Positive selection alone is sufficient for whole genome differentiation at the early stage of speciation process in the fall armyworm

Kiwoong NAM (ki-woong.nam@inra.fr)
INRA Centre de Montpellier https://orcid.org/0000-0003-3194-8673
Sandra Nhim
Universite de Montpellier
Stéphanie Robin
Inria
Anthony Bretaudeau
Inria
Nicolas Nègre
Universite de Montpellier
Emmanuelle d'Alençon
INRA Centre de Montpellier

Research article

Keywords: speciation, whole genome differentiation, genome hitchhiking, genome-wide congealing, fall armyworm, Spodoptera frugiperda

Posted Date: December 7th, 2019

DOI: https://doi.org/10.21203/rs.2.18055/v1

Version of Record: A version of this preprint was published at BMC Evolutionary Biology on November 13th, 2020. See the published version at https://doi.org/10.1186/s12862-020-01715-3.

Positive selection alone is sufficient for whole genome differentiation at the early stage of speciation process in the fall armyworm

Kiwoong Nam[1*], Sandra Nhim[1], Stéphanie Robin[2,3], Anthony Bretaudeau[2,3], Nicolas Nègre[1], Emmanuelle d'Alençon[1]

[1] DGIMI, University of Montpellier, INRA, 34095, Montpellier, France

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

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

*Correspondance: Kiwoong NAM, ki-woong.nam@inra.fr

ABSTRACT

Background: The process of speciation inherently involves the transition from genetic to genomic differentiation. In the absence of a geographic barrier, the whole genome differentiation may occur only when the homogenizing effect of recombination is overcome across the whole genome. The fall armyworm is observed as two sympatric strains with different host-plant preferences across the entire habitat. These two strains exhibit a very low level of genetic differentiation across the whole genome, suggesting that whole genome differentiation occurred at an early stage of speciation. In this study, we aim at identifying critical evolutionary forces responsible for the whole genome differentiation in the fall armyworm.

Results: We found that these two strains exhibit a low level of genomic differentiation (Fst = 0.0176), while 91.3% of 10kb windows have genetically differentiated sequences (Fst > 0). We observed that a genomic reduction in migration rate due to combined effects of mild positive selection and genetic linkages to selectively targeted loci are responsible for the whole genome

differentiation. Phylogenetic analysis shows that positive selection generates the whole genome differentiation by sub-setting of variants in one strain from the other.

Conclusions: From these results, we concluded that positive selection alone is sufficient for whole genome differentiation during the process of speciation. This study demonstrates that the propensity of adaptation alone determines the speciation events, suggesting that adaptive evolution is a single critical driving force for species diversity.

Keywords: speciation, whole genome differentiation, genome hitchhiking, genome-wide congealing, fall armyworm, Spodoptera frugiperda

BACKGROUND

The process of speciation inherently involves whole genome differentiation (WGD). This process is initiated from the differentiation of a few loci by positive selection and then continued by the enlargement of differentiated loci to a whole genome sequence[1]. If a geographic barrier ceases a gene flow between a pair of diverging populations, WGD may occur readily through the accumulation of mutations by selection or drift. However, if such geographic barriers do not exist, the transition from genetic to genomic differentiation is impeded by gene flow between a pair of diverging populations, because recombination in hybrids constantly homogenizes the DNA sequences between a pair of diverging populations[2]. Therefore, a special condition is necessary to overcome this homogenizing effect of gene flow across a whole genome. Speciation models involving such a special condition can be classified into four groups. First, special genome structures have been proposed to be a driver of WGD, such as the close genetic distance between loci targeted by positive selection and loci with chromosomal rearrangement[3, 4] or the clustering of selectively targeted loci within a genome[5, 6]. Second, if positive selection is sufficiently strong (i.e, $s > m$[7] or $s > r$[8], where s, m, and r are selection coefficient, migration rate, and recombination rate, respectively), the diverging effect of selection dominates the homogenizing effect of gene flow, enabling WGD. Third, if many loci are targeted by positive selection, the combined effect of selection can be sufficiently strong to reduce a migration rate across the whole genome, enabling WGD (termed genome hitchhiking model[9, 10]). Lastly, if selective sweeps target a very large number of loci, the average distance from a locus to a nearest selectively targeted locus decreases, and WGD may occur by genetic linkage to the targets[11].

The contribution of special genome structures has been empirically supported from a wide range of species, in the way of chromosomal rearrangements[3, 12–16] or the clustering of selectively targeted loci[5, 17, 18]. These examples explain the genetic differentiation of loci that are genetically linked to a special genome structure. Therefore, the genetic differentiation of completely

unliked loci (such as different chromosomes) remains unexplained. Isolation by adaptation, a positive correlation between a genetic difference and adaptive divergence[19, 20], has been presented as a support for genome hitchhiking[21]. Though genome hitchhiking is expected to generate the pattern of isolation by adaptation, it is still unclear whether genome hitchhiking initiates or reinforces genetic differentiation in cases of isolation by adaptation. In short, the cause of WGD during the speciation process is still largely unknown in natural conditions.

The fall armyworm, *Spodoptera frugiperda* (Lepidoptera: Noctuidae), exists as two sympatric strains, corn strain (sfC) and rice strain (sfR), named after their preferred host-plants, throughout their entire habitats[22]. The proportion of hybrids between these two strains can reach 16%[23], implying a frequent gene flow. These two strains have different mating time and sexual pheromone blends[24–26], implying pre-mating reproductive isolation. Hybrids that are generated from lab strains have reduced fitness compared with parental strains[27], and reciprocal transplant experiments show differential fitness between original hosts and alternative hosts[28], implying the existence of post-zygotic reproductive isolation. The existence of reproductive isolation implies these two strains have experienced a speciation process.

We observed in a previous study that the two strains collected in Mississippi exhibit genetic differentiation across almost entire genome sequences, implying the presence of WGD, while the extent is very low (average Fst = 0.019)[29]. This level of WGD is one of the lowest among all reported cases, to our knowledge, implying that WGD may occur at much earlier stages of the process of speciation than previously thought. This pattern of genetic differentiation is potentially a snapshot that two diverging groups just entered the phase of WGD at an early stage of speciation, and making the fall armyworm an ideal model species to pinpoint minimal evolutionary forces responsible for WGD. This pattern of WGD can be generated by a single event of very strong

positive selection. Alternatively, mild positive selection targeting many loci can also generate this pattern. Lastly, special genomic structures may contribute to this WGD.

In this study, we aim at identifying minimal evolutionary forces responsible for WGD at the early stage of speciation in the fall armyworm based on population genomics approach. We tested the role of strong or mild positive selection in WGD. We also tested the contribution of special genome structures to the WGD. Finally, we inferred the evolutionary history of the transition from genetic to genomic differentiation.

RESULTS

Genome assembly and SNV identification

Since the existing reference genome sequence by Gouin et al. is fragmented[29], we generated a new genome assembly with 27.5X and 33.1X coverages PacBio reads from sfC and sfR, respectively. The resulting assemblies are 384Mb and 380Mb in size for sfC and sfR, respectively (Table 1). These sizes are closer to the expectation by flow cytometry (396 ± 3 Mb)[29] than the assemblies by Gouin et al (438Mb and 371Mb for sfC and sfR, respectively). The contiguity of genome assemblies is also greatly improved, as N50 of our assembles is 900Mb and 1,129Mb for sfC and sfR, respectively, while that of Gouin et al. is 52.8kb and 28.5kb for sfC and sfR, respectively. The new assemblies have higher proportion of complete BUSCO (Benchmarking Universal Single-Copy Orthologs) genes[30] than Gouin et al. (1601/1658 = 96.6% and 1461/1658 = 88.1% for our assembly and Gouin et al, respectively, in sfC; 1616/1658 = 97.4% and 1551/1658 = 93.5% for our assembly and Gouin et al, respectively, in sfR), implying that the correctness is also improved. We identified 21,839 genes and 22,026 genes from sfC and sfR reference genomes, respectively. BUSCO analysis shows that the correctness of gene annotation is improved as well(Table S1).

Recently, Liu et al. generated two genome assemblies from a natural invasive population of the fall armyworm[31]. These two assemblies have higher contiguity than the assemblies generated in this study (Table 1). However, the assembly sizes are far greater than the expectation by flow-cytometry (532Mb - 544Mb and 396 ± 3 Mb for Liu et al. and the expected size, respectively), perhaps due to heterozygous positions. These two assemblies have lower numbers of complete and single-copy BUSCO genes (1442 and 1480) than our assembly (1572 and 1573), demonstrating lower correctness than ours. The assemblies by Liu et al. have much higher numbers of complete and *duplicated* BUSCO gene (134 and 97) than ours (29 and 43), suggesting that misassemblies may inflate the size of genome assemblies.

Resequencing data from nine female individuals from each of sfC and sfR collected in the wild[29] were mapped against this sfC assembly. One individual from sfR was excluded in the following analysis because of a particularly low mapping rate and an average read depth (denoted as R1, Gouin et al.[29]) (Fig. S2). 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, after strict filtering on both variant and non-variant sites. We first performed all analysis based on the mapping against sfC reference genomes (the results from the mapping against sfR is shown at the end of Results).

WGD between sfC and sfR

The genome-wide Fst calculated between sfC and sfR is 0.0176, which is comparable to our previous study (0.019)[29]. No randomized 500 groupings show higher Fst than the grouping according to strains (equivalent to $p < 0.002$), demonstrating significant WGD between strains[29]. This WGD can be either caused by a few loci with very high levels of differentiation or by many loci with low levels of differentiation. To test these two possibilities, we calculated Fst in 10kb

windows. Among total windows, 91.3% have Fst greater than 0 (Fig. 1), as shown in our previous study[29], demonstrating the existence of WGD.

The role of positive selection in WGD

To investigate the role of selection in WGD, we identified targets of positive selection from the outliers of genetic differentiation using haplotype scores[32]. If each of minimum 100 consecutive SNPs in a minimum 1kb has a significantly greater haplotype score than the rest of the genome ($p < 0.001$), we defined this locus as an outlier. In total, 433 outliers were identified from 170 scaffolds out of 1,000, and the length is 4,023bp on average. Fst at outliers is 0.104 on average (the leftmost point at Fig. 2A), which is much higher than the genomic average (0.0176). We calculated the distance from each nucleotide from the nearest outliers and sorted the nucleotides according to the distance. Then, 100 groups were generated according to the distance in a way that each group has comparable numbers of nucleotides. If the distance is less than 200kb, Fst decreases rapidly as the distance increases (Fig. 2A). If the distance is higher than 200kb, on the other hand, the change in Fst is negligible. This result shows that the effect of genetic linkage to selectively targeted sites is clearly observed only if the distance is shorter than 200kb. The total length of sequences within this distance (< 200kb) is 72,280,543bp, which corresponds to 18% of the genome assembly.

The Fst calculated along the distance is always greater than 0.012, even when the distance is higher than 2Mb. Fst calculated from scaffolds without outliers is 0.0161, which is significantly lower than that from scaffolds with outliers (Fst = 0.020; $p < 0.01$, bootstrapping test) but greater than zero (Fig. 2B). This result indicates that the genetic linkage to the outliers alone does not explain the WGD. Instead, the results demonstrate the existence of a genomic factor that increases the level of genetic differentiation across the whole genome without genetic linkage to selectively targeted sites. In other words, the result supports the existence of a genomic reduction in the migration rate between strains.

153 This genomic reduction can be caused either by strong positive selection targeting a small number
154 of loci or by the combined effects of mild positive selection because, in this case, migration rate
155 between sfC and sfR will be effectively reduced across the whole genome[9, 10]. First, we test the
156 presence of very strong positive selection targeting a small number of loci. The longest outlier,
157 which could be one of the strongest targets of selective sweeps, is 27,365bp in size. This outlier
158 occupies only 1.56% of the scaffold, meaning that the vast majority of sequences within this
159 scaffold does not have significantly different Fst from the genomic average. This result does not
160 support the existence of very strong positive selection targeting a small number of loci. Instead,
161 mild positive selection targeting many loci is the most likely reason for the reduction in migration
162 rate. This interpretation is in line with the genome hitchhiking model[9, 10].

163 The observed negative correlation between Fst and the distance from outliers shown in Figure 2A,
164 however, could be a byproduct of a correlation between Fst and the strength of background
165 selection[33]. Background selection may target all selectively constrained sequences in principle,
166 while the strength depends on the selection coefficients and the density of constrained sites[34]. If
167 outliers have a higher proportion of selectively constrained sites than the rest of the genome,
168 background selection alone could generate the observed negative correlation between Fst and the
169 distance from outliers. Thus, we tested if background selection could generate the observed
170 negative correlation between the distance and Fst, assuming that the exon density is a proxy of the
171 strength of background selection. Exons are overrepresented at the outliers compared with non-
172 outliers (233 outliers of 433 have exons; p=0.041, randomization test with 1,000 replications),
173 implying that the outliers may experience stronger background selection than the rest of genomic
174 loci. The exon density calculated in 100kb windows is negatively correlated with π (Spearman's ρ =
175 -0.211, $p < 2.2 \times 10^{-16}$) (Fig. 3A), showing the presence of background selection. Fst, however, is
176 not significantly correlated with exon density ($\rho = 0.021$, $p = 0.2032$) (Fig. 3B). The same trend is
177 observed when we calculated the correlations from scaffolds without outliers (Fig. S3). This result

demonstrates that background selection is not a dominating factor determining the local level of genetic differentiation. Thus, it is unlikely that background selection alone causes the observed positive correlation between Fst and distance to the outliers.

The role of special genome structures

According to the divergence hitchhiking model[5, 6], the clustering of selectively targeted loci within a genome is the main cause of WGD. If a locus is targeted by strong positive selection, then the effective rate of migration is reduced in this region, and following events of positive selection targeting sequences within this region may generate a long stretch of differentiated DNA. In this case, a long sequence with genetic differentiation is expected. The longest outliers, however, occupies only 1.56% of the scaffold, as shown above. Therefore, this model is unsupported from our observation.

We investigated the role of chromosomal rearrangements in WGD. We identified 1,254 loci with chromosomal inversions, 1,060bp in median length. We considered that a chromosomal rearrangement is strain-specific if the difference in allele frequency is higher than an arbitrarily chosen criterion, 0.75. The number of strain-specific inversions is only two. Fst calculated from these inversions is lower than zero (-0.063 and -0.064), showing that the contribution of chromosomal inversion to genetic differentiation is unsupported. The number of inter-scaffold rearrangement is 1,724, and only one of them has a difference in allele frequency higher than 0.75. Fst calculated from 10kb flanking sequences of the breaking points is lower than zero (-0.115 and -0.0783 at each side). Thus, it is unlikely that chromosomal rearrangement contributes to WGD.

From genetic to genomic differentiation

Then, we inferred evolutionary history from genetic to genomic differentiation between sfC and sfR. If selection targeting on an outlier initiates genetic differentiation, this outlier should have a higher level of an absolute level of genetic differentiation, which can be estimated from d_{XY}

statistics[35]. Four out of the 433 outliers have higher d_{XY} than the genomic average (FDR corrected $p < 0.05$) (Fig. S4). We identified three genes from these outliers, of which the corresponding function is unclear.

The rest of the outliers have lower π than the genomic average in both strains (Fig. S5), suggesting that selective sweeps accompanied by positive selection on these loci (see below for the possibility of background selection). The d_{XY} calculated from these outliers is lower than the genomic average (Fig. S6), showing that the outliers have a shorter coalescence time than the genomic average, which is expected in the presence of selective sweeps. Principal component analysis exhibits a clear grouping according to the strains at the outliers, while sfC shows much less genetic variability than sfR (Fig 4A). sfC has a lower π than sfR in these outliers ($p = 0.0007$; Wilcoxon rank sum test) (Fig. S5). These results show that positive selection generated different genotypes between strains and that sfC experienced stronger selective sweeps than sfR.

The outliers contain 275 protein-coding genes (Table S2). These protein-coding sequences include a wide range of genes important for the interaction with host-plants, such as P450, chemosensory genes, immunity gene, and oxidative stress genes[29] (Table S4). This result shows that positive selection on these host-plant genes potentially contributed to the transition from genetic to genomic differentiation.

WGD by subsetting variants

We inferred the genetic relationship among individuals of sfC and sfR. Principal component analysis shows that the genetic variation of sfR includes that of sfC (Fig. 4B). This result suggests the possibility that sfR is an ancestral status and that sfC was derived from the sfR. To test this possibility, we reconstructed a phylogenetic tree from whole genome sequences with *S. litura* as an outgroup. The resulting tree shows that sfC individuals constitute a monophyletic group within the sfR clade (Fig. 4C), supporting that the sfC was indeed derived from ancestral sfR. The genetic

relationship among sfC and sfR individuals was also analyzed from the ancestry coefficient, and distinct origins of sfC and sfR are unsupported (Fig. S7). This pattern is not observed from the outliers, which show that each sfC and sfR generates distinct groups (Fig. 4A). This result can be interpreted that positive selection targeting the outliers causes the WGD by sub-setting sfC variants from ancestral sfR variants.

The different genetic relationship between outliers and whole genome sequences indicates that selective sweeps accompanied with positive selection are mainly responsible for the outliers, instead of background selection. The observation that outliers have different genotypes between strains (Fig. 4A) is in line with positive selection by fixation of newly generated mutations. Background selection at recombination deserts can create different sequences only if two populations have *a priori* different genotypes across the whole genomes, which is incompatible with our observation (Fig. 4B). Thus, we consider that positive selection and selective sweeps are mainly responsible for the outliers.

We tested the possibility of an extreme case where both sfC and sfR have monophyly, but all identified sfR individuals except R6 in Fig. 3C are F1 hybrids between sfR females and sfC males. In this case, maternally-derived mitochondrial CO1 genes used to identify strains in this study[29] will have distinctly different sequences between R2 - R9 and C1 - C9, while paternally derived sequences will not show such a pattern. As all individuals analyzed in this study are females, the Z chromosomes were derived from males in the very previous generation (please note that Lepidoptera has ZZ females and ZW males). 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[36]*. The scaffold containing TPI gene is 3,688,019bp in length, and the number of variants is 201,075. The Fst calculated between sfC and sfR without R6 is 0.061, which is higher than the genomic average (0.0176). We calculated Fst from randomized groupings with 500

replicates, and only four replicates have Fst higher than 0.061, corresponding to the p-value equal to 0.008. This result demonstrates a significant genetic differentiation of paternally derived Z chromosomes between strains. Therefore, we exclude the possibility of the extreme case with F1 hybrids.

Using alternative reference genome assembly

Since we used the reference genomes from the sfC, the results might be systematically affected by ascertainment bias during the identification of variants. Thus, we performed the same analyses based on the mapping against sfR reference genomes. We observed the same trends between the mapping against the reference genome assemblies of sfC and sfR (Fig. S8-S16). The only exception is the list of genes in the outliers (Table S3, S4). If a gene is highly differentiated, the mapping rate is essentially low because of large differences in sequences between reference genomes and resequencing data. Then, the identification of genes within the outliers can be severely affected by the usage of reference genome sequences. We identified 423 outliers, which include nine outliers with higher d_{XY} than the genomic average (FDR corrected $p < 0.05$) (Fig. S12). These loci contain four protein-coding genes, including 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[37]. Glutamine synthetase 2 is important in activating the TOR pathway, which is the main regulator of cell growth in response to environmental changes to maintain fecundity in planthoppers[38]. This result raises the possibility that disruptive selection on female fecundity is responsible for initiating genetic differentiation between strains.

DISCUSSION

In this paper, we show in fall armyworms that mild positive selection contributed to WGD at the early stage of the speciation process by the effect of genetic linkages to selectively targeted sites

and by a genomic reduction in migration rate. We do not find the contribution of very strong positive selection or special genomic structures in WGD, suggesting that mild positive selection appears to be sufficient for the early stages of the speciation process by WGD.

Once a WGD pattern is generated, the rate of genetic differentiation across the whole genome can be accelerated by following positive selection until the end of the speciation process (Fig. 5A). Theoretical studies[7, 39] show that if the number of targets is higher than a certain threshold, targeted loci have a synergistic effect in increasing linkage disequilibrium among targets, consequently resulting in the acceleration of WGD. The non-linear dynamics of WGD according to the number of accumulated selectively targeted loci were termed genome-wide congealing[40]. The WGD between sfC and sfR proposes the possibility that any following positive selection may accelerate the WGD until the end of the speciation process by providing genome-wide WGD, in line with the genome-wide congealing model. In other words, positive selection alone may be sufficient for speciation during the entire process of speciation, first by generating WGD due to a genomic reduction in migration rate and genetic linkages to selectively targeted sites as shown in this paper, and second by by accelerating WGD by the synergistic effect of linkage disequilibrium at selectively targeted loci. We do not argue that sfC and sfR will evolve into two species in the future. Instead, we argue that a condition for the accelerating WGD is made in these two strains by positive selection. This explanation is in contrast with ‘genic view of speciation’[1] (Fig. 5B). In this model, speciation is initiated from the differentiation from a few loci, and continued by the enlargement of differentiated loci, and finished by WGD. Therefore, WGD is passively generated at the end of a speciation process according to this model.

The level of WGD can be different among different geographical populations in the fall armyworm. Therefore, the fall armyworm can be optimal model species to study the process of WGD by providing snapshots of different phases of a speciation process, ranging the initiation of genetic

differentiation, the initiation of WGD, and the acceleration of WGD. Attempts to identify these phases, often termed ‘speciation continuum’, are typically based on closely related multiple species[41, 42]. However, different species may have experienced very different evolutionary histories. Thus, studying a single species at varying levels of genetic differentiation may shed light on the exact process of WGD.

Our observation that sfC was derived from sfR is different from previous reports, which show that sfC and sfR are sister groups between each other across the wide range of geographic populations based on the mitochondrial DNA [43, 44]. From our data, the mitochondrial genomes have almost completely differentiated between strains (Fst=0.920), and sfC and sfR appear to be the sister group of each other (Fig. S17 - S19), confirming the previous reports. The discrepancy between nuclear and mitochondrial patterns could be explained by the initiation of genetic differentiation promoted by the disruptive selection on female fecundity genes (see Results). We observed that the genetic differentiation between strains was initiated from the female fecundity gene. If the genetic differentiation between sfC and sfR was initiated by disruptive selection on female fecundity, maternal genealogy would show a bifurcating genetic transmission pattern between strains. As mitochondria are inherited through maternal lineage, the mitochondrial sequences will show the same bifurcating genetic transmission pattern as the female fecundity gene. Then, mitochondrial genomes will be differentiated more anciently than nuclear genomes. A molecular clock study shows that the mitochondrial genomes diverged between sfC and sfR two million years ago[43], which corresponds to 2×10^7 generations according to the observation from our insectarium (10 generations per year). Our simulation shows that the observed Fst (0.0176) is unlikely to be generated during this time with a wide range of migration rates (Fig. S20), confirming that the mitochondrial genome diverged more anciently than the nuclear genome. This explanation does not exclude the possibility of positive selection on mitochondrial genes.

Our observation that sfC was derived from sfR can explain the contradictory patterns of reduction in hybrid fitness. A group of studies reported that hybrids from female sfC and male sfR have lower fitness or reproductive success than that from female sfR and male sfC [23, 27, 45, 46]. On the other hand, another group of studies reported the opposite pattern[47–49]. We raise the possibility that the heterogeneous genetic background of sfR caused this contrasting pattern. If sfC was derived from sfR, the genetic distance between sfC and sfR individuals is different according to the used sfR individuals. Then, the pattern of fitness reduction in hybrids can also be different depending on the used sfR individuals, and even opposite patterns might be generated.

Several genetic markers have been proposed to identify strains, including mitochondrial CO1[50], sex chromosome FR elements [51], and Z-linked TPI[36]. We found that FR elements are a reliable marker to identify strains (Fig. S21). The concordance of identified strains between mitochondrial CO1 and TPI can be as low as 74%[36]. TPI is included in the gene list within the genomic island of differentiation, suggesting that linked selection is responsible for the differentiation of TPI gene. d_{XY} from TPI (0.0345) is slightly lower than the genomic average (0.0384 with 0.0383-0.0386 of 95% confidence interval), showing the genetic differentiation of TPI postdates the genomic average. Considering our observation that sfC was derived from sfR, the pattern of genetic variation of TPI gene might differ depending upon the genetic distance between sfC and analyzed sfR individuals (Fig. S22). Therefore, we propose to use mitochondrial markers for the identification of strains.

CONCLUSIONS

We demonstrated that positive selection is a sufficient evolutionary force for WGD at the early stage of speciation. Species with a large population size may have a higher adaptive evolutionary rate than that with a small population size because larger populations have a higher influx of mutations, of which a proportion has beneficial effects (hard sweeps). In addition, large populations have a higher proportion of existing variants that can be beneficial in the future depending upon

environmental changes (soft sweeps)[52]. A positive correlation between adaptive evolutionary rates and population sizes has been reported from a wide range of taxa[53, 54] (but see Galtier[55]). Moreover, a positive correlation between the size of populations and the strength of selective sweeps was reported[56]. If the propensity of positive selection determines speciation, and if large populations experience a higher frequency of positive selection, species with large population size might have a higher potential to be speciated depending upon the disruption of fitness landscape due to environmental changes, proposing a close link between the taxonomic diversity among species and the genetic diversity within species.

METHODS

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

We generated assemblies from Illumina paired-end sequences[29] (166X and 308 X coverage for sfC and sfR, respectively) using platanus[57]. Then, errors in PacBio were corrected using Ectools[58], and uncorrected reads were discarded. The remaining reads are 8,918,141,742bp and

11,005,855,683bp for sfC and sfR, respectively. The error-corrected reads were used to assemble genome sequences using SMARTdenovo[59]. The paired-end Illumina reads were mapped against the genome assemblies using bowtie2[60], and corresponding bam files were generated. We improved the genome assemblies with these bam files using pilon[61].

For the genome assemblies of sfC, both Illumina paired-end and mate-pair reads were mapped the genome assemblies using bwa[62], and scaffolding was performed using BESST[63]. Since only paired-end libraries were generated from sfR in our previous study[29], we used only paired-end sequences to perform scaffolding for sfR. The gaps were filled using PB-Jelly[64]. The correctness of assemblies was assessed using insect BUSCO (insecta_odb9)[30].

Then, protein-coding genes were annotated from the genome sequences using MAKER[65]. First, repetitive elements were masked using RepeatMasker[66]. Second, *ab initio* gene prediction was performed with protein-coding sequences from two strains in *S. frugiperda[29]* 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 sfC[67] and locally assembled transcriptome from RNA-Seq data from 11 samples using Trinity[68] for sfR. Third, two gene predictors, SNAP[69] and Augustus[70], were trained to improve gene annotations. 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 sfC and sfR in *S. frugiperda* is used to identify variants. Low-quality nucleotides (Phred score < 20) and adapter sequences in the reads were removed using AdapterRemoval[71]. 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 tool[72]. Variants were called using samtools mpileup[73] only from the mappings with Phred mapping score higher than 30. Then, we discarded all called positions unless a genotype is determined from all individuals and variant calling score is higher than 40. We also discarded variants if the read depth is higher than 3,200 or lower than 20.

We used vcftools to calculate population genetics statistics, such as π and Fst[74]. 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 flashpca[75]. For ancestry coefficient analysis, we used sNMF[76] 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 AAF[77]. As an outgroup, we used simulated fastq files from the reference genomes of *S. litura[78]* using genReads[79] with an error rate equal to 0.02. Reads were mapped against the mitochondrial genome (KM362176) using bowtie2[60] to generate the mitochondrial phylogenetic tree, and variants were called using samtools mpileup[73]. 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 prank[80]. The phylogenetic tree was reconstructed from this new alignment using FastME[81] with 1,000 bootstrapping.

The outliers of genetic differentiation were identified from hapFLK scores calculated from hapflk software[32]. As the computation was not feasible with the whole genome sequences, we randomly divided sequences in the genome assemblies into eight groups. Fst distributions from these eight groups were highly similar between each other (Fig. S1). P-values showing the statistical

significance of genetic differentiation were calculated from each position using scaling_chi2_hapflk.py in the same software package.

The reference genome and gene annotation are available from the BioInformatics Platform for Agroecosystem Arthropods together with the genome browser (www.bipaa.genouest.org/sp/spodoptera_frugiperda). This data can be found at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) as well (project id: PRJEB29161). Resequencing data is available from NCBI Sequence Read Archive, and the corresponding project ID is PRJNA494340.

LIST OF ABBREVIATIONS

BUSCO: Benchmarking Universal Single-Copy Orthologs

CO1: Cytochrome c oxidase subunit I

DNA: Deoxyribonucleic acid

FDR: False discovery rate

NPRL2: Nitrogen permease regulator 2-like

sfC: Corn strain in Spodoptera frugiperda

sfR: Rice strain in Spodoptera frugiperda

TOR pathway: target of rapamycin pathway

TORC1: Target of rapamycin complex 1

TPI: Triosephosphate isomerase

WGD: Whole genome differentiation

DECLARATIONS

- Ethics approval and consent to participate: NA
- Consent for publication: NA

- Availability of data and materials: All data used in this study is available in NCBI SRA (PRJEB29161) and BIPAA (www.bipaa.genouest.org/sp/spodoptera_frugiperda).
- Competing interests: The authors declare that they have no competing interests
- Funding:
    1. A grant from the Department of Santé des Plantes et Environnement at Institut national de la recherche given to KN (adaptivesv): High-performance computation
    2. A grant from the French National Research Agency given to ED (ANR-12-BSV7-0004-01): The generation of Pac-Bio data
    3. A grant from Institut Universitaire de France given to NN: Resequencing from Mississippi population
- Authors' contributions:
    1. KN: Conceptualization; Formal Analysis; Funding Acquisition, Investigation, Project Administration, Writing the paper
    2. SN: Preparation of gDNA for PacBio sequencing
    3. SR: Data Curation
    4. AB: Data Curation; gene annotation
    5. NN: Funding Acquisition; Generating NGS data
    6. EA: Funding Acquisition; Preparation of PacBio reads

    All authors have read and approved the manuscript.
- Acknowledgments: We acknowledge Kazuei Mita (Southwest University in Chongqing, China) for the permission to start using the whole genome sequencing data from Spodoptera litura before publication.

REFERENCES

1. Wu C-I. The genic view of the process of speciation. Journal of Evolutionary Biology. 2001;14:851–865.

2. Felsenstein J. Skepticism towards Santa Rosalia, or why are there so few kinds of animals? Evolution. 1981;35:124–138. http://www.jstor.org/stable/2407946. Accessed 21 Oct 2016.

3. Noor MAF, Grams KL, Bertucci LA, Reiland J. Chromosomal inversions and the reproductive isolation of species. PNAS. 2001;98:12084–8. doi:10.1073/pnas.221274498.

4. Feder JL, Nosil P, Flaxman SM. Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. Front Genet. 2014;5:295.

5. Via S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. Philos Trans R Soc Lond, B, Biol Sci. 2012;367:451–60.

6. Via S, West J. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Mol Ecol. 2008;17:4334–45.

7. Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. Mol Ecol. 2014;23:4074–88.

8. Barton NH. Gene flow past a cline. Heredity. 1979;43:333–9. doi:10.1038/hdy.1979.86.

9. Feder JL, Gejji R, Yeaman S, Nosil P. Establishment of new mutations under divergence and genome hitchhiking. Philos Trans R Soc Lond B Biol Sci. 2012;367:461–74. doi:10.1098/rstb.2011.0256.

10. Feder JL, Nosil P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution. 2010;64:1729–47.

11. Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity (Edinb). 1986;57:357–76.

12. Noor MAF, Garfield DA, Schaeffer SW, Machado CA. Divergence between the Drosophila pseudoobscura and D. persimilis genome sequences in relation to chromosomal inversions. Genetics. 2007;177:1417–28.

13. Lowry DB, Willis JH. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biol. 2010;8.

14. Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature. 2011;477:203–6. doi:10.1038/nature10341.

15. Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, et al. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 2014;344:1410–4.

16. Stevison LS, Hoehn KB, Noor MAF. Effects of inversions on within- and between-species recombination and divergence. Genome Biol Evol. 2011;3:830–41.

17. Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, et al. Genomics of rapid incipient speciation in sympatric threespine stickleback. PLOS Genet. 2016;12:e1005887. doi:10.1371/journal.pgen.1005887.

18. Ma T, Wang K, Hu Q, Xi Z, Wan D, Wang Q, et al. Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. PNAS. 2018;115:E236–43. doi:10.1073/pnas.1713288114.

19. Nosil P, Funk DJ, Ortiz-Barrientos D. Divergent selection and heterogeneous genomic divergence. Mol Ecol. 2009;18:375–402.

20. Nosil P, Egan SP, Funk DJ. Heterogeneous genomic differentiation between walking-stick ecotypes: “isolation by adaptation” and multiple roles for divergent selection. Evolution. 2008;62:316–36.

21. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends in Genetics. 2012;28:342–50. doi:10.1016/j.tig.2012.03.009.

22. Pashley DP. Host-associated genetic differentiation in fall armyworm (Lepidoptera: Noctuidae): a sibling species complex? Annals of the Entomological Society of America. 1986;79:898–904.

23. Prowell DP, McMichael M, Silvain J-F. Multilocus genetic analysis of host use, introgression, and speciation in host strains of fall armyworm (Lepidoptera: Noctuidae). Annals of the Entomological Society of America. 2004;97:1034–44. doi:10.1603/0013-8746(2004)097[1034:MGAOHU]2.0.CO;2.

24. Hänniger S, Dumas P, Schöfl G, Gebauer-Jung S, Vogel H, Unbehend M, et al. Genetic basis of allochronic differentiation in the fall armyworm. BMC Evolutionary Biology. 2017;17:68. doi:10.1186/s12862-017-0911-5.

25. Schöfl G, Heckel DG, Groot AT. Time-shifted reproductive behaviours among fall armyworm (Noctuidae: Spodoptera frugiperda) host strains: evidence for differing modes of inheritance. Journal of Evolutionary Biology. 2009;22:1447–59. doi:10.1111/j.1420-9101.2009.01759.x.

26. Unbehend M, Hänniger S, Meagher RL, Heckel DG, Groot AT. Pheromonal divergence between two strains of Spodoptera frugiperda. Journal of chemical ecology. 2013;39:364–376. http://link.springer.com/article/10.1007/s10886-013-0263-6. Accessed 13 Sep 2016.

27. Dumas P, Legeai F, Lemaitre C, Scaon E, Orsucci M, Labadie K, et al. Spodoptera frugiperda (Lepidoptera: Noctuidae) host-plant variants: two host strains or two distinct species? Genetica. 2015;143:305–16. doi:10.1007/s10709-015-9829-2.

28. Orsucci M, Mone Y, Audiot P, Gimenez S, Nhim S, Nait-Saidi R, et al. Transcriptional plasticity evolution in two strains of Spodoptera frugiperda (Lepidoptera: Noctuidae) feeding on alternative host-plants. bioRxiv. 2018;:263186. doi:10.1101/263186.

29. Gouin A, Bretaudeau A, Nam K, Gimenez S, Aury J-M, Duvic B, et al. Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda , Noctuidae) with different host-plant ranges. Scientific Reports. 2017;7:11816. doi:10.1038/s41598-017-10461-4.

30. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. doi:10.1093/bioinformatics/btv351.

31. Liu H, Lan T, Fang D, Gui F, Wang H, Guo W, et al. Chromosome level draft genomes of the fall armyworm, Spodoptera frugiperda (Lepidoptera: Noctuidae), an alien invasive pest in China. bioRxiv. 2019;:671560. doi:10.1101/671560.

32. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics. 2013;193:929–941.

33. Charlesworth B. The effects of deleterious mutations on evolution at linked sites. Genetics. 2012;190:5–22. doi:10.1534/genetics.111.134288.

34. Durrett R. Natural Selection. Probability Models for DNA Sequence Evolution. 2nd Ed. Springer Science & Business Media; 2008.

35. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23:3133–57.

36. Nagoshi RN. The fall armyworm Triosephosphate Isomerase (Tpi) gene as a marker of strain identity and interstrain mating. Ann Entomol Soc Am. 2010;103:283–92. doi:10.1603/AN09046.

37. Wei Y, Lilly MA. The TORC1 inhibitors Nprl2 and Nprl3 mediate an adaptive response to amino-acid starvation in Drosophila. Cell Death Differ. 2014;21:1460–8.

38. Jacinto E, Hall MN. TOR signalling in bugs, brain and brawn. Nature Reviews Molecular Cell Biology. 2003;4:117–26. doi:10.1038/nrm1018.

39. Barton NH. What role does natural selection play in speciation? Philos Trans R Soc Lond B Biol Sci. 2010;365:1825–40. doi:10.1098/rstb.2010.0001.

40. Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. J Hered. 2014;105:810–20. doi:10.1093/jhered/esu038.

41. Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, et al. Transitions between phases of genomic differentiation during stick-insect speciation. Nat Ecol Evol. 2017;1:82.

42. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013;23:1817–28. doi:10.1101/gr.159426.113.

43. Kergoat GJ, Prowell DP, Le Ru BP, Mitchell A, Dumas P, Clamens A-L, et al. Disentangling dispersal, vicariance and adaptive radiation patterns: a case study using armyworms in the pest genus Spodoptera (Lepidoptera: Noctuidae). Mol Phylogenet Evol. 2012;65:855–70.

44. Dumas P, Barbut J, Ru BL, Silvain J-F, Clamens A-L, d'Alençon E, et al. Phylogenetic molecular species delimitations unravel potential new species in the pest genus Spodoptera Guenée, 1852 (Lepidoptera, Noctuidae). PLOS ONE. 2015;10:e0122407. doi:10.1371/journal.pone.0122407.

45. Pashley DP, Martin JA. Reproductive incompatibility between host strains of the Fall Armyworm (Lepidoptera: Noctuidae). Ann Entomol Soc Am. 1987;80:731–3. doi:10.1093/aesa/80.6.731.

46. Nagoshi RN, Meagher R. Fall armyworm FR sequences map to sex chromosomes and their distribution in the wild indicate limitations in interstrain mating. Insect Mol Biol. 2003;12:453–8.

47. Kost S, Heckel DG, Yoshido A, Marec F, Groot AT. A Z-linked sterility locus causes sexual abstinence in hybrid females and facilitates speciation in Spodoptera frugiperda. Evolution. 2016;70:1418–27.

48. Whitford F, Quisenberry SS, Riley TJ, Lee JW. Oviposition preference, mating compatibility, and development of two fall armyworm strains. The Florida Entomologist. 1988;71:234–43. doi:10.2307/3495426.

49. Groot AT, Marr M, Heckel DG, Schöfl G. The roles and interactions of reproductive isolation mechanisms in fall armyworm (Lepidoptera: Noctuidae) host strains. Ecological Entomology. 2010;35:105–18. doi:10.1111/j.1365-2311.2009.01138.x.

50. Pashley DP. Host-associated differentiation in armyworms (Lepidoptera: Noctuidae): An allozymic and mtDNA perspective. Electrophoretic studies on agricultural pests. 1989.

51. Lu YJ, Kochert GD, Isenhour DJ, Adang MJ. Molecular characterization of a strain-specific repeated DNA sequence in the fall armyworm Spodoptera frugiperda (Lepidoptera: Noctuidae). Insect Mol Biol. 1994;3:123–30.

52. Hermisson J, Pennings PS. Soft sweeps molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169:2335–2352. http://www.genetics.org/content/169/4/2335.short. Accessed 10 Feb 2016.

53. Lanfear R, Kokko H, Eyre-Walker A. Population size and the rate of evolution. Trends in Ecology & Evolution. 2014;29:33–41. doi:10.1016/j.tree.2013.09.009.

54. Gossmann TI, Keightley PD, Eyre-Walker A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol. 2012;4:658–67.

55. Galtier N. Adaptive protein evolution in animals and the effective population size hypothesis. PLOS Genet. 2016;12:e1005774. doi:10.1371/journal.pgen.1005774.

56. Nam K, Munch K, Mailund T, Nater A, Greminger MP, Krützen M, et al. Evidence that the rate of strong selective sweeps increases with population size in the great apes. PNAS. 2017;114:1613–8. doi:10.1073/pnas.1605660114.

57. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24:1384–95. doi:10.1101/gr.170720.113.

58. Gurtowski J. ectools: tools for error correction and working with long read data. Python. 2017. https://github.com/jgurtowski/ectools.

59. Ruan J. smartdenovo: Ultra-fast de novo assembler using long noisy reads. C. 2017. https://github.com/ruanjue/smartdenovo.

60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

61. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE. 2014;9:e112963. doi:10.1371/journal.pone.0112963.

62. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–595. https://academic.oup.com/bioinformatics/article-abstract/26/5/589/211735.

63. Sahlin K, Chikhi R, Arvestad L. Assembly scaffolding with PE-contaminated mate-pair libraries. Bioinformatics. 2016;32:1925–32. doi:10.1093/bioinformatics/btw064.

64. Rizk G, Gouin A, Chikhi R, Lemaitre C. MindTheGa

75. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33:2776–2778.

76. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. Genetics. 2014;196:973–83. doi:10.1534/genetics.113.160572.

77. Fan H, Ives AR, Surget-Groba Y, Cannon CH. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. 2015;16:522.

78. Cheng T, Wu J, Wu Y, Chilukuri RV, Huang L, Yamamoto K, et al. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. Nature Ecology & Evolution. 2017;1:1747–56. doi:10.1038/s41559-017-0314-4.

79. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLOS ONE. 2016;11:e0167047. doi:10.1371/journal.pone.0167047.

80. Löytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014;1079:155–70.

81. Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol. 2015;32:2798–800. doi:10.1093/molbev/msv150.

467

Table 1. The comparison of contiguity and correctness.

		Corn strain		Rice strain		Liu et al.	
	Statistics	New assembly	Gouin et al	New assembly	Gouin et al	Male	Female
Contiguity	Assembly size	384,358,373	437,873,304	379,902,278	371,020,040	543,659,128	531,931,622
	number of sequences	1,000	41,577	1,054	29,127	21,840	27,258
	N50	900,335	52,781	1,129,192	28,526	14,162,803	13,967,093
	L50	124	1,616	91	3,761	16	17
	N90	196,225	3,545	165,330	6,422	6,440	5,122
	L90	450	18,789	421	13,881	3,030	5,612
Correctness (BUSCO)	Complete	1,601	1,461	1,616	1,551	1,576	1577
	Complete and single-copy	1,572	1,276	1,573	1,518	1,442	1,480
	Complete and duplicated	29	185	43	33	134	97
	Fragmented	21	127	11	69	45	48
	Missing	36	70	31	38	37	33
	Total	1,658	1,658	1,658	1,658	1,658	1658

FIGURE LEGENDS

Figure 1. **Whole genome differentiation between strains** The histogram was made from Fst calculated in 10kb windows. The red vertical bar indicates Fst = 0.

Figure 2. **The effect of genetic linkage to selectively targeted loci on Fst** A. Fst calculated according to the distance from the nearest outlier of genetic differentiation. The left-most point corresponds to Fst from the outliers. The solid red curve is fitted smooth-spline with df = 5, and the red dotted curves are 95% confidence intervals with 1,000 bootstrapping. The vertical dotted line indicates the distance equal to 200kb. B. The barplot shows Fst from the scaffolds with outliers and without outliers.

Figure 3. **Testing background selection** The relationship of exon density with π (A) and Fst (B).

Figure 4. **Subsetting of sfC variants from ancestral sfR variants** Principal component analysis from the outliers (A) and from the whole genome (B). The red and blue dots represent sfC and sfR, respectively. C. A phylogenetic tree reconstructed from the whole genome.

Figure 5. **Speciation models concerning whole genome differentiation**. The process of speciation initiates from genetic differentiation between population a and population b, and finishes when these two populations are evolved to species a and species b with completely differentiated genomes. A. According to genome hitchhiking[9, 10] and genome-wide congealing model[39, 40], positive selection targeting many loci cause whole genome differentiation with a low extent by the combined effect of mild positive selection. Following positive selection rapidly accelerate the rate of whole genome differentiation by the synergistic effect of linkage disequilibrium across the whole genome until whole genome sequences are completely differentiated. In this model, whole genome differentiation is generated at the early stage of a speciation process. B. According to the genic view of speciation[1], the fully differentiated loci are progressively enlarged or additional fully differentiate loci are generated until whole genome sequences are differentiated. In this model, whole genome differentiation is generated at the end of a speciation process.

Figures

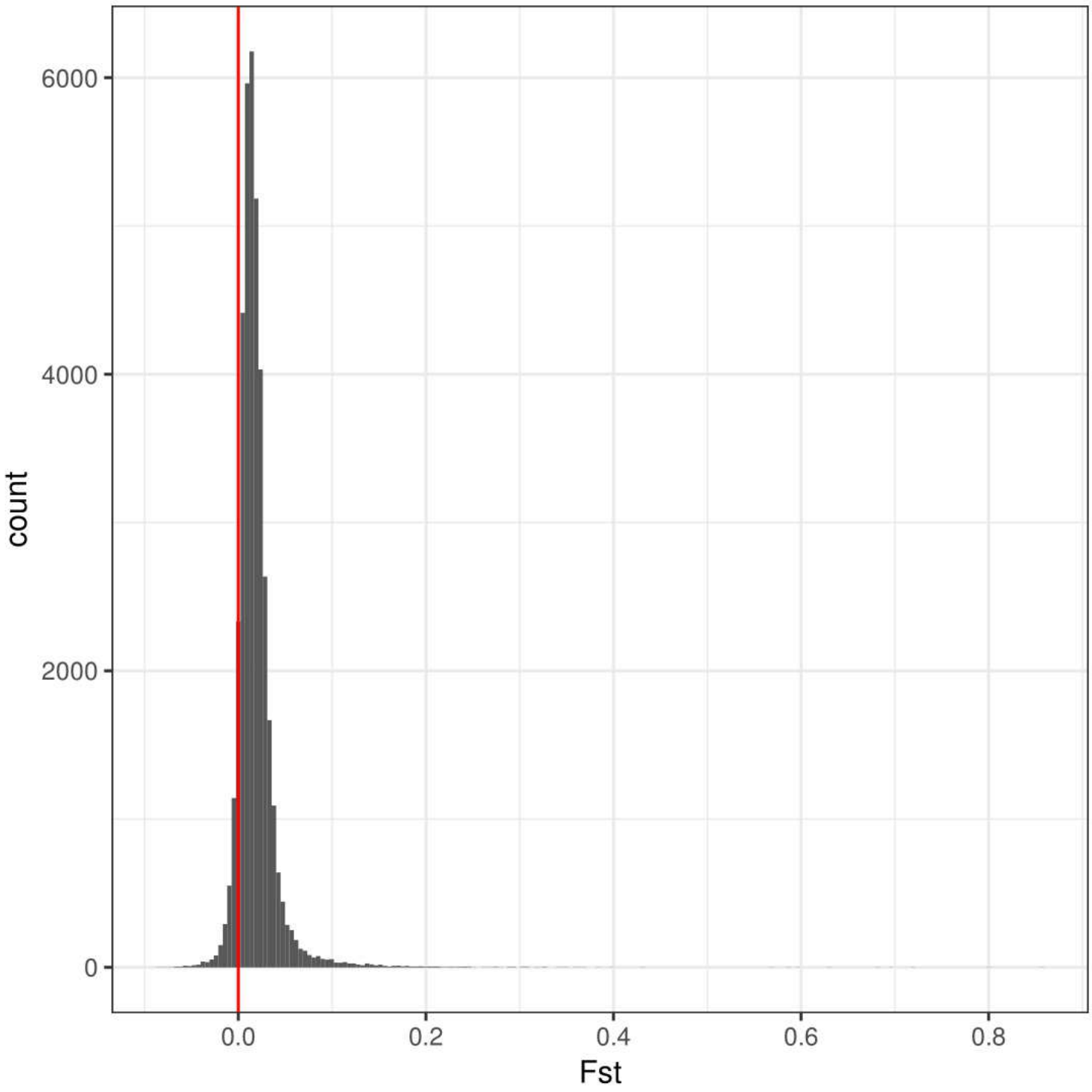

Figure 1

Whole genome differentiation between strains The histogram was made from Fst calculated in 10kb windows. The red vertical bar indicates Fst = 0.

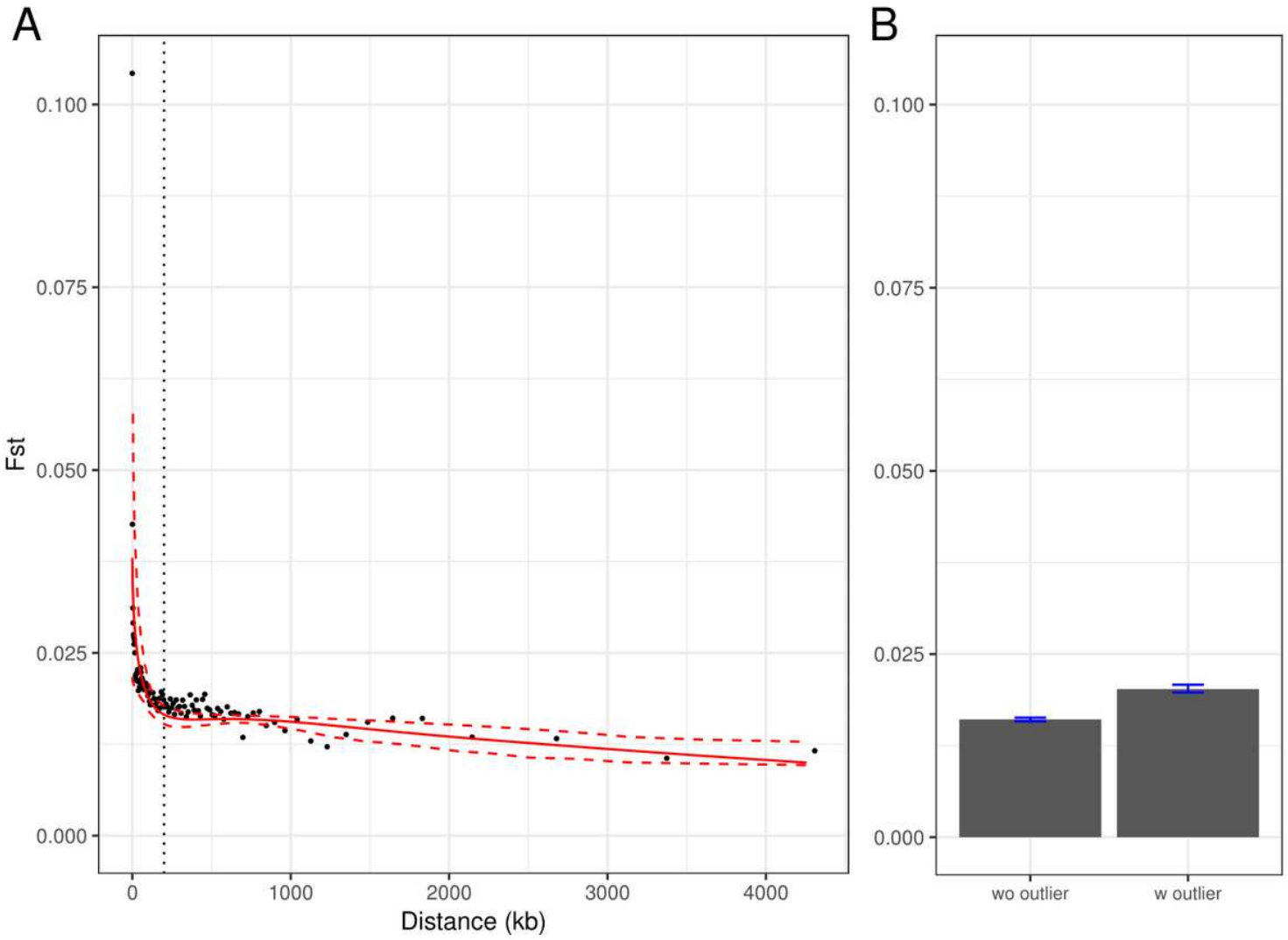

Figure 2

The effect of genetic linkage to selectively targeted loci on Fst A. Fst calculated according to the distance from the nearest outlier of genetic differentiation. The left-most point corresponds to Fst from the outliers. The solid red curve is fitted smooth-spline with df = 5, and the red dotted curves are 95% confidence intervals with 1,000 bootstrapping. The vertical dotted line indicates the distance equal to 200kb. B. The barplot shows Fst from the scaffolds with outliers and without outliers.

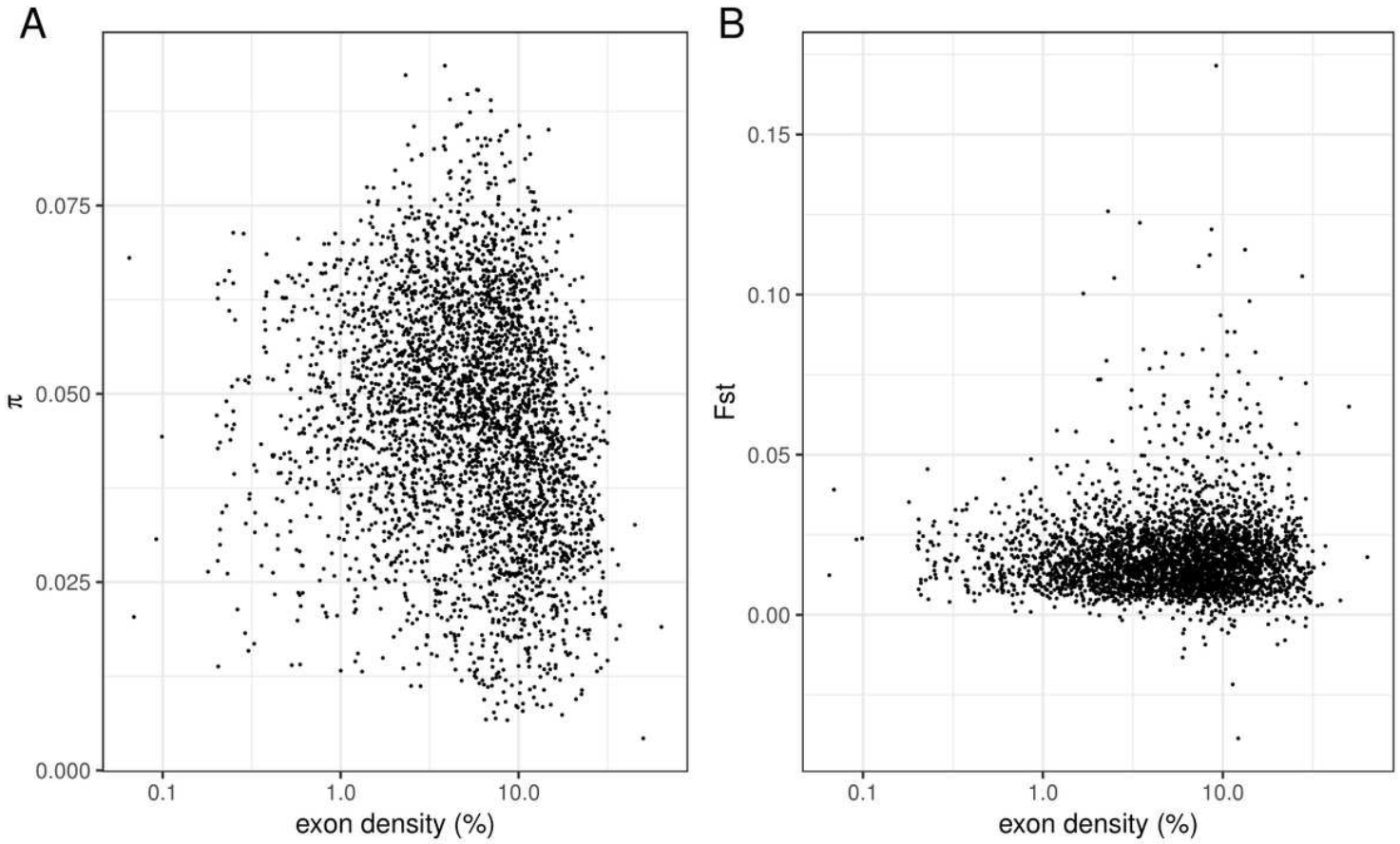

Figure 3

Testing background selection The relationship of exon density with π (A) and Fst (B).

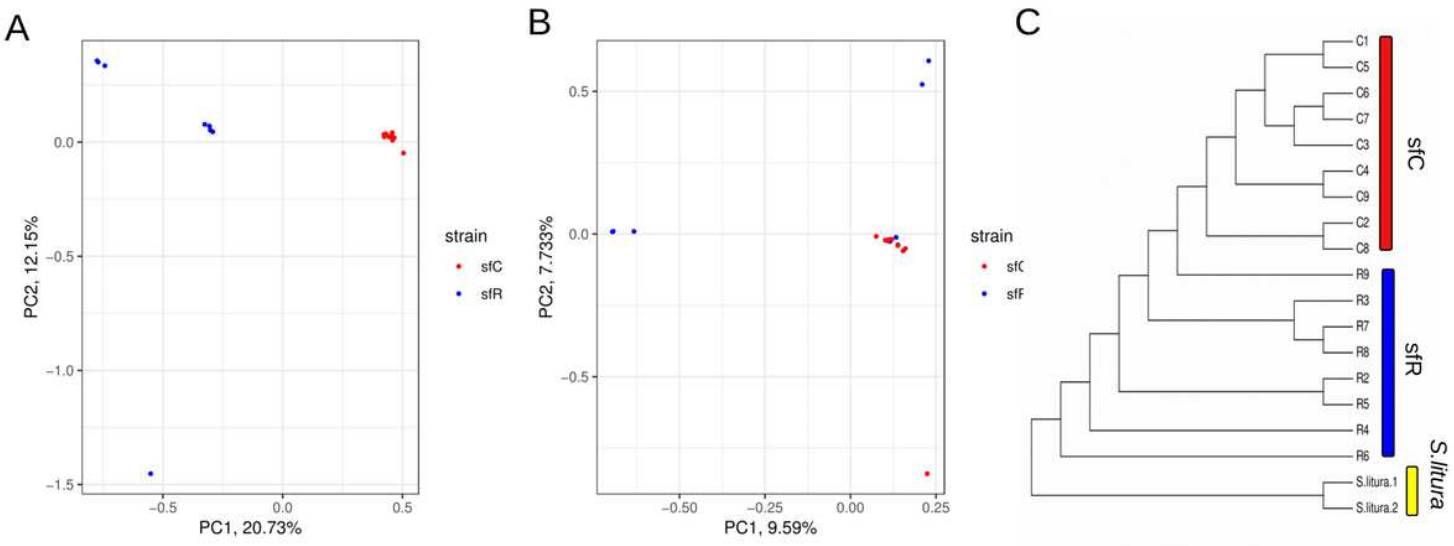

Figure 4

Subsetting of sfC variants from ancestral sfR variants Principal component analysis from the outliers (A) and from the whole genome (B). The red and blue dots represent sfC and sfR, respectively. C. A phylogenetic tree reconstructed from the whole genome.

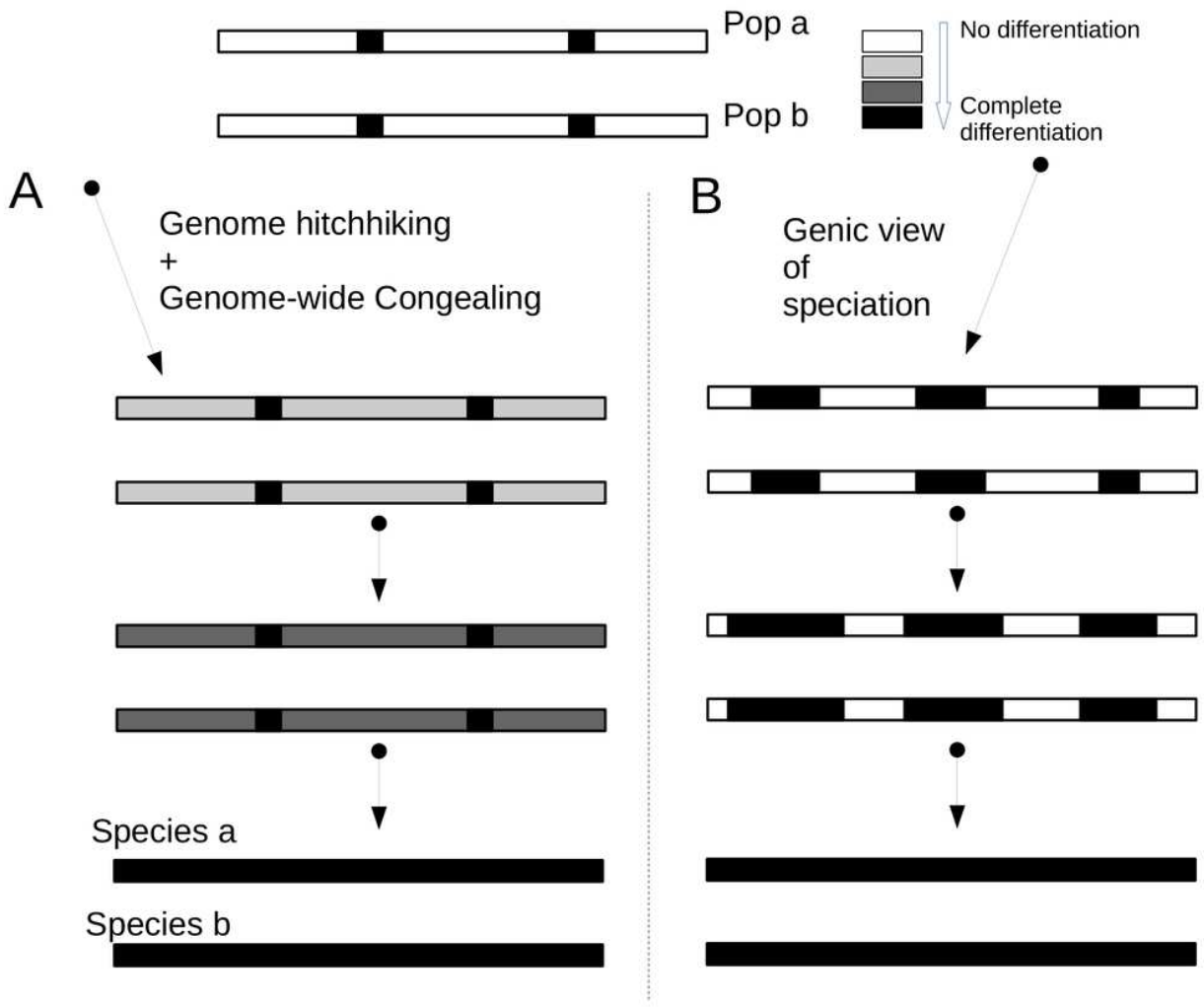

Figure 5

Speciation models concerning whole genome differentiation. The process of speciation initiates from genetic differentiation between population a and population b, and finishes when these two populations are evolved to species a and species b with completely differentiated genomes. A. According to genome hitchhiking[9, 10] and genome-wide congealing model[39, 40], positive selection targeting many loci cause whole genome differentiation with a low extent by the combined effect of mild positive selection. Following positive selection rapidly accelerate the rate of whole genome differentiation by the synergistic effect of linkage disequilibrium across the whole genome until whole genome sequences are completely differentiated. In this model, whole genome differentiation is generated at the early stage of a speciation process. B. According to the genic view of speciation[1], the fully differentiated loci are progressively enlarged or additional fully differentiate loci are generated until whole genome sequences are differentiated. In this model, whole genome differentiation is generated at the end of a speciation process.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FS20.png
- FS7.png
- FS21.png
- SI.pdf
- FS15.png
- FS16.png
- FS17.png
- FS18.png
- FS1.png
- FS3.png
- FS2.png
- FS6.png
- FS22.png
- FS5.png
- FS4.png
- FS14.png
- FS19.png
- FS12.png
- FS9.png
- FS11.png
- FS10.png
- FS8.png
- FS13.png