



Genomic differentiation is initiated without physical linkage among targets of divergent selection in Fall armyworms

Ki Woong Nam, Sandra Nhim, Stéphanie Robin, Anthony Bretaudeau, Nicolas Nègre, Emmanuelle d'Alençon

► To cite this version:

Ki Woong Nam, Sandra Nhim, Stéphanie Robin, Anthony Bretaudeau, Nicolas Nègre, et al.. Genomic differentiation is initiated without physical linkage among targets of divergent selection in Fall armyworms. 2018. hal-02788771

HAL Id: hal-02788771

<https://hal.inrae.fr/hal-02788771>

Preprint submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 Genomic differentiation is initiated without physical linkage among targets 2 of divergent selection in Fall armyworms

3
4 Kiwoong Nam^{1*}, Sandra Nhim¹, Stéphanie Robin^{2,3}, Anthony Bretaudeau^{2,3}, Nicolas Nègre¹,
5 Emmanuelle d'Alençon¹
6

7 ¹DGIMI, INRA, Univ. Montpellier, 34095, Montpellier, France

8 ²INRA, UMR-IGEPP, BioInformatics Platform for Agroecosystems Arthropods, Campus Beaulieu,
9 Rennes, 35042, France

10 ³INRIA, IRISA, GenOuest Core Facility, Campus de Beaulieu, Rennes, 35042, France

11 * corresponding author (ki-woong.nam@inra.fr)

12 ABSTRACT

13 The process of speciation involves the differentiation of whole genome sequences. Gene flow
 14 between population impedes this process because recombination in hybrids homogenizes
 15 sequences. Accumulating empirical cases demonstrate that speciation indeed occurs in the presence
 16 of gene flow, and several speciation models have been proposed to explain the process of whole
 17 genome differentiation. Polymorphism patterns from a pair of very recently diverged taxa may
 18 provide insightful information to identify critical evolutionary forces enabling genomic
 19 differentiation. The Fall armyworm, *Spodoptera frugiperda* is observed as two sympatric strains,
 20 corn strain, and rice strain, named from their preferred host plants, throughout the entire range of
 21 habitats. The difference in host-plant ranges suggests a possibility of ecological divergent selection.
 22 In this study, we analyzed whole genome sequences from these two strains from Mississippi based
 23 on population genetics approaches and *de novo* genome assembling to study initial steps toward
 24 genomic differentiation. The genomic F_{st} is low (0.017), while 91.3% of 10kb windows have F_{st}
 25 greater than 0, suggesting genome-wide differentiation with a low extent. Principal component
 26 analysis and phylogenetic analysis show that corn strains were derived from ancestral rice strains
 27 and these two strains experienced population expansion with a greater extent in corn strains. We
 28 identified only three strain-specific chromosomal rearrangements, and the role these rearrangements
 29 in genomic differentiation is not supported. We identified 423 and 433 outliers of genetic
 30 differentiation between strains from the mappings against the reference genomes of corn and rice
 31 strains, respectively. Among them, four and nine outliers have a higher level of absolute sequence
 32 divergence (d_{XY}) than genomic average from these two mappings, and these outliers contain genes
 33 related to female fecundity. The rest of outliers have a reduced level of genetic diversity suggesting
 34 signatures of selective sweeps. In these outliers, corn strains have diverged genotypes from rice
 35 strains, and this divergence is observed only from the flanking sites in which the distance from the
 36 nearest outliers is less than 1kb, implying that physical linkage among outliers is unimportant for
 37 genomic differentiation. Gene density is negatively correlated with nucleotide diversity, but not
 38 with F_{st} . This result suggests that, while the level of local genetic diversity is affected by the
 39 strength of selection, selection is not a primary source of local variation in genomic differentiation,
 40 and genomic reduction in migration rate is the most likely reason for genomic differentiation. From
 41 these results, we proposed that in *S. frugiperda* divergence female fecundity traits caused the
 42 initiation of genetic difference and that following divergent selection targeting many loci results in
 43 the reduction in genomic migration rate, which creates genome-wide genetic differentiation. This
 44 explanation is in line with the genome hitchhiking model.

45

46 INTRODUCTION

47 Gene flow impedes the process of speciation because recombination in hybrids homogenizes
48 sequences between populations¹. An exceptional condition is, therefore, necessary to overcome the
49 homogenizing effect of gene flow (reviewed in ²). As speciation processes inherently involve
50 genomic differentiation by reproductive barriers, generated through collective or sequential actions
51 of evolutionary forces³, how the homogenizing effect of recombination is overcome throughout
52 whole genomes is a key issue to understand the speciation process⁴. Accumulating empirical reports
53 show that speciation indeed occurs in the presence of gene flow⁵, implying that the homogenizing
54 effect of recombination can be effectively overcome.

56 Divergent selection is one of the main players occurring during the process of speciation. If
57 selection is sufficiently strong (i.e, $s > m^6$ or $s > r^7$, where s , m , and r are selection coefficient,
58 migration rate, and recombination rate, respectively), the effect of selection dominates that of gene
59 flow and recombination, thus genomic differentiation may not be hampered by gene flow. If
60 selection is weak ($s < m$ and $s < r$), other conditions are necessary for genomic differentiation.
61 Physical linkage among the targets might be responsible for genomic differentiation, as selective
62 sweeps⁸ increase in the level of genetic differentiation at sites physically linked to the targets of
63 divergent selection. For example, if divergent selection targets a large number of loci, then the
64 average physical distance from a neutral locus to the targets decreases, thus whole genome
65 sequences can be differentiated by the concerted actions of divergent selection⁹. In another
66 speciation model, termed divergence hitchhiking, if a locus is targeted by strong divergent selection,
67 then the effective rate of migration is reduced in this region, and following events of divergent
68 selection targeting sequences within this region may generate a long stretch of differentiated DNA
69 (up to several Mb)^{10,11}. Population-specific chromosomal rearrangements can also contribute to the
70 process of speciation because recombination is inhibited in hybrids^{12–15} and physical linkage
71 between targets of divergent selection and the loci with a chromosomal rearrangement may create
72 long genomic regions with differentiation¹⁶. Whole genome sequences may be also differentiated
73 without physical linkage among targets of selection. According to the genome hitchhiking model, if
74 divergent selection targets many loci, then genome-wide migration rate is effectively reduced, and
75 whole genome sequences can be differentiated^{17,18}.

77 If the number of targeted loci is sufficiently high, genomic differentiation may occur rapidly. The
78 loci targeted by population-specific divergent selection may have correlated allele frequencies, and
79 corresponding linkage disequilibrium among targets will be then generated^{6,19,20}. Theoretical
80 studies^{6,19} show that if the number of targets is higher than a certain threshold, targeted loci have a
81 synergistic effect in increasing linkage disequilibrium among targets, thus genomic differentiation is
82 consequently accelerated. This non-linear dynamics of genomic differentiation according to the
83 number of occurred selection events were termed genome-wide congealing²¹. It should be noted that
84 any diversifying factors, including divergent selection, background selection, and assortative
85 mating²², may contribute to genome-wide congealing. Thus, the critical question of how genomic
86 differentiation occurs in the presence of gene flow is the condition for the transition to the phase of
87 genome-wide congealing. For example, divergence hitchhiking may provide a condition for post-
88 genome-wide congealing phase⁴. Alternatively, genome-wide reduction in migration rate (genome
89 hitchhiking) or chromosomal rearrangement may contribute to this condition as well.

91 Divergence hitchhiking model has been supported by pea aphids¹⁰, stickleback²³, and poplar²⁴.
92 However, as Feder and Nosil demonstrated¹⁸, long differentiated sequences can be observed only
93 from a specific condition, when effective population size (Ne) and migration rate are low ($Ne =$
94 $1,000$, $m = 0.001$), and selection is very strong ($s = 0.5$). Isolation by adaptation, a positive
95 correlation between a genetic difference and adaptive divergence^{25,26}, has been presented as a

support for genome hitchhiking, which indeed causes isolation by adaptation⁴. However, it is still unclear whether genome hitchhiking initiates or reinforces genetic differentiation in cases of isolation by adaptation.

The Fall armyworm, *Spodoptera frugiperda*, (Lepidoptera, Noctuidae) is a pest species observed as two sympatric strains, corn strain (sfC hereafter) and rice strain (sfR) named from their preferred host-plants, throughout North and South American continents²⁷. Based on maker-genotyping, these two strains appear to have different DNA sequences^{27,28}. In a wide geographical range in North America, 16% of individuals were reported to be hybrids²⁹, suggesting frequent gene flow between strains. In our previous study, we observed that these two strains have a weak but significant genomic differentiation ($F_{st} = 0.019$, $p < 0.005$), and that the differentiated loci were distributed across the genome³⁰. Low level of genomic differentiation and widespread occurrence of hybrids make these two strains an ideal system to explore critical evolutionary forces for genomic differentiation in the presence of gene flow. Whole genome differentiation between sfC and sfR might involve both premating reproductive isolation through assortative mating^{31–33}, or postmating reproductive isolation by ecological divergent selection, or by reduced hybrid fertility³⁴.

In this study, we aim at identifying evolutionary forces that are responsible for genomic differentiation between sfC and sfR at the very initial stage of the speciation process. Using resequencing data generated in our previous study³⁰, we test the role of several evolutionary events in genomic differentiation, including chromosomal rearrangements, physical linkages among targeted loci, and genomic reduction in migration rate. The results presented here allow us to infer the evolutionary history explaining the genomic differentiation between strains in *S. frugiperda*.

RESULTS

In order to accurately detect signatures of genome divergence, it is important to have a contiguous reference genome assembly. The reference genome assemblies for sfC and sfR generated from our previous study contain 41,577 and 29,127 scaffolds, respectively³⁰ (Table 1). To improve the reference genome sequences we performed *de novo* genome assembly from Pac-bio reads (27.5X and 33.1X coverages for sfC and sfR, respectively). Errors in these reads were corrected by Illumina assemblies, which were generated from the reads used in our previous study³⁰. The Pac-bio reads were assembled using SmartDenovo³⁵, and scaffolding was performed using Illumina paired-ends and mate-pairs used in our previous study. The resulting assemblies are now closer to the expected genome sizes, 396 ± 3 Mb, estimated by flow cytometry³⁰ (Table 1). Moreover, the contiguity is also significantly improved, as N50 is 900kb and 1,129kb for corn and rice reference genome sequences, respectively. The numbers of sequences are 1,000 and 1,054 for sfC and sfR, respectively. BUSCO analysis³⁶ shows that the correctness is also increased, especially for the sfC (Supplementary Table 1). The numbers of identified protein-coding genes are 21,839 and 22,026 for sfC and sfR, respectively. BUSCO analysis shows that gene annotation is also improved, especially for sfC (supplementary Table 2).

Resequencing data from nine female individuals from each of corn and rice strains collected in the wild³⁰ were mapped against these two nuclear reference genome assemblies using bowtie2³⁷ with very exhaustive search parameters (see methods). As one individual from rice strain has a particularly low mapping rate and an average read depth (denoted as R1, Gouin et al³⁰) (Supplementary Figure 1), we excluded this individual from the following analysis. Variants were called using samtools mpileup³⁸, and we performed stringent filtering by discarding all sites unless Phred variant calling score is higher than 40 and genotypes are determined from every single individual. The numbers of variants are 48,981,416 from 207,415,852 bp and 49,832,320 from 205,381,292 bp from the mapping against sfC and sfR reference genomes, respectively. As analyses

from the resequencing data might be affected by ascertainment bias, we performed all analyses based on these two reference genomes. We present the results only from the sfC reference genomes in the main text unless mentioned specifically. We show the results from the sfR reference genome in the supplementary information (Supplementary Figure 14-21).

The genome-wide F_{st} calculated between sfC and sfR is 0.017, which is comparable to our previous study (0.019)³⁰. As this low level of differentiation could be caused by chance, we calculated F_{st} from randomized groupings with 500 replications. We observed that no randomized grouping has higher F_{st} than the grouping according to strains (equivalent to $p < 0.002$), thus we concluded that the genomic sequences are significantly differentiated between strains, as we did in our previous study³⁰. This genomic differentiation can be either caused by a few loci with a very high level of differentiation or by many loci with a low level of differentiation. To test these two possibilities, we calculated F_{st} in 10 kb window. Among total windows, 91.3% of these windows have F_{st} greater than 0 (Figure 1), supporting the latter explanation. The low level of genetic differentiation implies that these two strains do not experience genome-wide congealing yet.

Genetic relationships among individuals were inferred using principal component analysis (PCA). The result shows that sfR has a higher genetic variability among individuals than sfC, and we hypothesized that sfC was derived from ancestral sfR (Figure 2a). To test this hypothesis, we reconstructed a phylogenetic tree using assembly-free approach³⁹ with *S. litura*⁴⁰ as an outgroup. The resulting tree shows that sfC individuals constitute a monophyletic group, implying that the sfC was indeed derived from ancestral sfR (Figure 2b). The pattern of the phylogenetic tree is subtly different from that of PCA. The phylogenetic tree shows that sfC has monophyly, implying that the sfC individuals were derived from a single individual. However, the result from PCA does not support the single origin of sfC. This discrepancy is perhaps caused by an incomplete lineage sorting in the ancestry of sfC or by frequent gene flow between sfC and sfR. However, we cannot exclude a possibility that statistical artifacts, such as long-branch attractions⁴¹. The genetic relationship among individuals was also analyzed from ancestry coefficient⁴², and we observed that distinct origins of sfC and sfR are not supported (Supplementary Figure 2).

We tested the possibility of an extreme case where both sfC and sfR have monophyly, but all identified sfR individuals except R6 on Figure 2b are F1 hybrids between sfR females and sfC males. In this case, maternally-derived mitochondrial CO1 genes used to identify strains in this study³⁰ will have distinctly different sequences between R2-R9 and C1-C9, while paternally derived sequences will not show such a pattern between these two groups except R6. As all individuals analyzed in this study are females, the Z chromosomes were derived from males in the very previous generation. Thus, we tested significant genetic differentiation of Z chromosomes between sfC and sfR without R6. TPI gene is known to be linked to Z chromosomes in *S. frugiperda*⁴³, and we observed this gene from Contig269 by blasting. This contig is 3,688,019bp in length, and the number of variants is 201,075. The F_{st} calculated between sfC and sfR without R6 is 0.061, which is higher than the genomic average (0.017). We calculated F_{st} from randomized groupings with 500 replicates, and only four replicates have F_{st} higher than 0.061, corresponding p-value equal to 0.008. This result demonstrates a significant genetic differentiation of paternally derived Z chromosomes between strains identified by mitochondrial sequence, and we exclude the possibility of the extreme case with F1 hybrids.

We inferred changes in N_e from two statistics, π and Watterson's θ . Watterson's θ represent more recent levels of genetic diversity than π . The calculated π is 0.043 and 0.044 for sfC and sfR, respectively. The π is not significantly different between sfC and sfR ($p=0.27$, permutation test with 100 randomizations). The calculated Watterson's θ is 0.064 and 0.061 for sfC and sfR,

196 respectively, and sfC has higher Watterson's θ than sfR ($p < 0.01$). This result indicates that both
197 sfC and sfR experienced population expansion with a greater extent in sfC, possibly due to higher
198 fitness in sfC.

199
200 Chromosomal rearrangements specific to a single population can cause a genetic differentiation
201 because recombination is inhibited in hybrids^{12,13,15}. Thus, we estimated the role of chromosomal
202 rearrangements in genomic differentiation by identifying the propensity of strain-specific
203 chromosomal rearrangements. Using BreakDancer⁴⁴ we identified 1,254 loci with chromosomal
204 inversions, with 1,060bp in median sequence length. We considered that a chromosomal
205 rearrangement is strain-specific if the difference in allele frequency is higher than an arbitrarily
206 chosen criterion, 0.75. F_{st} calculated from these inversions are lower than zero (-0.063 and -0.064),
207 meaning that the contribution of chromosomal inversion to genetic differentiation is not supported.
208 The number of inter-scaffold rearrangement is 1,724, and only one of them has a difference in allele
209 frequency higher than 0.75. F_{st} calculated from 10kb flanking sequences of the breaking points is
210 lower than zero (-0.115 and -0.0783 at each side). Thus, we excluded the possibility that
211 chromosomal rearrangement is a principal cause of genomic differentiation.

212
213 Then, we test the possibility that selection is responsible for genomic differentiation from outliers of
214 genetic differentiation. We used correlated haplotype score⁴⁵ to estimate the level of genetic
215 differentiation between strains. If each of minimum 100 consecutive SNPs in minimum 1kb has a
216 significantly greater haplotype score than the rest of the genome ($p < 0.001$), we defined this locus
217 as an outlier. As the mapping rate of reads against highly differentiated sequences is necessarily low,
218 the identification of outliers can be severely affected by the usage of reference genome sequences.
219 Therefore, here we present the results from both corn and rice reference genome sequences (refC
220 and refR). In total, 433 outliers at 170 scaffolds and 423 outliers at 148 scaffolds were identified
221 from the mappings against refC and refR, respectively. The average length of these outliers is
222 4,023bp and 4,095bp for refC and refR, respectively. The longest outlier is 27,365bp and 33,110bp
223 in length for refC and refR, respectively. These outliers occupy only small fractions of the scaffolds
224 (1.56% and 1.82% for refC and refR, respectively), suggesting that extremely strong selective
225 sweeps are not supported. Thus, it is unlikely that very strong selection targeting these regions
226 causes whole genome differentiation.

227
228 We test the possibility of the divergence hitchhiking¹⁰, a hypothesis that a strong selection creates
229 DNA sequences with reduced local migration rate, and following selection events within this
230 sequence generates a long stretch of DNA sequence with an elevated level of genetic differentiation.
231 According to this speciation model, lowly differentiated sequences between highly differentiated
232 sequences are generated by ancestral polymorphisms, rather than gene flow¹⁰. Thus, these lowly
233 differentiated sequences between highly differentiated sequences will show clustered ancestry maps
234 according to the extant strains, whereas the rest of lowly differentiated sequences in the genome
235 will not show such a clustering. From the scaffolds with the outliers, we identified lowly
236 differentiated sequences (hapflk score < 1 , Supplementary Figure 3 to see the histogram of all
237 positions at these scaffolds), 154,163bp and 273,797bp in total size from refC and refR,
238 respectively. Then, sNMF software was used to infer ancestry coefficients⁴². Figure 3 shows that sfC
239 and sfR have different ancestry at outliers, while the lowly differentiated sequences within the
240 scaffolds with outliers do not show any apparent clustering according to extant strains. Thus,
241 divergence hitchhiking is not supported by our data.

242
243 If a genetic locus is resistant against gene flow from the beginning of genetic differentiation, this
244 sequences is expected to show a higher level of absolute genetic divergence, which can be estimated
245 from d_{XY} statistics⁴⁶. We observed that four out of the 433 outliers from refC and nine out of the 423

outliers from refR have higher d_{XY} than genomic average (FDR corrected $p < 0.05$) (Supplementary Figure 4, 5). We denote these outliers as genomic islands of divergence in this paper. These genomic islands of divergence contain three and four protein-coding genes from refC and refR, respectively. These genes include NPRL2 and Glutamine synthetase 2. NPRL2 is a down-regulator of TORC1 activity, and this down-regulation is essential in maintaining female fecundity during oogenesis in response to amino-acid starvation in *Drosophila*⁴⁷. Glutamine synthetase 2 is important in activating TOR pathway, which is the main regulator of cell growth in response to environmental changes to maintain fecundity in plant hoppers⁴⁸. This result raises the possibility that disruptive selection on female fecundity is responsible for initiating genetic differentiation between strains. The function of the other five genes is unclear, thus other traits might be important in initiating genomic differentiation as well.

If genetic differentiation is initiated by selection on female fecundity, mitochondrial genomes will show a higher level of absolute level of sequence divergence than nuclear genome because mitochondrial genomes are transmitted only through the maternal lineage. We performed mapping all reads against mitochondrial genomes (KM362176) and identified 371 variants from 15,230bp. The result from PCA shows that, contrary to the nuclear pattern, sfC and sfR individuals fall into two distinct groups (Figure 4a). Ancestry coefficient analysis shows that each of two strains has a distinct ancestry (Figure 4b) (see Supplementary Figure 6 to find a correlation between K and cross entropy). To generate a mitochondrial phylogenetic tree, we extracted sequences of *S.frugiperda* from mitochondrial Variant Call Format file, and we created a multiple sequence alignment together with the mitochondrial genome sequence of *S.litura* (KF701043). Then, a phylogenetic tree was reconstructed using the minimum evolution approach⁴⁹. The tree shows that sfC and sfR are a sister group of each other (Figure 4c). This mitochondrial pattern is also observed from other studies in *S.frugiperda*^{28,30,50}. We excluded a possibility that strong linked selection on mitochondrial genomes alone causes the different phylogenetic pattern between nuclear and mitochondrial genomes because in this case the topology is expected to be unchanged while only relative lengths of ancestral branches to tips are different between nuclear and mitochondrial trees (Supplementary Figure 7). Instead, this pattern can be explained by an ancient divergence of mitochondrial genomes, which is followed by a gradual genetic differentiation of nuclear genomes.

A molecular clock study shows that the mitochondrial genomes diverged between sfC and sfR two million years ago²⁸, which corresponding 2×10^7 generations according to the observation from our insectarium (10 generations per year). Assuming that the N_e is 4×10^6 for both strains, the number of generations during this mitochondrial divergence time is five times of N_e . We performed a simple forward simulation⁵¹ with a wide range of migration rate to test this divergence time can explain the level of observed genetic differentiation ($F_{st} = 0.017$). No simulation generates F_{st} equal or lower than 0.017 (Supplementary Figure 8), supporting that mitochondrial genomes diverged more anciently than nuclear genomes.

We investigated the role of the rest of outliers, denoted by genomic islands of differentiation in this paper. Genomic islands of differentiation have much lower π than the genomic average in both strains (Supplementary Figure 9), and sfC has a lower π than sfR ($p = 0.0007$; Wilcoxon rank sum test), suggesting that the genomic islands of differentiation were targeted by linked selection, as a form of selective sweeps⁸ or background selection⁵², with a greater extent in sfC. d_{XY} calculated from genomic islands of differentiation is on average lower than the genomic average (Supplementary Figure 10), suggesting that these sequences were targeted by linked selection targeted after the split between sfC and sfR. PCA from genomic islands of divergence and genomic islands of differentiation shows that these two types of genomic islands have a clear grouping according to strains (Figure 5), which was observed from mitochondrial genomes (Figure 4a) but

not from nuclear genomes (Figure 2a). Interestingly, the sequences of genomic islands of divergence have comparable genetic variability between sfC and sfR, whereas sfC has a lower genetic variability in the sequence of genomic islands of differentiation than sfR. From these results, we concluded that the sfC diverged from sfR by linked selection.

We investigated the role of physical linkage by performing PCA with varying distances to the nearest genomic island of differentiation. When the distance is less than 1kb, genetic variations of sfC individuals are included within the range of genetic variation of sfR individuals (PC1 of the leftmost panel at Figure 6), while divergence of sfC from sfR is also supported (PC2 of the leftmost panel at Figure 6). If the distance is higher than 1kb, the divergence of sfC from sfR is not observed (Figure 6), suggesting that the effect of physical linkage to genomic islands of differentiation disappears rapidly as the distance increases. The short linkage disequilibrium in a species with large N_e is expected from a theoretical analysis¹⁸, and reported from empirical cases^{53,54}. These results show that physical linkages among targets of linked selection are not the primary cause of genomic differentiation.

Then, we tested a possibility of genome hitchhiking^{17,18}, a hypothesis stating that genomic differentiation is caused by a genome-wide reduction in migration rate due to many loci under selection. If the strength of selection determines the level of genetic differentiation, a positive correlation between F_{st} and the strength of selection is expected. Alternatively, if a genomic reduction in migration rates dominates the effect of selection, this correlation is not expected. We assume that the exon density is a proxy for the strength of selection. Exon densities calculated in 100kb window are negatively correlated with π (Spearman's $\rho = -0.211$, $p < 2.2 \times 10^{-16}$) (Figure 7), showing that the local genetic diversity pattern is affected by selection. F_{st} , however, is not significantly correlated with exon density ($\rho = 0.021$, $p = 0.2032$) (Figure 7). This result supports the hypothesis that a genomic reduction in migration rate dominates the variation of genetic differentiation.

In principle, both selective sweeps and background selection may target these genomic islands of differentiation as linked selection. Background selection may cause genetic differentiation between populations only if these two populations are *a priori* differentiated by a geographical separation or a tight physical linkage to a target of selective sweeps. As sfC and sfR are sympatrically observed and the physical linkage among genomic islands of differentiation is not supported as shown above, we assume that selective sweeps are mainly responsible for the genomic islands of differentiation and inferred traits under adaptive evolution from the function of genes within genomic islands of differentiation. These islands contain 275 and 295 protein-coding genes from refC and refR, respectively (the full list can be found from Supplementary Table 3-4). These protein-coding sequences include a wide range of genes important for the interaction with host-plants, such as P450, chemosensory genes, esterase, immunity gene, and oxidative stress genes³⁰ (Table 2), suggesting that ecological divergent selection is important for genomic differentiation. Interestingly, *cyc* gene, which plays a key role in circadian clock⁵⁵, is also included in the list of the potentially adaptively evolved genes. Thus, divergence selection on *cyc* may be responsible for pre-mating reproductive isolation due to allochronic mating time^{31,32}.

A QTL study shows that genetic variations in *vrlle* gene can explain differentiated allochronic mating behavior in *S.frugiper*³¹. This gene is not found in the outliers. F_{st} calculated from a 10kb window containing this gene is 0.017 and 0.016 for refC and refR, respectively, which is similar to genomic average (0.017). Thus, it appears that this gene does not have a direct contribution to genomic differentiation.

DISCUSSION

In this study, we showed that genetic differentiation between strains in *S. frugiperda* is initiated by the divergence of genes associated with female fecundity from the gene list in the genomic islands of divergence (Figure 8 to see a possible evolutionary scenario of genetic differentiation between sfC and sfR). Afterward, divergent selection targeting many loci appears to reduce the genome-wide migration between strains, which have low but significant genome-wide genetic differentiation, in line with the genome hitchhiking model^{17,18}. The physical linkage among targets of linked selection appears to be unimportant for genomic differentiation in *S. frugiperda*. We observed that genomic islands of differentiation contain genes associated with interaction with host-plants, thus the adaptive evolution of this ecological trait appears to promote genomic differentiation between strains. A circadian gene (*cyc*) is also found from a genomic island of differentiation, and it is unclear whether this gene is associated with the assortative mating due to allochronic mating patterns in *S. frugiperda*. If this is true, both divergent selection and assortative mating generate genomic differentiation by a genomic reduction in migration rate between strains, since assortative mating generates the same footprints on DNA sequences as divergent selection.

The heterozygosity of these strains is unprecedented high, as the calculated π is 0.043-0.044. In two other Noctuid pests, *S. litura* and *Helicoverpa armigera*, π calculated from multiple populations across their distribution area ranges from 0.0019 to 0.016⁴⁰, and from 0.008 to 0.01⁵⁶, respectively. *Heliconius melpomene*, a butterfly species, has π between 0.021 and 0.029⁵⁷. To explain the extremely high level of heterozygosity in *S. frugiperda*, we first checked the possibility that a considerable proportion of identified variants is false positives. We performed additional filterings, on the top of applied ones, by including additional 12 criteria. These additional filterings discarded only 34 out of 48,981,416 and 17 out of 49,832,320 variants from the mapping against refC and refR, respectively. Thus, we exclude the possibility that false positives caused the high level of heterozygosity. We inferred past demographic history using pairwise sequentially Markovian coalescent⁵⁸ based on assumptions that generation time is the same with lab strains at our insectarium (10 generation/yr) and mutation rate is the same with *H. melpomene* (2.9×10^{-9} /site/generation)⁵⁹. Extremely rapid population expansions were inferred from both two strains (N_e was increased from 9.6×10^5 to 1.2×10^7) between 10 mya and 100 mya (Supplementary Figure 11). A possible explanation of this rapid expansion is the merge of genetically diverged ancestral populations by hybridization. In this scenario (Figure 8), two populations were separated by geographical barriers and genetically differentiated. At some moment, the geographical barriers were removed, and these populations started to be merged by hybridization. As the merged population maintains a large proportion of variants, this population has a high level of heterozygosity. This population is extant sfR. Afterward, a group of sfR started to diverge by ecological divergent selection, and assortative mating and this group became the extant sfC.

The pattern of genomic differentiation can be different among geographic populations. For example, pairs of different geographical populations may have different levels of genomic differentiation (F_{st}). The genomic islands of differentiation can be also different if a proportion of divergent selection is specific to a single geographical population, thus it is worthwhile to test if the same loci are identified as genomic islands of divergence across diverse geographic populations. If levels of genomic differentiation vary among different geographical populations in *S. frugiperda*, it might be possible to find a pair of strains that enter to a phase of genome-wide congealing. Attempts to find the process towards complete genomic differentiation, often called ‘speciation continuum’, are typically based on closely related multiple species^{60,61}. However, different species may have experienced very different evolutionary histories. Thus, studying a single species with

395 varying levels of genetic differentiation might shed light on the exact process of genomic
396 differentiation.

397
398 Several genetic markers have been proposed to identify strains, including mitochondrial CO1⁶², sex
399 chromosome FR elements⁶³, and Z-linked TPI⁴³. We found that FR elements are a reliable marker to
400 identify strains (Supplementary Figure 12). TPI is included in the gene list within the genomic
401 island of differentiation, and d_{xy} from TPI (0.0345) is slightly lower than genomic average (mean is
402 0.0384 with 0.0383-0.0386 of 95% confidence interval). Thus, the genetic differentiation of TPI
403 appears to occur after the initiation of genetic differentiation between sfC and sfR. The concordance
404 of identified strains between mitochondrial CO1 and TPI can be as low as 74% (Table 5 at ⁴³), and
405 this imperfect concordance might be due to the different divergence time. Thus, we propose to use
406 mitochondrial markers to identify strains for unambiguous strain identification.

407
408 The process of speciation proposed in this study can be further tested based on insect rearing or lab
409 experiments (such as CRISPR/CAS9). For example, we proposed in this study that female fecundity
410 could be a key trait that initiated genetic differentiation between strains because genes associated
411 with this trait appears to have a resistance against gene flow. The reason for this resistance can be a
412 reduction in hybrid fitness, and we can test this possibility by insect-rearing. We also raise a
413 possibility in this paper that *cyc* gene might be associated with allochronic mating behavior, and we
414 can test this possibility using CRISPR/CAS9 experiment as well. These future studies will shed
415 light on the relationship between genotypes and phenotypes that plays critical roles in the process of
416 speciation.

417 418 METHOD

419 We extracted high molecular weight DNA using MagAttract© HMW kit (Qiagen) from one pupa of
420 sfC and two pupae of sfR with a modification of the original protocol to increase the yield. The
421 quality of extraction was assessed by checking DNA length (> 50kb) on 0.7% agarose gel
422 electrophoresis, as well as pulsed-field electrophoresis using the Rotaphor (Biometra) and gel
423 containing 0.75% agarose in 1X Loening buffer, run for 21 hours at 10°C with an angle range from
424 120 to 110° and a voltage range from 130 to 90V. DNA concentration was estimated by fluorimetry
425 using the QuantiFluor Kit (Promega), 9.6 µg and 8.7 µg of DNA from sfC and sfR, respectively,
426 which was used to prepare libraries for sequencing. Single-Molecule-Real-Time sequencing (12
427 SMRT cells per strain, equivalent to expected genome coverage of 20x) was performed using a
428 PacBio RSII (Pacific Biosciences) with P6-C4 chemistry at the genomic platform Get-PlaGe,
429 Toulouse, France (<https://get.genotoul.fr/>). The total throughput is 11,017,798,575bp in 1,513,346
430 reads and 13,259,782,164bp in 1,692,240 reads for sfC and sfR, respectively. The average read
431 lengths are 7,280bp and 7,836bp for sfC and sfR, respectively.

432 We generated assemblies from Illumina paired-end sequences³⁰ (166X and 308 X coverage for sfC
433 and sfR, respectively) using platanus⁶⁴. Then, errors in PacBio were corrected using Ectools⁶⁵, and
434 uncorrected reads were discarded. The remaining reads are 8,918,141,742bp and 11,005,855,683bp
435 for sfC and sfR, respectively. The error-corrected reads were used to assemble genome sequences
436 using SMARTdenovo³⁵. The paired-end Illumina reads were mapped against the genome
437 assemblies using bowtie2³⁷, and corresponding bam files were generated. We improved the genome
438 assemblies with these bam files using pilon⁶⁶.

439 For the genome assemblies of sfC, both Illumina paired-end and mate-pair reads were mapped the
440 genome assemblies using bwa⁶⁷, and scaffolding was performed using BESST⁶⁸. Since only paired-
441 end libraries were generated in our previous study³⁰, we used only paired-end sequences to perform
442 scaffolding for sfR. The gaps were filled using PB-Jelly⁶⁹. The correctness of assemblies was
443 assessed using insect BUSCO (insecta_odb9)³⁶.

Then, protein-coding genes were annotated from the genome sequences using MAKER⁷⁰. First, repetitive elements were masked using RepeatMasker⁷¹. Second, *ab initio* gene prediction was performed with protein-coding sequences from two strains in *S.frugiperda*³⁰ and *Helicoverpa armigera* (Harm_1.0, NCBI ID: GCF_002156995), as well as insect protein sequences from *Drosophila melanogaster* (BDGP6) and three Lepidoptera species, *Bombyx mori* (ASM15162v1), *Melitaea cinxia* (MelCinx1.0), and *Danaus plexippus* (Dpv3) in ensemble metazoa. For transcriptome sequences, we used reference transcriptome for sfC⁷² and locally assembled transcriptome from RNA-Seq data from 11 samples using Trinity⁷³ for sfR. Third, two gene predictors, SNAP⁷⁴ and Augustus⁷⁵, were trained and gene annotations were improved. Multiple trainings of the gene predictors do not decrease Annotation Edit Distance Score, thus we used the gene annotation with only one training. Fourth, we discarded all gene prediction if eAED score is greater than 0.5.

Paired-end Illumina resequencing data from nine individuals from each of corn and rice strains in *S.frugiperda* is used to identify variants. Low-quality nucleotides (Phred score < 20) and adapter sequences in the reads were removed using AdapterRemoval⁷⁶. Then, reads were mapped against reference genomes using bowtie2, with very exhaustive local search parameters (-D 25 -R 5 -N 0 -L 20 -i S,1,0.50), which is more exhaustive search than the -very-sensitive parameter preset. Potential PCR or optical duplicates were removed using Picard tool⁷⁷. Variants were called using samtools mpileup³⁸ only from the mappings with Phred score higher than 30. Then, we discarded all called positions unless a genotype is identified from all individuals and variant calling score is greater than 40. We also discarded variants if the read depth is higher than 3,200 or lower than 20.

We used vcf tools to calculate population genetics statistics, such as π and F_{st} ⁷⁸. Watterson's θ and d_{XY} were calculated using house-perl scripts. To estimate the genetic relationship among individuals, we first converted VCF files to plink format using vcftools, then PCA was performed using flashpca⁷⁹. For ancestry coefficient analysis, we used sNMF⁴² with K values ranging from 2 to 10, and we chose the K value that generated the lowest cross entropy.

Phylogenetic tree of the nuclear genome was generated using AAF³⁹. As an outgroup, we used simulated fastq files from the reference genomes of *S.litura* using genReads⁸⁰ with an error rate equal to 0.02. Reads were mapped against the mitochondrial genome (KM362176) using bowtie2³⁷ to generate the mitochondrial phylogenetic tree, and variants were called using samtools mpileup³⁸. From the mitochondrial VCF file, a multiple sequence alignment was generated using house-perl script. Then, the whole mitochondrial genome from *S.litura* (KF701043) was added to this multiple sequence alignment, and a new alignment was generated using prank⁸¹. The phylogenetic tree was reconstructed from this new alignment using FastME⁴⁹ with 1,000 bootstrapping.

The outliers of genetic differentiation were identified from hapFLK scores calculated from hapflk software⁴⁵. As the computation was not feasible with the whole genome sequences, we randomly divided sequences in the genome assemblies into eight groups. F_{st} distributions from these eight groups were highly similar between each other (Supplementary Figure 13). P-values showing the statistical significance of genetic differentiation were calculated from each position using scaling_chi2_hapflk.py in the same software package.

DATA AVAILABILITY

The reference genome and gene annotation are available from BioInformatics Platform for Agroecosystem Arthropods together with the genome browser (<https://bipaa.genouest.org/is/>). This data can be found at European Nucleotide Archive (<https://www.ebi.ac.uk/ena>) as well (project id:

493 PRJEB29161). Resequencing data is available from NCBI Sequence Read Archive. Corresponding
494 project ID is PRJNA494340.

495

496 ACKNOWLEDGMENTS

497 This work was partially supported by a grant from SPE department at INRA (adaptivesv) for K.N.,
498 by a grant from the French National Research Agency (ANR-12-BSV7-0004-01;

499 <http://www.agence-nationale-recherche.fr/>) for E.A., and by a grant from Institut Universitaire de
500 France for N.N.

501

502 REFERENCE

1. Felsenstein, J. Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution* **35**, 124–138 (1981).
2. Bolnick, D. I. & Fitzpatrick, B. M. Sympatric speciation: models and empirical evidence. *Annu. Rev. Ecol. Evol. Syst.* **38**, 459–487 (2007).
3. Wu, C.-I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
4. Feder, J. L., Egan, S. P. & Nosil, P. The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012).
5. Nosil, P. Speciation with gene flow could be common. *Mol. Ecol.* **17**, 2103–2106
6. Flaxman, S. M., Wacholder, A. C., Feder, J. L. & Nosil, P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* **23**, 4074–4088 (2014).
7. Barton, N. H. Gene flow past a cline. *Heredity* **43**, 333–339 (1979).
8. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
9. Barton, N. & Bengtsson, B. O. The barrier to genetic exchange between hybridising populations. *Heredity* **57**, 357–376 (1986).
10. Via, S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 451–460 (2012).
11. Via, S. & West, J. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.* **17**, 4334–4345 (2008).
12. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
13. Butlin, R. K. Recombination and speciation. *Mol. Ecol.* **14**, 2621–2635 (2005).

14. Noor, M. A. F., Grams, K. L., Bertucci, L. A. & Reiland, J. Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci.* **98**, 12084–12088 (2001).
15. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
16. Feder, J. L., Nosil, P. & Flaxman, S. M. Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. *Front. Genet.* **5**, 295 (2014).
17. Feder, J. L., Gejji, R., Yeaman, S. & Nosil, P. Establishment of new mutations under divergence and genome hitchhiking. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 461–474 (2012).
18. Feder, J. L. & Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evol. Int. J. Org. Evol.* **64**, 1729–1747 (2010).
19. Barton, N. H. What role does natural selection play in speciation? *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 1825–1840 (2010).
20. Schilling, M. P. *et al.* Transitions from single- to multi-locus processes during speciation with gene flow. *Genes* **9**, (2018).
21. Feder, J. L. *et al.* Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. *J. Hered.* **105**, 810–820 (2014).
22. Kopp, M. *et al.* Mechanisms of assortative mating in speciation with gene flow: connecting theory and empirical Research. *Am. Nat.* **191**, 1–20 (2017).
23. Marques, D. A. *et al.* Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLOS Genet* **12**, e1005887 (2016).
24. Ma, T. *et al.* Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proc. Natl. Acad. Sci.* **115**, E236–E243 (2018).
25. Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402 (2009).
26. Nosil, P., Egan, S. P. & Funk, D. J. Heterogeneous genomic differentiation between walking-stick ecotypes: ‘isolation by adaptation’ and multiple roles for divergent selection. *Evol. Int. J. Org. Evol.* **62**, 316–336 (2008).
27. Pashley, D. P. Host-associated genetic differentiation in fall armyworm (Lepidoptera: Noctuidae): a sibling species complex? *Ann. Entomol. Soc. Am.* **79**, 898–904 (1986).

28. Kergoat, G. J. *et al.* Disentangling dispersal, vicariance and adaptive radiation patterns: a case study using armyworms in the pest genus *Spodoptera* (Lepidoptera: Noctuidae). *Mol. Phylogenet. Evol.* **65**, 855–870 (2012).
29. Prowell, D. P., McMichael, M. & Silvain, J.-F. Multilocus genetic analysis of host use, introgression, and speciation in host strains of fall armyworm (Lepidoptera: Noctuidae). *Ann. Entomol. Soc. Am.* **97**, 1034–1044 (2004).
30. Gouin, A. *et al.* Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda* , Noctuidae) with different host-plant ranges. *Sci. Rep.* **7**, 11816 (2017).
31. Hänniger, S. *et al.* Genetic basis of allochronic differentiation in the fall armyworm. *BMC Evol. Biol.* **17**, 68 (2017).
32. Schöfl, G., Heckel, D. G. & Groot, A. T. Time-shifted reproductive behaviours among fall armyworm (Noctuidae: *Spodoptera frugiperda*) host strains: evidence for differing modes of inheritance. *J. Evol. Biol.* **22**, 1447–1459 (2009).
33. Unbehend, M., Hänniger, S., Meagher, R. L., Heckel, D. G. & Groot, A. T. Pheromonal divergence between two strains of *Spodoptera frugiperda*. *J. Chem. Ecol.* **39**, 364–376 (2013).
34. Dumas, P. *et al.* *Spodoptera frugiperda* (Lepidoptera: Noctuidae) host-plant variants: two host strains or two distinct species? *Genetica* **143**, 305–316 (2015).
35. Ruan, J. *smartdenovo: Ultra-fast de novo assembler using long noisy reads*. (2017).
36. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
38. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. Fan, H., Ives, A. R., Surget-Groba, Y. & Cannon, C. H. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* **16**, (2015).
40. Cheng, T. *et al.* Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat. Ecol. Evol.* **1**, 1747–1756 (2017).

41. Huelßenbeck, J. P. & Hillis, D. M. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42**, 247–264 (1993).
42. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983 (2014).
43. Nagoshi, R. N. The fall armyworm Triosephosphate Isomerase (Tpi) gene as a marker of strain identity and interstrain mating. *Ann. Entomol. Soc. Am.* **103**, 283–292 (2010).
44. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
45. Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* **193**, 929–941 (2013).
46. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
47. Wei, Y. & Lilly, M. A. The TORC1 inhibitors Nprl2 and Nprl3 mediate an adaptive response to amino-acid starvation in *Drosophila*. *Cell Death Differ.* **21**, 1460–1468 (2014).
48. Jacinto, E. & Hall, M. N. TOR signalling in bugs, brain and brawn. *Nat. Rev. Mol. Cell Biol.* **4**, 117–126 (2003).
49. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
50. Dumas, P. *et al.* Phylogenetic molecular species delimitations unravel potential new species in the pest genus *Spodoptera* Guenée, 1852 (Lepidoptera, Noctuidae). *PLOS ONE* **10**, e0122407 (2015).
51. Haller, B. C. & Messer, P. W. SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).
52. Charlesworth, B. The effects of deleterious mutations on evolution at linked sites. *Genetics* **190**, 5–22 (2012).
53. Song, S. V., Downes, S., Parker, T., Oakeshott, J. G. & Robin, C. High nucleotide diversity and limited linkage disequilibrium in *Helicoverpa armigera* facilitates the detection of a selective sweep. *Heredity* **115**, 460–470 (2015).

54. Sved, J. A., Cameron, E. C. & Gilchrist, A. S. Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. *PLOS ONE* **8**, e69078 (2013).
55. Rutila, J. E. *et al.* Cycle is a second bHLH-PAS clock protein essential for circadian rhythmicity and transcription of *Drosophila* period and timeless. *Cell* **93**, 805–814 (1998).
56. Anderson, C. J. *et al.* Hybridization and gene flow in the mega-pest lineage of moth, *Helicoverpa*. *Proc. Natl. Acad. Sci.* **115**, 5034–5039 (2018).
57. Martin, S. H. *et al.* Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics* **203**, 525–541 (2016).
58. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
59. Keightley, P. D. *et al.* Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Mol. Biol. Evol.* **32**, 239–243 (2015).
60. Riesch, R. *et al.* Transitions between phases of genomic differentiation during stick-insect speciation. *Nat. Ecol. Evol.* **1**, 82 (2017).
61. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013).
62. Pashley, D. P. Host-associated differentiation in armyworms (Lepidoptera: Noctuidae): An allozymic and mtDNA perspective. *Electrophor. Stud. Agric. Pests* (1989).
63. Lu, Y. J., Kochert, G. D., Isenhour, D. J. & Adang, M. J. Molecular characterization of a strain-specific repeated DNA sequence in the fall armyworm *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *Insect Mol. Biol.* **3**, 123–130 (1994).
64. Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
65. Gurtowski, J. *ectools: tools for error correction and working with long read data*. (2017).
66. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
67. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

68. Sahlin, K., Chikhi, R. & Arvestad, L. Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics* **32**, 1925–1932 (2016).
69. Rizk, G., Gouin, A., Chikhi, R. & Lemaitre, C. MindTheGap : integrated detection and assembly of short and long insertions. *Bioinformatics* btu545 (2014).
doi:10.1093/bioinformatics/btu545
70. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
71. RepeatMasker Home Page. Available at: <http://www.repeatmasker.org/>. (Accessed: 13th October 2017)
72. Legeai, F. *et al.* Establishment and analysis of a reference transcriptome for *Spodoptera frugiperda*. *BMC Genomics* **15**, 704 (2014).
73. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
74. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
75. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
76. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
77. *picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.* (Broad Institute, 2018).
78. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
79. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
80. Stephens, Z. D. *et al.* Simulating next-generation sequencing datasets from empirical mutation and sequencing models. *PLOS ONE* **11**, e0167047 (2016).
81. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).

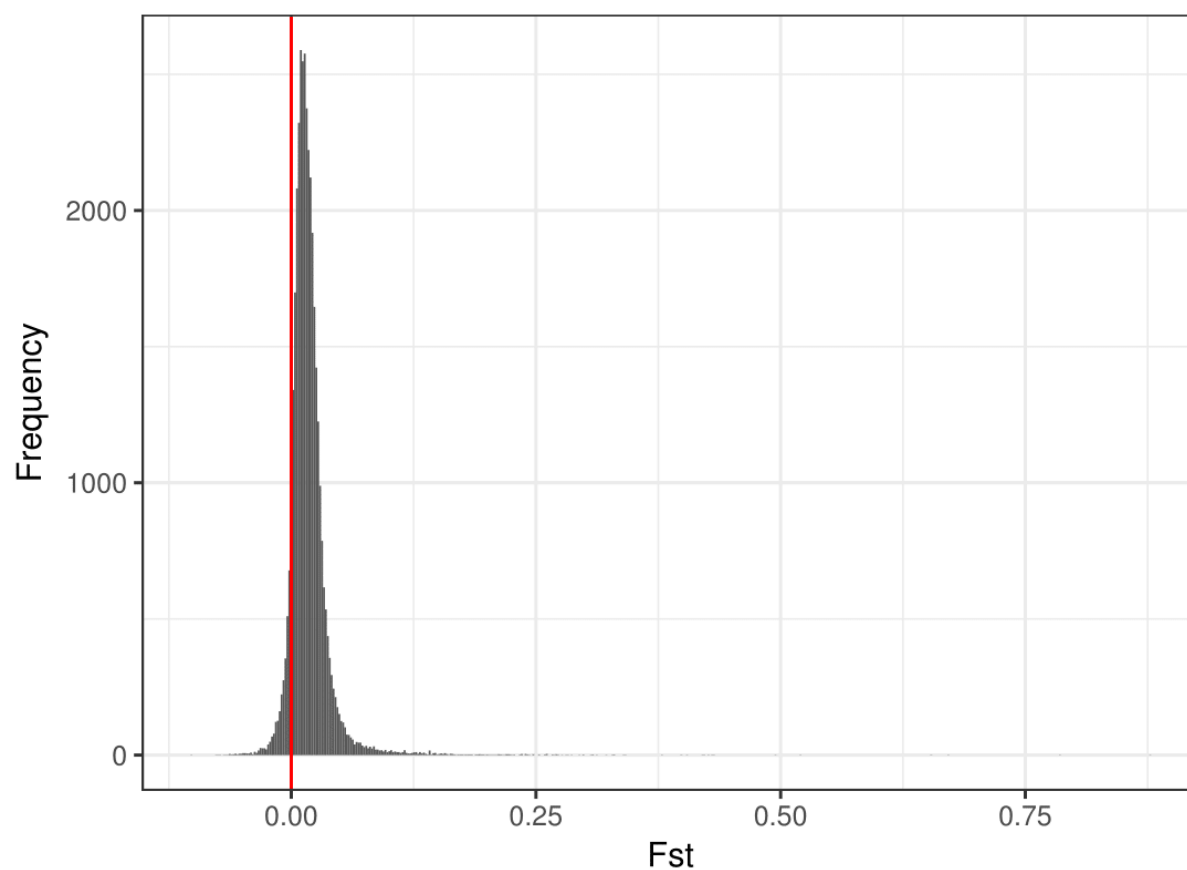
Table 1. Summary statistics of genome assemblies produced in this study (New assembly) and published assembly³⁰ from corn and rice strains.

	Corn strain		Rice strain	
statistics	New assembly	Gouin et al	New assembly	Gouin et al
Assembly size	384,358,373	437,873,304	379,902,278	371,020,040
number of sequences	1,000	41,577	1,054	29,127
Longest sequence (bp)	5,279,935	943,242	7,849,854	314,108
Shortest sequence (bp)	8,866	888	10,636	500
N50	900,335	52,781	1,129,192	28,526
L50	124	1,616	91	3,761
N90	196,225	3,545	165,330	6,422
L90	450	18,789	421	13,881
%GC	36.3432	35.0770	36.3724	36.0741
%N	0.0689	2.5989	0.0006	0.0352

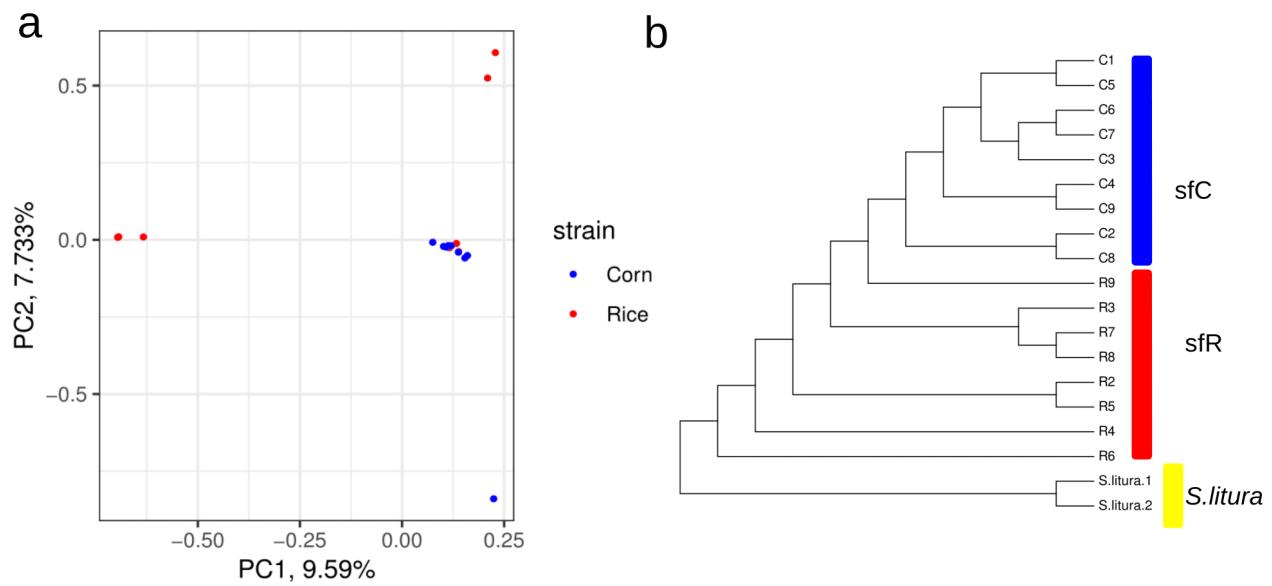
520 Table 2. The number of genes within genomic islands of differentiation that are potentially
521 associated with interactions with host-plants.

522

Functions	refC	refR
Chemosensory	3	3
Immunity	1	0
Oxidative stress	10	9
Development	4	4
P450	1	3
Circadian Signaling	0	1
Esterase	0	2
Serine Protease	0	1

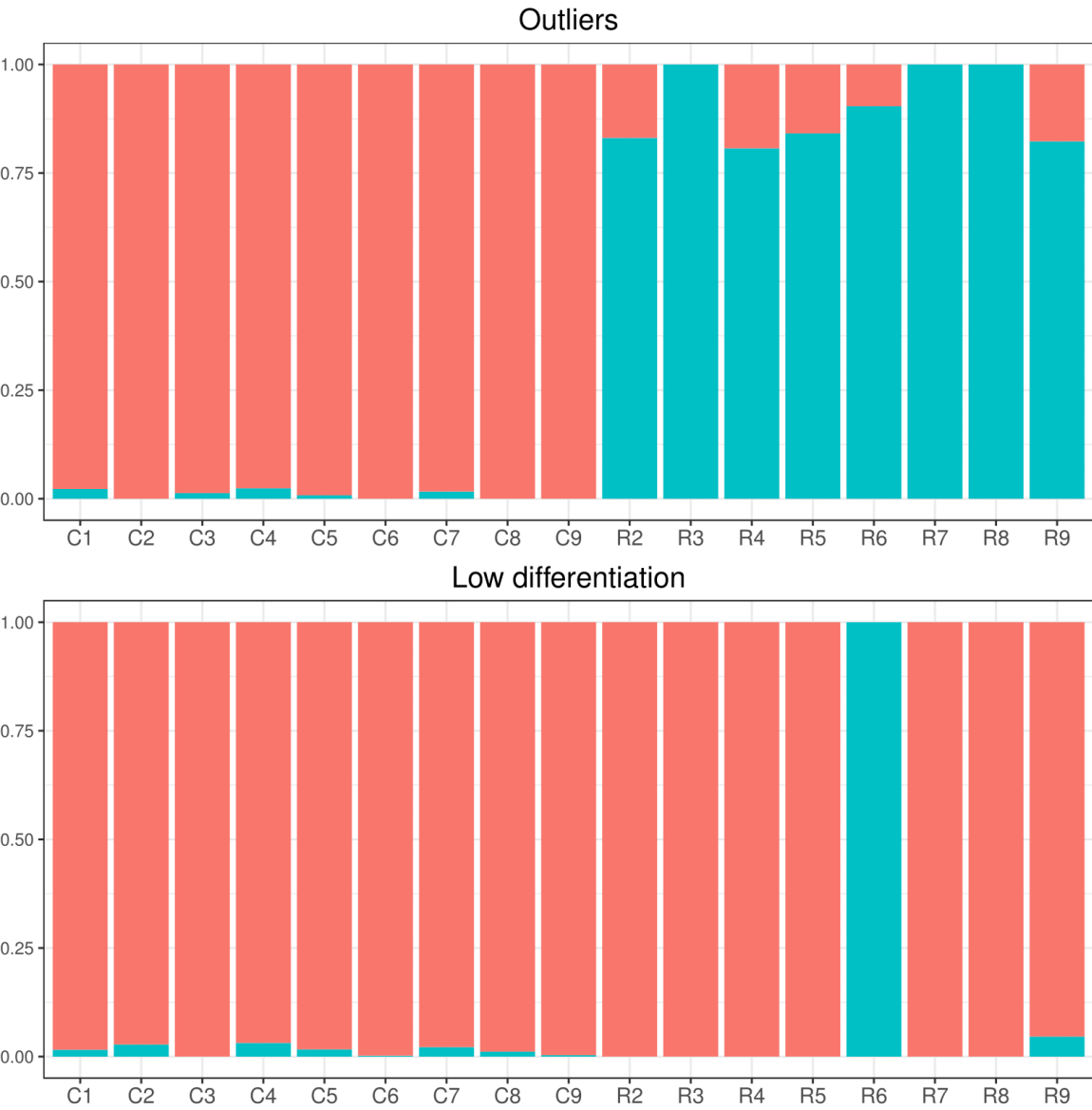


524 **Figure 1. The distribution of Fst calculated in 10 kb window** The red vertical line indicates Fst =
525 0, which means no genetic differentiation between corn and rice strains.
526

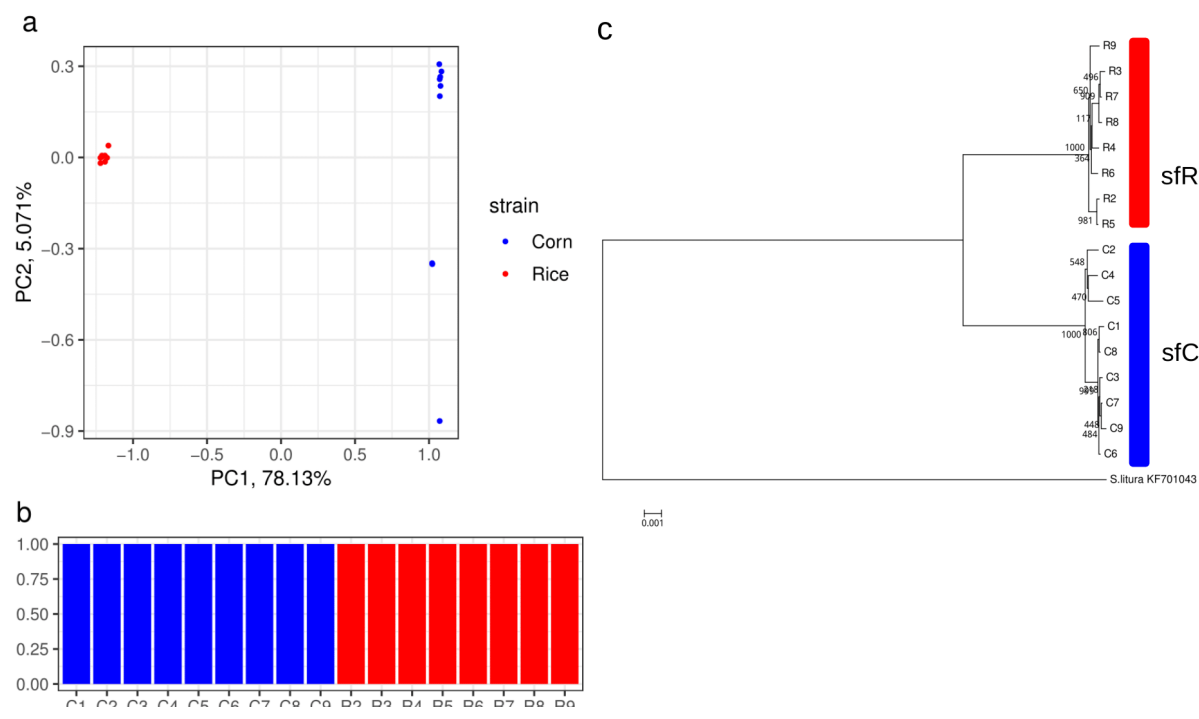


527 **Figure 2. Genetic relationship between corn and rice strains** a) The result from principal
528 component analysis. The red and blue circles represent individuals from corn and rice strains,
529 respectively. b) Phylogenetic tree reconstructed using AAF approach.

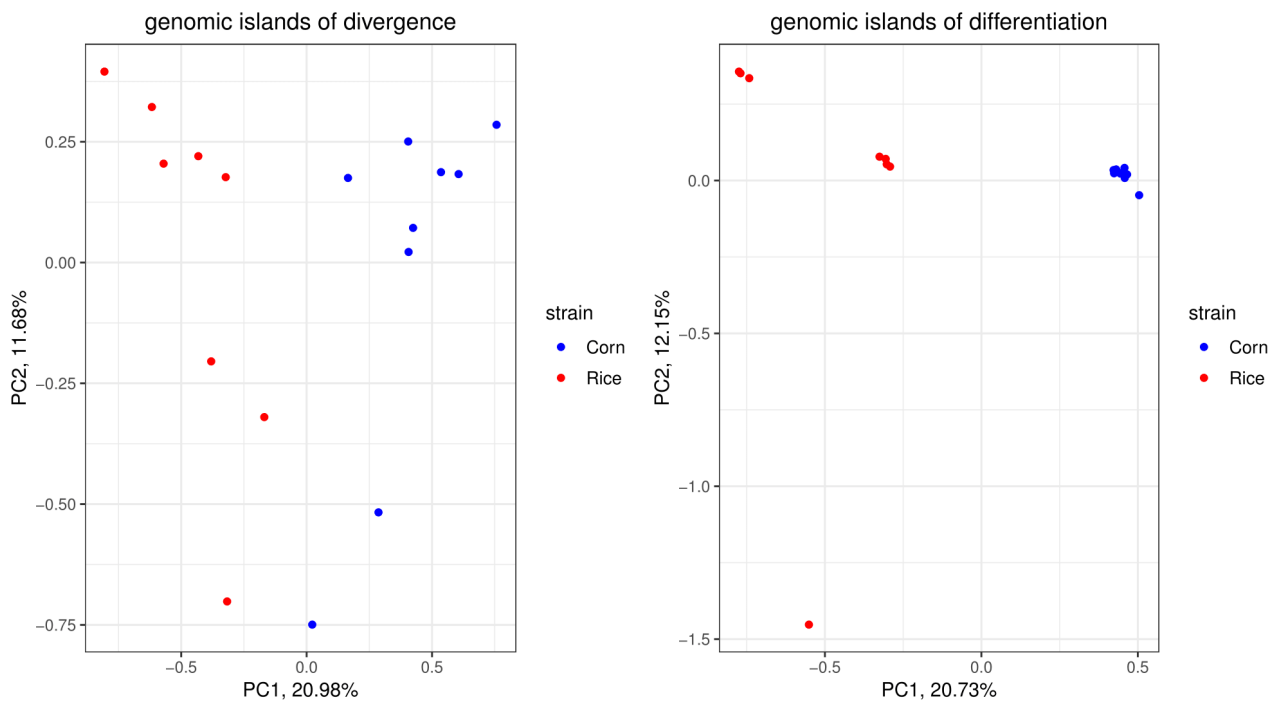
530



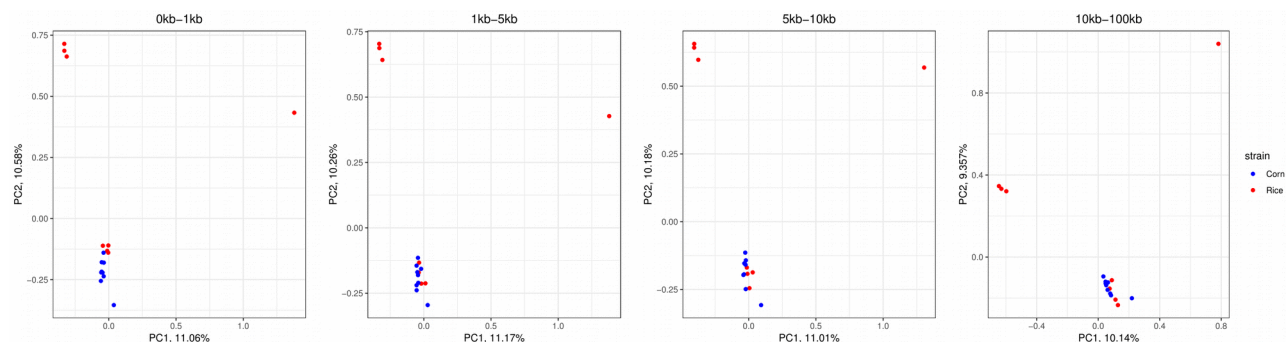
531 **Figure 3. Testing the divergence hitchhiking model.** Ancestry coefficient calculated from the
532 outliers of genetic differentiation (upper) and lowly differentiated sequences (hapflk score < 1,
533 154,163bp in size) (bottom).
534



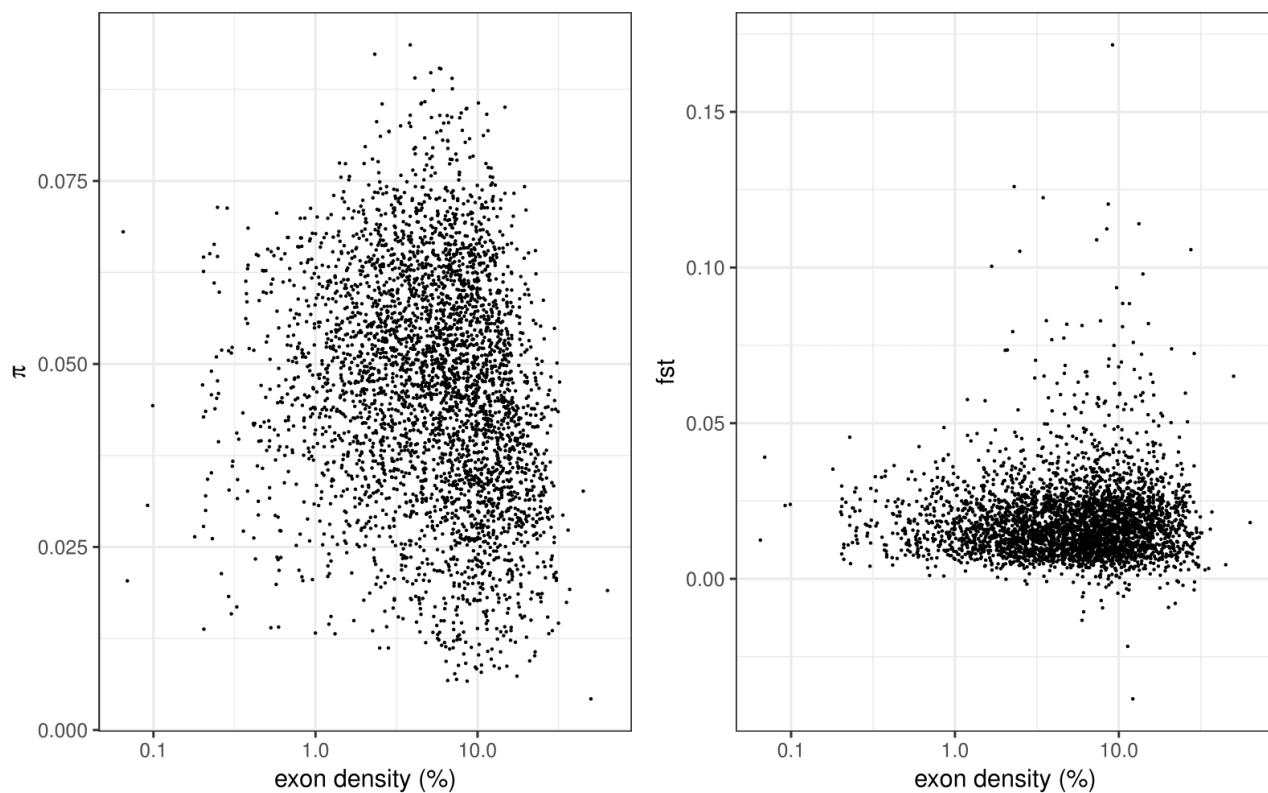
536 **Figure 4. Genetic relationship between corn and rice strains in the mitochondrial genome.** a)
537 The result from principal component analysis. The red and blue circles represent individuals from
538 sfC ad sfR, respectively. b) Ancestry coefficient results at K = 2. c) Phylogenetic tree reconstructed
539 using minimum evolution approach.



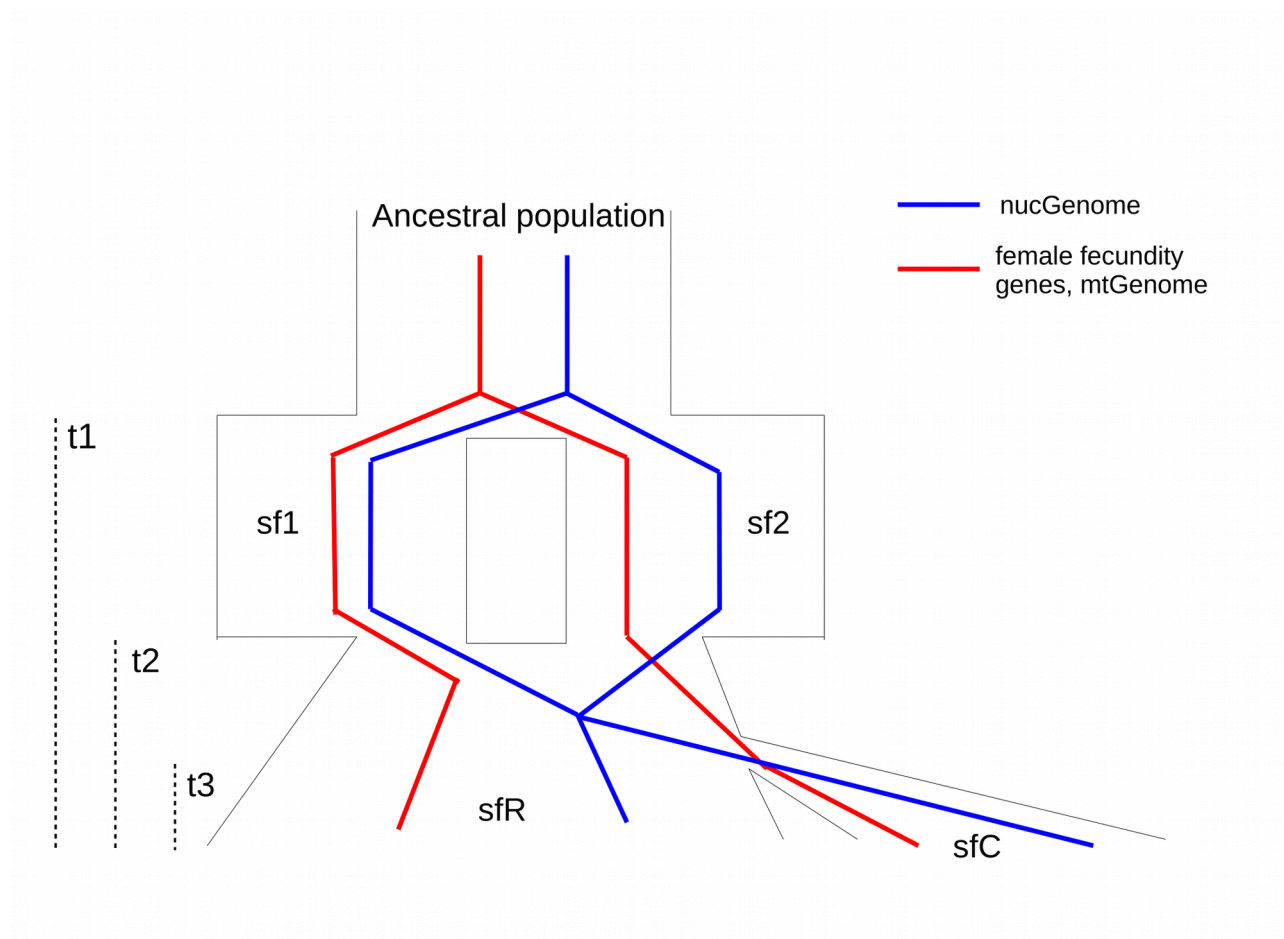
540 **Figure 5. Genetic relationship among individuals in outliers of genetic differentiation** The result
541 of principal component analysis from genomic islands of divergence (left), which have higher level
542 of both relative level of genetic differentiation (hapflk score) and absolute level of genetic
543 divergence (d_{XY}), and genomic islands of differentiation (left), which have higher level of genetic
544 differentiation (hapflk score) only.



545 **Figure 6. The effect of physical linkage to the genomic islands of genetic differentiation** The
546 result of principal component analysis at varying distances from the nearest the genomic islands of
547 genetic differentiation. The result is based on the mappings against refC. See Supplementary Figure
548 20 for the result based on the mapping against refR.
549



551 **Figure 7. The effect of selection on local variation of diversity and differentiation** Plots showing
552 the correlation of exon density with π (left) and F_{st} (right) calculated from 100kb windows, based
553 on the mapping against refC. See Supplementary Figure 21 for the result based on the mapping
554 against refR.



555 **Figure 8. A possible evolutionary scenario of genetic differentiation** The average genealogy of
556 mitochondrial genomes and female fecundity genes (red lines) as well as nuclear genomes (blue
557 lines) are depicted. In this scenario, an ancestral population was split into two populations, sf1 and
558 sf2, at t1. At t2, two populations were merged by hybridization and extant sfR was generated.
559 However, local gene flow between sf1 and sf2 was inhibited at female fecundity genes because
560 hybrids of these genes had a reduction in fitness. Thus, the genealogy of the female fecundity genes
561 remained separated and sequences were kept diverging. The genealogy of mitochondrial genomes is
562 the same with the female fecundity genes because of selection on females and maternal inheritance.
563 After t3, divergent selection targeting many genes caused a genetic differentiation according to the
564 sequences of mitochondrial genomes and female fecundity genes by reducing genomic migration
565 rate, and extant sfC was generated.