Contribution of GC content to differentiated region formation along a speciation continuum

Fushi Ke[1,2,#], Liette Vasseur[3]

[1] Department of Biological Sciences, National University of Singapore, Singapore, Singapore

[2] School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, China

[3] Department of Biological Sciences, Brock University, St. Catharines, ON, Canada

For correspondence: fjkfs@163.com

Fushi Ke: https://orcid.org/0000-0002-5837-7260

Liette Vasseur: https://orcid.org/0000-0001-7289-2675

Abstract

Genome features can interact with external evolutionary processes and are involved in the formation of highly differentiated regions (i.e., genomic islands) potentially containing adaptation and speciation loci. However, GC content that can elevate both regional recombination and mutation has not been investigated in evolving lineages to test its effects on formation of heterozygous differentiation landscape. Here, we employed the genomes of 499 Asian honey bee (*Apis cerana*), with a widely-distributed central lineage diverged and admixed with its peripheral lineages at both population genetic and phylogenetic levels, to investigate the effects of GC content on mutation accumulation and lineage divergence along the speciation continuum. We find that genomic islands are generally with low GC content (i.e., lower-than-average), and with lower-than-average divergence (d_{xy}) initially to higher-than-average ones at deeper divergence, corresponding to the suggested evolution models transforming from linked selection to divergent selection. When comparing with high-GC regions however, we find higher mutation load indicating inefficient natural selection in low-GC regions, which thus support the important role of low GC content instead of evolutionary process (e.g., natural selection) in genomic island formation. Finally, low-GC regions possess higher proportion of lineage-specific polymorphisms than high-GC regions due to limited recombination and admixture under the gene-flow scenario, and reconciliate discordance between mitochondrial and nuclear phylogenies in *A. cerana*. Our results shed light on the contribution of polymorphisms in low-GC regions to divergent region formation along the speciation continuum and their application in reconstruction of intraspecific phylogeny, especially in lineages with gene flow.

Keywords

Eastern honey bee; low GC content; speciation continuum; mutation load; recombination

Introduction

Genomic islands (with elevated F_{ST}) are resistant to gene flow between incipient lineages and potentially involved in adaptation and reproductive isolation (Turner et al. 2005). Understanding the formation of these divergent peaks is imperative in defining how new species arise (Han et al. 2017; Wang et al. 2019; Wolf and Ellegren 2017). While modern sequencing technologies has facilitated our understanding of heterogeneous genomic landscape in many species (Han et al. 2017; Ravinet et al. 2017; Vijay et al. 2016; Wang et al. 2016, 2019; Wolf and Ellegren 2017), the causes of genomic islands against the troughs of differentiation across the genome remains controversial (Cruickshank and Hahn 2014). Increasing evidence shows that genomic islands may emerge by processes unrelated to speciation (Cruickshank and Hahn 2014; Guerrero and Hahn 2017; Ravinet et al. 2017; Wolf and Ellegren 2017), with multiple evolutionary models being proposed for the formation of genomic islands in evolving lineages (Burri 2017; Cruickshank and Hahn 2014; Guerrero and Hahn 2017; Han et al. 2017; Ravinet et al. 2017; Vijay et al. 2016; Wang et al. 2016, 2019; Wolf, Ellegren 2017). Linked selection can increase genomic differentiation (F_{ST}) while reducing nucleotide divergence (d_{xy}) in genomic islands compared with background regions (Charlesworth et al. 1993; Irwin et al. 2016; Ravinet et al. 2017; Burri 2017). On the contrary, divergent selection coupling with gene flow forms genomic islands with both elevated F_{ST} and d_{xy} (Feder et al. 2012; Nosil et al. 2009; Wolf and Ellegren 2017). Without gene flow, sorting of ancient polymorphism has been suggested to be the cause for the formation of genomic islands with high F_{ST} and d_{xy} values (Guerrero and Hahn 2017; Noor and Bennett 2009). Genome features, such as gene density, recombination, and mutation rate, may influence the efficiency and extent of these non-mutually exclusive evolutionary processes and contribute to the formation of heterozygous differentiation landscape along the speciation continuum (Burri 2021; Castellano et al. 2020; Leroy et al. 2021; Ravinet et al. 2017).

GC content (proportion of guanine (G) and cytosine (C) nucleotides) can elevate both regional recombination (Charlesworth et al., 2018; Fullerton et al. 2001; Kiktev et al. 2018; Meunier et al. 2004; Pessia et al. 2012) and mutation (Fryxell and Moon 2005; Kiktev et al. 2018). Unlike

other genome features (e.g., gene density) however, few have investigated GC content in chronologically explicit lineages. At the beginning of divergence, due to low mutation rate in low (i.e., GC content lower-than-average, the same hereinafter) GC regions (Kiktev et al. 2018), we may expect lower level of nucleotide diversity (π) and divergence (d_{xy}) in these regions than those in high GC (i.e. GC content higher-than-average, the same hereinafter) regions. Over time, due to inefficient natural selection (low recombination rate), accumulated mutations in low GC regions may reach similar d_{xy} levels to high GC regions, where high mutation rate can be counteracted by more efficient selection (high recombination rate, Roze 2021). At deeper divergence (e.g., phylogenetic timescale), low GC regions can possess more mutations and show higher d_{xy} values than high GC regions due to the combined effects of accumulated mutations and limited recombination. Under the gene-flow scenario, since gene flow is largely restricted in low recombination regions (Wong and Filatov et al. 2023), over time, d_{xy} pattern in low GC regions should be similar with the no-gene-flow condition. Gene flow may occur in these regions only over deeper divergence due to breakdown of linkage (Cutter and Choi 2010; Noor and Bennett 2009). On the contrary, high GC regions may have lower d_{xy} values compared with counterparts in no-gene-flow circumstance due to genetic homogeneity between lineages. Along the speciation continuum, due to constant limited mutation and recombination in low GC regions, we may expect that low GC regions can always be with higher differentiation (F_{ST}) along the speciation continuum, because of low nucleotide diversity (Cruickshank and Hahn 2014) and restricted genetic admixture (Wong and Filatov et al. 2023). Meanwhile, GC content can play an important role in mutation accumulation and lineage divergence, resulting in lower-than-average d_{xy} initially to higher-than-average d_{xy} at deeper divergence. Nevertheless, this hypothesis needs to be tested with diverging lineage pairs at both population genetic and phylogenetic timescales.

Apis cerana is a widespread cavity-nesting honeybee group (Ji et al. 2020; Qiu et al. 2023; Radloff et al. 2010; Su et al. 2023). Its distribution in Asia ranges from Pakistan to Japan, and extends from the adjacent regions of Siberia in the North to its southern distribution in Indonesia, from where the bee swarms invaded the Oceania (Koetz et al. 2013). There are four putative species in this group based on the inference of genome-wide single nucleotide

polymorphisms (SNPs) (Su et al. 2023): *A. cerana* (Asian Mainland and Indian black, Lineage M), *A. cerana* (Sundaland, Lineage S), *A. cerana* (Indian yellow, Lineage I), and *A. cerana* (Philippines, Lineage P). The mainland lineage (Lineage M) is the most widely-distributed and formed several peripheral populations with varied levels of divergence at the surrounding of its central population (Ji et al. 2020; Qiu et al. 2023). The central population of Lineage M admixes with the Lineage M peripheral populations and introgresses with the other *A. cerana* species in the South (Lineage S) (Ji et al. 2023), forming multiple admixed lineage pairs diverged at both population genetic and phylogenetic timescales. Here, we used resequencing genomes of 499 *A. cerana* and six *A. mellifera* individuals (outgroup), with lineages diverged across the speciation continuum, to investigate the contribution of GC content to genome-wide differentiation landscape and mutation accumulation over time. We found low GC regions were usually with higher genomic differentiation (e.g., genomic islands) due to reduced mutation and recombination. These regions had lower d_{xy} initially, gradually accumulating lineage-specific mutations that led to higher d_{xy} compared with high GC regions. Nevertheless, high mutation loads in low GC regions at different investigated timescales indicated this pattern was not likely contributed by natural selection but common genomic features. Finally, conflict between mitochondrial phylogeny (Qiu et al. 2023) and its nuclear counterpart (Ji et al. 2020) in *A. cerana* Lineage M was likely due to the above-mentioned admixture between adjacent lineages. They were reconciliated with polymorphisms in low GC regions for a better taxonomy of honeybee lineages and further conservation.

117 **Methods**

118 ***2.1 Sequencing data processing, SNP calling, and variant filtering***

119 A total of 507 *Apis cerana* workers were initially sampled across its range, including individuals
120 from China (331), India (14), Japan (105), Malaysia (26), Myanmar (2), North Korea (1),
121 Pakistan (5), Singapore (1), South Korea (6), Thailand (11) and Viet Nam (5) (Chen et al. 2018;
122 Ji et al. 2020; Qiu et al. 2023; Wakamiya et al. 2023). Six *A. mellifera* individuals were employed
123 as the outgroup individuals (Harpur et al. 2014). After preprocessing, eight *A. cerana*
124 individuals with low sequencing depth or with unknown sampling sites were removed,
125 resulting in a total of 505 individuals (499 *A. cerana* and six *A. mellifera* individuals) for further
126 analysis (Supplementary Table 1).

127

128 All sequencing reads were downloaded from National Center for Biotechnology Information
129 (NCBI) Sequence Read Archive (SRA) (PRJNA592293, PRJNA418874, PRJNA870246,
130 PRJNA216922), or DNA Data Bank of Japan (DDBJ) database (PRJDB14080). Clean reads were
131 trimmed and filtered using Trimmomatic (v 0.39, Bolger et al. 2014). Reads with a minimum
132 length of 50bp were kept for analysis. We further individually mapped clean reads to the *A.*
133 *cerana* genome (GCF_001442555.1, Park et al. 2015) by using bwa (v 0.7.17, Li 2013).
134 Duplicates were marked by Picard (v 2.24.0, http://broadinstitute.github.io/picard/). A
135 standard pipeline of variant calling based on Genome Analysis Toolkit (GATK 4.2.4.0, McKenna
136 et al. 2010) was applied with HaplotypeCaller for indel realignment and individual variant
137 calling, and GenotypeGVCFs for joint variant genotyping with all samples. We further
138 performed hard filtering based on suggested parameters in GATK by using VariantFiltration
139 command. This dataset, including non-variant sites, was used for calculation of population
140 genetic parameters, such as nucleotide diversity (π) and divergence (d_{xy}).

141

142 We also used Vcftools (v 0.1.16, Danecek et al. 2011) in SNP hard filtering when investigating
143 genetic structure and reconstructing phylogeny. In genetic structure analysis, bi-allelic SNPs
144 with a maximum missing value of 20% and a minor allele frequency of 0.05 were isolated, after
145 which one SNP from a window of 20kb DNA sequence was kept for analysis. For phylogeny

reconstruction, we kept all bi-allelic SNPs for the analysis.

2.2 Genetic structure and population genetic parameter calculation

We used ADMIXTURE (v 1.3.0, Alexander et al. 2009) for genetic structure analyses and performed each K from 2 to 20 with 30 replicates. The best K value was identified according to cross-entropy criterion. The outgroup individuals (*A. mellifera*) were excluded for the genetic clustering analysis. To characterize environments of each sample, we also extracted dataset of monthly climate and climatic water balance for global terrestrial surfaces from 1996 to 2015 (high-spatial resolution, 1/24°, ~4-km) (Abatzoglou et al. 2018). We then performed Principal Component Analysis (PCA) with twelve monthly averages of the twelve variables (a total of 144 monthly variables, Supplementary table 1).

Based on the dataset with non-variant sites, we calculated π, d_{xy}, F_{ST} in a 100 kb window size within and among our identified lineages using Python scripts downloaded from https://github.com/simonhmartin/genomics_general. We removed all individuals that had more than 20% admixed ancestry from the other lineage, resulting in 24 (Aba), 20 (Bomi), 224 Central, 26 (Hainan), 16 (Ind_Pak), 105 (Japan), 27 (Lineage S), 23 (Northeast), 21 (Qinghai) and 11 (Taiwan). Six *A. mellifera* (i.e. Mel, outgroup) were also included in parameter calculation. Windows with less than 1000 genotyped sites (variant and invariant) were removed from the analysis. Regional GC-content and coding sequence percentage were also calculated based on the *A. cerana* genome sequence (GCF_001442555.1). For each lineage, calculation of recombination rate in each 100 kb window was conducted by using FastERR (v 2.0, Gao et al. 2016).

To estimate the divergence of population / lineage pairs, we calculated a corrected estimate of sequence divergence (d_a) in each window by subtracting mean polymorphism of two lineages (π) from the total divergence (i.e., d_{XY}) (Cruickshank and Hahn 2014). The d_a statistic is a substitute of divergent time (T, in generations) as they can be transformed by the equation: $T = d_a / (2\mu)$, where μ equals to 5.27×10^{-9} (Wallberg et al. 2014). Since Central

lineage of the *A. cerana* Lineage M was closely related with almost all other defined lineages, we used Central as the invariable lineage in divergence estimation (e.g., Central vs. Aba, Central vs. Bomi, Central vs. Qinghai).

2.3 Characterization and correlation analysis of summary statistics in different regions

The pilot analysis and visualization of genome-wide parameters showed distinct pattern of differentiation in high GC and low GC regions (Figure 2A, Supplementary Figure 1). Therefore, we divided the genome-wide windows into high GC (higher than average GC content) and low GC (lower or equal to average GC content) categories. Difference in parameter values in high GC and low GC windows were compared and tested based on Wilcoxon rank sum test in ggsignif R package (v 0.6.4, Ahlmann-Eltze and Patil 2021). We further standardized per-window F_{ST} in each cluster pair to a Z score (Han et al. 2017) based on the formula: Z-F_{ST} = (window F_{ST} − mean F_{ST}) / standard deviation F_{ST}. We then compared difference of parameter values in the genomic islands (windows with Z- F_{ST} ⩾2 from each lineage pair) and background regions (windows with Z- F_{ST} < 2) with Wilcoxon rank sum test.

The Pearson's correlation coefficient for intrapopulation summary statistics between populations (e.g. π_i vs. π_j), within populations (e.g. π_i vs. ρ_i), and interpopulation statistics of cluster pairs (e.g. π_i vs. F_{STij}) were calculated and performed in R package corrplot (v 0.94, Wei and Simko 2021). We further conducted correlation analysis in high GC and low GC regions, respectively.

2.4 Investigation of mutation load

We investigated synonymous and non-synonymous mutations in high GC and low GC regions in each lineage. Bi-allelic SNPs were extracted from each lineage and annotated with SNPeff (v 5.1d, Cingolani et al. 2012). In each lineage, we calculated dN/dS (non-synonymous /

203 synonymous SNPs) ratio that represents magnitude of mutation load. In addition, numbers
204 of SNPs in low GC regions divided by high GC regions were calculated in each functional
205 group (MODIFIER, LOW, MODERATE, and HIGH) in the defined lineages. To further
206 characterize whether the mutation loads were shared or lineage-specific, mutation load in
207 three highland lineages (i.e., Aba, Bomi, and Qinghai) that colonized Qinghai-Tibet Plateau
208 independently were investigated. Mutation loads in lineage pair with admixture (e.g., Central
209 and Hainan) were further investigated to estimate the effects of gene flow on lineage-specific
210 and shared mutations in low GC as well as high GC regions.
211

212 ### *2.5 Nuclear phylogeny reconstruction*

213 For phylogeny reconstruction, we divided the genome into multiple 200 kb windows, and
214 reconstructed window phylogeny by using IQ-TREE 2 (v 2.2.2.7, Minh et al. 2020). In model
215 testing, we employed +ASC model to include ascertainment bias correction as SNP data did
216 not contain constant sites (Minh et al. 2020). Following the manual, we also appended -bnni
217 to the regular UFBoot command to reduce the risk of overestimating branch support
218 (Hoang et al. 2018). Since genomic regions with high recombination rate and low
219 recombination rate showed differentiated efficiency in tree reconstruction (e.g., Pease and
220 Hahn 2013), and our analysis showed distinct pattern of recombination and mutation
221 accumulation in high GC and low GC regions (Supplementary Figure 2, 3), we divided
222 genomic windows into high GC (above genome-wide average GC content) and low GC
223 (below genome-wide average GC content) categories, and generated summary tree in
224 ASTRAL III (v 5.7.1, Zhang et al. 2018) based on window trees in each category.

225 **Results**

226 ***3.1 Genetic structure and divergence of Apis cerana***

227 The optimal K value was 14 in the resequencing of 499 *A. cerana* individuals (Figure 1). *A.*
228 *cerana* Sundaland individuals (samples from Malaysia and Singapore, Lineage S) were first
229 separated from the mainland species (Lineage M) due to long-term divergence. Island
230 lineages from the mainland species (e.g. Taiwan, Hainan and Japan) and populations that
231 colonized extreme habitats (e.g. high elevation: Ind_Pak, Bomi, Aba and Qinghai; or high
232 latitude: Northeast) gradually separated. These lineages had clear genetic ancestry and
233 geographical distribution (Figure 1). Environmentally differentiated habitats also contributed
234 to the formation of two clusters in the Central lineage, and four clusters in the Japan
235 lineage. These clusters displayed intensive gene flow with adjacent ones and were difficult
236 to distinguish from each other. We thus grouped the two clusters from Central and the four
237 from Japan into one lineage, respectively. In total, we identified ten lineages based on
238 genomic clustering and geography and performed the following analysis with this grouping
239 scheme. The Central lineage (of Lineage M) was widely distributed and showed varied
240 extent of genetic admixture with peripheral lineages of Lineages M and S (Figure 1).
241 Clustering of bioclimatic variables could not clearly separate these lineages with some
242 extant of overlapping, suggesting a combination of phylogeny, genetic drift, and natural
243 selection in the formation of these lineages.
244
245 We further calculated divergent times of each lineage from the Central lineage based on a
246 corrected estimate of sequence divergence (d_a) (Cruickshank and Hahn 2014). Time nodes
247 from seven lineage pairs (i.e. t1 to t7, Supplementary Table 2) were chosen and used in a
248 time series to investigate changes of genomic differentiation landscape and characterize
249 genomic polymorphisms. Divergent time calculation in Central with *A. mellifera* pair (i.e.,
250 Central *v.s.* Mel, t7) was consistent with the estimated ~ 7 million years (Wallberg et al.
251 2014), which supported the reliability of divergent time estimation based on d_a. Estimated
252 divergent times of the other time nodes were: 123 thousand years ago (kya) in Central *v.s.*
253 Qinghai (t1), 151 kya in Central *v.s.* Bomi (t2), 186 kya in Central *v.s.* Hainan (t3), 295 kya in

Central *v.s.* Japan (t4), 471 kya in Central *v.s.* Taiwan (t5), 674 kya in Central *v.s.* Lineage S (t6). We further divided genomic regions into high GC (above genome-wide average GC content) and low GC (below genome-wide average GC content) categories and calculated d_a for each category. We found a consistent pattern of increased divergent time from t1 to t7 in low GC windows, while not in high GC regions (Supplementary Table 2). This was consistent with the idea that higher recombination could accelerate gene flow between divergent lineages in high GC regions, while not in low GC regions due to limited recombination.

3.2 Landscape of genomic variation and differentiation in chronological lineage pairs

Investigating genomic variation landscape over time (Figure 2A, Supplementary Figure 1), we found a distinct pattern of nucleotide divergence (i.e. d_{xy}) in high GC and low GC regions. In low GC regions, a trend of increased divergence from t1 to t7, with d_{xy} values in t6 and t7 higher than average (Figure 2A, Supplementary Figure 1). In high GC regions however, divergences of different lineage pairs over time were indistinguishable between lineage pairs when compared with low GC regions in the genome-wide context (Figure 2A). We further compared nucleotide divergence in low GC and high GC regions over time and found significantly lower d_{xy} values in low GC regions from t1 to t4 than in high GC regions (Figure 2B, Mann– Whitney U test, $p < 0.001$). In t5, d_{xy} values were similar and showed no significant difference between low CG and high GC regions. In t6 and t7 however, d_{xy} values in low GC regions were significantly higher than in high GC regions (Figure 2B, Mann–Whitney U test, $p < 0.001$).

Correlations between d_{xy} and genetic differentiation (F_{ST}) (Figure 2C) also switched over time in low GC regions, with a negative correlation from t1 to t5, then positive in t6 and t7. This pattern was only found in low GC regions (Figure 2C). Similar patterns were also found in correlations of d_{xy} with recombination rate (rho) and GC content (Supplementary Figure 2A, B). We also found positive correlations between nucleotide diversity (π) / mutation load and

GC content in low GC regions while negative / limited correlations in high GC regions (Supplementary Figure 2), a pattern also identified between diversity parameters with recombination rate (rho) (Supplementary Figure 2). These patterns supported the effects of GC content on nucleotide diversity and mutation loads by regulation of regional recombination and mutation. Consistently negative correlations between nucleotide diversity (π) and gene density were detected across the genome, as well as in high GC and low GC regions (Supplementary Figure 2), supporting widespread linked selection in the functional elements.

Genomic differentiation (F_{ST}) in low GC regions were always higher and recombination rate (rho) and nucleotide diversity (π) lower than those in high GC regions (Supplementary Figure 3), thus supporting our initial hypothesis about the contribution of low GC content in formation of genomic islands. We further characterized genomic islands (normalized F_{ST}, Z-F_{ST} ⩾ 2) over time and found a significantly lower GC content in genomic islands compared with genomic backgrounds (Z- F_{ST} < 2) (Supplementary Figure 4). Being consistent with the pattern found in low GC regions, recombination rate and nucleotide diversity in genomic islands were significantly lower than those in genomic backgrounds (Supplementary Figure 4). d_{xy} comparisons of genomic islands and backgrounds also switched over time, with lower-than-average d_{xy} values initially (t1 to t5) to higher-than-average ones at t6 and t7 (Supplementary Figure 4). This pattern supported the proposed model of genomic island formation transformed from linked selection (i.e., high F_{ST} and low d_{xy}) to divergent selection (i.e., high F_{ST} and high d_{xy}) over deeper divergence.

3.3 Characterization of polymorphisms in low GC and high GC regions

Patterns of accumulated polymorphism (e.g., mutations) were differentiated in low GC and high GC regions (Figure 3 A, B; Supplementary Figure 2E), with synonymous (dS) and non-synonymous (dN) sites positively correlated with GC content in low GC regions, and negatively correlated in high GC regions (Supplementary Figure 2). The differentiated pattern was due to limited recombination in low GC regions (Supplementary Figure 3) which

312 could facilitate the accumulation of mutations, a pattern that could not be found in high GC
313 regions with higher recombination. The consistently positive correlation in low GC regions
314 indicated these regions were able to accumulate more mutations and divergence than high
315 GC regions in deeper divergence, although there are lower mutation rates in these regions.
316
317 We further calculated dN/dS values and found they were significantly higher in low GC
318 regions than in high GC regions. We divided mutation load into variants with different
319 impacts (MODIFIER, LOW, MODERATE, and HIGH), and calculated ratio of SNP numbers in
320 low GC over high GC regions. We found that low GC regions had significantly higher ratios
321 of non-neutral variants (LOW, MODERATE, and HIGH) than neutral ones (MODIFIER) in all
322 defined lineages (Figure 3B). This pattern supported the idea that the genome architecture
323 (i.e. GC content) was involved in mutation accumulation and lineage differentiation
324 regardless of external factors such as natural selection.
325
326 We further investigated ratio of polymorphisms that were lineage-specific and shared in
327 different lineages. With three lineages (Bomi, Aba and Qinghai) that independently
328 colonized highland habitats and were with no genetic admixture (Figure 1), we determined
329 that most of the polymorphisms (e.g. mutation load) were unique to each lineage and
330 showed no percentage difference in high GC and low GC regions (Figure 3C). We further
331 investigated lineage-specific and shared polymorphisms in lineage pair with genetic
332 admixture. With Central v.s. Hainan as an example, we found more lineage-specific
333 polymorphism than shared was only in low GC regions (Figure 3D).
334

335 ### *3.4 Reconstruction of A. cerana nuclear phylogeny*

336 After characterizing the landscape of genomic variation in high GC and low GC regions, we
337 reconstructed phylogeny of *A. cerana* based on nuclear polymorphism in low GC regions
338 because they were not or little affected by gene flow and with more lineage-specific
339 polymorphisms (Figure 3D). We got a nuclear topology similar to mitochondrial phylogeny
340 (Figure 4 and Supplementary Figure 5, Qiu et al. 2023), with lineages from Taiwan and

Hainan being basal lineages of *A. cerana* Lineage M (Qiu et al. 2023). The phylogeny based on polymorphism in high GC regions however was influenced by gene flow, with admixed lineages clustering together (Supplementary Figure 6). Individuals showing hybrid ancestry were always situated between lineages (e.g. Central v.s. derived lineage, Supplementary Figure 6 and Figure 1) and had higher polymorphism than the individuals within the derived lineages (Supplementary Figure 6).

347 # Discussion

348 The formation of heterozygous differentiation landscape is suggested to result from the
349 combined effects of external evolutionary processes and genome features (Ravinet et al.
350 2017). Here, we used hundreds of eastern honey bee genomes along the speciation
351 continuum to investigate the influences of GC content on mutation accumulation and
352 lineage divergence by conducting correlation analysis of population genetic parameters and
353 genomic features, and characterization of different regions across the genome. We found
354 significantly lower GC content in genomic islands compared with background regions
355 (Supplementary Figure 4). This suggests that lower mutation rate (Fryxell and Moon 2005;
356 Kiktev et al. 2018) and limited recombination (Charlesworth et al. 2018; Kiktev et al. 2018;
357 Pessia et al. 2012) in these regions could contribute to lower nucleotide diversity that leads
358 to elevated differentiation (Cruickshank and Hahn 2014). These regions were with lower-
359 than-average divergence (d_{xy}) at population genetic timescales to higher-than-average d_{xy}
360 at phylogenetic levels. This corresponds to the suggested evolution models transformed
361 from "sweep-before-differentiation" and background selection (i.e. linked selection,
362 Charlesworth et al. 1993; Irwin et al. 2016; Ravinet et al. 2017), to divergent selection with
363 gene flow (Feder et al. 2012; Nosil et al. 2009; Wolf and Ellegren 2017) or sorting of ancient
364 polymorphisms (Guerrero and Hahn 2017; Han et al. 2017; Noor and Bennett 2009; Wolf
365 and Ellegren 2017). In addition, correlations between d_{xy} and recombination rate (rho) in low
366 GC regions were negative from t1 to t5 but positive in t6 and t7. A similar trend was also
367 found in the correlations between d_{xy} and F_{ST}. This pattern is consistent with widespread
368 linked selection in low recombination regions (Ravinet et al. 2017), while involvement of
369 divergent selection due to the breakdown of linkage in these regions can occur over deeper
370 divergence (Cutter and Choi 2010; Noor and Bennett 2009).
371
372 Few studies have reported a similar pattern (e.g., Liu et al. 2022) likely because few analyze
373 lineage pairs at both population genetic and phylogenetic timescales in a single system.
374 Many studies report one dominant model in genomic island formation, with contribution of
375 external processes and genomic features. For example, recurrent selection is reported to be

involved in the formation of genomic islands in population pairs of Green-backed Tit (*Parus monticolus*) (Jiang et al. 2023), Zhou et al. (2024) indicate that divergent selection contributes to regionally elevated differentiation in incipient species of Plover (*Charadrius alexandrines* and *C. dealbatus*) (Zhou et al., 2024). Nevertheless, these genomic islands were both contributed by reductions in recombination (Jiang et al. 2023; Zhou et al. 2024), a genome feature that could be shaped by regional GC content.

We further investigated synonymous (dS) and non-synonymous (dN) polymorphisms, and mutation load (dN/dS) in both low GC and high GC regions (Figure 4). The results showed higher mutation loads (i.e. higher dN and dN/dS ratio) in low GC regions than high GC regions, suggesting inefficient natural selection (Arbuthnott and Rundle 2012; Bertorelle et al. 2022; Leroy et al. 2021) in these regions compared with the rest of the genome. Need to note that this pattern was not only in peripheral lineages with lower effective population size but also in the central lineage and Lineage S, indicating less effects of genetic drift on accumulation of high mutation load in low GC regions (Bertorelle et al. 2022). The contribution of external evolutionary processes (e.g., linked selection, divergent selection) and genome features to regional variation and recombination thus needs to be reconsidered. The reduced nucleotide diversity of the genomic islands may not result from a reduced effective population size by selection (Feder and Nosil, 2009; Turner and Hahn 2010), genetic drift or both (Burri 2021). It might be shaped by local genomic features including low GC content, where lower mutation rate governs the accumulations of less random mutations (Castellano et al. 2020), and reduced recombination protects low GC regions from natural selection (Leroy et al. 2021) and gene flow (Wong and Filatov 2023) along the speciation continuum.

In contrast to low GC regions, higher recombination in high GC regions may facilitate more efficient natural selection (Roze 2021) and genetic admixture (Wong and Filatov 2023). This forms lower mutation load and reduces divergence in these regions at lineage pairs of both intraspecific and interspecific levels, a pattern influenced by gene flow and linked selection (e.g., Ke et al. 2022; Stankowski et al. 2019; Wang et al. 2022; Zhang et al., 2022). Therefore,

admixed ancestry in high GC regions could blur nuclear phylogeny of two diverging lineages in gene-flow scenario and cause conflict between nuclear and cytoplasmic DNA topologies (e.g., Ji et al. 2020; Qiu et al. 2023). The conflict however could be reconciled by polymorphisms in low GC regions with randomly accumulated and lineage-specific mutations, and lower recombination (Pease and Hahn, 2013).

The efficiency and extant of external evolutionary processes (e.g., nature selection) were largely determined by regional GC content based on our investigation of mutation load and lineage divergence at different genome regions along the speciation continuum. In previous studies, while correlations of genomic variation and divergence in evolving lineages are commonly found in different working systems (e.g., Cuevas et al 2022; Vijay et al. 2016), with contributions likely from conserved genomic architecture (Burri et al. 2015; Renaut et al. 2014; Shang et al. 2023), the influences of GC content that can regulate both regional mutation and recombination are not investigated. We thus suggest that further studies should be conducted in broader taxa to understand the formation of heterozygous differentiation landscape by incorporating effects of GC content on mutation accumulation and lineage differentiation.

Acknowledgments

We would like to thank the Information Technology Services, HKU, for offering research computing facilities in the computations of this project.

Data and code availability

All sequencing data could be retrieved from public data base. Software and packages for analysis in this project are all available online.

Author contributions

Conceptualization, F.S.K.; methodology and analysis, F.S.K.; Writing, F.S.K. and L.V.

References

Abatzoglou J.T., Dobrowski S.Z., Parks S.A. and Hegewisch K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data*, 5(1): 1-12.

Ahlmann-Eltze C. and Patil I., 2021. ggsignif: R package for displaying significance brackets for 'ggplot2'.

Alexander D. H., Novembre J. and Lange K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19: 1655-1664.

Arbuthnott D. and Rundle H.D., 2012. Sexual selection is ineffectual or inhibits the purging of deleterious mutations in *Drosophila melanogaster*. *Evolution*, 66: 2127-2137.

Bertorelle G., Raffini F., Bosse M., Bortoluzzi C., Iannucci A., Trucchi E., Morales H.E. and van Oosterhout C., 2022. Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, 23(8): 492-503.

Bolger A.M., Lohse M. and Usadel B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15): 2114-20.

Burri R., Nater A., Kawakami T., Mugal C.F., Olason P.I., Smeds L., Suh A., Dutoit L., Bureš S., Garamszegi L.Z., Hogner S., Moreno J., Qvarnström A., Ružić M., Sæther S.A., Sætre G.P., Török J. and Ellegren H., 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula flycatchers*. *Genome Research*, 25(11): 1656-65.

Burri R., 2017. Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 1(3): 118-131.

Burri R., 2021. Evolution: small populations, low recombination, big trouble?. *Current Biology*, 31(6): R282-R284.

Castellano D., Eyre-Walker A. and Munch K., 2020. Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other Homininae. *Genome Biology and Evolution*, 12(1): 3550-3561.

Charlesworth B., Morgan M.T. and Charlesworth D., 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4): 1289-303.

Charlesworth, D., Zhang, Y., Bergero, R., Graham, C., Gardner, J. and Yong, L., 2020. Using GC content to compare recombination patterns on the sex chromosomes and autosomes of the guppy, *Poecilia reticulata*, and its close outgroup species. *Molecular Biology and Evolution*, *37*(12): 3550-3562.

Chen C., Wang H., Liu Z., Chen X., Tang J., Meng F. and Shi W., 2018. Population genomics provide insights into the evolution and adaptation of the eastern honey bee (*Apis cerana*). *Molecular Biology Evolution*, 35: 2260-2270.

Cingolani P., 2012. Variant annotation and functional prediction: SnpEff. In *Variant calling: methods and protocols*: 289-314. New York, NY: Springer US.

Cruickshank T.E. and Hahn M.W., 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13): 3133-3157.

Cuevas A., Eroukhmanoff F., Ravinet M., Sætre G.P. and Runemark A., 2022. Predictors of genomic differentiation within a hybrid taxon. *PLoS Genetics*, 18(2): e1010027.

Cutter A.D. and Choi J.Y., 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Research*, 20(8): 1103-1111.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T. and McVean G., and 1000 Genomes Project Analysis Group., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156-2158.

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

Feder J.L., Egan S.P. and Nosil P., 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7): 342-350.

Fryxell K.J. and Moon W.J., 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution*, 22(3): 650-658.

Fullerton S.M., Bernardo Carvalho A. and Clark A.G., 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Molecular Biology and Evolution*, 18(6): 1139-1142.

Gao F., Ming C., Hu W., and Li H., 2016. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3: Genes|genomes|genetics*, 6: 1563-1571.

Guerrero R. F., and Hahn M.W., 2017. Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, 26: 5362-5368.

Han F., Lamichhaney S., Grant B. R., Grant P. R., Andersson L., and Webster M.T., 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, 27: 1004-1015.

Harpur B.A., Kent C.F., Molodtsova D., Lebon J.M., Alqarni A.S., Owayss A.A. and Zayed A., 2014. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences*, *111*(7): 2614-2619.

Hoang D.T., Chernomor O., Von Haeseler A., Minh B.Q. and Vinh L.S., 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2): 518-522.

Irwin D.E., Milá B., Toews D.P., Brelsford A., Kenyon H.L., Porter A.N., Grossen C., Delmore K.E., Alcaide M. and Irwin J.H., 2018. A comparison of genomic islands of differentiation across three young avian species pairs. *Molecular Ecology*, 27(23): 4839-4855.

Ji Y., Li X., Ji T., Tang J., Qiu L., Hu J., Dong J., Luo S., Liu S., Frandsen P.B. and Zhou X., 2020. Gene reuse facilitates rapid radiation and independent adaptation to diverse habitats in the Asian honeybee. *Science Advances*, 6(51): eabd3590.

Jiang Z., Song G., Luo X., Zhang D., Lei F. and Qu Y., 2023. Recurrent selection and reduction in recombination shape the genomic landscape of divergence across multiple population pairs of Green-backed Tit. *Evolution Letters*, 7(2): 99-111.

Ke F., Vasseur L., Yi H., Yang L., Wei X., Wang B. and Kang M., 2022. Gene flow, linked selection, and divergent sorting of ancient polymorphism shape genomic divergence landscape in a group of edaphic specialists. *Molecular Ecology*, 31(1): 104-118.

Kiktev D.A., Sheng Z., Lobachev K.S. and Petes T.D., 2018. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 115(30): E7109-E7118.

Koetz A.H., 2013. Ecology, behaviour and control of *Apis cerana* with a focus on relevance to the Australian incursion. *Insects*, 4(4): 558-592.

Leroy T., Rousselle M., Tilak M.K., Caizergues A.E., Scornavacca C., Recuerda M., Fuchs J., Illera J.C., De Swardt D.H., Blanco G. and Thébaud C., 2021. Island songbirds as windows into evolution in small populations. *Current Biology*, 31(6): 1303-1310.

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

Liu Y., Yu W., Wu B. and Li J., 2022. Patterns of genomic divergence in sympatric and allopatric speciation of three Mihoutao (Actinidia) species. *Horticulture Research*, 9: uhac054.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. and DePristo M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9): 1297-1303.

Meunier J. and Duret L., 2004. Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution*, 21(6): 984-990.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A. and Lanfear R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5): 1530-1534.

Noor M.A. and Bennett S.M., 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, 103(6): 439-444.

Park D., Jung J.W., Choi B.S., Jayakodi M., Lee J., Lim J., Yu Y., Choi Y.S., Lee M.L., Park Y. and Choi I.Y., 2015. Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. *BMC Genomics*, 16: 1-16.

Pease J.B. and Hahn M.W., 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution*, 67(8): 2376-84.

Pessia E., Popa A., Mousset S., Rezvoy C., Duret L. and Marais G.A., 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, 4(7): 675-682.

Qiu L., Dong J., Li X., Parey S.H., Tan K., Orr M., Majeed A., Zhang X., Luo S., Zhou X. and Zhu C., 2023. Defining honeybee subspecies in an evolutionary context warrants strategized conservation. *Zoological Research*, 44(3): 483.

Radloff S.E., Hepburn C., Hepburn H.R., Fuchs S., Hadisoesilo S., Tan K., Engel M.S., Kuznetsov V., 2010. Population structure and classification of *Apis cerana*. *Apidologie*, 41: 589-601.

563 Ravinet M., Faria R., Butlin R.K., Galindo J., Bierne N., Rafajlović M., Noor M.A.F., Mehlig B.
564 and Westram A.M., 2017. Interpreting the genomic landscape of speciation: a road
565 map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30(8): 1450-
566 1477.
567 Renaut S., Owens G.L. and Rieseberg L.H., 2014. Shared selective pressure and local
568 genomic landscape lead to repeatable patterns of genomic divergence in
569 sunflowers. *Molecular Ecology*, 23(2): 311-324.
570 Roze D., 2021. A simple expression for the strength of selection on recombination
571 generated by interference among mutations. *Proceedings of the National*
572 *Academy of Sciences*, 118(19): p.e2022805118.
573 Shang H., Field D.L., Paun O., Rendón-Anaya M., Hess J., Vogl C., Liu J., Ingvarsson P.K., Lexer
574 C. and Leroy T., 2023. Drivers of genomic landscapes of differentiation across a
575 Populus divergence gradient. *Molecular Ecology*, 32(15): 4348-4361.
576 Smith D.R., 2011. Asian honey bees and mitochondrial DNA. In: Hepburn HR, Radloff SE,
577 editors. Honeybees of Asia. Berlin (Germany): Springer Verlag; 69 – 93.
578 Stankowski S., Chase M.A., Fuiten A.M., Rodrigues M.F., Ralph P.L. and Streisfeld M.A., 2019.
579 Widespread selection and gene flow shape the genomic landscape during a
580 radiation of monkeyflowers. *PLoS Biology*, 17(7): e3000391.
581 Su Y.C., Chiu Y.F., Warrit N., Otis G.W. and Smith D.R., 2023. Phylogeography and species
582 delimitation of the Asian cavity-nesting honeybees. *Insect Systematics and*
583 *Diversity*, 7(4): 5.
584 Turner T. L., Hahn, M. W., and Nuzhdin S. V., 2005. Genomic islands of speciation in
585 *Anopheles gambiae*. *PLoS Biology*, 3: e285.
586 Turner T.L. and Hahn M.W., 2010. Genomic islands of speciation or genomic islands and
587 speciation?. *Molecular ecology*, 19(5): 848-850.
588 Vijay N., Bossu C.M., Poelstra J.W., Weissensteiner M.H., Suh A., Kryukov A.P., and Wolf J.B.,
589 2016. Evolution of heterogeneous genome differentiation across multiple contact
590 zones in a crow species complex. *Nature Communications*, 7: 13195.
591 Wakamiya T., Kamioka T., Ishii Y., Takahashi J.I., Maeda T., and Kawata M., 2023. Genetic
592 differentiation and local adaptation of the Japanese honeybee, *Apis cerana*
593 japonica. *Ecology and Evolution*, 13(10): e10573.
594 Wallberg A., Han F., Wellhagen G., Dahle B., Kawata M., Haddad N., Simões Z.L., Allsopp
595 M.H., Kandemir I., De la Rúa P., Pirk C.W., Webster M.T., 2014. A worldwide survey
596 of genome sequence variation provides insight into the evolutionary history of the
597 honeybee Apis mellifera. *Nature Genetics*, 46(10): 1081-8.
598 Wang L, Liu S, Yang Y, Meng Z, Zhuang Z. 2022 Linked selection, differential introgression
599 and recombination rate variation promote heterogeneous divergence in a pair of
600 yellow croakers. *Molecular Ecology*, 31(22): 5729-5744.
601 Wang, B., Mojica, J.P., Perera, N., Lee, C.R., Lovell, J.T., Sharma, A., Adam, C., Lipzen, A., Barry,
602 K., Rokhsar, D.S. and Schmutz, J., 2019. Ancient polymorphisms contribute to
603 genome-wide variation by long-term balancing selection and divergent sorting in
604 *Boechera stricta*. *Genome biology*, 20: 1-15.

605 Wang J., Street N. R., Scofield, D. G., and Ingvarsson P. K., 2016. Variation in linked selection
606 and recombination drive genomic divergence during allopatric speciation of
607 European and American aspens. *Molecular Biology and Evolution*, 33: 1754-1767.
608 Wei T. and Simko V., 2021. R package 'corrplot': Visualization of a Correlation Matrix.
609 (Version 0.92), https://github.com/taiyun/corrplot
610 Wolf J.B.W. and Ellegren H., 2017. Making sense of genomic islands of differentiation in light
611 of speciation. *Nature Reviews Genetics*, 18: 87-100.
612 Wong E.L. and Filatov D.A., 2023. The role of recombination landscape in species
613 hybridisation and speciation. *Frontiers in Plant Science*, 14: 1223148.
614 Zhang C., Rabiee M., Sayyari E. and Mirarab S., 2018. ASTRAL-III: polynomial time species
615 tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, 19: 15-
616 30.
617 Zhang H., Zhang X., Wu G., Dong C., Liu J. and Li M., 2023. Genomic divergence and
618 introgression among three Populus species. *Molecular Phylogenetics and*
619 *Evolution*, 180: 107686.
620 Zhou W., Zhang N., Huang K., Lin H., Tu J., Zheng C., Que P., Chiang C.Y., Martinez J.,
621 Naerhulan H. and Székely T., 2024. Divergent selection in low recombination
622 regions shapes the genomic islands in two incipient shorebird species. *Molecular*
623 *Biology and Evolution*, 41(2): msae006.

Figure Legends

Figure 1. Genetic structure and bioclimate variable clustering of mainland lineage *Apis cerana* (Lineage M). Additional samples from Malaysia and Singapore (i.e., Sundaland lineage, Su et al., 2023; Lineage S) were included. The optimal K=14 was selected based on cross-validation error in ADMIXTURE. Two and four genetic clusters were identified from Central and Japan lineages, respectively. The 144 monthly bioclimate variables can be found in Supplementary Table 1.

Figure 2. Characterization of genome-wide differentiation landscape in low GC and high GC regions. **(A)** Landscape of genomic differentiation over time (t1 to t6) varied along with regional GC content. Landscape of genomic differentiation from t1 to t7 in Supplementary Figure 1; **(B)** Patterns of nucleotide divergence (d_{xy}) in low GC and high GC regions shift over divergent time. Wilcoxon rank sum test was conducted with values between low GC and high GC regions. Empty circles represent significantly lower values in low GC than high GC regions, while black circles indicate low GC regions with higher values; **(C)** Correlation coefficients of nucleotide divergence (d_{xy}) and genetic differentiation (F_{ST}) varied over divergent time in different genome regions.

Figure 3. Profile of variants in low GC and high GC regions. **(A)** Higher dN/dS (ratio of nonsynonymous over synonymous mutations) in low GC regions than high GC regions (Paired t test, $p < 0.001$). **(B)** Higher ratio of deleterious mutations than neutral mutants. Paired t tests were conducted between ratio of different mutant types. *** indicates $p < 0.001$. Types of putative variant impact could refer to SNPeff (Cingolani et al., 2012); **(C)** More lineage-specific mutations than the shared in both low GC and high GC regions of isolated lineages. Aba, Bomi and Qinghai highland lineages represent three independent colonization to Qinghai-Tibet Plateau and are with no admixture with each other (Figure 1). **(D)** Ratio of lineage-specific and shared mutation load differentiated in low GC and high GC regions in lineage pair with gene flow. Lineage pair Central with Hainan as an example. Number of SNPs in each comparison in bracket.

Figure 4. Reconstruction of *Apis cerana* phylogeny based on nuclear polymorphisms in low GC regions. Note that the topology was consistent with the mitochondrial phylogeny of *A. cerana* Lineage M (Qiu et al. 2023). Bootstrap values and Individual IDs can refer to phylogeny in Supplementary Figure 5; Het/Hom indicated ratio of homozygous sites divided by heterozygous sites in an individual.

Supplementary Table 1. Sample information and bioclimate variables of *Apis cerana*.

Supplementary Table 2. Divergent time estimations between Central lineage of *Apis cerana* Lineage M with other seven lineages.

Figure 1.

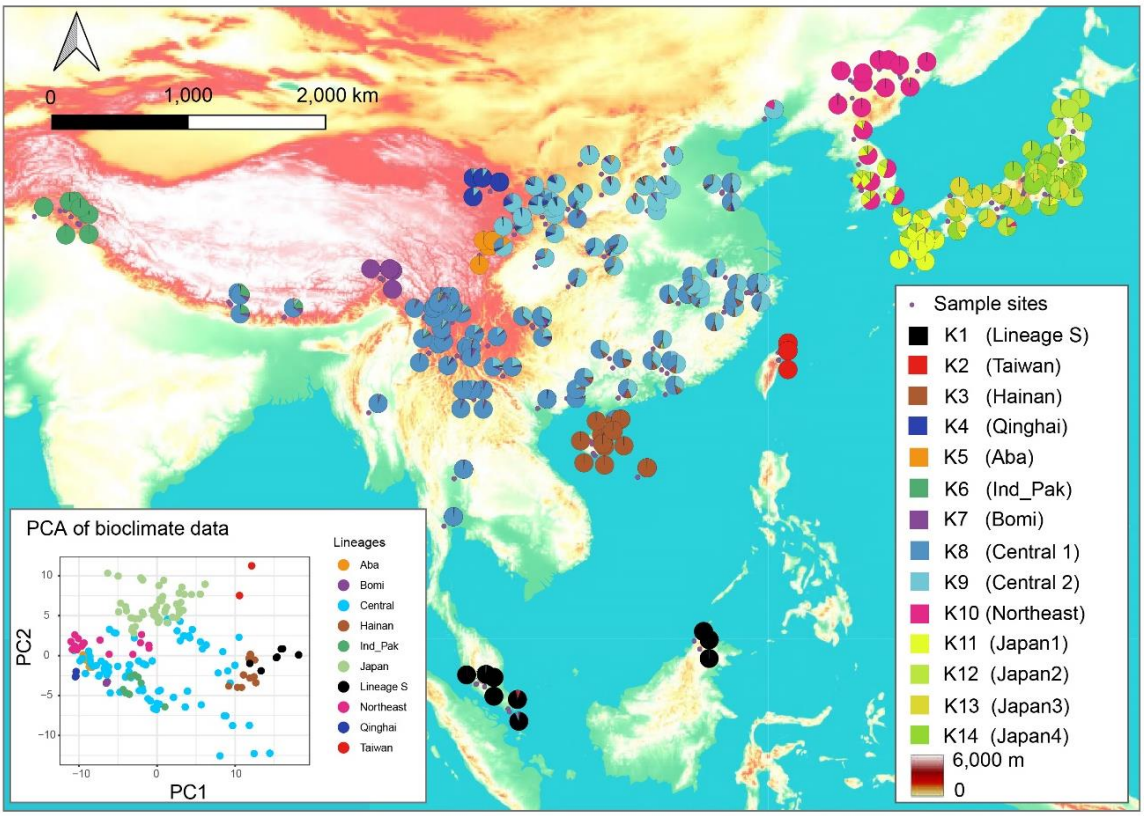

0
1,000
2,000 km
Sample sites
K1 (Lineage S)
K2 (Taiwan)
K3 (Hainan)
K4 (Qinghai)
K5 (Aba)
K6 (Ind_Pak)
K7 (Bomi)
K8 (Central 1)
K9 (Central 2)
K10 (Northeast)
K11 (Japan1)
K12 (Japan2)
K13 (Japan3)
K14 (Japan4)
6,000 m
0
PCA of bioclimate data
Lineages
Aba
Bomi
Central
Hainan
Ind_Pak
Japan
Lineage S
Northeast
Qinghai
Taiwan
PC1
PC2

Figure 2.

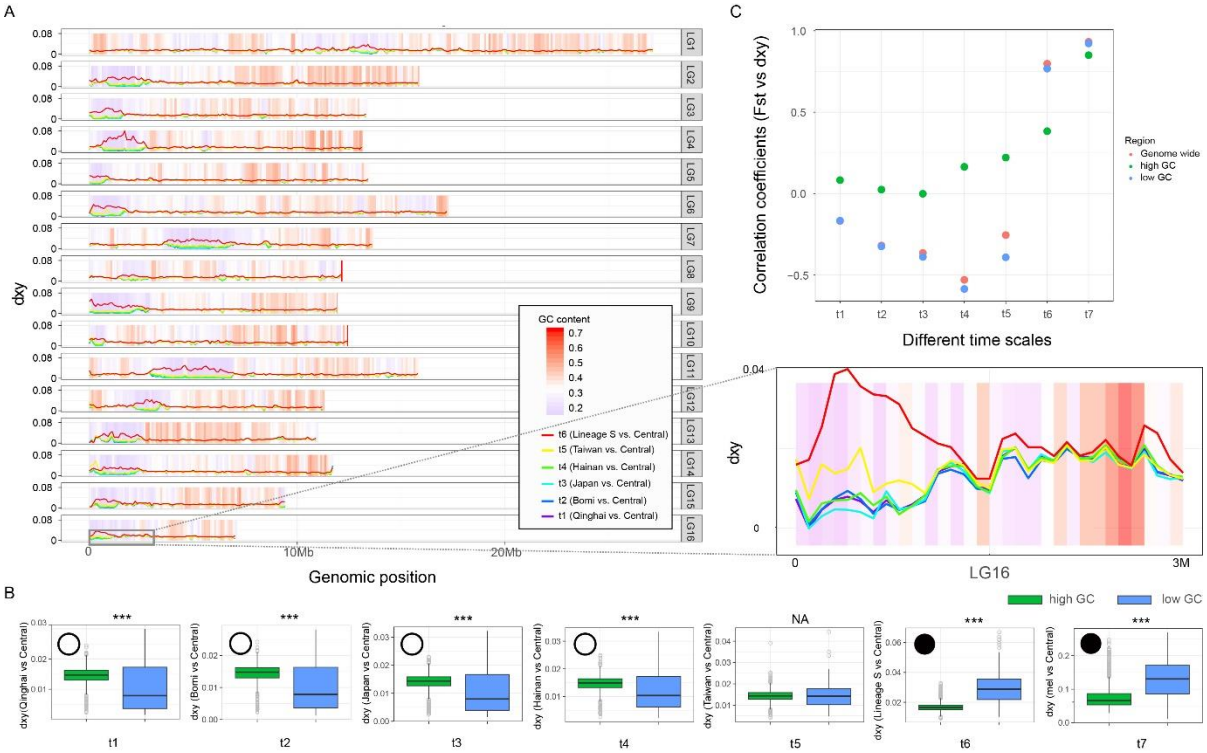

A
dxy
Genomic position
GC content
t6 (Lineage S vs. Central)
t5 (Taiwan vs. Central)
t4 (Hainan vs. Central)
t3 (Japan vs. Central)
t2 (Bomi vs. Central)
t1 (Qinghai vs. Central)
C
Correlation coefficients (Fst vs dxy)
Different time scales
Region
Genome wide
high GC
low GC
LG16
B
high GC
low GC
t1
t2
t3
t4
t5
t6
t7

Figure 3.

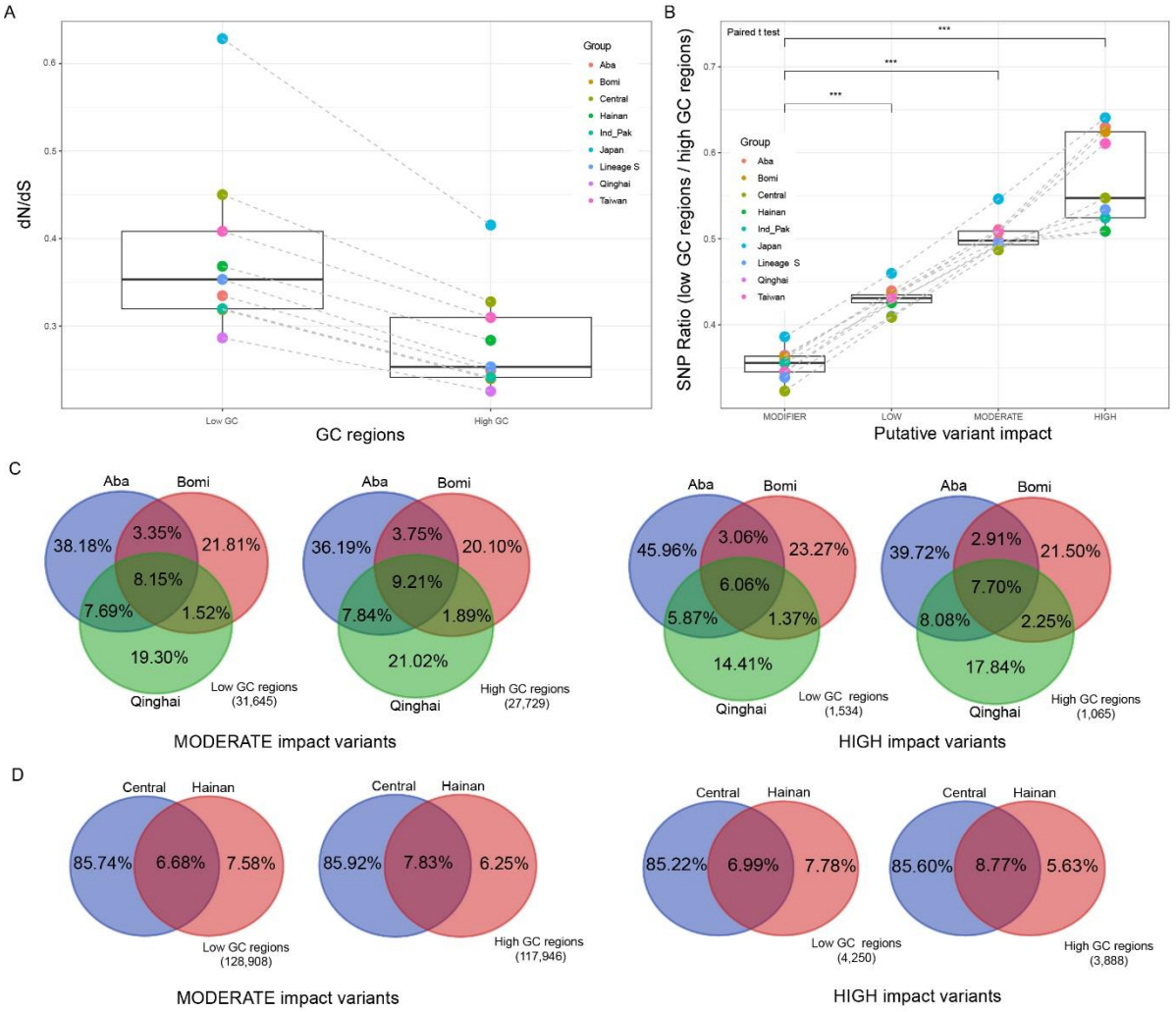
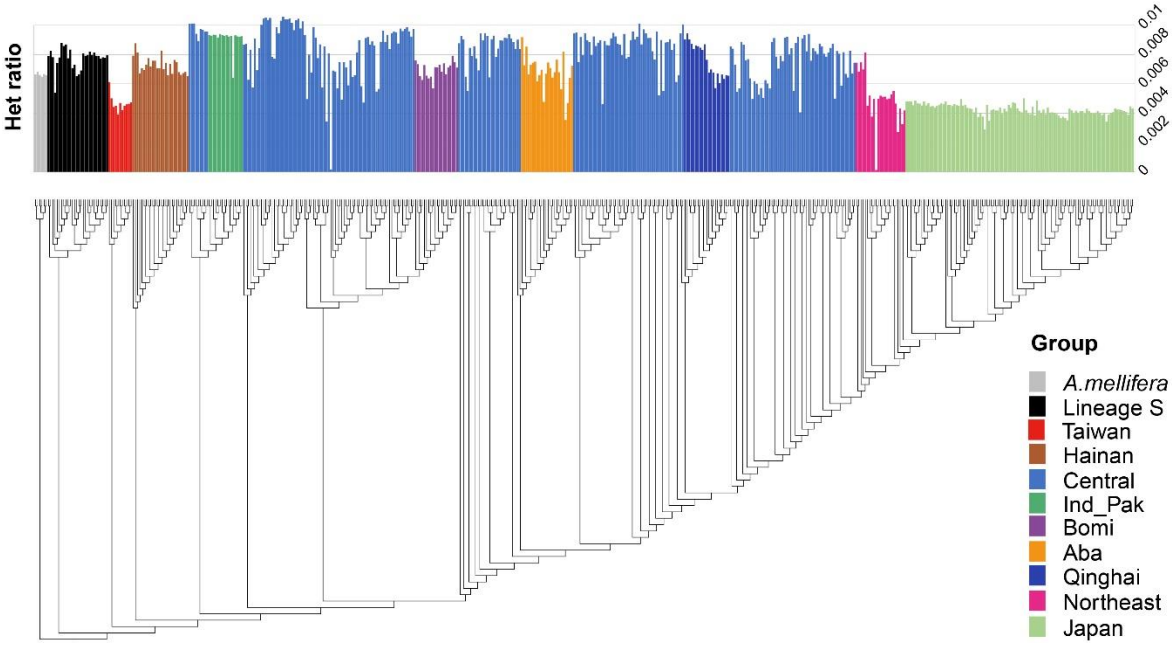
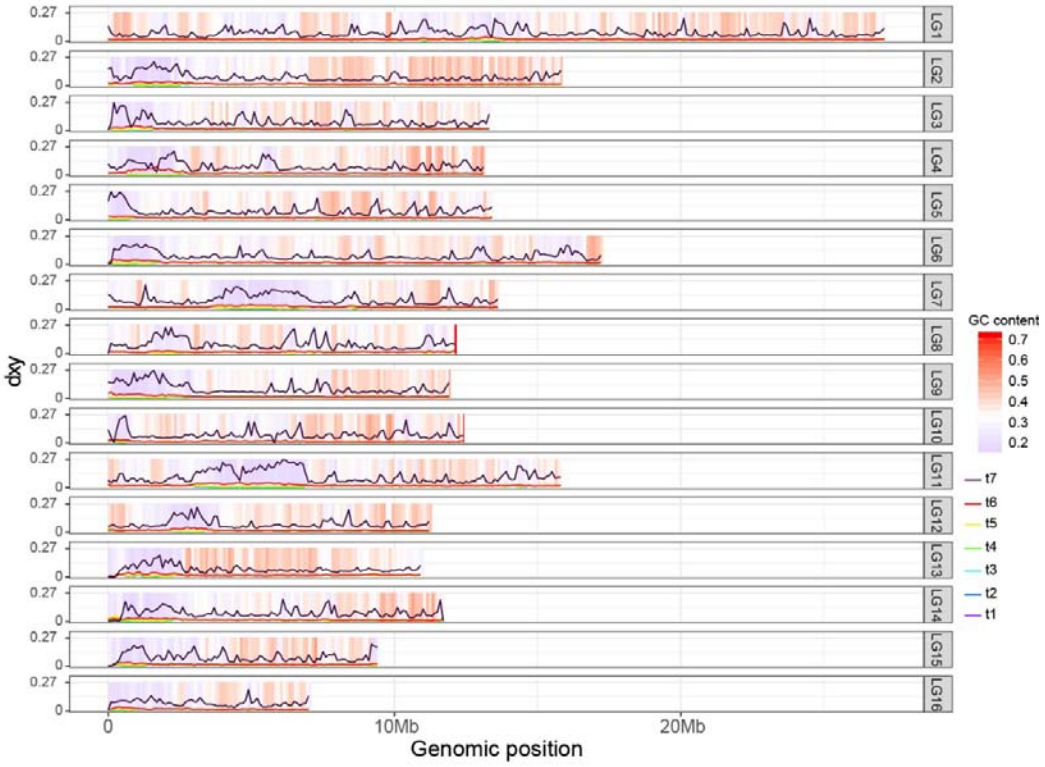
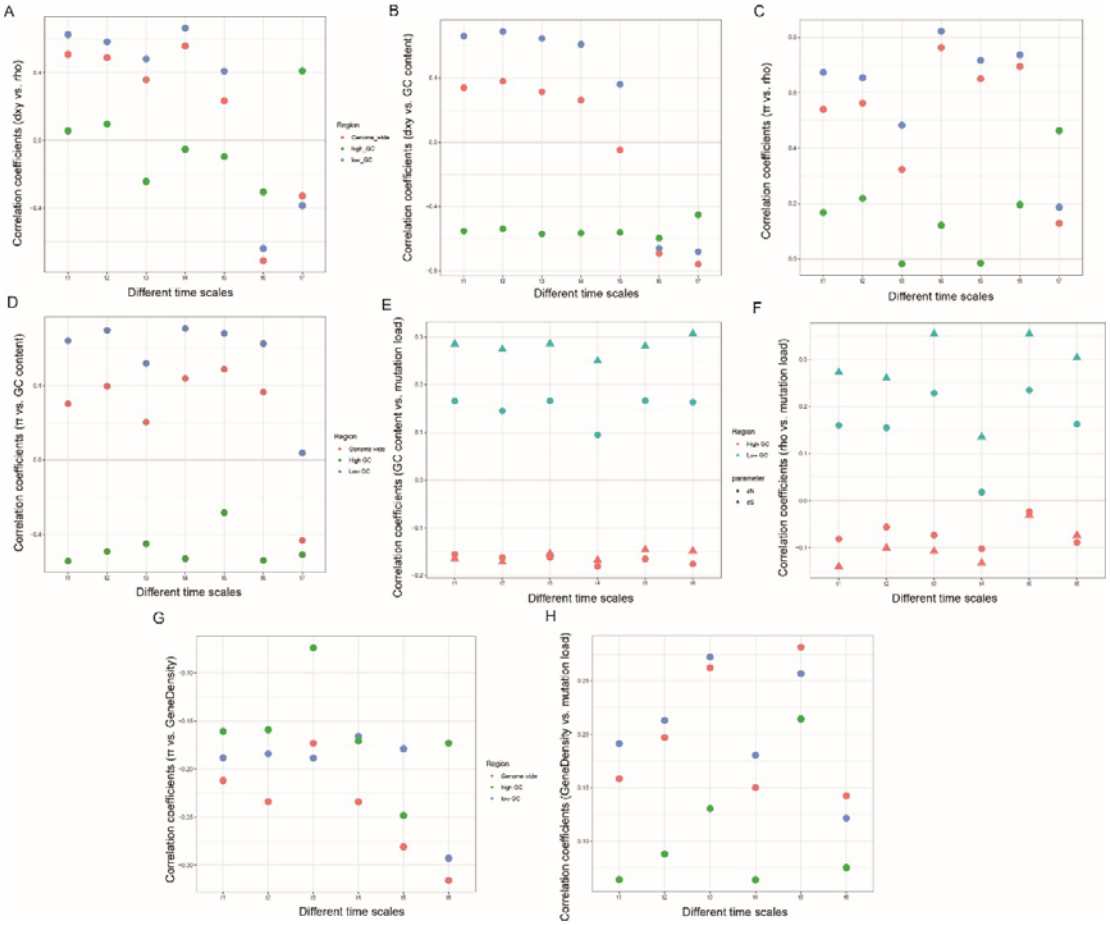

A
dN/dS
0.6
0.5
0.4
0.3
Low GC
High GC
GC regions
Group
Aba
Bomi
Central
Hainan
Ind_Pak
Japan
Lineage S
Qinghai
Taiwan
B
SNP Ratio (low GC regions / high GC regions)
Paired t test
***
0.7
0.6
0.5
0.4
MODIFIER
LOW
MODERATE
HIGH
Putative variant impact
C
Aba
Bomi
38.18%
3.35%
21.81%
8.15%
7.69%
1.52%
19.30%
Qinghai
Low GC regions
(31,645)
36.19%
3.75%
20.10%
9.21%
7.84%
1.89%
21.02%
High GC regions
(27,729)
MODERATE impact variants
45.96%
3.06%
23.27%
6.06%
5.87%
1.37%
14.41%
Low GC regions
(1,534)
39.72%
2.91%
21.50%
7.70%
8.08%
2.25%
17.84%
High GC regions
(1,065)
HIGH impact variants
D
Central
Hainan
85.74%
6.68%
7.58%
Low GC regions
(128,908)
85.92%
7.83%
6.25%
High GC regions
(117,946)
MODERATE impact variants
85.22%
6.99%
7.78%
Low GC regions
(4,250)
85.60%
8.77%
5.63%
High GC regions
(3,888)
HIGH impact variants

Fig 4.

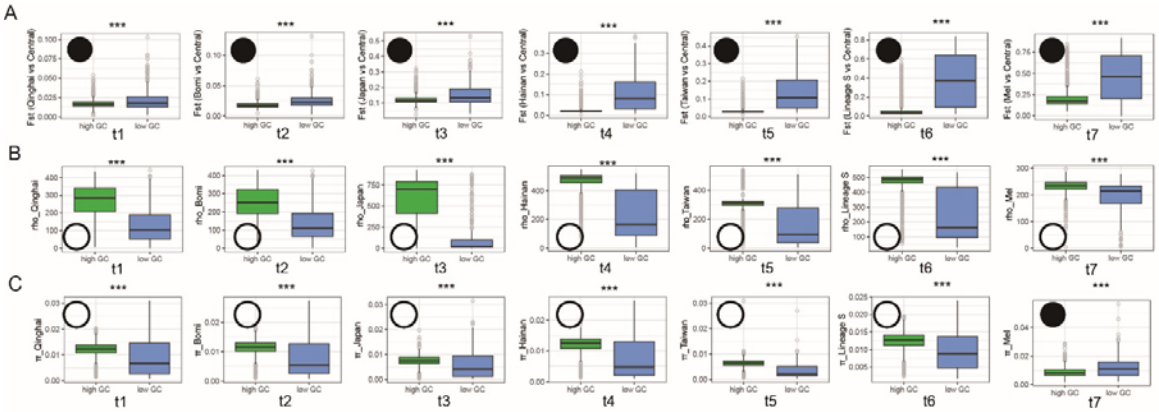

Het ratio
0.01
0.008
0.006
0.004
0.002
0
Group
A.mellifera
Lineage S
Taiwan
Hainan
Central
Ind_Pak
Bomi
Aba
Qinghai
Northeast
Japan



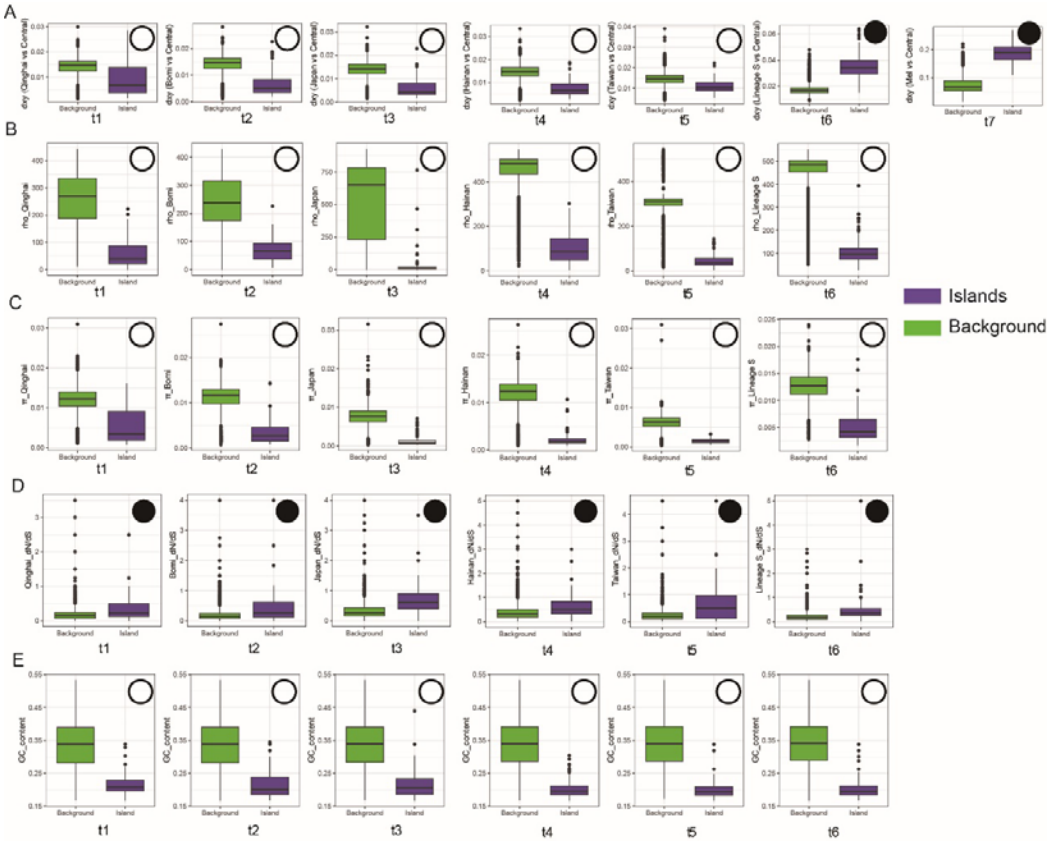
Supplementary Figure 1. Landscape of genomic differentiation over time varied along with regional GC content. Central *v.s.* Qinghai (t1), Central *v.s.* Bomi (t2), Central *v.s.* Hainan (t3), Central *v.s.* Japan (t4), Central *v.s.* Taiwan (t5), Central *v.s.* Lineage S (t6), and Central *v.s.* Mel (t7).



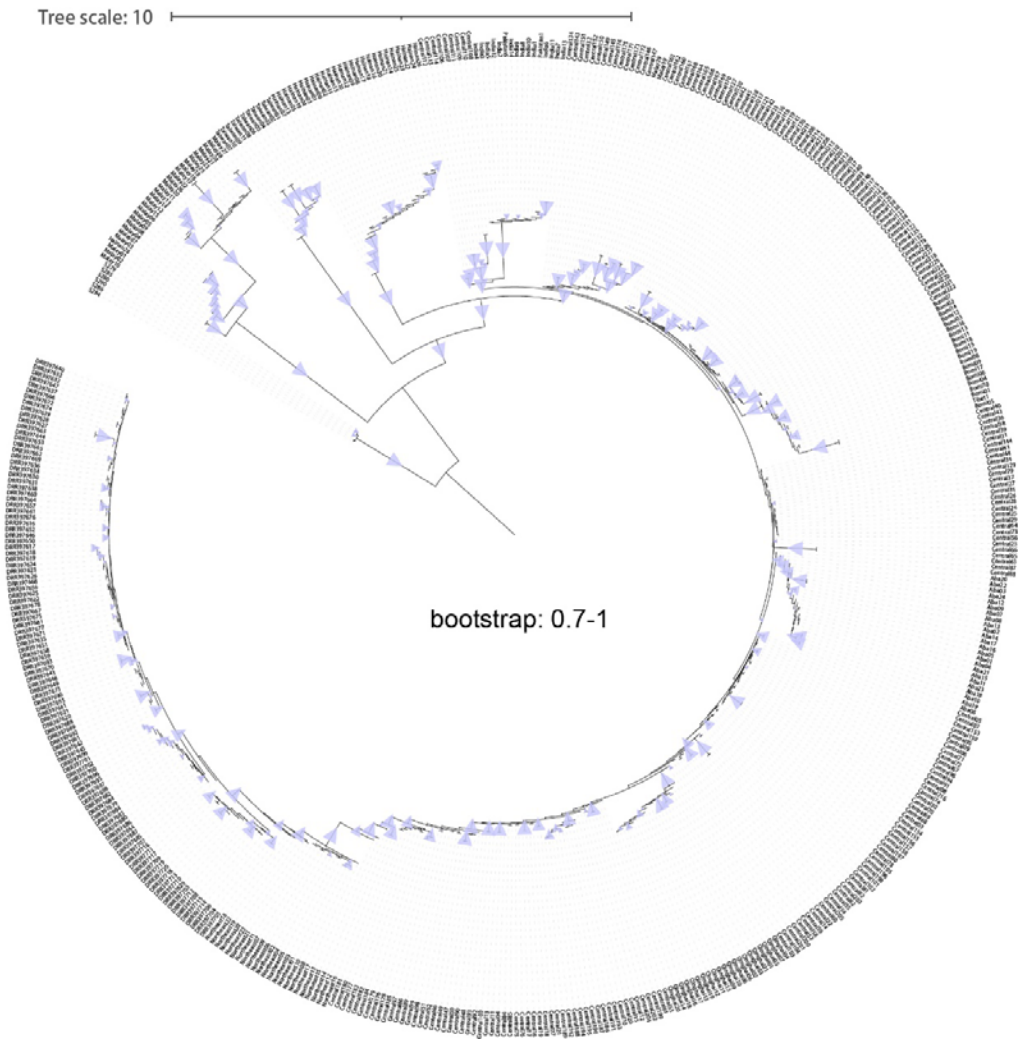
Supplementary Figure 2. Correlation coefficients of genome-wide parameters over divergent time.



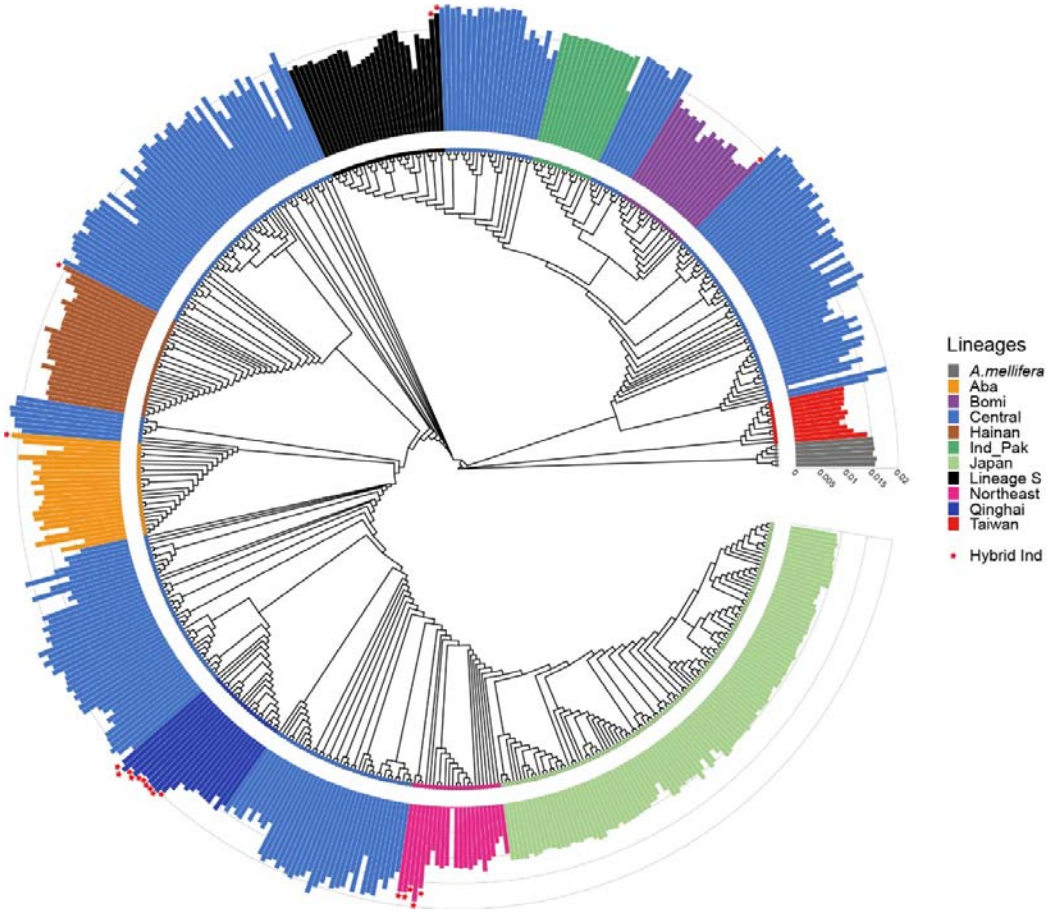
Supplementary Figure 3. Characterization of low GC and high GC regions over divergent time. **(A)** F_{ST}; **(B)** recombination (rho); **(C)** nucleotide diversity (π). Wilcoxon rank sum test was conducted with values between low GC and high GC regions. Empty circles represent significantly lower values in low GC than high GC regions, while black circles indicate low GC regions with higher values.



Supplementary Figure 4. Characterization of genomic islands and background regions over divergent time. **(A)** d_{xy}; **(B)** recombination (rho); **(C)** nucleotide diversity (π); **(D)** dN/dS; **(E)** GC content. Wilcoxon rank sum test was conducted with values between genomic islands and background regions. Empty circles represent significantly lower values in islands than background regions, while black circles indicate islands with higher values.



Supplementary Figure 5. Reconstruction of *Apis cerana* phylogeny based on nuclear polymorphisms in low GC regions. Bootstrap values over 0.7 were presented and showed with blue triangles.



Supplementary Figure 5. Reconstruction of *Apis cerana* phylogeny based on nuclear polymorphisms in high GC regions. Het/Hom indicate ratio of homozygous v.s. heterozygous sites in an individual. The hybrid individuals referred to genetic structure in Figure 1.