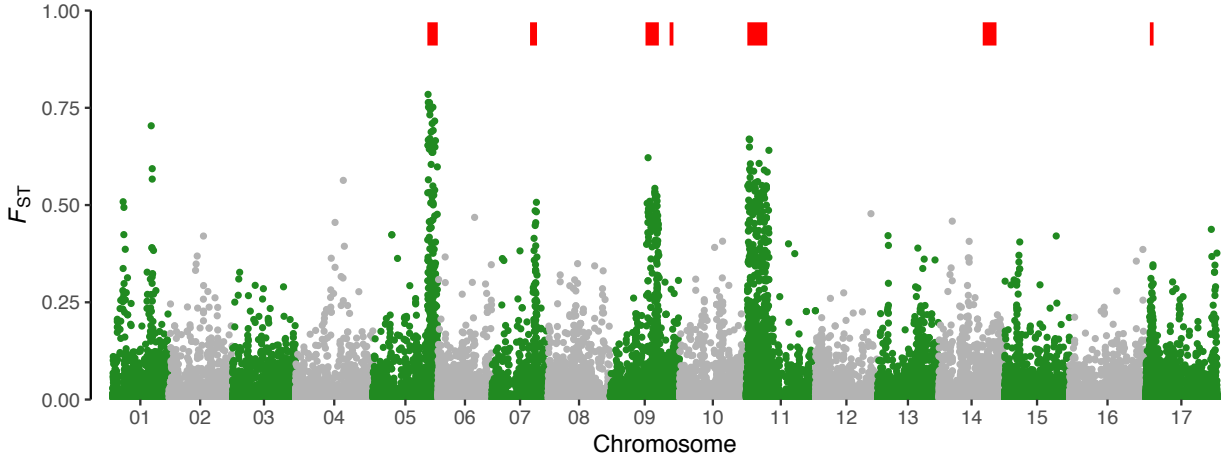
TABLE 2 Bayes factors of genome-environment association analyses with coverage and soil data for putative inversions treated as single loci. Asterisks indicate Bayes factors above significance thresholds computed with simulated POD samples

Variable	pet05.01	pet07.01	pet09.01	pet09.02	pet11.01	pet14.01	pet17.01
Grass	6.423*	-1.917	-4.443	-8.937	7.064*	-10.441	8.333*
Forb	13.942*	-5.516	-3.309	-8.141	-7.581	-7.601	-4.286
Debris	14.286*	-4.332	-4.853	-9.645	-4.219	-10.423	0.572
Cover	17.758*	-0.64	-2.146	-8.735	1.473	-9.978	11.316*
Cover PC1	20.125*	-0.471	-1.188	-7.905	0.016	-9.672	7.337*
Total N	4.216	6.822*	5.223*	2.62	9.635*	-6.687	-7.267
NO3-N	4.85	7.413*	4.64	2.319	10.517*	-6.482	-7.171
Ca	0.13	8.1*	-5.029	-6.64	-4.792	-6.115	-7.023
P	-1.725	4.957*	-3.374	-2.521	0.819	-8.546	-8.808
S	-3.023	4.078*	-4.504	-5.88	4.324*	-9.177	-8.886
Soil PC2	6.048*	0.5	1.364	-2.347	8.469*	-8.699	-5.121

Only the environmental variables with a significant association with at least one putative inversion are shown.

899 **FIGURES**

900



901

902 **FIGURE 1** Weir and Cockerham's F_{ST} between dune and non-dune ecotypes, as well as location
903 of putative inversions (indicated by red bars on top)

904

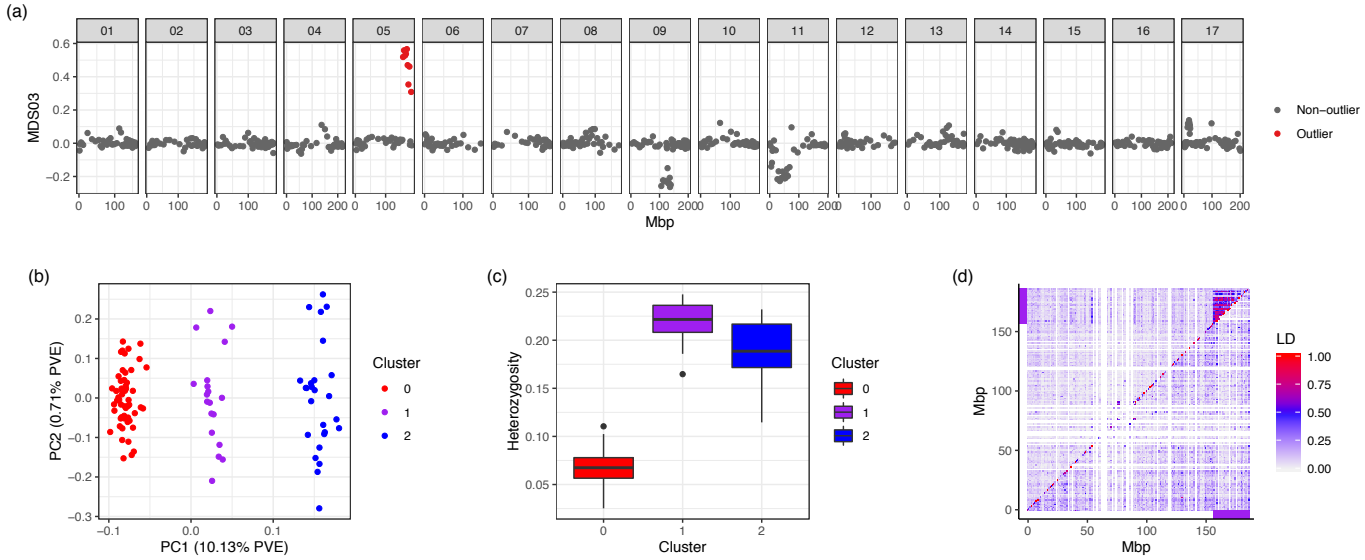


FIGURE 2 Characterization of the MDS outlier region on chromosome 5 (pet05.01). (a) Genome plot of corresponding MDS values across 17 reference chromosomes. Each dot represents a window of 50 SNPs, and outlier windows are highlighted in red. (b) PCA based on SNPs from outlier region. Three clusters identified using k-means clustering correspond to two homozygote groups (blue and red) and a heterozygote group (purple). (c) Heterozygosity for each of the groups identified in PCA. (d) LD plot for chromosome 5. Upper triangle with all individuals and lower triangle with only individuals homozygous for the more common orientation. SNPs were summarized and the second highest R^2 values were presented in 1 Mbp windows. Purple bars represent the location of the inversion

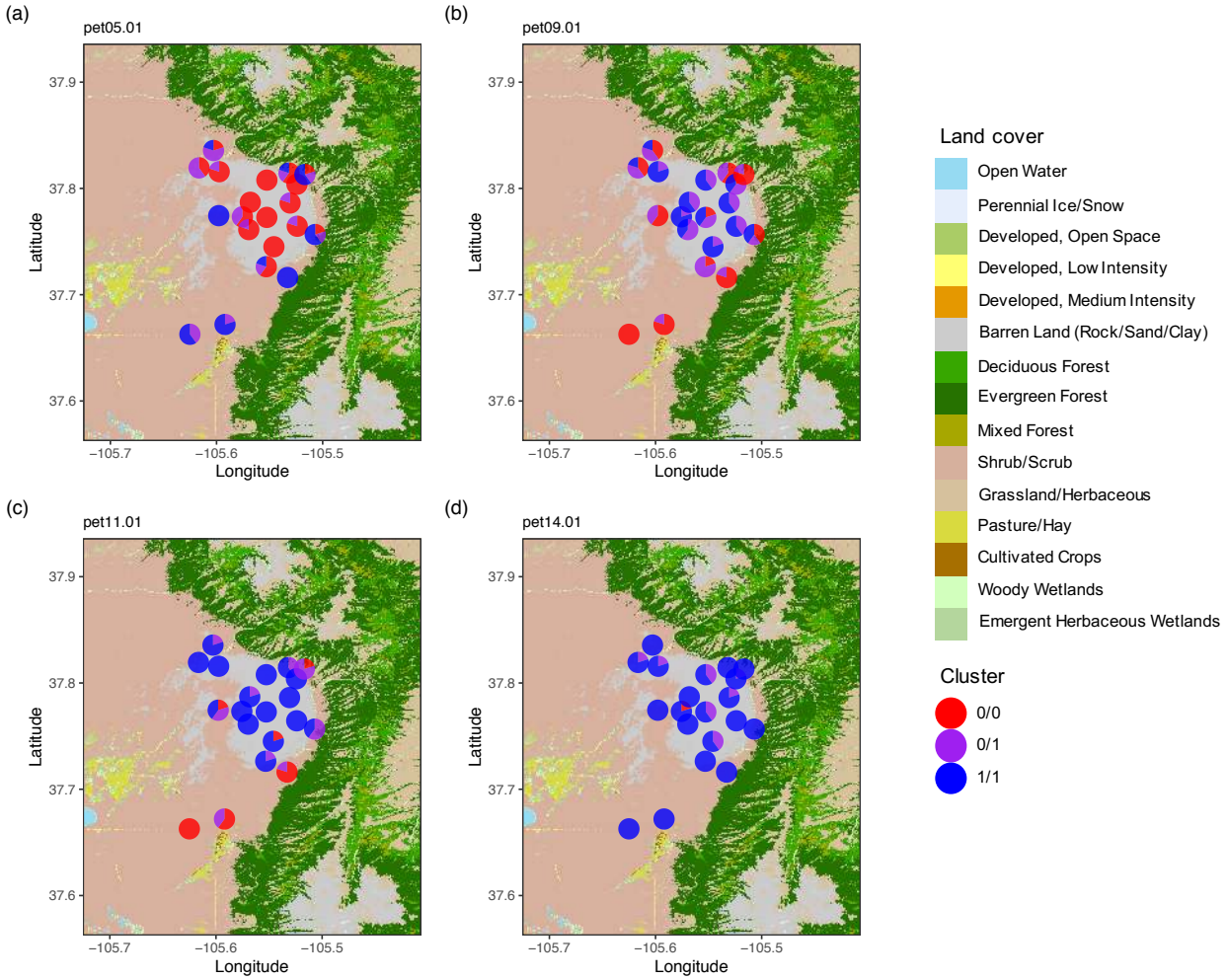


FIGURE 3 Map of Great Sand Dune National Park showing genotype distributions of (a) pet05.01, (b) pet09.01, (c) pet11.01 and (d) pet14.01. Genotypes are based on k-means cluster assignment in PCA. One of the haplotypes (inversion orientations) is more commonly found in dunes, which are represented by barren land surrounded by shrubby habitat in the map. Land cover classification downloaded from Multi-Resolution Land Characteristics Consortium (https://www.mrlc.gov/) at 30-m resolution

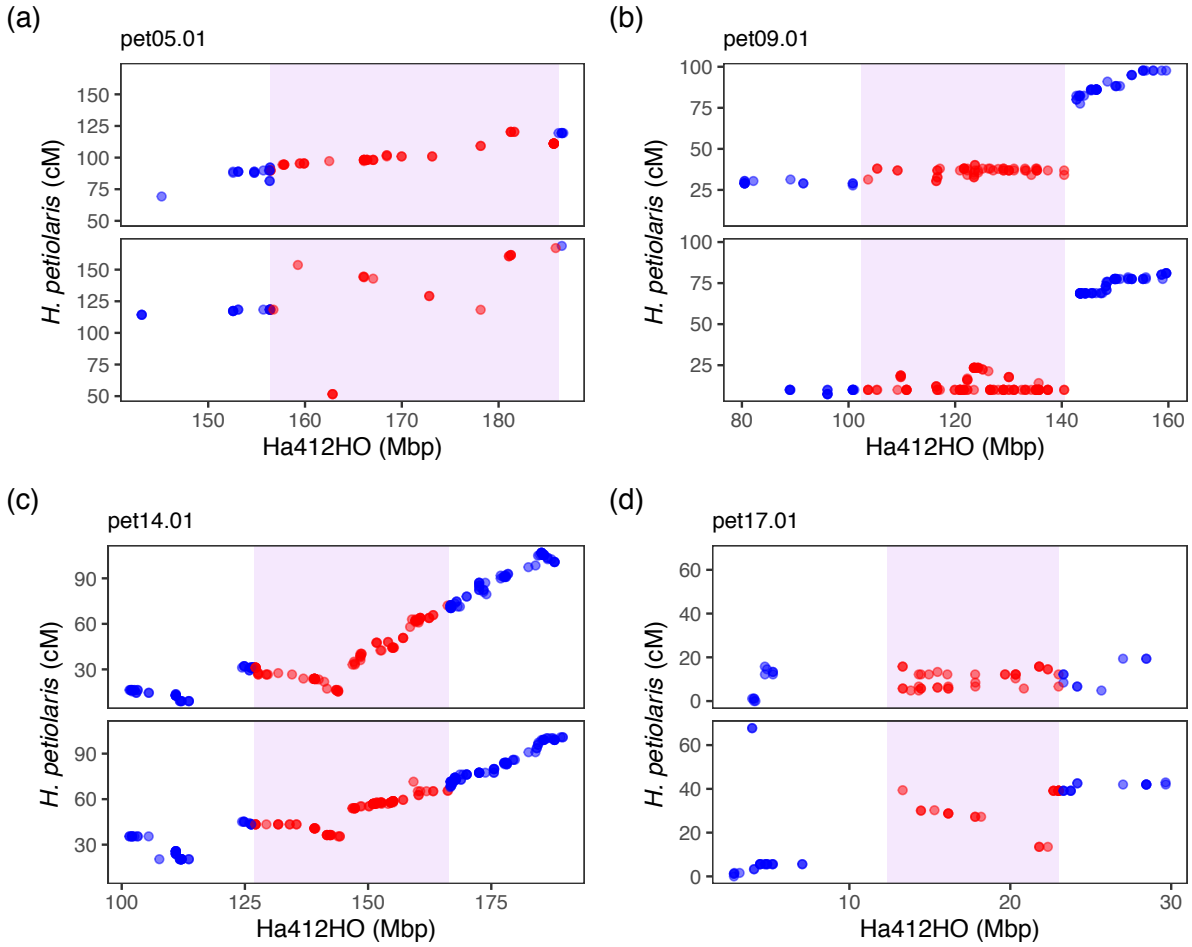


FIGURE 4 Genetic map comparisons for (a) pet05.01, (b) pet09.01, (c) pet11.01 and (d) pet17.01. Maps for non-dune (top panels) and dune (bottom panels) are plotted relative to the HA412HOv2 reference genome. Regions identified by lostruct and the markers that fall within them are highlighted in violet. Different patterns of marker orders are shown: reverse ordering between ecotypes for pet05.01 (a); recombination suppression in both maps for pet09.01 (b); similar forward ordering for pet14.01 (c); as well as recombination suppression in one map and reverse ordering in another for pet17.01(d)

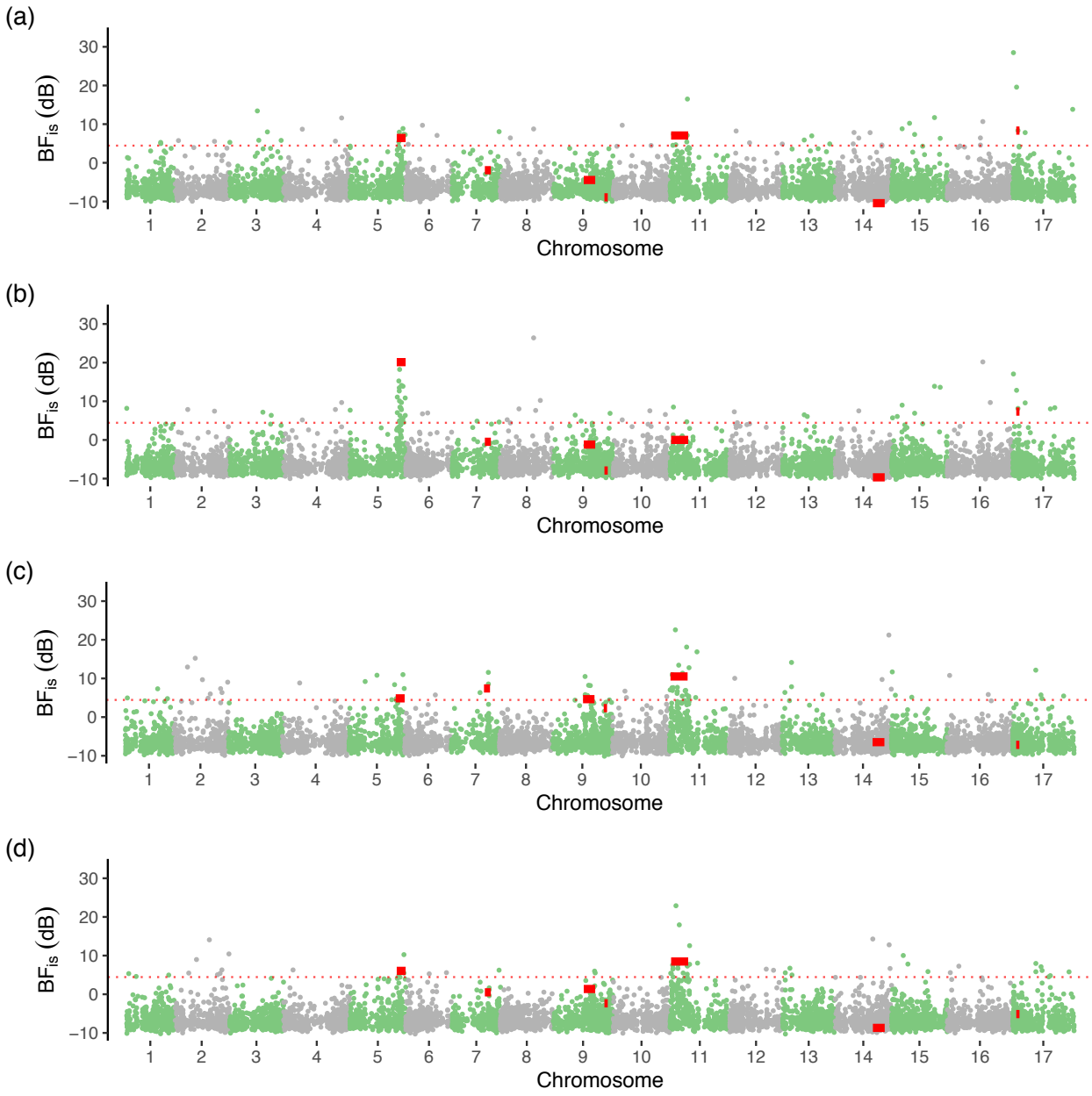


FIGURE 5 Genome-environment association for (a) % grasses, (b) coverage PC1, (c) soil NO_3 nitrogen and (d) soil PC2. Bayes factors (BF_{is}, in deciban unit) was estimated using the importance sampling estimator approach in BayPass. SNPs on different reference chromosomes are represented in alternate colors. The locations and BF_{is} values of 7 putative inversions are indicated by red solid bars. Red horizontal dashed lines represent 1% significance thresholds computed from simulated samples