Article

Patterns of genomic divergence in sympatric and allopatric speciation of three Mihoutao (Actinidia) species

Yongbo Liu*, Wenhao Yu, Baofeng Wu and Junsheng Li

State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, 8 Dayangfang, Beijing 100012, China

*Corresponding author: E-mail: liuyb@craes.org.cn

Abstract

Isolation by geographic distance is involved in the formation of potential genomic islands and the divergence of genomes, which often result in speciation. The mechanisms of sympatric and allopatric speciation associated with geographic distance remain a topic of interest to evolutionary biologists. Here, we examined genomic divergence in three Actinidia species from large-scale sympatric and allopatric regions. Genome sequence data revealed that hexaploid Actinidia deliciosa originated from Actinidia chinensis and supported the speciation-with-gene-flow model in sympatric regions. The common ancestor of Actinidia setosa and A. deliciosa migrated from the mainland to the Taiwan Island ~2.91 Mya and formed A. setosa ~0.92 Mya, and the speciation of A. setosa is consistent with the divergence-after-speciation model with selective sweeps. Geographic isolation resulted in population contraction and accelerated the process of lineage sorting and speciation due to natural selection. Genomic islands contained genes associated with organ divergence and speciation because of limited gene flow between populations and drive population differentiation [5].

Understanding the process of speciation and post-speciation ecological adaptation has been advanced by studies of genomic polymorphism and divergence among the genomes of closely related species [6, 7], and genomic islands of divergence containing functional variants undergoing speciation have been widely observed [8]. Such genomic islands are generated by variation in gene flow between loci, ancestral diverged haplotypes, recurrent background selection with genomic recombination, and ecological adaptation [8–10]. The identification of genomic islands of population divergence or speciation requires information on ancient polymorphisms, gene flow, and natural selection on traits [8, 11], as genomic divergence varies with evolutionary history, length of isolation, and adaptation to local post-speciation selection [12].

Genomic data provide evidence for the role of tectonic evolution in speciation and bioregion formation [10, 13, 14]. This study aimed to understand the genomic basis for speciation of three closely related wild Actinidia species (Actinidia chinensis, Actinidia delicosa and Actinidia setosa) and to elucidate their adaptation to local selection in nature, employing population genomic resequencing. The Actinidia genus (named Mihoutao in Chinese) originated in China and is now mainly distributed from the Qinling Mountains to the Hengduan Mountains. The geographic distribution of A. chinensis and A. delicosa has sympatric regions, with the former predominating in the eastern part of China and the latter in the west, whereas A. setosa, an endemic species, is restricted to the island of Taiwan and is allopatric to the other two species. Island and mountain topography separate biological populations and accelerate population differentiation and speciation because of limited gene flow between populations [14–17]. Thus, the three wild kiwifruit species provide ideal subjects for investigating the role of geographic isolation in sympatric and allopatric speciation.
The phylogenetic relationships of the three species are still controversial. *A. chinensis* has diploid and tetraploid wild individuals, whereas *A. delicosa* exhibits complex ploidy in nature, ranging from tetraploid to octoploid [18]. *A. setosa* has only diploid individuals. *A. chinensis* and *A. delicosa* have been described as two distinct species based on morphological features such as hair types, over-wintering buds, and mature fruit flesh color [19]. *A. setosa* was once considered to be a variety of *A. chinensis* [19, 20], but it was subsequently reclassified as a distinct species [21]. Huang et al. (2014) reclassified them as three varieties of a single species [22], although morphological intermediates and hybrid forms between *A. delicosa* and *A. chinensis* have been found in nature [23]. Chloroplastic and mitochondrial DNA sequence data suggest that *A. delicosa* and *A. chinensis* represent a species complex from which *A. setosa* diverged [25].

**Materials and methods**

**Plants and sampling**

The three species *A. chinensis*, *A. delicosa*, and *A. setosa* are closely related but show distinct morphological differences; the main difference is the trichomes in fruits, leaves, and stems. *A. chinensis* fruits are subglobose and tomentose or glabrous, *A. delicosa* have subglobose to cylindrical fruits that are densely hispid, and *A. setosa* fruits are subglobose or ellipsoidal and densely hispid [25]. We sampled the three Actinidia species in nature from six *A. chinensis* populations, three *A. delicosa* populations, and one *A. setosa* population sampled from three mountains on the Taiwan Island (Table S1). Leaves from more than 500 individual plants were sampled randomly. The ploidy of individual plants was measured by flow cytometry 1–2 days after collecting fresh leaves (Supplementary Fig. S1) [26]. In addition, *Actinidia arguta* was sampled from Shaanxi province as an outgroup.

**Library preparation**

We extracted genomic DNA from 187 samples of three species from nine natural populations (1.5 μg per sample). Following the manufacturer’s recommendations, DNA sequencing libraries were generated using a TruSeq Nano DNA HT sample preparation kit (Illumina, USA). DNA samples were fragmented by sonication to a size of 350 bp, and further PCR amplification was performed to end-polish, A-tail, and ligate the DNA fragments with the full-length adapter for Illumina sequencing. The PCR products were purified (AMPure XP system), and the size distributions of the libraries were analyzed and quantified using an Agilent 2100 Bioanalyzer and real-time PCR, respectively.

**Genome sequencing and quality control**

Whole genomes of 187 samples were sequenced on the Illumina HiSeq 2500 platform, generating 1346 Gb of raw data. We removed low-quality paired-end reads (>10 nt aligned to the adaptor, allowing for ≤10% mismatches; ≥10% unidentified nucleotides [N]; >50% bases with phred quality <5; putative PCR duplicates) and retained 935.5 Gb of high-quality paired-end reads (Q20 ≥ 94.2% and Q30 ≥ 86.7%) (Table S1). After removing low-quality samples, a total of 139 samples (including one outgroup) with high-quality data were used for subsequent analyses, representing 12–15 plants per population (Table S1). Genome alignment of the *Actinidia* sequences indicated an average depth of 7.7-fold and coverage of 93.3% relative to the reference genome of *A. chinensis* (the heterozygous kiwifruit “Hongyang”) (Table S2) [27].

**Read mapping and SNP calling**

The remaining high-quality paired-end reads were mapped to the diploid *A. chinensis* “Hongyang” reference genome [27] using BWA (Burrows-Wheeler Aligner) version 0.7.8 [28] with the parameters “-mem -t 4 -k 32 -M” [28]. SAMtools was used to remove duplicate reads [29]. After alignment, single nucleotide polymorphisms (SNPs) were called on a population scale using a Bayesian approach implemented in SAMtools (version 1.3). We then calculated genotype likelihoods and allele frequencies from reads for each individual at each genomic location using a Bayesian approach. SNPs were identified using the “mpileup” command with the parameters “-q 1 -C 50 -t SP -t DP -m 2 -F 0.002”. To exclude SNP-calling errors caused by incorrect mapping or InDels, SNPs were filtered (coverage depth ≥2 and ≤50, RMS mapping quality ≥20, maf ≥0.05, miss ≤0.1). Consequently, 3,556,911 high quality SNPs remained after filtering from the initial collection of 44,549,362 raw SNPs.

To solve the difference in SNPs from diploid and polyploid populations, we have called SNPs with the GATK software setting HaplotyperCaller “—ploidy 4” for tetraploid populations and “—ploidy 6” for hexaploid populations. However, most software packages, including ADMIXTURE, PLINK, and VCFtools, cannot deal with multiallelic SNPs but only with biallelic SNPs, which limits the next analysis. We only obtained an NJ tree based on multiallelic SNPs for polyploid populations after low-quality SNP filtering, and the tree was the same as that constructed from biallelic SNPs. This may indicate that the analysis of biallelic SNPs is reliable for polyploid populations, as previous studies have suggested [30]. In addition, we analyzed genetic structure using only homozygous or heterogenous biallelic SNPs of tetraploid and hexaploid individuals (data not shown), and the results were similar to those obtained using all SNPs. Thus, the next analysis was performed using all biallelic SNPs. SNPs were annotated based on comparisons to the reference genome using ANNOVAR [31].

**Phylogenetic tree and population structure**

Individual-based maximum-likelihood (ML) and neighbor-joining (NJ) trees were constructed based on the
p-distance using TreeBeST v1.9.2 software [32]. The population genetic structure was analyzed using an expectation maximization algorithm implemented in frappe [33]. The number of assumed genetic clusters K was set from 2 to 8, with 10,000 iterations for each run. Principal component analysis (PCA) was performed to evaluate the genetic structure of populations using GCTA software [34].

Genetic diversity and LD

Nucleotide diversity (θπ) of each population was calculated using a sliding-window approach (40-kb windows with 20-kb increments). We calculated the correlation coefficient (r²) of alleles to estimate linkage disequilibrium (LD) using Haploview [35]. The average r² value was calculated for pairwise markers in 500-kb windows and averaged across the whole genome, and LD decay figures were drawn using an R script.

Demographic history and gene flow

To investigate the demographic history of the three species, the effective population size (Ne) of the three Actinidia species was estimated over the last 10 million years using a pairwise sequentially Markovian coalescence (PSMC) model [36]. Parameters were set to “-N N30 -t15 -r5 -p’4 þ 25*2 þ 4 þ 6’”. The mutation rate (μ) was set to 5.4 × 10⁻⁹ per base per generation [37], and the generation time (g) was set to six years because kiwifruit begin to bear fruit at about three years of age.

We inferred the demographic history of the three species using the diffusion approximation for demographic inference (θaαί) approach [38]. We used the model θ = 4 × μ × Nref × L, where μ was the mutation rate (5.4 × 10⁻⁹) and L was the number of SNPs. To ensure selective neutrality, only the SNPs within intergenic regions of autosomal chromosomes were used. We evaluated model fitting and optimized model selection using a composite-likelihood ratio test. The optimal model was selected with the highest log-likelihood value from five constructed divergence models.

The divergence time of these natural Actinidia populations in China was estimated under a Bayesian approach using BEAST v2.3.4 [39] according to a neutral mutation rate (μ = 5.4 × 10⁻⁹). MCMC runs were performed for 10,000,000 generations, with sampling every 1000 generations following a burn-in of the initial 50% of cycles. We examined the sampling adequacy and convergence of the chains to a stationary distribution using Tracer v1.5. TreeAnnotator summarized a post burn-in tree and produced a maximum clade credibility chronogram showing mean estimated divergence time with 95% HPD intervals [40].

A population-level admixture analysis was conducted to detect historical gene flow between wild A. chinensis populations using TreeMix v1.1.24 [41] with the command “-i input -bootstrap -k 10000 -m migration events -o output”. The populations were regarded as candidates around which potential migration edges were added, generating new arrangements of the ML tree accounting for migration events [41]. From one to ten migration events were gradually added to the ML tree until 98% of the variance between the populations could be explained.

Identification of genomic windows and selected regions

To identify genome-wide selective sweeps associated with natural adaptation, we measured genome-wide variation for the pairwise comparisons of the three species (A. chinensis, A. deliciosa, and A. setosa) and effects of geographic isolation on speciation. Geographic isolation included three scenarios, i.e. oceanic island (TWS + ZSC as one oceanic population vs. TPC + YAC + SYC + JZD + ZLD as one mainland population), lake island (JSC island vs. YMC land), and mountain isolation (four diploid populations, TPC, YAC, SYC, and YMC). A sliding window approach (40-kb window and 20-kb step size) was used to identify selected regions associated with specific adaptations. The genome-wide distribution of FST and Tajima’s D was calculated using VCFtools [42]. Absolute divergence (DXY) was calculated from sample allele frequency posterior probabilities at each site and was then averaged over each 20-kb window using an in-house python script.

Population-scaled recombination rates (ρ = 4Ne c) were estimated for diploid species between each pair of SNPs and then weight-averaged over each 20-kb window using the program LDhelmet v1.10 [43] with default parameters (100,000 burn-in iterations, 1,000,000 Markov chain iterations, and a block penalty of 50). We only used retained windows with at least 20 SNPs after filtering SNPs with MAF > 5% to minimize the effects of rare variants [9].

Genomic regions with significant high FST values (Z-FST > 2, corresponding to a top 5% level) were identified as highly divergent regions and designated genomic islands. Z-FST was defined by the formula Z-FST = (FST × FST')/std-FST (Han et al. 2017), where FST was a per-window estimate, and FST' and std-FST were the mean and standard deviation of FST across windows. To identify regions in the genome linked to potential lineage-specific adaptations, genetic diversity, the number of SNPs, Tajima’s D, and DXY in genomic islands were compared with genomic background for any pairwise comparisons. We compared characteristics of genomic islands in the pairwise comparisons.

We considered the windows with the top 5% FST and log2 (θπ ratio) values simultaneously as candidate outliers under strong selective sweeps for pairwise comparisons of the three species and geographic isolation scenarios. All outlier windows were assigned to corresponding SNPs and genes.

ANOVA and randomization analysis were used to detect differences in DXY between genomic islands and the rest of the genome for each pairwise comparison using R software.
Annotation analysis of selection regions
We annotated genes in outlier genomic regions using the A. chinensis "Hongyang" reference genome. Enrichment analyses of these genes were conducted using the Gene Ontology (GO) [44] and KEGG (http://www.genome.jp/kegg) databases. Significant enriched gene function was tested using a false discovery rate (FDR) with a corrected binomial distribution probability approach at the level of P < 0.0545. The functions of the candidate genes were inferred from the NCBI (http://www.ncbi.nlm.nih.gov) and PIR (http://www.uniprot.org) databases.

GWAS analysis
A total of 3,556,911 SNPs (MAF ≥0.05; missing rate ≤0.1, depth ≥2) were used in GWAS for the trichome trait. The trichome trait was defined here as a presence or absence trait; A. chinensis had a value of "0" (absence) and A. delicosa a value of "1" (presence). To reduce the risk of false positives from population structure, we conducted the analysis with the genome-wide efficient mixed-model association (GEMMA) software package [45]. For mixed linear model analysis, we used the following equation:

\[ y = X\alpha + S\beta + K\mu + \epsilon. \]

where \( \alpha \) and \( \beta \) are fixed effects representing marker effects and non-marker effects, respectively; \( y \) represents phenotype; and \( \mu \) represents unknown random effects. \( X, S, \) and \( K \) are the incidence matrices for \( \alpha, \beta, \) and \( \mu, \) respectively, and \( \epsilon \) is a vector of random residual effects. The \( S \) matrix for population-structure correction was built up using the top three PCs. We built up the \( K \) matrix using the matrix of simple matching coefficients.

Results
Genome resequencing and population structure
After stringent quality filtering, we identified a total of 3,556,911 high-quality SNPs from the initial collection of 44,549,362 raw SNPs. The number of SNPs varied among the populations and ploidy races, and the percentage of heterogeneous SNPs in hexaploid and tetraploid populations was higher than that in diploid populations, with the lowest percentage in the TWS population (Table S3).

Based on phylogenetic analysis, individuals from the same population clustered together, and each of the three species clustered into a different subclade (Fig. 1A). A. setosa (TWS) had short clade branch lengths. The ZSC population (tetraploid A. chinensis) and the SMD population (tetraploid A. delicosa) clustered together and were located in the middle of diploid TIC and TPC + YAC + SYNC.

Principal component analysis (PCA) of SNP genotypes among all individuals clearly revealed a history of population divergence (Fig 1B). The first principal component (PC1; variance explained = 35.7%) separated A. setosa (TWS) from the other two species, and the second principal component (PC2; variance explained = 27.1%) separated hexaploid A. delicosa (JZD and ZLD) from the other lineages. The geographically isolated populations (eastern TIC vs. central SYC + YAC + TPC) within A. chinensis were separated by the third principal component (variance explained = 13.5%). The PCA results were consistent with the clustering analyses that showed geographic populations (best K = 6) (Fig. 1C). Three clusters corresponded to the three species (K = 3), and when K = 2, two clusters were A. chinensis and A. setosa, and the hexaploid A. delicosa included the other two species information (Fig. 1C). The tetraploid A. delicosa (SMD) had a genetic background from other populations in cases of all K values, indicating that it is likely to be an intermediate phenotype produced in the heterozygote, which may be attributed to shared ancestral polymorphism or introgression events (Fig. 1C).

Demographic history
The effective population size (\( Ne \)) of A. chinensis and A. setosa began to decrease at ~1 Mya, experienced an increase at ~0.3–0.5 Mya, and then experienced a bottleneck event at ~100 kya (Fig. 2A) around the Marine Isotope Stage 5 (MIS, 80–130 kya) during the last major interglacial period. A. delicosa increased rapidly after ~100 kya and exhibited the largest historical \( Ne \) values (Fig. 2A). The smallest historical \( Ne \) in A. setosa was consistent with its restricted island distribution in Taiwan.

A Bayesian analysis revealed that the last common ancestor of the three species was present ~20.2 Mya (Fig. 2B), consistent with one leaf fossil record of Actinidia (20–26 Mya, during the Miocene Epoch) [46]. The divergence time between A. chinensis and the other two species was ~10.1 Mya (Fig. 2B), and A. delicosa and A. setosa diverged ~0.92 Mya. The degree of gene flow between A. delicosa and A. setosa was the highest, followed by that between A. delicosa and A. chinensis, and the lowest was between A. chinensis and A. setosa (Fig. 2B).

At the population scale, BEAST analysis showed that the TWS population diverged from the inland populations ~2.91 Mya (95% confidence interval 2.87–2.95) (Fig. 2C). The two hexaploid populations ZLD and JZD diverged 0.76 Mya and together diverged from the other populations 2.10 Mya. The divergence times between these diploid populations were ~1.13–1.69 Mya (Fig. 2D). TreeMix analysis showed gene flow in local regions, from TIC (diploid) to SMD (tetraploid) in eastern China, and from north to south, including from TPC to TIC and then to TWS (Fig. 2C).
Genomic islands of divergence

The landscape of the weighted $F_{ST}$ distribution varied in the three pairwise species comparisons (Kolmogorov–Smirnov test, $P < 2.2e^{-16}$) (Fig. S3, Table S4). Genetic divergence ($F_{ST}$) was highly heterogeneous along genomes (Fig. S4A) and negatively related to nucleotide diversity ($\pi$) and Tajima’s D (Fig. S5; Fig. 3A). Genomic regions with $Z$-transformed $F_{ST}$ values ($Z-F_{ST}$) $\geq 2$, corresponding to 4–5% of windows, were considered to be genomic islands for pairwise comparisons. We identified 56 common genomic islands among the three pairwise comparisons. These genomic islands showed reduced $\pi$ and Tajima’s D values, indicating that linked selection had acted on these islands (Table S5).

We observed significantly elevated $D_{XY}$ in genomic islands compared with the rest of the genomes in pairwise comparisons of the three species ($P < 0.001$; randomization test) (Fig. 3B; Table S5), irrespective of whether the comparisons were between sympatric (A. chinensis and A. deliciosa) or allopatric species (A. setosa vs. A. chinensis or A. deliciosa). Genomic islands showed a significantly reduced recombination rate ($\rho$) compared with the rest of the genome for the three pairwise comparisons ($P < 2.2e^{-16}$, Mann–Whitney U test; Table S5).

To characterize the genetic basis of this speciation, we conducted selective sweeps to identify genes with the top 5% of $F_{ST}$ (genomic islands) and $\log_2 (\theta/\pi)$ ratio simultaneously (Fig. S4B). Among a total of 719 genes in these genomic islands under positive selection, 543 genes were identified (Table S6). Of these, 55 genes were shared among the three comparisons, which indicates the presence of diverged haplotypes in their ancestral population possibly as a result of strong divergent selection. No highly enriched GO terms or KEGG pathways were found for the three pairwise comparisons. Some genes were associated with local adaptation (e.g. WRKY 40), organ development (e.g. ZFP, bHLH, MYB), and reproductive isolation (e.g. FPA, KAN2, MS1) (Table S6).

Divergence of isolation

The genetic diversity ($\theta/\pi$) of oceanic and lake islands was not significantly different from those of land populations (Table S7). $F_{ST}$ values of oceanic isolation (0.2–0.4) were significantly higher than the value for lake isolation (<0.2) (Fig. S6). The four mountain populations of diploid A. chinensis did not show an isolation-by-distance effect of $F_{ST}$ (Table S7).
Figure 2. Speciation and demographic history of the three Actinidia species. (A) PSMC analysis results for Ne over the last 10^7 years. (B) ""∂_a∂_i"" analysis showing the demographic history of the three Actinidia species (CH, A. chinensis; DE, A. delicosa; SE, A. setosa) from ~2.02 x 10^7 years ago to the present. Species divergences occurred at 1.01 x 10^7 and 9.24 x 10^5 years ago. The average number of migrants per year between species is shown between the black arrows. (C) Phylogenetic network of the inferred relationships among the nine natural populations. Arrows indicate migration events, and a spectrum of heat colors indicates the migration weights of the migration events. The scale bar shows ten times the average standard error of the entries in the sample covariance matrix. (D) Divergence times between natural populations estimated using BEAST analysis.
Figure 3. The heterogeneity of genomic divergence in the three species. (A) Correlations of absolute divergence (Dxy), genetic divergence (FST), Tajima’s D (Tjm), and nucleotide diversity (Pi) for the three pairwise comparisons of A. chinensis (C), A. deliciosa (D), and A. setosa (S). Blue and red circles indicate a positive or negative correlation, respectively. The color intensity and circle size are proportional to the Spearman’s correlation coefficient. (B) Values of Dxy in genomic islands compared with the background genome in the three pairwise comparisons of A. chinensis–A. deliciosa (top), A. chinensis–A. setosa (middle), and A. deliciosa–A. setosa (bottom). The "ou" represents genomic islands, and "bg" represents the background genome.
Figure 4. Genomic regions with strong selective signals in natural populations of the three species. (A) Tajima’s D values in oceanic island (lower left), mainland (upper left), lake island (lower right), and neighbor land (upper right) populations. (B) Decay of linkage disequilibrium (LD) in the natural populations, with one line per population. (C) Distribution of log2(θπ ratios) and FST values calculated in 40-kb sliding windows with 20-kb increments between the oceanic group (TWS and ZSC) and the mainland group (TPC + YAC + SYC + JZD + ZLD). The data points in blue and green (corresponding to the top 5% of the empirical log2(θπ ratios) ratio distribution and the top 5% of the empirical FST distribution) are genomic regions under selection.

The Tajima’s D values of oceanic island and lake island populations significantly differed from 0 (>0), indicating positive selection or a bottleneck effect in oceanic islands (Fig. 4A). Moreover, ZSC showed a higher LD among loci than other populations (P < 10^{-16}, Mann–Whitney U test) (Fig. 4B), reflecting relatively higher inbreeding and genetic drift in the island population and skewed allele frequency spectra. The three mountain populations, ZLD, JZD, and YAC, showed higher LD than the other populations (except ZSC) (Fig. 4B).

Oceanic isolation showed 1006 genomic windows and lake isolation showed 910 windows, with 76 common windows. These genomic islands showed increased levels of Tajima’s D and nucleotide diversity (π) in oceanic island populations, whereas genomic islands had reduced Tajima’s D values and nucleotide diversity for lake isolation (Table S5). The values of absolute divergence D_{XY} were positively correlated with FST for oceanic isolation but not for lake isolation (Fig. S7). These results indicate the effect of geographic isolation on genomic divergence is related to the time length of isolation.

D_{XY} was elevated in genomic islands compared with the rest of the genomes in the lake island comparison (P < 0.001), whereas D_{XY} decreased in genomic islands in the oceanic island comparison (P < 0.001) (Fig. S8; Table S5). The decreased D_{XY} values indicate that such genomic islands resulted from recurrent background selection or selective sweeps because long-term selection reduced genetic variation in the most recent common ancestor (MRCA).

To examine the genetic basis of geographic isolation associated with speciation, we identified genes undergoing selective sweeps using the top 5% of FST values and θπ ratio cutoffs (Fig. 4C). Among a total of 780 genes under positive selection, 623 genes were identified (Table S8). A total of 11 identified genes were shared among the oceanic and lake isolation comparisons. One highly enriched GO term, “single-organism developmental process” (GO:0044767, biological process), was identified (7 genes, corrected P = 0.026) for oceanic isolation (Fig. S9), and the significantly enriched KEGG pathway “selenocompound metabolism” (2 genes, corrected P = 0.08) was associated with oceanic isolation (Fig. S10). No significantly enriched GO terms or KEGG pathways were found for lake isolation.

GWAS analysis for the trichome trait
Trichome morphology differs among the three species. A genome-wide association study (GWAS) for the trichome trait was performed to test the local adaptation
of the three species, and it revealed 2268 SNPs and 2033 genes for the trichome trait (Table S9). The selective sweep analysis for oceanic isolation and the trichome trait GWAS revealed 20 common genes (Table S10). The trichome trait was significantly associated with an interval on chromosome (Chr) 19, as well as Chr 24 regions of the reference genome (Fig. S11, Table S11). The highest associated SNPs on Chr 19 and Chr 24 corresponded to two genes, Ach19g305841 (phosphoglycerate kinase, PGK) and Ach24g141021 (DNA/RNA-binding protein KIN17), respectively (Fig. 5).

GWAS analysis for the trichome trait revealed genes involved in trichome initiation, formation, and development, such as zinc finger proteins (ZFPs), TCP proteins, bHLH and WD40 proteins, MYB transcription factors, DUF/TBL1 (trichome birefringence-like) proteins, cellulose synthase, and in ecological adaptation via biotic and abiotic stress resistance, such as members of the WRKY gene family, PERK1, auxin response factors and induced proteins, and genes in the ethylene (EIN3) and gibberellin (GA) phytohormone pathways (Table S9).

Discussion
Our population genomic analysis suggests heterogeneous patterns of genomic divergence in the sympatric and allopatric speciation of three Actinidia species. The genomic islands of divergence that distinguish the three Actinidia species most likely derived from gene flow in the sympatric regions, lineage sorting of ancient polymorphisms, and ecological adaptation under selective sweeps. Geographic isolation plays a key role in the formation of genomic islands during speciation, which varies with the time length of divergence.

Our results indicate the common ancestor of A. deliciosa and A. setosa diverged from A. chinensis at \( \sim 10.1 \) Mya. A. chinensis and A. deliciosa interbreed in sympatric regions but maintain distinct morphologies. A previous study found that A. deliciosa has a family of repeat sequences not present in diploid A. chinensis, which suggests that it may be an allopolyploid species and that other species were involved in the origin of A. deliciosa [47]. In the current study, we found that A. deliciosa diverged from A. chinensis, but there was insufficient information to support its allopolyploid or autoploid origin. Diploid A. chinensis is probably itself a cryptic polyploid or rediploidized paleopolyploid because its basic chromosome number (n = 29) is high [18, 48]. Most chromosomes of A. chinensis have Robertsonian-like centromeric translocations [49]. Polyploidy is one of the important mechanisms of speciation, and approximately
70% of flowering plants have undergone polyploidy events during their evolutionary history [50], which increases their adaptive capability. Indeed, we found that the effective population size of *A. deliciosa* increased rapidly, which presumably was advantageous for ecological adaptation. Morphological intermediates and hybrid forms between *A. deliciosa* and *A. chinensis* have been found in nature [23], and the two species have had extensive gene flow and shared ancestral variation, even across ploidy levels [51].

Our results suggest that *A. setosa* is derived from the last common ancestor shared with *A. deliciosa* in mainland China ∼2.91 Mya and formed a separate species in the Taiwan Island at ∼0.92 Mya. The divergence time (∼2.91 Mya) of *A. setosa* was less than the formation time (∼5 Mya) of the Taiwan Island [52]. A low effective population size (Ne) and recombination rate (ρ) probably contributed to the low degree of *A. setosa* genetic diversity [53]. Liang and Ferguson (1985) proposed that *A. setosa* is a separate species [21], and Chat et al. (2004) found that *A. setosa* diverged from *A. chinensis* and *A. deliciosa* based on chloroplastic and mitochondrial DNA sequences [24]. We conclude that the formation of *A. setosa* resulted from its low effective population size under island isolation, with minimal gene flow, reduced recombination, and increased genetic drift due to a high hitchhiking effect or selective sweeps [54].

The level of divergence (FST) between the three species was highly heterogeneous along the genome, probably owing to demographic processes and genetic drift [11]. Population contraction due to geographic isolation (*A. setosa*) or population expansion due to gene flow and introgression (*A. deliciosa*) resulted in rapid stochastic loss or fixation of alleles (haplotypes) [55]. However, absolute divergence (DXY) was elevated in genomic islands for both sympatric speciation (*A. chinensis*–*A. deliciosa*) and allopatric speciation (*A. deliciosa*–*A. setosa* and *A. chinensis*–*A. setosa*). Elevated DXY in genomic islands resulted from differential gene flow between genomic regions or from the presence of anciently diverged haplotypes [8]. In sympatric speciation under the speciation-with-gene-flow model, genetic exchange via hybridization and introgression can result in a genetic admixture, which elevates absolute and relative divergence in genomic islands [56]. In allopatric divergence without gene flow, genomic islands are generally formed by lineage sorting [8]. Divergence resulting from geographic isolation leads to lineage sorting and the retention of ancestral polymorphisms, with the accumulation of species-specific genetic changes and the sorting of ancestral variation among lineages [57]. Moreover, natural selection (selective sweeps) in the isolation of the Taiwan Island accelerated lineage sorting in accordance with the divergence-after-speciation model [10, 12], as there was no recurrent interspecific hybridization and introgression during allopatric speciation. *A. chinensis* occurs mostly at altitudes between 200 and 1200 m, whereas *A. deliciosa* is usually found at 800–1400 m and *A. setosa* at 1300–2600 m [48]. We conclude that selective sweeps decreased the genetic diversity (π) of *A. setosa* in the Taiwan Island but fixed the variation of neutral loci through a hitchhiking effect and increased genetic diversity in genomic islands, which exhibited a low recombination rate compared with the background genome. Low recombination regions are generally associated with a high hitchhiking effect or selective sweeps [54]. This suggests a key role for genomic islands in plant speciation, associated with geographic isolation that presents a barrier to gene flow and recombination and thus accelerates divergence under selective sweeps [58].

Interestingly, we found reduced DXY values in genomic islands of oceanic island populations (TWS + ZSC) compared with mainland populations, indicating ongoing background selection and recurrent selective sweeps due to linked selection in the genomic islands of isolated populations [8]. This suggests that habitat isolation resulted in strong natural selection stresses with low genetic diversity and positive high Tajima’s D in oceanic island populations. Inbreeding caused by geographic isolation leads to the fixation and deletion of alleles and subsequent loss of genetic variance [59, 60]. Island populations are often genetically isolated from the mainland, and most island populations are small, thus facilitating random genetic drift [57].

Our results showed that the genetic effects of geographic isolation varied with the time length of isolation, over 60 years for the thousand-island lake isolation, 7000–9000 years for the Zhoushan Archipelago, and ∼5 Mya for the Taiwan Island [52]. Long-term oceanic isolation led to decreased genetic diversity (π) of populations in the oceanic islands and diverged (FST) from mainland populations, whereas short-term lake isolation did not. A new species (*A. setosa*) was formed on the Taiwan Island, and a tetraploid population of *A. chinensis* was generated in the Zhoushan Archipelago. Geographic isolation is one of the main reasons for population differentiation and eventual speciation, with limited gene flow among populations [15, 61–63]. As an example, Mes et al. [64] found that ca. 85% of the species in genus *Aeonium* resulted from geographic isolation. We suppose that the Zhoushan Archipelago population derived from the SMD population because the Zhoushan Archipelago and the Siming mountains belonged to the Tiantai Mountain Chain before formation of the Zhoushan Archipelago as a result of rising sea levels. The high LD of the ZSC population indicates that individuals experienced positive selection or selective sweeps in the isolated Zhoushan Archipelago, which may explain the maintenance of high divergence between species despite extensive gene flow [65]. In addition, genetic differentiation was not significantly related to geographic distance in four diploid *A. chinensis* populations. The dioecy characteristics of *A. chinensis* can decrease its inbreeding rate and counteract the negative effects of geographic isolation [66]. Gene flow can weaken the differentiation effects of geographic isolation [67–70]. The differential effects of geographic
isolation on genetic diversity resulted from the time length of isolation and local ecological adaptation, which probably cause a time-delayed extinction debt [71].

Genes in genomic islands have been associated with local adaptation, organ development, and reproductive isolation, and they are likely to experience morphological and ecological divergence between lineages due to geographic isolation [65]. For example, MS1 (PHD finger protein MALE STERILITY 1), a nuclear transcriptional activator, participates in tapetal development and pollen wall biosynthesis. FPA regulates flowering time via an independent day-length pathway, and the KANADI gene family regulates the development of the carpel and the outer integument of the ovule [72]. In addition, a large proportion of genes were transcription factors, e.g. WRKY transcription factors involved in the response to pathogens [73]. In the current study, some genes associated with isolation divergence were related to trichome initiation, such as members of the TBL (Trichome Birefringence-like) gene family that participate in the synthesis and deposition of secondary cell wall cellulose, and C2H2-like zinc finger protein (ZFP) genes that regulate trichome initiation, branching, and shoot maturation [74]. MYB, bHLH69, and WD3 genes are associated with trichome differentiation, and DUF/TBL genes contribute to secondary wall cellulose synthesis [75]. Similarly, a WD40-bHLH-MYB regulatory mechanism plays a key role in controlling trichome formation and development [76–78]. In addition, proteins in the plant-specific transcription factor families TCP, YUC, and NAC modulate cell proliferation and differentiation, regulate development, and may be involved in trichome cell differentiation and leaf development and morphology [11, 79]. Our GWAS analysis also identified PGK1 and KIN17 genes, which are significantly associated with the trichome trait. KIN17 participates in the control of plant growth and development in response to oxidative stress and UV radiation [80]. PGK is important for glycolysis (EMP pathway) and the tricarboxylic acid cycle, and PGK1 is associated with cold responses and cell wall metabolism [81, 82]. These results suggest that the trichome trait is involved in stress responses, e.g. enhancing cold tolerance by promoting low levels of anaerobic respiration and high levels of aerobic respiration to reduce sugar consumption, which may explain the high elevation distribution of Actinidia deliciosa plants with trichomes.

In conclusion, our study provides insights into the phylogenetic relationships, speciation histories, and corresponding genomic divergence of three Actinidia species and sheds light on the mechanisms by which geographic isolation influences speciation through its effects on population size and gene flow among populations. Regardless of the large-scale or small-scale biogeographic patterns, the three species showed post-speciation adaptation to the local environment, with the trichome trait phenotype and related genes. The results distinguished the mechanisms of genomic divergence for sympatric and allopatric speciation processes of the three species, involving gene flow, ancestral polymorphisms, and selective sweeps. Our study suggests that geographic isolation represents an important factor in speciation with selective sweeps or background selection, and it highlights the significance of the time length of divergence.

Acknowledgements

We thank Liu F., Han W., and local farmers for their assistance in field sampling and Prof. Cheng F. and Dr. Yang Y. for their assistance in the analysis of recombination rate. This work was financially supported by the Biodiversity Survey, Observation and Assessment of the Ministry of Ecology and Environment, China (2019HJ2096001006) and the General Research Project of CRAES, China (No. 2016YSKY-08).

Author contributions

Y.L. designed the study. Y.L., W.Y., B.W., and J.L. performed the experiment. Y.L. wrote the manuscript. All authors read and approved the manuscript.

Data availability

Data will be uploaded after acceptance.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary data

Supplementary data is available at Horticulture Research Journal online.

References

1. Veltsos P, Ritchie MG. Sympatric Speciation. In: eLS. Chichester: John Wiley & Sons, Ltd, 2011.

2. Coyne JA. Sympatric speciation. Curr Biol. 2007;17:787–8.

3. Fahrig L. Ecological responses to habitat fragmentation per se. Annu Rev Ecol Evol Syst. 2017;48:1–23.

4. Sexton JP, Hangartner SB, Hoffmann AA. Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution. 2014;68:1–15.

5. Rangel-Villalobos H, Martinez-Sevilla VM, Martinez-Cortés G et al. Importance of the geographic barriers to promote gene drift and avoid pre- and post-Columbian gene flow in Mexican native groups: evidence from forensic STR loci. Am J Phys Anthropol. 2016;160:298–316.

6. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 2016;18:87–100.

7. Van Doren BM, Campagna L, Helm B et al. Correlated patterns of genetic diversity and differentiation across an avian family. Mol Ecol. 2017;26:3982–97.

8. Han F, Lamichhaney S, Grant BR et al. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin’s finches. Genome Res. 2017;27:1004–15.
1. Wang B et al. Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in Boechera stricta. Genome Biol. 2019;20:126.

2. Stankowski S, Chase MA, Fuiiten AM et al. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. PLoS Biol. 2019;17:e3000391.

3. Ma T, Wang K, Hu Q et al. Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. Proc Natl Acad Sci U S A. 2018;115:E236–43.

4. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic sorting in wide variation by long-term balancing selection and divergent gene flow. Mol Ecol. 2014;23:3133–57.

5. Hazzi NA, Moreno JS, Ortiz-Moviliva C, Palacio RD. Biogeographic regions and events of isolation and diversification of the endemic biota of the tropical Andes. Proc Natl Acad Sci U S A. 2018;115:7985–90.

6. Wang Q, Zhang ML, Yin LF. Genetic diversity and population differentiation of Mammillaria pectinifera cactus biochem Syst Ecol. 2014;39:1–12.

7. Yang F-S, Qin A-L, Li Y-F, Wang X-Q. Great genetic differentiation of Actinidia chinensis among populations of the endemic cactus Mammillaria pectinifera in a changing landscape in the Tehuacán Valley, Mexico. Gen Mol Res. 2014;13:73.

8. Wang Q, Zhang ML, Yin LF. Genetic diversity and population differentiation of Capparis spinosa (Capparaceae) in northwestern China. Biochem Syst Ecol. 2016;66:1–7.

9. Yang-F, Qin A-L, Li Y-F, Wang X-Q. Great genetic differentiation among populations of Meconopsis integrifolia and its implication for plant speciation in the Qinghai-Tibetan plateau. PLoS One. 2012;7:e37196.

10. Crowhurst RN, Gleadle AP, MacRae EA et al. Analysis of expressed sequence tags from Actinidia: applications of a cross species EST database for gene discovery in the areas of flavor, health, color and ripening. BMC Genomics. 2008;9:351.

11. Liang C, Francisco AR. Emendation of the latin name of Actinidia chinensis Pl. Var Hispida C F Liang Guihua. 1984;4:181–2.

12. Liang C. Classification of Actinidia. J Syst Evol. 1975;13:32–5.

13. Liang C, Francisco AR. Revision of infraspecific taxa of Actinidia chinensis planc. Guihua. 1985;5:71–2.

14. Huang H. Actinidia: Classification, Resources, Domestication and Cultivation. Beijing: Science China Press, 2014.

15. He Z-C, Zhong Y, Liu H-T, Tang X-H. Quantitative taxonomic analyses of Actinidia (Actinidiaceae) in China based on micro-morphological characters of foliar trichomes. J Syst Evol. 2000;38:121–36.

16. Chat J, Jáuregui B, Petit RJ, Nadot S. Reticulate evolution in monkeyflowers. PLoS One. 2015;9:877–90.

17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.

18. Viellea AJ, Severin J, Ureta-Vidal A et al. EnsemblComparative Gen-Trees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19:327–35.

19. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005;28:289–301.

20. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

21. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.

22. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.

23. Liang Z, Duan S, Sheng J et al. Whole-genome resequencing of 472 Viitis accessions for grapevine diversity and demographic history analyses. Nat Commun. 2019;10:1190.

24. An L, Zhou Z, Yan A, Gan Y. Progress on trichome development regulated by phytohormone signaling. Plant Signal Behav. 2014;6:1599–62.

25. Bouckaert R, Heled J, Kühnert D et al. BEAST 2: a software platform for bayesian evolutionary analysis. PLoS Comput Biol. 2014;10:e1003537.

26. Heffelfrich P, Rieb E, Abrami G et al. TreeAnnotator: Versatile Visual Annotation of Hierarchical Text Relations. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA), 2018.

27. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8:e1002967.

28. Danecek P, Auton A, Abecasis G et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

29. Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. PLoS Genet. 2012;8:e1003090.

30. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.

31. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–4.

32. Cui ZX. Actinidia in China. Shandong, China: Shandong Science and Technology Press, 1993.

33. Crowhurst RN, Gardner RC. A genome-specific repeat sequence of Actinidia deliciosa var. deliciosa genome highlights the challenges associated with draft genomes and gene prediction in plants. BMC Genomics. 2018;19:257.

34. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.

35. Ferguson AR, Huang H. Genetic resources of kiwifruit: domestication and breeding. Hortic Rev. 2007;33:1–121.

36. Pilkington SM, Crowhurst R, Hilario E et al. A manually annotated Actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. BMC Genomics. 2018;19:257.

37. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.

38. Ferguson AR, Huang H. Genetic resources of kiwifruit: domestication and breeding. Hortic Rev. 2007;33:1–121.

39. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.

40. Ferguson AR, Huang H. Genetic resources of kiwifruit: domestication and breeding. Hortic Rev. 2007;33:1–121.

41. Pilkington SM, Crowhurst R, Hilario E et al. A manually annotated Actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. BMC Genomics. 2018;19:257.
53. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016;17:422–33.
54. Kaplan N, Hudson R, Langley C. The hitchhiking effect revisited. Genetics. 1989;123:887–99.
55. Aris-Brosou S, Excoffier L. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol Biol Evol. 1996;13:494–504.
56. Choi JY, Purugganan M, Stacy EA. Divergent selection and primary gene flow shape incipient speciation of a riparian tree on Hawaii island. Mol Biol Evol. 2020;37:695–710.
57. Chiang Y-C, Hung K-H, Moore S-J et al. Paraphyly of organelle DNAs in Cycas sect. Asiorientales due to ancient ancestral polymorphisms. BMC Evol Biol. 2009;9:161.
58. Tavares H, Whibley A, Field DL et al. Selection and gene flow shape genomic islands that control floral guides. Proc Natl Acad Sci U S A. 2018;115:11006–11.
59. Bijlsma R, Loeschcke V. Genetic erosion impedes adaptive responses to stressful environments. Evol Appl. 2012;5:117–29.
60. Lienert J. Habitat fragmentation effects on fitness of plant populations – a review. J Nat Conserv. 2004;12:53–72.
61. Browne I, Karubian J. Habitat loss and fragmentation reduce effective gene flow by disrupting seed dispersal in a neotropical palm. Mol Ecol. 2018;27:3055–69.
62. Gao Y, Ai B, Kong H et al. Geographical pattern of isolation and diversification in karst habitat islands: a case study in the Primulina eburnea complex. J Biogeogr. 2015;42:2131–44.
63. Gao Y, Ai B, Kong H et al. Geographical pattern of isolation and diversification in karst habitat islands: a case study in the Primulina eburnea complex. J Biogeogr. 2015;42:2131–44.
64. Wiberg R, Scobie A, Ahara S et al. Dispersal constraints for the conservation of the grassland herb Thymus pulegioides L. in a highly fragmented agricultural landscape. Conserv Genet. 2015;16:765–76.
65. Liu et al.