Contribution of local recombination and AT-biased mutations to differentiated region formation in *Apis cerana*

Fushi Ke[1,#], Liette Vasseur[2]
[1] Independent researcher, Guangzhou, China
[2] 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 architecture can interact with evolutionary processes and be involved in the formation of differentiated regions potentially containing adaptation and speciation loci. Regions of low GC content can be linked to low mutation and recombination rates. These effects on the formation of heterozygous differentiation landscapes have not been investigated. Here, we explored mutation accumulation and group divergence along a speciation continuum using 499 genomes of *Apis cerana*, with a widely distributed Central group diverged with its peripheral groups at both population genetic and phylogenetic timescales. We found that the differentiated regions had generally lower recombination and GC content than the rest of the genome, with lower-than-average divergence (d_{xy}) initially to higher-than-average ones at deeper timescale. Most rare alleles (~80%) are AT-biased that derived from GC, resulting in lower AT-biased mutations in low-GC regions. Multiple regression analysis further shows that reduced mutation and recombination rates are associated with decreased diversity, particularly in regions with low GC content. Moreover, in all *A. cerana* groups, higher mutation load and less efficient selection in low-GC regions compared with high-GC regions suggest that restricted recombination is important in mutation accumulation (e.g., AT-biased) and group divergence. This pattern explains the increased d_{xy} in low-GC regions over evolutionary time. Finally, low-GC regions possess higher proportion of group-specific polymorphisms, which reconciliate discordance between mitochondrial and nuclear phylogenies in *A. cerana*. Our results highlight the contribution of genome architectures to the formation of differentiation landscapes along divergent groups, emphasizing caution regarding loci identified solely from differentiated regions.

Keywords

Eastern honeybee; low GC; mutation rate; mutation load; recombination

38 **Significance Statement**

39 This study shows the low nucleotide diversity in differentiated regions could be
40 simply attributed to genome architectures (e.g., gBGC, local recombination and
41 mutation rates). Further studies should consider the effects of genome architecture
42 on understanding the formation of heterozygous differentiation landscape along
43 diverging groups and identification of adaptive loci.

Introduction

Genomic islands are genomic regions with reduced nucleotide diversity (π) and elevated genetic differentiation (F_{ST}) that may potentially be 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 have facilitated our understanding of heterogeneous genomic landscapes 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 adaptation and speciation (Cruickshank and Hahn 2014; Guerrero and Hahn 2017; Ravinet et al. 2017; Wolf and Ellegren 2017). Authors (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) have proposed multiple evolutionary models to explain the formation of genomic islands in evolving groups. Recurrent selection can increase F_{ST} while reducing nucleotide divergence (d_{xy}) in genomic islands (Charlesworth et al. 1993; Irwin et al. 2016; Ravinet et al. 2017; Burri 2017). Selection in allopatry can generate regions with elevated genetic differentiation (with reduced nucleotide diversity), while not affecting their nucleotide divergence (Cruickshank & Hahn, 2014; Irwin et al., 2018). On the other hand, divergent selection coupling with gene flow may lead to the formation of genomic islands with both elevated F_{ST} and d_{xy} (Feder et al. 2012; Wolf and Ellegren 2017). Genomic islands with high F_{ST} and d_{xy} can also arise from ancient polymorphism, especially if gene flow is absent (Guerrero and Hahn 2017; Noor and Bennett 2009). In addition, genome architectures, such as gene density, recombination, and mutation rate, can all 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).

Unlike other genome architectures, few have investigated the contribution of

74 GC content (proportion of guanine (G) and cytosine (C) nucleotides) on mutation
75 accumulation and genetic differentiation along diverging groups. While GC content
76 may not be a direct driver of differentiation or divergence, positive correlation
77 between GC content and regional recombination rate has been reported from yeast
78 to human (e.g. Meunier et al. 2004; Pessia et al. 2012; Kiktev et al. 2018). Regions
79 with high-GC content are more susceptible to double-strand DNA breaks (i.e.,
80 recombination) than low-GC ones (Petes 2001; Kiktev et al. 2018; Bouwman et al.,
81 2023). On the contrary, recombination can drive the fixation of G and C alleles
82 through GC-biased gene conversion (gBGC) (Charlesworth et al., 2020; Fullerton et al.
83 2001; Meunier et al. 2004; Pessia et al. 2012).

84 Through generation of AT mutations (Hershberg and Petrov, 2010; Lynch 2007),
85 GC content may increase the local mutation rate (Fryxell and Moon 2005; Kiktev et al.
86 2018, Long et al., 2018). Rare, derived alleles can be used as a proxy of mutation rate
87 since natural selection has yet to act on them in the short evolutionary time
88 (Coventry et al., 2010; Keinan and Clark, 2012). In *Apis mellifera* genome, for
89 example, elevated C/G to A/T mutations account for over 80% of rare, derived alleles
90 (derived allele frequency < 1%) (Wallberg et al., 2015). Regional high GC would
91 therefore contribute to elevated local mutation rate. Simulation results in *Drosophila*
92 *melanogaster* show that local mutation rate is a crucial factor interacting with other
93 forces, such as recombination, to shape nucleotide diversity and thus differentiation
94 across the genome (Barroso and Dutheil 2023). Generation of more AT alleles can be
95 further counteracted by gBGC that leads to fixation of more GC alleles (e.g., Wallberg
96 et al., 2015). High recombination (e.g., in high GC regions) can lead to more efficient
97 purified selection by removing deleterious mutations (Hill and Robertson, 1966;
98 MacPherson et al., 2021). These effects may be prominent in genomes with high
99 recombination, such as the honeybee genome, which has an average rate of 19
100 cM/Mb (centimorgans per megabase pair)—approximately 10 times higher than that
101 of most eukaryotes (Beye et al., 2006). Nevertheless, the interplay between three
102 processes—the generation of AT-biased mutations, the fixation of GC alleles via gBGC,
103 and the removal of deleterious mutations—varies significantly across the genome

104 due to the heterozygous genome architectures (e.g., local GC content and
105 recombination). This interplay may contribute to GC dynamics across the genome
106 and formation of heterozygous differentiation landscape along the speciation
107 continuum.

108 *Apis cerana* is a widespread cavity-nesting honeybee group (Ji et al. 2020; Qiu et
109 al. 2023; Radloff et al. 2010; Su et al. 2023). Its distribution in Asia ranges from
110 Pakistan to Japan and extends from the adjacent regions of Siberia in the North to its
111 southern distribution in Indonesia, from where the bee swarms invaded Oceania
112 (Koetz et al. 2013). There are four putative species in this group based on the
113 inference of genome-wide single nucleotide polymorphisms (SNPs) (Su et al. 2023): *A.*
114 *cerana* (Asian Mainland and Indian black, Lineage M), *A. cerana* (Sundaland, Lineage
115 S), *A. cerana* (Indian yellow, Lineage I), and *A. cerana* (Philippines, Lineage P). The
116 mainland lineage (Lineage M) is the most widely distributed species and formed
117 several peripheral groups/populations with varied levels of divergence at the
118 surrounding of its Central population (Ji et al. 2020; Qiu et al. 2023). The Central
119 population of Lineage M admixes with the Lineage M peripheral populations and
120 introgresses with the other *A. cerana* species in the South (Lineage S) (Ji et al. 2023),
121 forming multiple admixed group pairs diverged at both population genetic and
122 phylogenetic timescales.

123 Here, we used resequencing genomes of 499 *A. cerana* and six *A. mellifera*
124 individuals (outgroup), with groups diverged over evolutionary time, to investigate
125 the contribution (direct or indirect) of GC content to genome-wide differentiation
126 landscape and mutation accumulation over time. Across the genome, GC content
127 (negative), recombination (positive) and mutation rates (positive) are the best
128 variables explaining genome-wide variation of π. When dividing the genome into
129 regions with different GC content, we found that low GC regions were always with
130 lower π and higher F_{ST} (i.e., genomic islands) than the rest of the genome. These
131 regions had few AT-biased mutations (the most dominant mutation type, from
132 ancestral G/C to derived A/T alleles) due to lower GC content, which suggested
133 lower mutation rate. In addition, lower recombination rates in these regions are

134 associated with reduced effects of genetic reshuffling on π. Groups at earlier stages
135 of divergence (i.e., population genetic timescale) showed low d_{xy} but high F_{ST} in low
136 GC regions, which meant even a small number of different mutations between
137 groups could cause high F_{ST} if π was low. The groups that were more diverged
138 exhibited elevated levels of d_{xy} due to the build-up of group-specific mutations (such
139 as fixed AT alleles), leading to d_{xy} values at the phylogenetic timescale (between
140 species) to be greater than the genome-wide average. Higher mutation loads and
141 less efficient selection in low GC regions of all investigated *A. cerana* groups
142 supported the important role of restricted local recombination in variant
143 accumulation and group divergence in these regions. Conflict between mitochondrial
144 phylogeny (Qiu et al. 2023) and its nuclear counterpart (Ji et al. 2020) in *A. cerana*
145 Lineage M could be due to the above-mentioned admixture between adjacent
146 groups. They were reconciled with polymorphisms in low GC regions, where more
147 group-specific variants accumulated. These group-specific variants were easily found
148 in regions with elevated differentiation (i.e., F_{ST}) and divergence (i.e., d_{xy}, Central *vs*.
149 Lineage S) but possibly not related to adaption and speciation processes.

Results

Genetic structure and divergence of Apis cerana

In ADMIXTURE analysis, the optimal K value (i.e., genetic clusters) was 14 when investigating the genomic polymorphisms of 499 *A. cerana* individuals (Figure 1, Supplementary Table 1). *A. cerana* Sundaland individuals (samples from Malaysia and Singapore, Lineage S) first separated from the mainland species (Lineage M) due to long-term divergence (Su et al. 2023). Island groups (e.g. Taiwan, Hainan and Japan) and populations that colonized extreme habitats (e.g. high elevation: Ind_Pak, Bomi, Aba and Qinghai; or high latitude: Northeast) gradually separated. These groups had clear genetic ancestry and geographical distribution (Figure 1). Environmentally differentiated habitats also contributed to the formation of two clusters in the Central group, and four clusters in the Japan group (Figure 1). These clusters displayed intensive gene flow with adjacent ones and were difficult to distinguish from each other (Figure 1). Based on geography, we grouped the two clusters from Central and the four from Japan into one group. In total, we identified ten groups based on genomic clustering and geography. We performed the following analysis with this grouping scheme. It is worth noting that the grouping scheme is not associated with subspecies identification in the eastern honeybee (e.g., Qiu et al., 2022), but for the convenience of data analysis and illustration. It is important to note that the investigated genomic variation and differentiation pattern, as well as divergence time estimation (see below) when combining two Central clusters as one group (Central group) was consistent with the pattern using either one or the other cluster. Clustering of bioclimatic variables could not clearly separate these groups with some overlapping, suggesting a combination of phylogeny, genetic drift, and natural selection in the formation of these groups. The Central group (of Lineage M) was widely distributed and showed varied extent of genetic admixture with peripheral groups of Lineage M and Lineage S (Figure 1). It was thus used as an invariable group for divergence estimation in a time series (see below).

179 Estimated divergence times of group pairs within *A. cerana* Lineage M were 502
180 thousand years (Kyr) in Central *vs.* Japan, 521 Kyr in Central *vs.* Qinghai, 653 Kyr in
181 Central *vs.* Bomi, 700 Kyr in Central *vs.* Hainan, 972 Kyr in Central *vs.* Taiwan
182 (Supplementary Table 2). The divergence was consistent for different group pairs
183 whether using individuals from Southern or Northern part of Central group except
184 for the Central with Japan pair, where individuals from Northern part of Central
185 group showed shallower divergence (Supplementary Table 2). The gradually
186 divergent pattern of five groups with Central group was consistent with the
187 topologies based on nuclear (see below) and mitochondrial polymorphisms (Qiu et
188 al., 2022), thus supporting the robustness of divergence time estimation. These time
189 nodes within population genetic timescale (t1: Central *vs.* Japan; t2: Central *vs.*
190 Qinghai; t3: Central *vs.* Bomi; t4: Central *vs.* Hainan; t5: Central *vs.* Taiwan) and two
191 additional ones from the phylogenetic timescale (t6: Central *vs.* Lineage S; t7: Central
192 *vs*. *A. mellifera*) were chosen (Supplementary Table 2) and used in a time series to
193 investigate changes of genomic differentiation landscape and characterize genomic
194 polymorphisms.
195
196 ***Landscape of genomic variation and differentiation of diverging group pairs***
197 By investigating genomic variation (nucleotide diversity, π) and differentiation
198 (genetic differentiation, F_{ST} and nucleotide divergence, d_{xy}) landscape over time
199 (Figure 2A, Supplementary Figure 1-3), we found a distinct pattern between high GC
200 and low GC regions, not just between centromere and non-centromeric regions.
201 Taking d_{xy} in low GC regions, for example, divergence increased from t1 to t7, with d_{xy}
202 values in t6 and t7 higher than the genome-wide average (Figure 2A, Supplementary
203 Figure 1). In high GC regions, however, divergences over time were indistinguishable
204 among the different group pairs when compared with low GC regions in the
205 genome-wide context (Figure 2A). We further compared nucleotide divergence in
206 low GC and high GC regions over time and found significantly lower d_{xy} values in low
207 GC regions from t1 to t4 than in high GC regions (Figure 2B, Mann-Whitney U test, *p*

208 < 0.001). In t5, t6 and t7, however, d_{xy} values in low GC regions were significantly
209 higher than in high GC regions (Figure 2B, Mann– Whitney U test, $p < 0.001$).
210
211 Correlations between d_{xy} and F_{ST} (Figure 2C) also switched over time in low GC
212 regions, with a negative correlation from t1 (r = -0.43) to t5 (r = -0.14), then positive
213 in t6 (r = 0.75) and t7 (r = 0.61). This pattern was not found in high GC regions (Figure
214 2C, all with positive r values). Correlation between d_{xy} and recombination rate (rho)
215 in low GC shifted pattern while not in high GC regions (Supplementary Figure 4A),
216 with positive r values from t1 (0.41) to t5 (0.26) and negative r values in t6 (-0.64)
217 and t7 (-0.48), and a similar trend found in correlations between d_{xy} and GC content
218 (Supplementary Figure 4B). In addition, we found both nucleotide diversity (π) and
219 mutation load showed positive correlations with GC content in low GC regions, but
220 negative correlations in high GC regions (Figure 3; Supplementary Figure 4).
221 Consistent negative correlations between nucleotide diversity (π) and gene density
222 were detected across the genome, as well as in high GC and low GC regions
223 (Supplementary Figure 4), supporting widespread linked selection in the functional
224 elements.
225
226 Genomic differentiation (F_{ST}) values in low GC regions were always higher and
227 recombination rate (rho) and nucleotide diversity (π) values lower than those in high
228 GC regions (Supplementary Figure 5), thus supporting the linkage of low GC content
229 with formation of genomic islands. We further characterized genomic islands
230 (normalized F_{ST}, Z-F_{ST} ≥ 2) over time and found a significantly lower GC content in
231 genomic islands compared with genomic backgrounds (Z- F_{ST} < 2) (Supplementary
232 Figure 6). Being consistent with the pattern found in low GC regions, recombination
233 rate and nucleotide diversity in genomic islands were significantly lower than those
234 in genomic backgrounds (Supplementary Figure 6). Comparisons of d_{xy} in genomic
235 islands and backgrounds also varied over time, with lower-than-average d_{xy} values
236 initially (t1 to t5) to higher-than-average at t6 and t7 (Supplementary Figure 6). This
237 pattern supported the proposed model of genomic island formation changing from

238 recurrent selection (i.e., high F_{ST} and low d_{xy}) to divergent selection (i.e., high F_{ST} and
239 high d_{xy}) over deeper divergence.
240
241 ***Contribution of recombination and AT-biased derived alleles to nucleotide diversity***
242 ***and group divergence***
243 We further investigated the effects of genome architectures on nucleotide diversity.
244 Rare, derived alleles (loci with derived allele frequency ≤ 1%, DAF_001) were
245 mutations that arose recently and were used here as a proxy for mutation rate
246 because natural selection had not yet worked on them (e.g., Coventry et al., 2010;
247 Keinan and Clark, 2012). Correlation analysis showed that GC content was positively
248 correlated with DAF_001 in both low and high GC regions across the genome (Figure
249 3). In addition, we found most of the mutations (~80%) were derived A/T alleles from
250 ancestral C/G ones (Supplementary Figure 7). This pattern supported the
251 predominant role of GC content on newly generated mutations (i.e., AT-biased)
252 across the genome (Hershberg and Petrov, 2010; Lynch 2007) with fewer mutations
253 in low GC regions (Figure 3). This process could contribute to the low nucleotide
254 diversity in low GC regions and therefore the presence of genomic islands
255 (Supplementary Figure 2, 3, 5, 6).
256
257 To further distinguish the effects of different genomic architectures on genomic
258 variation, we divided the genome into five regions (region 1 being low to region 5
259 being high GC) with equal-width intervals of GC content. We investigated the
260 contribution of each genome architecture to nucleotide diversity using a regression
261 analysis (Table 1). Four genomic architectures were included in the analysis: DAF_001,
262 GC content, recombination PC1, and gene density. Principal component 1 (PC1) was
263 extracted from recombination rate (explaining 74.91% of the total variance) and
264 nucleotide diversity (explaining 86.02% of the total variance) across six groups in *A.*
265 *cerana* Lineage M. Regression analysis across the five GC regions showed that the
266 models explained between 40.3% (region 3) and 85.7% (region 1) of the variance in
267 nucleotide diversity (Table 1). DAF_001 (i.e., proxy of mutation rate) had the

strongest and significantly positive effect in all regions, with coefficients ranging from 0.4222 (region 4) to 0.6985 (region 3; all with $p < 0.001$). GC content showed a consistent negative association in each region, with the strongest value in region 4 ($\beta = -0.5802$, $p < 0.001$). The effect is pronounced because gBGC promotes the fixation of GC alleles, counteracting a genome-wide bias toward AT mutations ($\beta=-2.341$, $p < 0.001$). Recombination rate was generally positively associated, especially in region 1 ($\beta = 0.4214$, $p < 0.001$) and region 2 ($\beta = 0.3638$, $p < 0.001$), but negatively associated in region 3 ($\beta = -0.1075$, $p < 0.01$). Gene density was negatively associated in regions 1–5 (e.g., region 2: $\beta = -0.2496$, $p < 0.001$). Based on the relationships DAF_001 $\propto \mu$ and $\rho = 3N_e c$ (where N_e is the effective population size, μ is the mutation rate, and c is the recombination rate, 3 is due to haplodiploid sex-determination system in honeybee), our models support important and independent roles for both mutation rate (proxied by DAF_001) and recombination rate (proxied by population scaled recombination rate) in shaping nucleotide diversity. Their contributions are particularly pronounced in low-GC regions (Table 1).

Following Wallberg et al. (2015), we further categorized mutation types in *A. cerana* groups using *A. mellifera* as the outgroup. There were four types of mutation S2S (strong to strong) (from C/G in *A. mellifera* to C/G in *A. cerana*), S2W (strong to weak) (from C/G to A/T), W2S (weak to strong) (from A/T to C/G), and W2W (weak to weak) (from A/T to A/T). In *A. cerana* lineage M (Figure 4A), relative site frequency spectrum of the four mutation types showed that an initially high S2W contributed by GC coupling with a fixation bias towards W2S mutations, consistent with a strong GC-biased gene conversion (gBGC) (Wallberg et al., 2015). We found, however, a higher percentage of nearly-fixed S2W mutations in the peripheral groups compared to the Central group and Lineage M, at both population genetic and phylogenetic levels (Figure 4B). This finding supports the contribution of fixed AT alleles to group divergence. But, the relative contribution of fixed AT alleles to group divergence varied among groups. For instance, the percentage differentiation of nearly fixed AT alleles is amplified by genetic drift (e.g., Japan vs. Lineage M Central Lineage) and

298 divergence time (Lineage S vs. Lineage M, Figure 4B, C). In several groups (e.g.,
299 Qinghai, Hainan, Lineage S, Figure 4B, C), we found more fixation of AT-biased alleles
300 in low GC regions (with lower AT mutations) compared with high GC regions (with
301 higher AT mutations), suggesting the involvement of other mechanisms, such as local
302 recombination and gBGC.
303
304 ***Profile of functional polymorphisms in low GC and high GC regions***
305 Patterns of accumulated polymorphism (e.g., mutations) differed between low GC
306 and high GC regions (Figure 5 A, B; Supplementary Figure 4E), with synonymous (dS)
307 and non-synonymous (dN) substitution rates positively correlated with GC content in
308 low GC regions and negatively correlated in high GC regions (Supplementary Figure
309 4). The differentiated pattern likely resulted from limited recombination and
310 inefficient purifying selection in low GC regions (Figure 3; Supplementary Figure 5
311 and 8), which could have facilitated the accumulation of mutations, a pattern that
312 could not be found in high GC regions with higher recombination (Figure 3). The
313 consistently positive correlation between mutations and GC content in low GC
314 regions indicated that these regions accumulated more mutations and divergence
315 than high GC regions at deeper divergence (Figure 2A, B), although fewer mutations
316 were found in these regions (Figure 3).
317
318 We calculated dN/dS and found significantly higher values in low GC regions
319 than in high GC regions (Figure 5A). We further divided the mutation load into
320 variants with different impacts (MODIFIER, LOW, MODERATE, and HIGH) and
321 calculated the ratio of SNP numbers in low GC over high GC regions. We found
322 significantly higher ratios of non-neutral variants (LOW, MODERATE, and HIGH) than
323 neutral ones (MODIFIER) in all defined groups (Figure 5B), which indicated low GC
324 regions could accumulate more impact SNPs than high GC regions due to inefficient
325 purifying selection. This pattern supported the idea that the genome architectures
326 (i.e. restricted recombination) was involved in mutation accumulation and

327 divergence in low GC regions of *A. cerana* groups due to Hill-Robertson interference
328 (Hill and Robertson, 1966).
329
330 We further investigated the ratio of polymorphisms that was group-specific or
331 shared among different groups in low GC and high GC regions. With three groups
332 (Bomi, Aba and Qinghai) that independently colonized highland habitats and had no
333 genetic admixture (Figure 1), we determined that most of the polymorphisms (e.g.
334 mutation load) were unique to each group and showed no percentage difference in
335 high GC and low GC regions (Figure 5C). For example, in MODERATE impact
336 polymorphism, Aba had 38.18% (12,082) and 19.19% (6,073) of unique and shared
337 variants in Low GC regions, and 36.19% (10,035) and 20.8% (5,768) in High GC
338 regions. We further investigated group-specific and shared polymorphisms in lineage
339 pairs with genetic admixture. With Central and Hainan group pair as an example, we
340 found more group-specific than shared polymorphism only in low GC regions (Figure
341 5D). For MODERATE impact polymorphism, Hainan had 7.58% (9,771) and 6.68%
342 (8,611) unique and shared variants in low GC regions, corresponding to 6.25% (7,372)
343 and 7.83% (9,235) in high GC regions.
344
345 ***Reconstruction of A. cerana nuclear phylogeny***
346 After characterizing the landscape of genomic variation in high GC and low GC
347 regions, we reconstructed the phylogeny of *A. cerana* based on nuclear
348 polymorphism in low GC regions with more group-specific polymorphisms (Figure
349 5D). We reconstructed a nuclear topology similar to the mitochondrial phylogeny
350 (Figure 6 and Supplementary Figure 9, mitochondrial phylogeny from Qiu et al. 2023),
351 with individuals from Taiwan and Hainan being basal groups of *A. cerana* Lineage M.
352 The phylogeny based on polymorphism in high GC regions was influenced by gene
353 flow, with admixed groups clustering together (Supplementary Figure 10). Individuals
354 showing hybrid ancestry were always situated between groups (e.g. Central vs.
355 derived group, Figure 1 and Supplementary Figure 10) and had higher polymorphism
356 than the individuals within the derived groups (Supplementary Figure 10).

357 **Discussion**

358 The formation of heterozygous differentiation landscape is suggested to result from
359 the combined effects of external evolutionary processes and genome architectures
360 (Ravinet et al. 2017). Here, we used 499 eastern honeybee genomes along the
361 speciation continuum to investigate the influences of GC content on mutation
362 accumulation and group differentiation by conducting correlation analysis of
363 population genetic parameters and genomic architectures, characterization of
364 different genomic regions, and regression analysis of genome architectures on
365 genomic variation. We found that the local GC content could elevate regional
366 mutations because GC to AT mutations accounted for around 80% of the newly
367 generated alleles (Supplementary Figure 7). This however could be counteracted by
368 GC-biased gene conversion (gBGC) that lead to fixation of GC alleles.. We also found
369 local GC content was positively correlated with recombination rate that further
370 interplayed with evolutionary forces and shaped mutation accumulation and group
371 differentiation across the genome.

372 Over the diverging groups, genomic islands have significantly lower GC content
373 and recombination compared with background regions (Supplementary Figure 4).
374 This is consistent with fewer mutations (i.e., lower GC) and limited recombination
375 that leads to lower nucleotide diversity and elevated differentiation (Cruickshank and
376 Hahn 2014). In addition, genomic islands with reduced nucleotide diversity were
377 shared across diverging groups of *A. cerana* (Supplementary Figure 2, 3), suggesting
378 the important role of common genome architectures (Barroso and Dutheil 2023).

379 The lower-than-average divergence (d_{xy}) in GC regions at population genetic
380 timescales (within species) to higher-than-average d_{xy} at phylogenetic levels
381 (between species) corresponds to the suggested evolution models transformed from
382 "sweep-before-differentiation" and background selection (i.e. recurrent selection,
383 Charlesworth et al. 1993; Irwin et al. 2016; Ravinet et al. 2017), to divergent selection
384 with gene flow (Feder et al. 2012; Nosil et al. 2009; Wolf and Ellegren 2017) or
385 sorting of ancient polymorphisms (Guerrero and Hahn 2017; Han et al. 2017; Noor
386 and Bennett 2009; Wolf and Ellegren 2017). Correlations between d_{xy} and

387 recombination rate (rho) in low GC regions were negative in population genetic
388 timescale but positive in phylogenetic timescale. A similar trend was also found in
389 the correlations between d_{xy} and F_{ST}. These results align with previously proposed
390 models of genetic island formation, which suggest that initial formation involves
391 recurrent selection in regions with low recombination (Ravinet et al. 2017). In
392 contrast, divergent selection due to the breakdown of linkage occurs over deeper
393 divergence (Cutter and Choi 2010; Noor and Bennett 2009). Few studies have
394 reported a similar pattern (increased d_{xy} values in genomic islands along the
395 speciation continuum, e.g., Liu et al. 2022, and shift pattern of parameter
396 correlations) likely because few have analyzed group pairs at both population genetic
397 (within species) and phylogenetic (between species) timescales in a single system.
398 Many studies report on one dominant model in genomic island formation, with the
399 contribution of external processes and genomic architectures. For example, recurrent
400 selection is reported to be involved in the formation of genomic islands in population
401 pairs of Green-backed Tit (*Parus monticolus*) (Jiang et al. 2023). Zhou et al. (2024)
402 indicate that divergent selection contributes to regionally elevated divergence and
403 genomic islands in incipient species of plover (*Charadrius alexandrines* and *C.*
404 *dealbatus*). However, reductions in recombination were needed in the formation of
405 these genomic islands in both cases (Jiang et al. 2023; Zhou et al. 2024).
406 Synonymous (dS) and non-synonymous (dN) polymorphisms, and mutation load
407 (dN/dS) in both low GC and high GC regions were also analyzed. The results showed
408 higher mutation loads (i.e. higher dN and dN/dS ratio) in low GC regions than high GC
409 regions, suggesting inefficient natural selection (Arbuthnott and Rundle 2012;
410 Bertorelle et al. 2022; Leroy et al. 2021) in these regions compared with the rest of
411 the genome. It is worth noting that this pattern was not only in peripheral groups
412 with lower genome-wide nucleotide diversity (e.g., Japan group) where it might be
413 attributed to genetic drift (Bertorelle et al. 2022). Instead, it was observed across all
414 *A. cerana* groups including *A. cerana* Lineage M Central group and Lineage S. This
415 broader finding supports the continuous effects of restricted recombination leading
416 to the accumulation of neutral polymorphisms and mutation load due to

417 Hill-Robertson interference (Hill and Robertson, 1966) in low GC regions. The relative
418 contribution of external evolutionary processes and genome architectures to
419 regional variation and differentiation thus needs to be reconsidered. The reduced
420 nucleotide diversity of the genomic islands may not result from natural selection
421 (Feder and Nosil 2009; Turner and Hahn 2010), genetic drift, or both (Burri 2021). It
422 could be independently contributed by local genomic architectures, where lower
423 mutation rate governs the accumulations of less random mutations (Castellano et al.
424 2020), and reduced recombination protects low GC regions from natural selection
425 (Hill and Robertson, 1966; Leroy et al. 2021) and gene flow (Wong and Filatov 2023)
426 along the speciation continuum. Therefore, the genomic islands identified in *A.*
427 *cerana* groups—whether at the population genetic timescale (characterized by high
428 F_{ST} and low d_{xy}) or the phylogenetic timescale (characterized by high F_{ST} and high
429 d_{xy})—could be largely explained by the influence of genome architectures (such as
430 low mutation and recombination rates) rather than by specific evolutionary
431 processes like recurrent or divergent selection.

432 In contrast to low GC regions, high recombination in high GC regions facilitate
433 natural selection (Roze 2021) and genetic admixture (Wong and Filatov 2023). This
434 leads to lower mutation load and reduces differentiation in these regions at group
435 pairs of both intraspecific and interspecific levels. Therefore, admixed ancestry in
436 high GC regions could blur nuclear phylogeny of two diverging groups in gene-flow
437 scenario and cause conflict between nuclear and cytoplasmic DNA topologies of *A.*
438 *cerana* (e.g., Ji et al. 2020; Qiu et al. 2023). The conflict, however, could be reconciled
439 by polymorphisms in low GC regions with restricted recombination (Pease and Hahn
440 2013). These findings supported differentiated recombination in high and low GC
441 regions related to regional nucleotide diversity and accumulation of mutation loads
442 (e.g., lead to lineage divergence) (Barroso and Dutheil 2023).

443 Our results suggested that genome architectures such as regional mutation
444 and recombination could shape mutation accumulation (e.g., lower nucleotide
445 diversity in low GC regions) and group divergence (e.g., more group-specific alleles in
446 low GC regions) at different genome regions along divergent groups. Genomic

447 studies often reveal correlations between variation and divergence in evolving
448 groups across various species (e.g., Cuevas et al., 2022; Vijay et al., 2016). These
449 patterns are likely influenced by conserved genomic architectures, such as
450 recombination (Burri et al., 2015; Renaut et al., 2014; Shang et al., 2023). The
451 influence of local mutations and gBGC in nucleotide diversity and genetic
452 differentiation has received less investigation (Boman et al., 2021). Due to the
453 significant effects of regional A/T biased mutation, gBGC, and recombination on
454 nucleotide diversity and fixation of alleles, we would suggest considering the effects
455 of local genome architectures in future analyses related to the formation of
456 heterozygous differentiation landscapes along diverging groups and identification of
457 adaptive loci based on summary parameters (e.g., F_{ST} and d_{xy}).

458 Genome architectures, such as variation of recombination rate (Booker et al.,
459 2020), have been widely reported to produce pseudo positives that could not be
460 easily distinguished from those generated in non-neutral processes in genome-scan
461 methods (e.g., Southcott & Kronforst, 2017). When explaining signals from
462 genome-scan methods, caution is advised, and stricter methods should be applied
463 when isolating adaptation and speciation loci. Several parameters have been
464 introduced to refine the analysis of population differentiation statistics and minimize
465 false positives. For F_{ST} statistic, ΔF_{ST} (Burri, 2017) and Population Branch Statistic (PBS)
466 (Yi et al., 2010), for example, have been proposed. These approaches leverage one or
467 several reference populations to reduce pseudo positive signals. When calculating
468 dxy or other linkage-based parameters like the haplotype homozygosity statistic
469 (Garud et al., 2015), the use of a control population can help eliminate false positives,
470 thus reducing burden in further functional validation efforts.

Materials & Methods

Sequencing data processing, SNP calling, and variant filtering

A total of 507 *Apis cerana* workers were initially sampled across its distribution range from previous studies, including individuals from China (331), India (14), Japan (105), Malaysia (26), Myanmar (2), North Korea (1), Pakistan (5), Singapore (1), South Korea (6), Thailand (11) and Viet Nam (5) (Chen et al. 2018; Ji et al. 2020; Qiu et al. 2023; Wakamiya et al. 2023). Six *A. mellifera* individuals (Harpur et al. 2014) were used as the outgroup individuals. After preprocessing, eight *A. cerana* individuals with elevated percentage of missing sites (≥20%) or with unknown sampling sites were removed, resulting in a total of 505 individuals (499 *A. cerana* and six *A. mellifera* individuals) to be used for the analysis (Supplementary Table 1).

All sequencing reads were downloaded from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (PRJNA592293, PRJNA418874, PRJNA870246, PRJNA216922), or DNA Data Bank of Japan (DDBJ) database (PRJDB14080). Clean reads were trimmed and filtered using Trimmomatic (v 0.39, Bolger et al. 2014). Reads with a minimum length of 50bp were kept for analysis. We further individually mapped clean reads to the *A. cerana* genome (GCF_001442555.1, Park et al. 2015) by using bwa (v 0.7.17, Li 2013). Duplicates were identified by Picard (v 2.24.0, http://broadinstitute.github.io/picard/). A standard pipeline of variant calling based on Genome Analysis Toolkit (GATK 4.2.4.0, McKenna et al. 2010) was applied with HaplotypeCaller for indel realignment and individual variant calling, and GenotypeGVCFs for joint variant genotyping in all samples with "--include-non-variant-sites and --heterozygosity 0.001". We further performed hard filtering based on suggested parameters in GATK by using VariantFiltration command (--filter-expression "SOR > 3.000", --filter-expression "FS > 60.000", --filter-expression "MQRankSum < -12.500", --filter-expression "ReadPosRankSum < -8.000", --filter-expression "ReadPosRankSum > 8.000", --filter-expression "MQ < 40.00", --filter-expression "QD < 2.00"). In addition, hard filtering was conducted with Vcftools (v 0.1.16, Danecek et al. 2011) using "—minDP 3 --min-meanDP 3

501 --max-meanDP 60 --minQ 30 --max-alleles 2 --max-missing 0.5 --remove-indels
502 --remove-filtered-all" to generate a high-quality dataset. This dataset, including
503 non-variant sites, was used for calculation of population genetic parameters,
504 including nucleotide diversity (π) and divergence (d_{xy}).
505
506 We also used Vcftools (v 0.1.16, Danecek et al. 2011) in SNP filtering when
507 investigating genetic structure and reconstructing phylogeny. In genetic structure
508 analysis, bi-allelic SNPs with a maximum missing value of 20% and a minor allele
509 frequency of 0.05 were extracted, after which one SNP from a window of 20-kb DNA
510 sequence was kept for analysis. We used a 20-kb window for thinning to remove
511 linkage of polymorphic sites in genetic structure inference. For window phylogeny
512 reconstruction with maximum likelihood method (see **2.5 Nuclear phylogeny**
513 **reconstruction**), we kept all bi-allelic SNPs in the entire dataset for the analysis.
514
515 ***Genetic structure and population genetic parameter calculation***
516 We used ADMIXTURE (v 1.3.0, Alexander et al. 2009) for genetic structure analyses
517 and performed each K from 2 to 20 with 30 replicates. The best K value was
518 identified according to cross-entropy criterion. The outgroup individuals (*A. mellifera*)
519 were excluded for the genetic clustering analysis. To characterize the environmental
520 conditions for each sample, we obtained monthly climatic variables (e.g.,
521 temperatures, precipitation) and water balance for global terrestrial surfaces from
522 1996 to 2015 (high-spatial resolution, 1/24°, ~4-km) (Abatzoglou et al. 2018). We
523 then performed a Principal Component Analysis (PCA) with twelve monthly averages
524 of the twelve variables (a total of 144 monthly variables, Supplementary Table 1) to
525 investigate environmental heterogeneity of the sampling locations.
526
527 To have an unbiased calculation of π and d_{xy}, we used the dataset with variant
528 and non-variant sites. We calculated π, d_{xy}, F_{ST} in a 100 kb window size within and
529 among our identified groups using Python scripts downloaded from
530 github.com/simonhmartin/genomics_general. We removed all individuals that had

531 more than 20% admixed ancestry from the other group, resulting in 24 (Aba), 20
532 (Bomi), 224 Central, 26 (Hainan), 16 (Ind_Pak), 105 (Japan), 27 (Lineage S), 23
533 (Northeast), 21 (Qinghai) and 11 (Taiwan). Six *A. mellifera* (i.e., Mel, outgroup) were
534 also included in parameter calculation. Windows with less than 1000 genotyped sites
535 (variant and invariant) were removed from the analysis. Regional GC-content and
536 coding sequence percentage were also calculated in 100 kb window based on the *A.*
537 *cerana* genome sequence (GCF_001442555.1). For each group, calculation of
538 recombination rate was conducted by using FastEPRR (v 2.0, Gao et al. 2016). Dataset
539 with variant sites was extracted with Vcftools (v 0.1.16, Danecek et al. 2011) and
540 phased using beagle (v 5.4, Browning et al., 2021) with default settings. We then
541 calculated population-scaled recombination rate of each group in FastEPRR in a
542 sliding window size of 100kb.
543
544 Within *A. cerana* Lineage M, we further estimated divergence times of five
545 groups (Bomi, Hainan, Japan, Qinghai and Taiwan) with the Central group based on
546 coalescent theory in two-two-outgroup (TTo) method (Sjödin et al., 2021), which is
547 relatively robust to migration and could alleviate biases due to drastic ancestral
548 population size change (Sjödin et al., 2021). The ancestral state of the sites was
549 inferred with *A. mellifera*, and *A. cerana* Lineage S was used as the outgroup in TTo
550 method. Since Central group of *A. cerana* Lineage M showed divergence in Southern
551 part and Northern part of the distributed region, we calculated and checked the
552 divergence time of each group with Central group using individuals from Southern
553 and Northern region of Central group. The mutation rate μ equals to 5.27×10^{-9} per
554 base pair following Wallberg et al. (2014) and generation time g to be 1 year (Ji et al.,
555 2020).
556
557 ***Characterization of summary statistics in different regions***
558 The pilot analysis and visualization of genome-wide parameters showed distinct
559 pattern of differentiation in high GC and low GC regions (Figure 2A, Supplementary
560 Figure 1). Therefore, we divided the genome-wide windows into high GC (higher than

561 average GC content) and low GC (lower or equal to average GC content) categories.
562 Difference in parameter values in high GC and low GC windows were compared and
563 tested based on Wilcoxon rank sum test in ggsignif R package (v 0.6.4, Ahlmann-Eltze
564 and Patil 2021). We further standardized per-window F_{ST} in each group pair to a Z
565 score (Han et al. 2017) based on the formula: Z-F_{ST} = (window F_{ST} − mean F_{ST}) /
566 standard deviation F_{ST}. We then compared the parameter values in the genomic
567 islands (windows with Z- F_{ST} ≥2 from each group pair) and background regions
568 (windows with Z- F_{ST} < 2) using a Wilcoxon rank sum test.
569
570 We conducted correlations of summary statistics to investigate the effects of
571 different evolutionary forces on formation of heterozygous genomic landscape. The
572 Pearson's correlation coefficient for intragroup summary statistics between groups
573 (e.g. π_i vs. π_j), within groups (e.g. π_i vs. rho_i), and intergroup statistics of group pairs
574 (e.g. π_i vs. F_{STij}) were calculated and performed in R package corrplot (v 0.94, Wei and
575 Simko 2021). The correlation analysis for each group pair was also performed in high
576 GC and low GC regions, respectively.
577
578 To investigate the effects of genome architectures on polymorphisms, we
579 grouped all individuals from *A. cerana* Lineage M and performed correlations
580 between GC content and recombination rate, gene density, and polymorphism of *A.*
581 *cerana* Lineage M, including rare derived alleles (with derived allele frequency ≤ 0.01,
582 DAF_001), and nucleotide diversity (π). We extracted principal component 1 (PC1) for
583 recombination rate (explained 74.91% variance) and nucleotide diversity (explained
584 86.02% variance) from six *A. cerana* Lineage M groups for analysis. In order to
585 capture variation across the genome instead of groups, we normalized each
586 parameter (i.e., recombination rate or nucleotide diversity) in each group before
587 analysis for PC1 extraction. To identify variables contributing to nucleotide diversity
588 (nucleotide diversity PC1), we further conducted regression analysis with genome
589 architectures (DAF_001, GC content, recombination PC1, and gene density) in
590 high-GC and low-GC regions, as well as in five divided regions based on equal-width

591 intervals of GC content.

592

593 The allele state (ancestral or derived) of each locus in *A. cerana* was inferred
594 with *A. mellifera* as the outgroup. The derived allele frequency of each locus in each
595 *A. cerana* group was then calculated with Vcftools (v 0.1.16, Danecek et al. 2011).
596 With *A. mellifera* as the outgroup, we further sorted and counted the number of four
597 mutation types in each *A. cerana* groups following Wallberg et al. (2014), including
598 S2S: strong to strong (from C/G in *A. mellifera* to C/G in *A. cerana*), S2W: strong to
599 weak (from C/G to A/T), W2S: weak to strong (from A/T to C/G), and W2W: weak to
600 weak (from A/T to A/T). We then investigated the relative site frequency spectrum of
601 four mutation types in each *A. cerana* group and along the GC gradient.

602

603 ***Investigation of mutation load***

604 Synonymous and non-synonymous mutations in high GC and low GC regions in each
605 group were investigated. Bi-allelic SNPs were extracted from each group and
606 functionally annotated with SNPeff (v 5.1d, Cingolani et al. 2012), a variant
607 annotation and effect prediction tool. In each group, we then calculated dN/dS
608 (non-synonymous / synonymous SNPs) ratio that represents magnitude of mutation
609 load. In addition, numbers of SNPs in low GC regions divided by high GC regions were
610 calculated in each functional group (MODIFIER, LOW, MODERATE, and HIGH) in the
611 defined groups. To further characterize whether the polymorphisms were shared or
612 group-specific, mutation load in three highland groups (i.e., Aba, Bomi, and Qinghai)
613 that colonized Qinghai-Tibet Plateau independently and with no admixture with each
614 other, was investigated. As a comparison, mutation loads in group pair with
615 admixture (i.e., Central and Hainan) were further investigated to estimate the effects
616 of gene flow on group-specific and shared variants in low GC as well as high GC
617 regions.

618

619 ***Nuclear phylogeny reconstruction***

620 For phylogeny reconstruction, we divided the genome into multiple 200-kb windows,

621 and reconstructed window phylogeny with maximum likelihood method by using
622 IQ-TREE 2 (v 2.2.2.7, Minh et al. 2020). In model testing, we employed +ASC model to
623 include ascertainment bias correction as SNP data did not contain constant sites
624 (Minh et al. 2020). We also appended -bnni to the regular UFBoot command to
625 reduce the risk of overestimating branch support according to Hoang et al. (2018).
626 Since genomic regions with high recombination rate and low recombination rate
627 showed differentiated efficiency in tree reconstruction (e.g., Pease and Hahn 2013),
628 and our analysis showed distinct pattern of recombination and mutation rates in high
629 GC and low GC regions (Figure 3, Supplementary Figure 4, 5), we divided genomic
630 windows into high GC (above genome-wide average GC content) and low GC (below
631 genome-wide average GC content) categories, and generated the species tree in
632 ASTRAL III (v 5.7.1, Zhang et al. 2018) based on window trees in each category.
633

Acknowledgments

We would like to thank the Information Technology Services in HKU and in NUS for offering research computing facilities in the computations of this project. We acknowledge the assistance of AI-Know (https://ai-know.nus.edu.sg/ai-chat) and DeepSeek (https://chat.deepseek.com/) in writing scripts used during the analysis of this research. We also thank Dr. Per Sjödin's help in divergence time calculation with TTo method.

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. Shell scripts were attached in supplementary files.

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.

Barroso, G.V. and Dutheil, J.Y., 2023. The landscape of nucleotide diversity in Drosophila melanogaster is shaped by mutation rate variation. *Peer Community Journal*, *3*.

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.

Beye M., Gattermeier I., Hasselmann M., Gempe T., Schioett M., Baines J.F., Schlipalius D., Mougel F., Emore C., Rueppell O. and Sirviö A., 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome research*, *16*(11), pp.1339-1344.

Bolger A.M., Lohse M. and Usadel B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15): 2114-2120.

Browning B. L., Tian X., Zhou Y., and Browning S.R., 2021. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet 108(10):1880-1890.

Boman J., Mugal C.F., Backström N., 2021. The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes. *Genome Biology and Evolution*., 13(5): evab064.

Booker T. R., Yeaman S., & Whitlock M. C., 2020. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Molecular Ecology*, 30, 1– 25.

Bouwman, B.A., Crosetto, N. and Bienko, M., 2023. A GC-centered view of 3D genome organization. *Current Opinion in Genetics & Development*, *78*, 102020.

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.

693 Burri R., 2021. Evolution: small populations, low recombination, big trouble? *Current*
694 *Biology*, 31(6): R282-R284.
695 Castellano D., Eyre-Walker A. and Munch K., 2020. Impact of mutation rate and
696 selection at linked sites on DNA variation across the genomes of humans
697 and other Homininae. *Genome Biology and Evolution*, 12(1): 3550-3561.
698 Charlesworth B., Morgan M.T. and Charlesworth D., 1993. The effect of deleterious
699 mutations on neutral molecular variation. *Genetics*, 134(4): 1289-303.
700 Charlesworth, D., Zhang, Y., Bergero, R., Graham, C., Gardner, J. and Yong, L., 2020.
701 Using GC content to compare recombination patterns on the sex
702 chromosomes and autosomes of the guppy, *Poecilia reticulata*, and its
703 close outgroup species. *Molecular Biology and Evolution*, *37*(12):
704 3550-3562.
705 Chen C., Wang H., Liu Z., Chen X., Tang J., Meng F. and Shi W., 2018. Population
706 genomics provide insights into the evolution and adaptation of the eastern
707 honey bee (*Apis cerana*). *Molecular Biology Evolution*, 35: 2260-2270.
708 Cingolani P., 2012. Variant annotation and functional prediction: SnpEff. In *Variant*
709 *calling: methods and protocols*: 289-314. New York, NY: Springer US.
710 Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson,
711 J.E., Rea, T.J., Muzny, D.M., Lewis, L.R. and Wheeler, D.A., 2010. Deep
712 resequencing reveals excess rare recent variants consistent with explosive
713 population growth. *Nature communications*, 1(1), 131.
714 Cruickshank T.E. and Hahn M.W., 2014. Reanalysis suggests that genomic islands of
715 speciation are due to reduced diversity, not reduced gene flow. *Molecular*
716 *Ecology*, 23(13): 3133-3157.
717 Cuevas A., Eroukhmanoff F., Ravinet M., Sætre G.P. and Runemark A., 2022.
718 Predictors of genomic differentiation within a hybrid taxon. *PLoS Genetics*,
719 18(2): e1010027.
720 Cutter A.D. and Choi J.Y., 2010. Natural selection shapes nucleotide polymorphism
721 across the genome of the nematode *Caenorhabditis briggsae*. *Genome*
722 *Research*, 20(8): 1103-1111.
723 Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker
724 R.E., Lunter G., Marth G.T., Sherry S.T. and McVean G., and 1000 Genomes
725 Project Analysis Group., 2011. The variant call format and VCFtools.
726 *Bioinformatics*, 27(15): 2156-2158.
727 Duret, L. and Galtier, N., 2009. Biased gene conversion and the evolution of
728 mammalian genomic landscapes. *Annual review of genomics and human*
729 *genetics*, *10*(1), pp.285-311.
730 Feder J.L. and Nosil P., 2009. Chromosomal inversions and species differences: when
731 are genes affecting adaptive divergence and reproductive isolation
732 expected to reside within inversions?. *Evolution*, 63(12): 3061-3075.
733 Feder J.L., Egan S.P. and Nosil P., 2012. The genomics of speciation-with-gene-flow.
734 *Trends in Genetics*, 28(7): 342-350.

735 Fryxell K.J. and Moon W.J., 2005. CpG mutation rates in the human genome are
736 highly dependent on local GC content. *Molecular Biology and Evolution*,
737 22(3): 650-658.
738 Fullerton S.M., Bernardo Carvalho A. and Clark A.G., 2001. Local rates of
739 recombination are positively correlated with GC content in the human
740 genome. *Molecular Biology and Evolution*, 18(6): 1139-1142.
741 Gao F., Ming C., Hu W., and Li H., 2016. New software for the fast estimation of
742 population recombination rates (FastEPRR) in the genomic era. *G3: Genes|*
743 *genomes| genetics*, 6: 1563-1571.
744 Guerrero R. F., and Hahn M.W., 2017. Speciation as a sieve for ancestral
745 polymorphism. *Molecular Ecology*, 26: 5362-5368.
746 Han F., Lamichhaney S., Grant B. R., Grant P. R., Andersson L., and Webster M.T., 2017.
747 Gene flow, ancient polymorphism, and ecological adaptation shape the
748 genomic landscape of divergence among Darwin's finches. *Genome*
749 *Research*, 27: 1004-1015.
750 Harpur B.A., Kent C.F., Molodtsova D., Lebon J.M., Alqarni A.S., Owayss A.A. and
751 Zayed A., 2014. Population genomics of the honey bee reveals strong
752 signatures of positive selection on worker traits. *Proceedings of the*
753 *National Academy of Sciences*, *111*(7): 2614-2619.
754 Hershberg, R. and Petrov, D.A., 2010. Evidence that mutation is universally biased
755 towards AT in bacteria. *PLoS genetics*, *6*(9), p.e1001115.
756 Hill, W.G. and Robertson, A., 1966. The effect of linkage on limits to artificial
757 selection. *Genetics Research*, *8*(3), pp.269-294.
758 Hoang D.T., Chernomor O., Von Haeseler A., Minh B.Q. and Vinh L.S., 2018. UFBoot2:
759 improving the ultrafast bootstrap approximation. *Molecular Biology and*
760 *Evolution*, 35(2): 518-522.
761 Irwin D.E., Milá B., Toews D.P., Brelsford A., Kenyon H.L., Porter A.N., Grossen C.,
762 Delmore K.E., Alcaide M. and Irwin J.H., 2018. A comparison of genomic
763 islands of differentiation across three young avian species pairs. *Molecular*
764 *Ecology*, 27(23): 4839-4855.
765 Ji Y., Li X., Ji T., Tang J., Qiu L., Hu J., Dong J., Luo S., Liu S., Frandsen P.B. and Zhou X.,
766 2020. Gene reuse facilitates rapid radiation and independent adaptation to
767 diverse habitats in the Asian honeybee. *Science Advances*, 6(51):
768 eabd3590.
769 Jiang Z., Song G., Luo X., Zhang D., Lei F. and Qu Y., 2023. Recurrent selection and
770 reduction in recombination shape the genomic landscape of divergence
771 across multiple population pairs of Green-backed Tit. *Evolution Letters*, 7(2):
772 99-111.
773 Keinan, A. and Clark, A.G., 2012. Recent explosive human population growth has
774 resulted in an excess of rare genetic variants. *science*, *336*(6082), 740-743.
775 Kiktev D.A., Sheng Z., Lobachev K.S. and Petes T.D., 2018. GC content elevates
776 mutation and recombination rates in the yeast Saccharomyces cerevisiae.
777 *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.

Long, H., Sung, W., Kucukyildirim, S. *et al.* 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution* 2, 237–240.

Lynch, M., 2007 The Origins of Genome Architecture, Ed. 1. Sinauer Associates, Sunderland, MA.

MacPherson, B., Scott, R., & Gras, R. 2021. Sex and recombination purge the genome of deleterious alleles: An Individual Based Modeling Approach. *Ecological Complexity*, *45*, 100910.

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.

Petes, T. 2001. Meiotic recombination hot spots and cold spots. *Nat Rev Genet* 2, 360–369.

821 Qiu L., Dong J., Li X., Parey S.H., Tan K., Orr M., Majeed A., Zhang X., Luo S., Zhou X.
822 and Zhu C., 2023. Defining honeybee subspecies in an evolutionary context
823 warrants strategized conservation. *Zoological Research*, 44(3): 483.
824 Radloff S.E., Hepburn C., Hepburn H.R., Fuchs S., Hadisoesilo S., Tan K., Engel M.S.,
825 Kuznetsov V., 2010. Population structure and classification of *Apis cerana*.
826 *Apidologie*, 41: 589-601.
827 Ravinet M., Faria R., Butlin R.K., Galindo J., Bierne N., Rafajlović M., Noor M.A.F.,
828 Mehlig B. and Westram A.M., 2017. Interpreting the genomic landscape of
829 speciation: a road map for finding barriers to gene flow. *Journal of*
830 *Evolutionary Biology*, 30(8): 1450-1477.
831 Renaut S., Owens G.L. and Rieseberg L.H., 2014. Shared selective pressure and local
832 genomic landscape lead to repeatable patterns of genomic divergence in
833 sunflowers. *Molecular Ecology*, 23(2): 311-324.
834 Roze D., 2021. A simple expression for the strength of selection on recombination
835 generated by interference among mutations. *Proceedings of the National*
836 *Academy of Sciences*, 118(19): p.e2022805118.
837 Southcott L., Kronforst M.R., 2017. A neutral view of the evolving genomic
838 architecture of speciation. *Ecology and Evolution*, 7(16):6358-6366.
839 Shang H., Field D.L., Paun O., Rendón-Anaya M., Hess J., Vogl C., Liu J., Ingvarsson P.K.,
840 Lexer C. and Leroy T., 2023. Drivers of genomic landscapes of
841 differentiation across a Populus divergence gradient. *Molecular Ecology*,
842 32(15): 4348-4361.
843 Sjödin, P., McKenna, J. and Jakobsson, M., 2021. Estimating divergence times from
844 DNA sequences. *Genetics*, *217*(4), p.iyab008.
845 Su Y.C., Chiu Y.F., Warrit N., Otis G.W. and Smith D.R., 2023. Phylogeography and
846 species delimitation of the Asian cavity-nesting honeybees. *Insect*
847 *Systematics and Diversity*, 7(4): 5.
848 Turner T. L., Hahn, M. W., and Nuzhdin S. V., 2005. Genomic islands of speciation in
849 *Anopheles gambiae*. *PLoS Biology*, 3: e285.
850 Turner T.L. and Hahn M.W., 2010. Genomic islands of speciation or genomic islands
851 and speciation?. *Molecular Ecology*, 19(5): 848-850.
852 Vijay N., Bossu C.M., Poelstra J.W., Weissensteiner M.H., Suh A., Kryukov A.P., and
853 Wolf J.B., 2016. Evolution of heterogeneous genome differentiation across
854 multiple contact zones in a crow species complex. *Nature Communications*,
855 7: 13195.
856 Wakamiya T., Kamioka T., Ishii Y., Takahashi J.I., Maeda T., and Kawata M., 2023.
857 Genetic differentiation and local adaptation of the Japanese honeybee,
858 *Apis cerana* japonica. *Ecology and Evolution*, 13(10): e10573.
859 Wallberg, A., Glémin, S. and Webster, M.T., 2015. Extreme recombination frequencies
860 shape genome variation and evolution in the honeybee, Apis
861 mellifera. *PLoS Genetics*, *11*(4), p.e1005189.
862 Wallberg A., Han F., Wellhagen G., Dahle B., Kawata M., Haddad N., Simões Z.L.,
863 Allsopp M.H., Kandemir I., De la Rúa P., Pirk C.W., Webster M.T., 2014. A
864 worldwide survey of genome sequence variation provides insight into the

evolutionary history of the honeybee Apis mellifera. *Nature Genetics*, 46(10): 1081-8.

Wang, B., Mojica, J.P., Perera, N., Lee, C.R., Lovell, J.T., Sharma, A., Adam, C., Lipzen, A., Barry, K., Rokhsar, D.S. and Schmutz, J., 2019. Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in *Boechera stricta*. *Genome biology*, 20: 1-15.

Wang J., Street N. R., Scofield, D. G., and Ingvarsson P. K., 2016. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Molecular Biology and Evolution*, 33: 1754-1767.

Wei T. and Simko V., 2021. R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92), https://github.com/taiyun/corrplot.

Wolf J.B.W. and Ellegren H., 2017. Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18: 87-100.

Wong E.L. and Filatov D.A., 2023. The role of recombination landscape in species hybridisation and speciation. *Frontiers in Plant Science*, 14: 1223148.

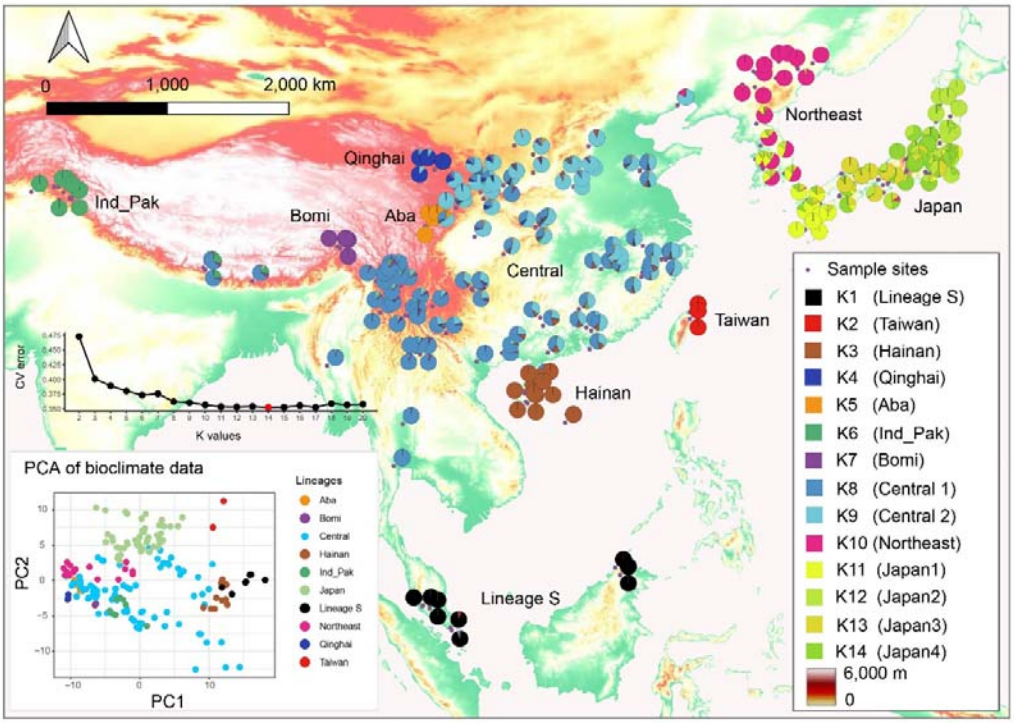
Yi X., Liang Y. U., Huerta- Sanchez E., Jin X., Cuo Z. X. P., Pool J. E., Xu X., Jiang H., Vinckenbosch N., Korneliussen T. S., Zheng H., Liu T., He W., Li K., Luo R., Nie X., Wu H., Zhao M., Cao H., … Wang J., 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329, 75– 78.

Zhang C., Rabiee M., Sayyari E. and Mirarab S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, 19: 15-30.

Zhou W., Zhang N., Huang K., Lin H., Tu J., Zheng C., Que P., Chiang C.Y., Martinez J., Naerhulan H. and Székely T., 2024. Divergent selection in low recombination regions shapes the genomic islands in two incipient shorebird species. *Molecular Biology and Evolution*, 41(2): msae006.

892 **Figures**

893 **Figure 1**



894

895 **Figure 1.** Genetic structure and bioclimate variable clustering of *Apis cerana* Lineage
896 M (mainland lineage). Additional samples from Malaysia and Singapore (i.e.,
897 Sundaland lineage (Lineage S); Su et al. 2023) were included. The optimal K=14
898 was selected based on cross-validation error in ADMIXTURE. Two and four
899 genetic clusters were identified from Central and Japan groups, respectively. The
900 144 monthly bioclimate variables can be found in Supplementary Table 1.

901

Figure 2

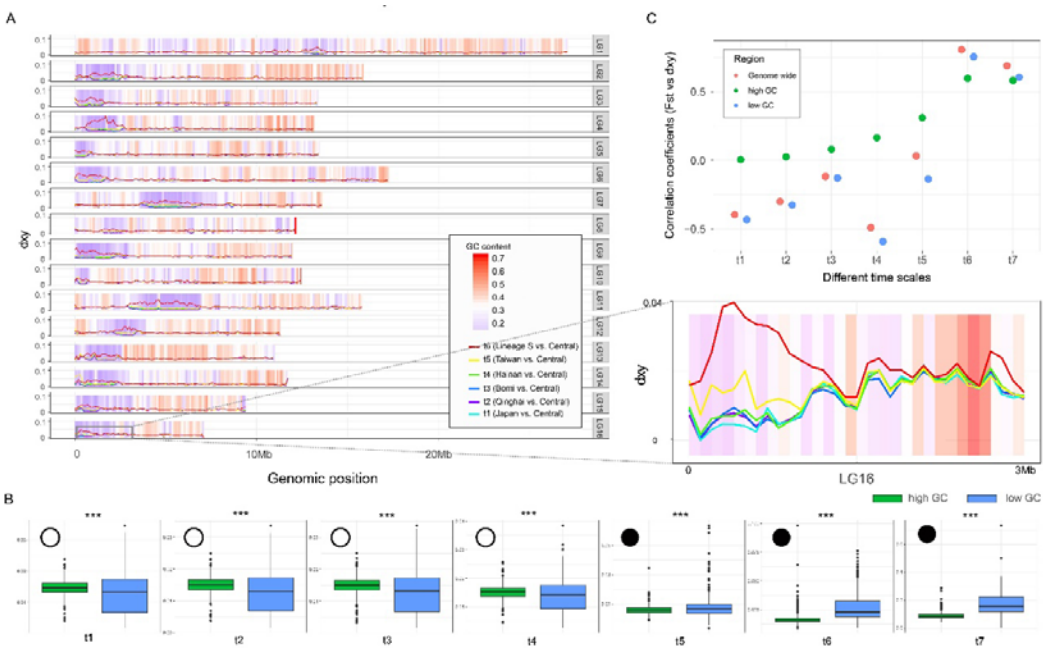
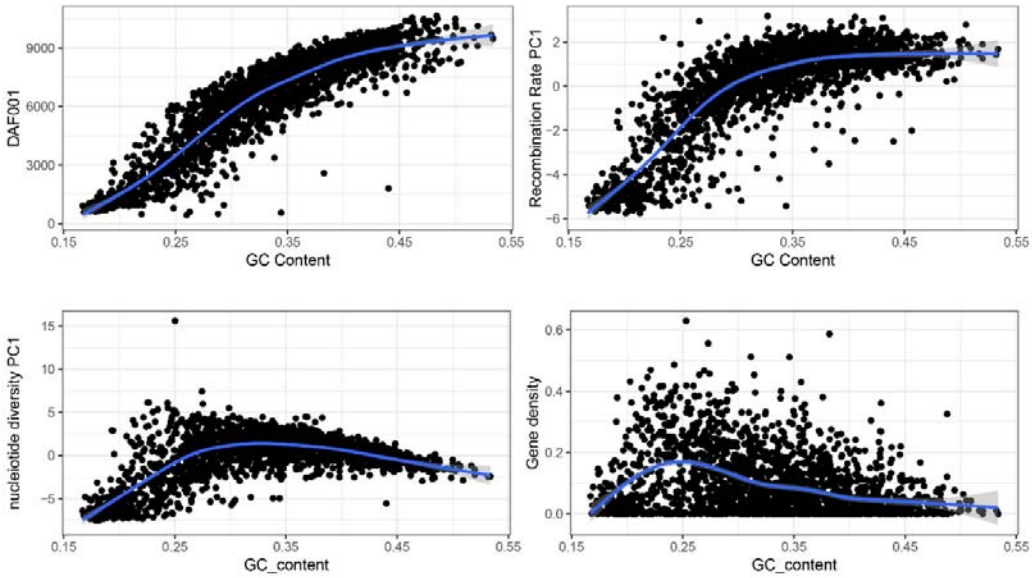


Figure 2. Characterization of heterozygous divergence (d_{xy}) landscape in high GC and low GC regions. **(A)** Landscape of d_{xy} over time (t1 to t6) varied along GC gradient. Landscape of d_{xy} from t1 to t7 in Supplementary Figure 1; **(B)** Patterns of d_{xy} in low GC and high GC regions shift over divergence time (i.e., progressing from t1 through t5 (population genetic timescale) to t6 and t7 (phylogenetic timescale); 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 d_{xy} and genetic differentiation (F_{ST}) varied over divergence time in different genomic regions.

916 **Figure 3**

917



918

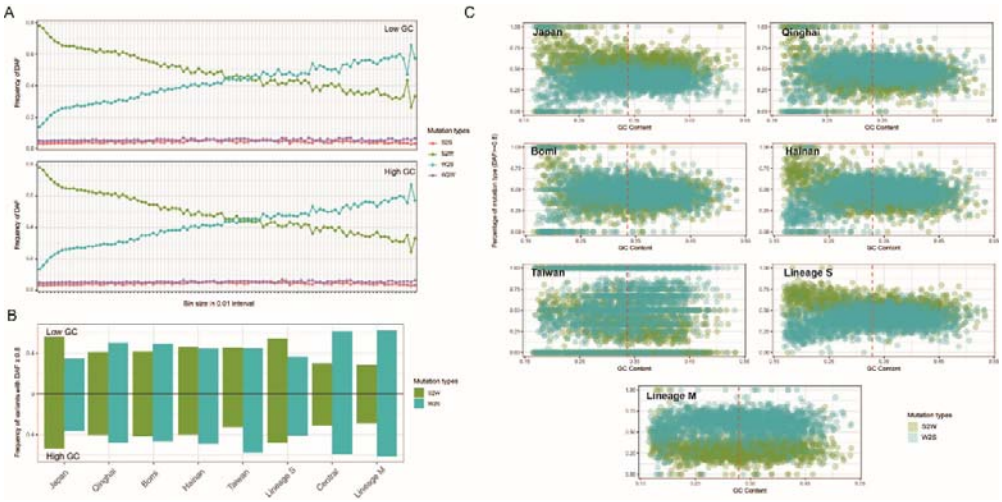
919 **Figure 3.** Correlation of GC content with different population genetic parameters

920 (DAF_001, nucleotide diversity PC1, recombination rate PC1 and gene density)

921 of *A. cerana* Lineage M. DAF_001 represents derived allele frequency (DAF) ≤
922 0.01. Nucleotide diversity PC1 (explained 86.02% variance) and recombination
923 rate PC1 (explained 74.91% variance) are the first principal component of
924 nucleotide diversity and recombination rate of six *A. cerana* Lineage M groups
925 (Japan, Qinghai, Bomi, Hainan, Taiwan and Central), respectively.

926

927 **Figure 4**



928

929 **Figure 4.** Variation of derived allele frequency (DAF) in low and high GC regions of
930 different *A. cerana* groups. **(A)** The relative site frequency spectrum of four
931 mutation types (bin size: 0.01, from 0-0.01 to 0.99-1) in *A. cerana* Lineage M
932 inferred with *A. mellifera* as the outgroup. S2S: strong to strong (from C/G in *A.*
933 *mellifera* to C/G in *A. cerana*), S2W: strong to weak (from C/G to A/T), W2S: weak
934 to strong (from A/T to C/G), and W2W: weak to weak (from A/T to A/T). **(B)**
935 Percentage differentiation of nearly fixed S2W and W2S alleles (DAF ≥ 0.8) in low
936 and high GC regions of different *A. cerana* groups. Japan, Qinghai, Bomi, Hainan,
937 Taiwan, Lineage S have higher percentage of S2W (C/G to A/T) type than Central
938 and Lineage M, especially in low GC regions. **(C)** Percentage of nearly fixed S2W
939 and W2S alleles in different *A. cerana* groups along the GC gradient. The red dash
940 line represents the mean GC content.

941

Figure 5

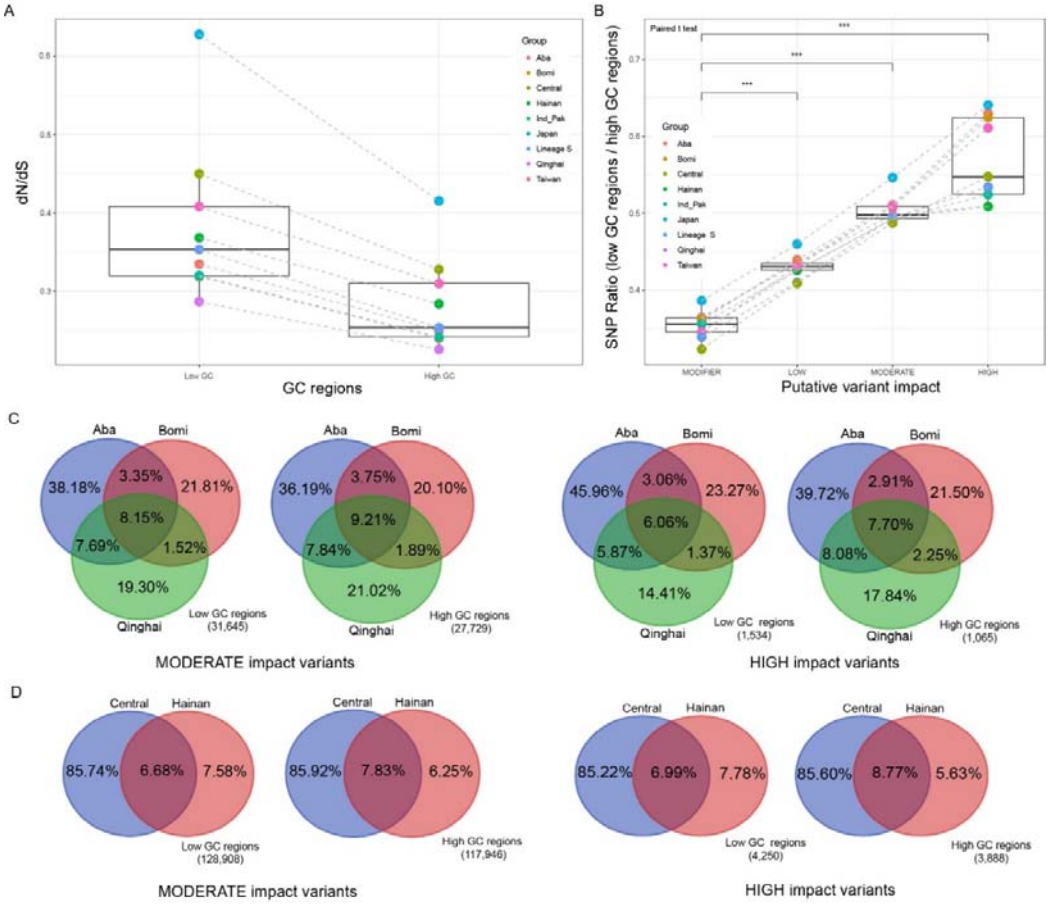
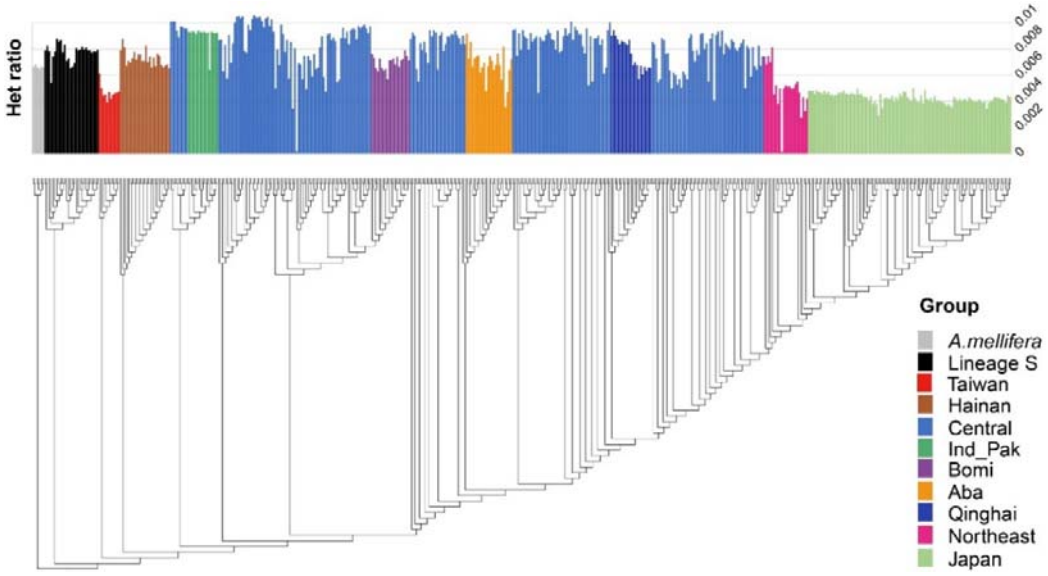


Figure 5. Profile of functional variants in high GC and low GC regions. **(A)** Higher dN/dS (ratio of nonsynonymous over synonymous mutations) in low GC regions than high GC regions (Paired t test, $p < 0.001$). **(B)** Higher ratio of putatively deleterious mutations than neutral mutants. Paired t 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)** Consistent pattern of more group-specific mutation load than the shared in both high GC and low GC regions of isolated groups. Aba, Bomi and Qinghai represent three independent colonization to Qinghai-Tibet Plateau and are with no admixture with each other (Figure 1). **(D)** Ratio of group-specific and shared mutation load differentiated in low GC and high GC regions in group pair with gene flow. Group pair Central vs. Hainan, as an example.

957 **Figure 6**



958

959 **Figure 6.** Reconstruction of *Apis cerana* phylogeny based on nuclear polymorphisms
960 in low GC regions. Bootstrap values and individual IDs could be referred to
961 phylogeny in Supplementary Figure 8; Het ratio indicates number of heterozygous
962 sites/number of homozygous sites in each individual.

963

964 **Table 1. Regression analysis of nucleotide diversity on genomic architectures.**

965

Regions	R^2	Adj.R^2	GC content	recombination	Gene density	DAF_001
Genome wide	0.6856	0.685	-2.341***	1.648***	-0.07695*	2.174***
Low-GC	0.7715	0.7706	-0.2402***	0.5548***	-0.0995***	0.5360***
High-GC	0.5543	0.5526	-0.9372***	0.0182(NA)	-0.0928***	0.5788***
region1	0.8465	0.8443	-0.0915*	0.4214***	-0.1700***	0.6128***
region2	0.644	0.6413	-0.2717***	0.3638***	-0.2496***	0.4904***
region3	0.4621	0.4591	-0.4429***	-0.1075**	-0.1495***	0.6985***
region4	0.4301	0.425	-0.5802***	0.1550***	-0.0091(NA)	0.4222***
region5	0.438	0.4054	-0.4521***	0.1446(NA)	0.1490(NA)	0.5408***

966

967 In addition to low-GC and high-GC regions, genomic windows (100 kb) were divided into five regions (region 1 -5,

968 from low to high GC) based on equal-width intervals of GC content. Nucleotide diversity (π) and

969 recombination (i.e., population-scaled recombination rate) are PC1s extracted from six *A. cerana* Lineage M

970 groups, respectively. DAF_001, as a proxy of mutation rate, is number of SNPs with rare and derived alleles

971 (daf <0.01) in Lineage M. DAF_001 $\propto \mu$ and recombination = 3 * Ne * c (Ne, effective population size; μ,

972 mutation rate; c, recombination rate). NA, *, **, *** indicate $p > 0.05$, $p < 0.05$, $p < 0.01$ and $p < 0.001$,

973 respectively.

Supplementary Table 1. Sample and sequencing information of *Apis cerana*.

Supplementary Table 2. Divergence time estimation between *A. cerana* Central group with other seven groups.

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

Supplementary Figure 2. Landscape of genomic variation (π) varied along with regional GC content.

Supplementary Figure 3. Landscape of genomic differentiation (F_{ST}) over time varied along with regional GC content. Central vs. Japan (t1), Central vs. Bomi (t3), and Central vs. Lineage S (t6).

Supplementary Figure 4. Correlation coefficients of genome-wide parameters over divergence time.

Supplementary Figure 5. Characterization of low GC and high GC regions over divergence time. **(A)** F_{ST}; **(B)** recombination (rho); **(C)** nucleotide diversity (π).

Supplementary Figure 6. Characterization of genomic islands and background regions over divergence time. **(A)** d_{xy}; **(B)** recombination (rho); **(C)** nucleotide diversity (π); **(D)** dN/dS; **(E)** GC content.

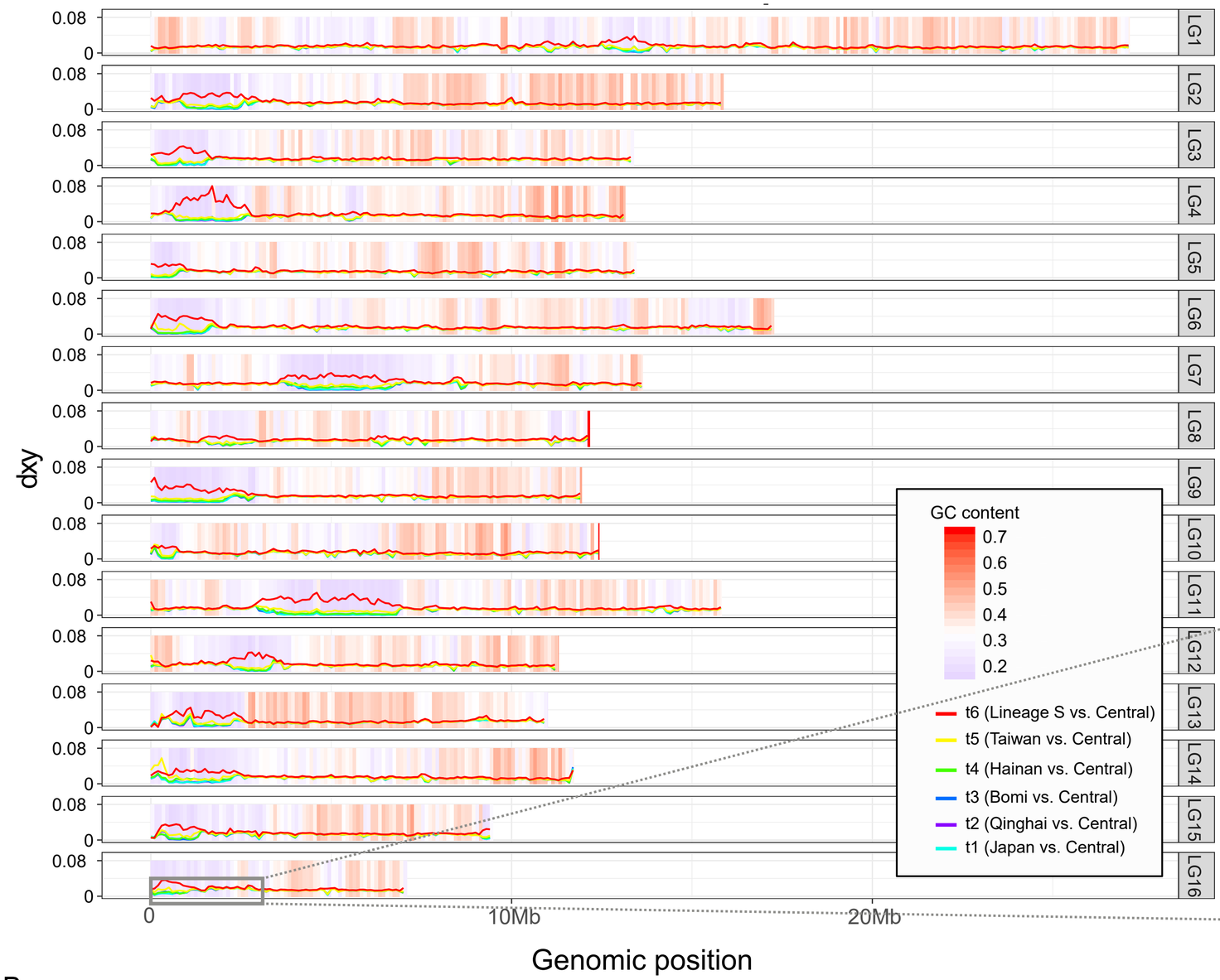
Supplementary Figure 7. Summary of mutation types in all derived alleles. Each panel represents loci within a specific range of derived allele frequency. The plot shows a high percentage of derived A/T alleles from ancestral C/G alleles (i.e. C/G_2_A/T), especially in loci with rare, derived alleles.

Supplementary Figure 8. Site frequency spectrum in different regions. Rare/common ratio of SNPs in the inset plot suggest more efficient selection in high GC regions than in low GC regions.

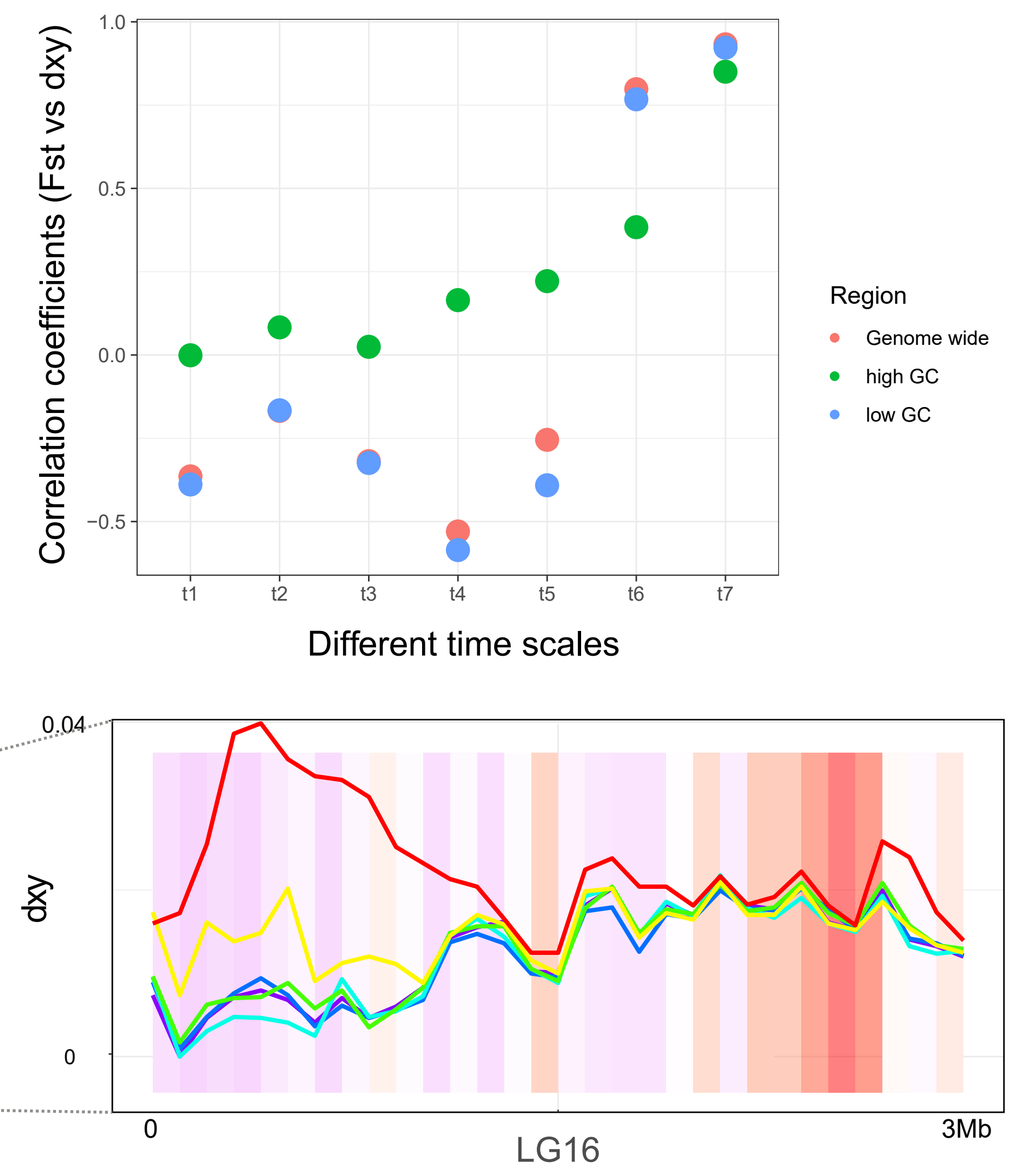
Supplementary Figure 9. Reconstruction of *Apis cerana* phylogeny based on nuclear polymorphisms in low GC regions. Bootstrap values over 0.7 were presented.

Supplementary Figure 10. Reconstruction of *Apis cerana* phylogeny based on nuclear polymorphisms in high GC regions. Het/Hom indicate ratio of homozygous vs. heterozygous sites in an individual. The hybrid individuals referred to genetic structure in Figure 1.

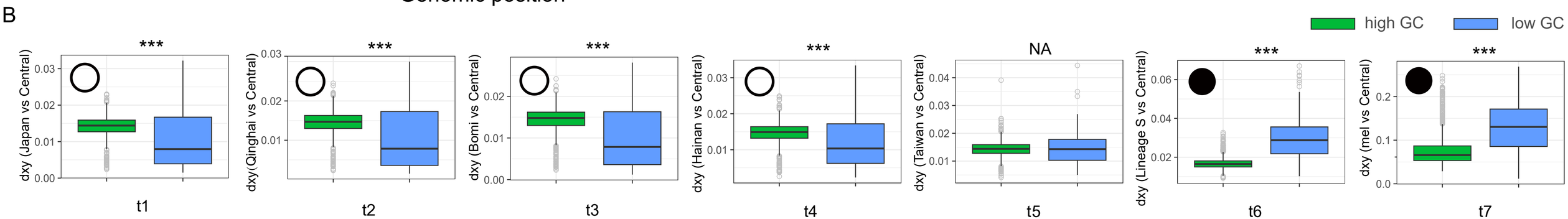
A

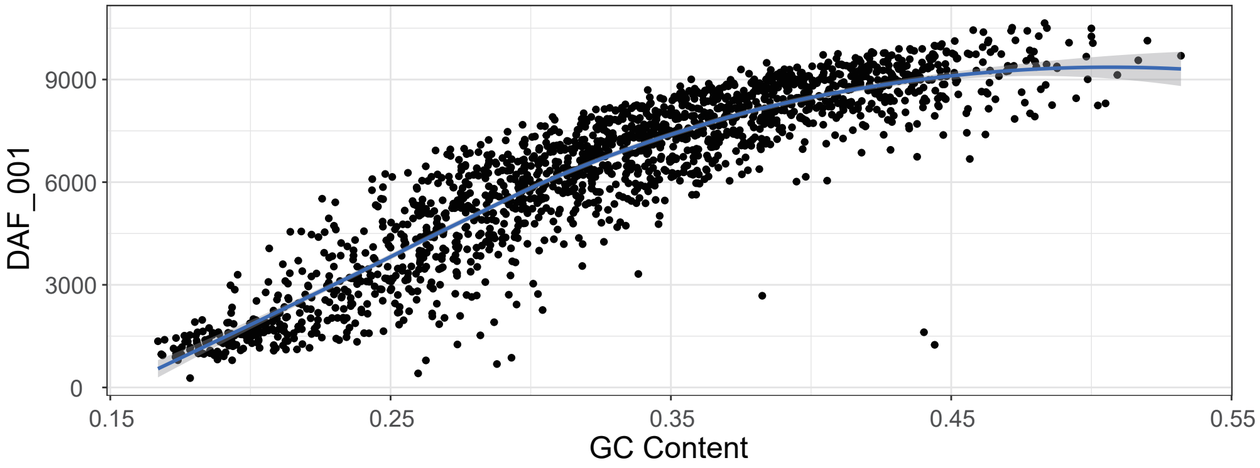

dxy
0.08
0
LG1
LG2
LG3
LG4
LG5
LG6
LG7
LG8
LG9
LG10
LG11
LG12
LG13
LG14
LG15
LG16
0
10Mb
20Mb
Genomic position
GC content
0.7
0.6
0.5
0.4
0.3
0.2
t6 (Lineage S vs. Central)
t5 (Taiwan vs. Central)
t4 (Hainan vs. Central)
t3 (Bomi vs. Central)
t2 (Qinghai vs. Central)
t1 (Japan vs. Central)

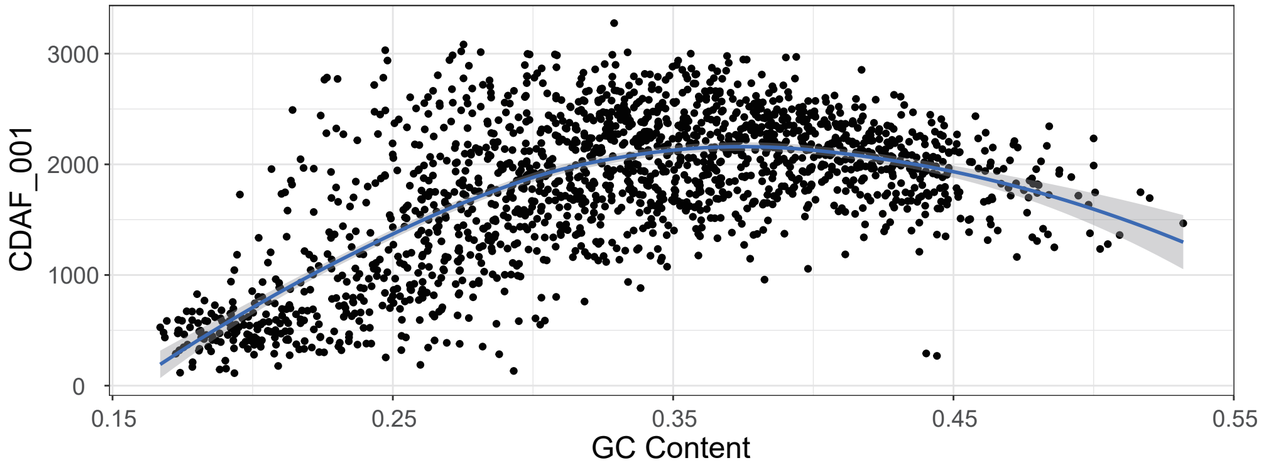
C

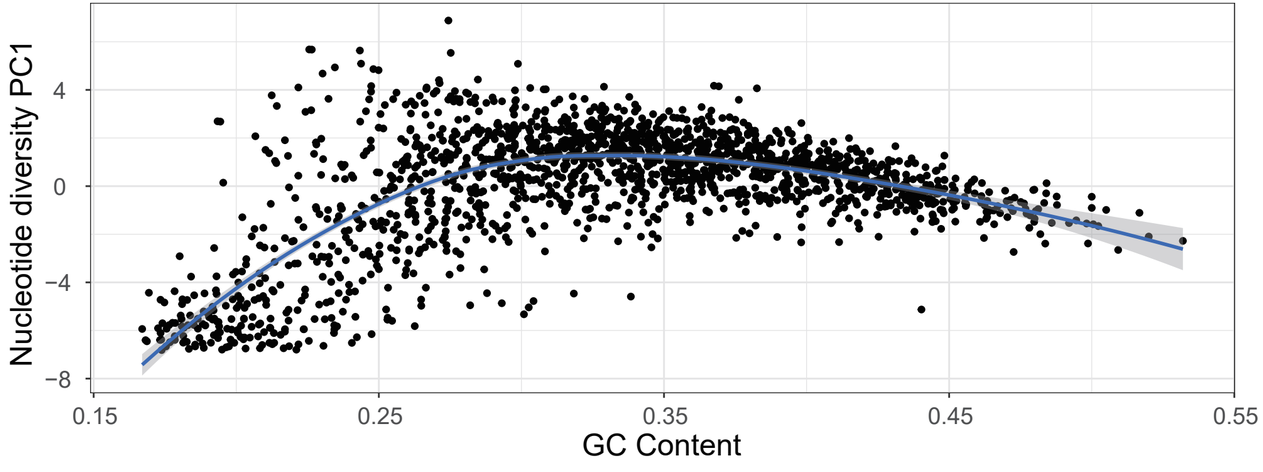

Correlation coefficients (Fst vs dxy)
1.0
0.5
0.0
−0.5
t1
t2
t3
t4
t5
t6
t7
Different time scales
Region
Genome wide
high GC
low GC
dxy
0.04
0
0
3Mb
LG16

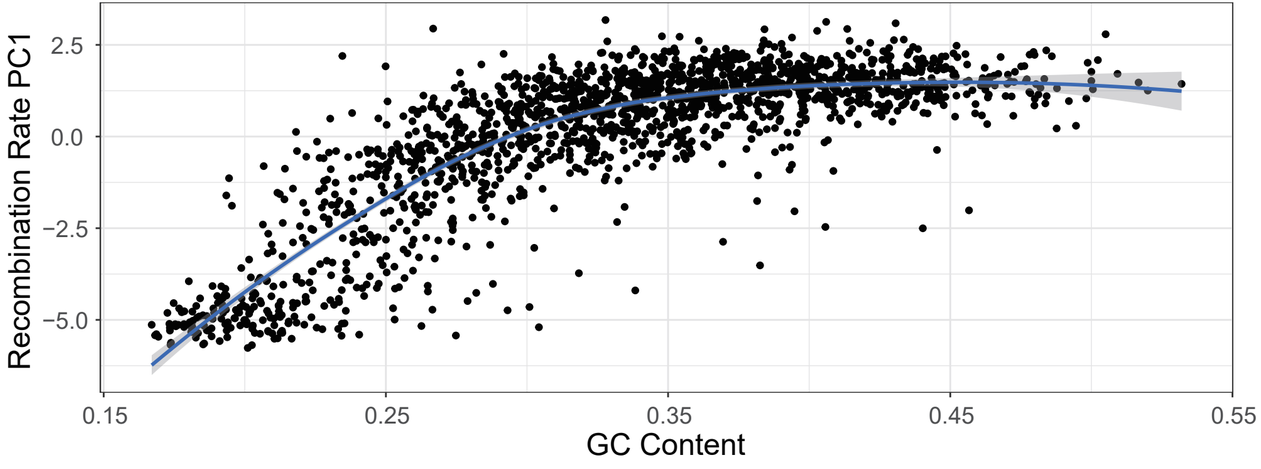
B

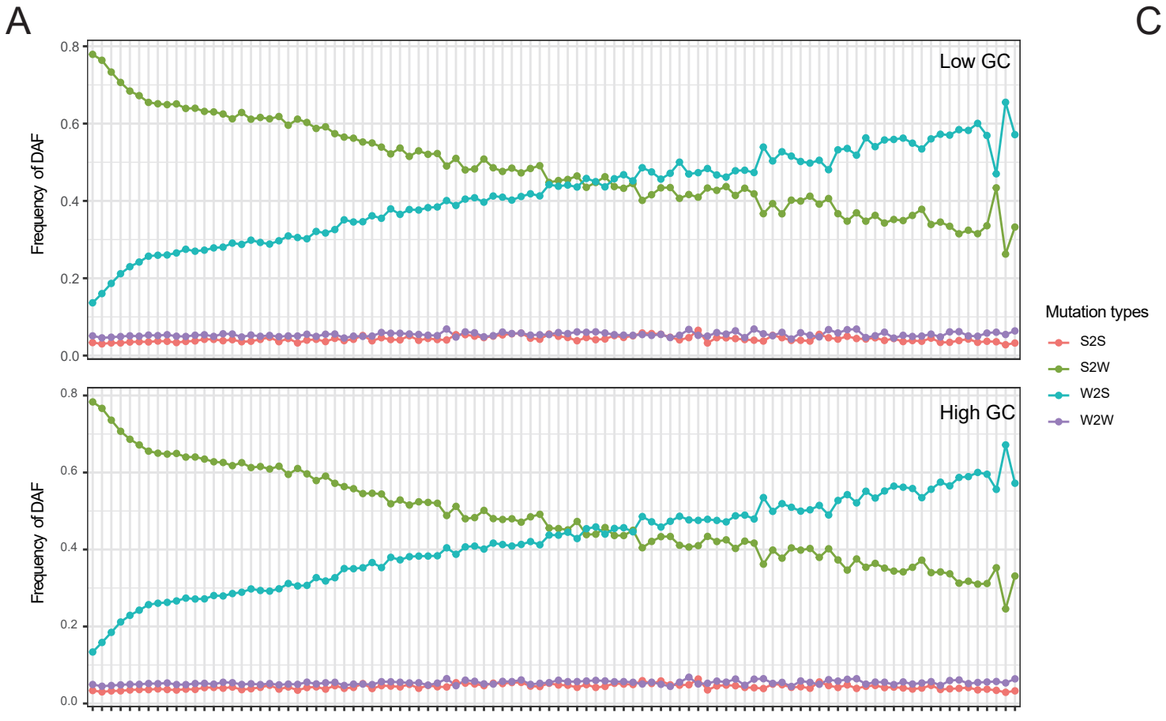

high GC
low GC
***
***
***
***
NA
***
***
dxy (Japan vs Central)
dxy(Qinghai vs Central)
dxy (Bomi vs Central)
dxy (Hainan vs Central)
dxy (Taiwan vs Central)
dxy (Lineage S vs Central)
dxy (mel vs Central)
t1
t2
t3
t4
t5
t6
t7

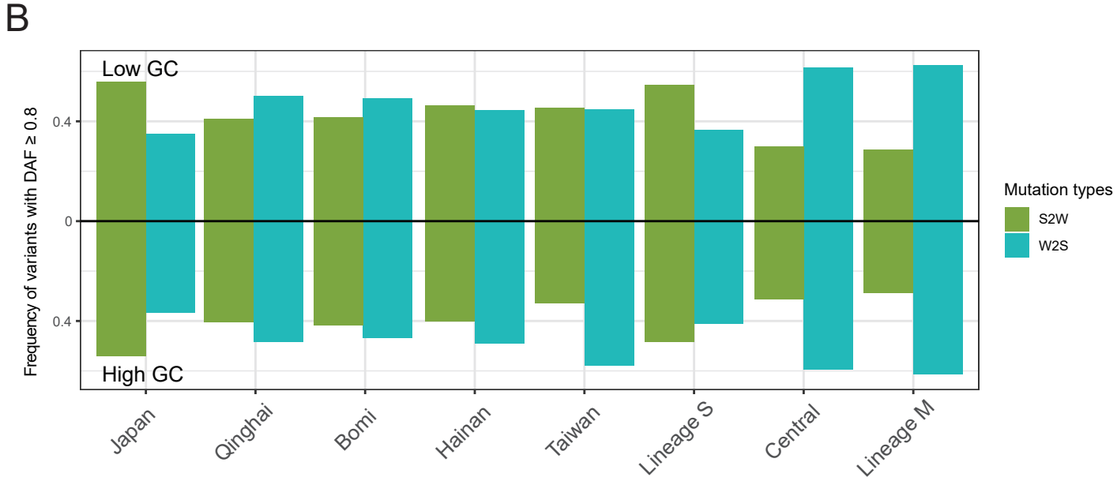

DAF_001
9000
6000
3000
0
0.15
0.25
0.35
0.45
0.55
GC Content

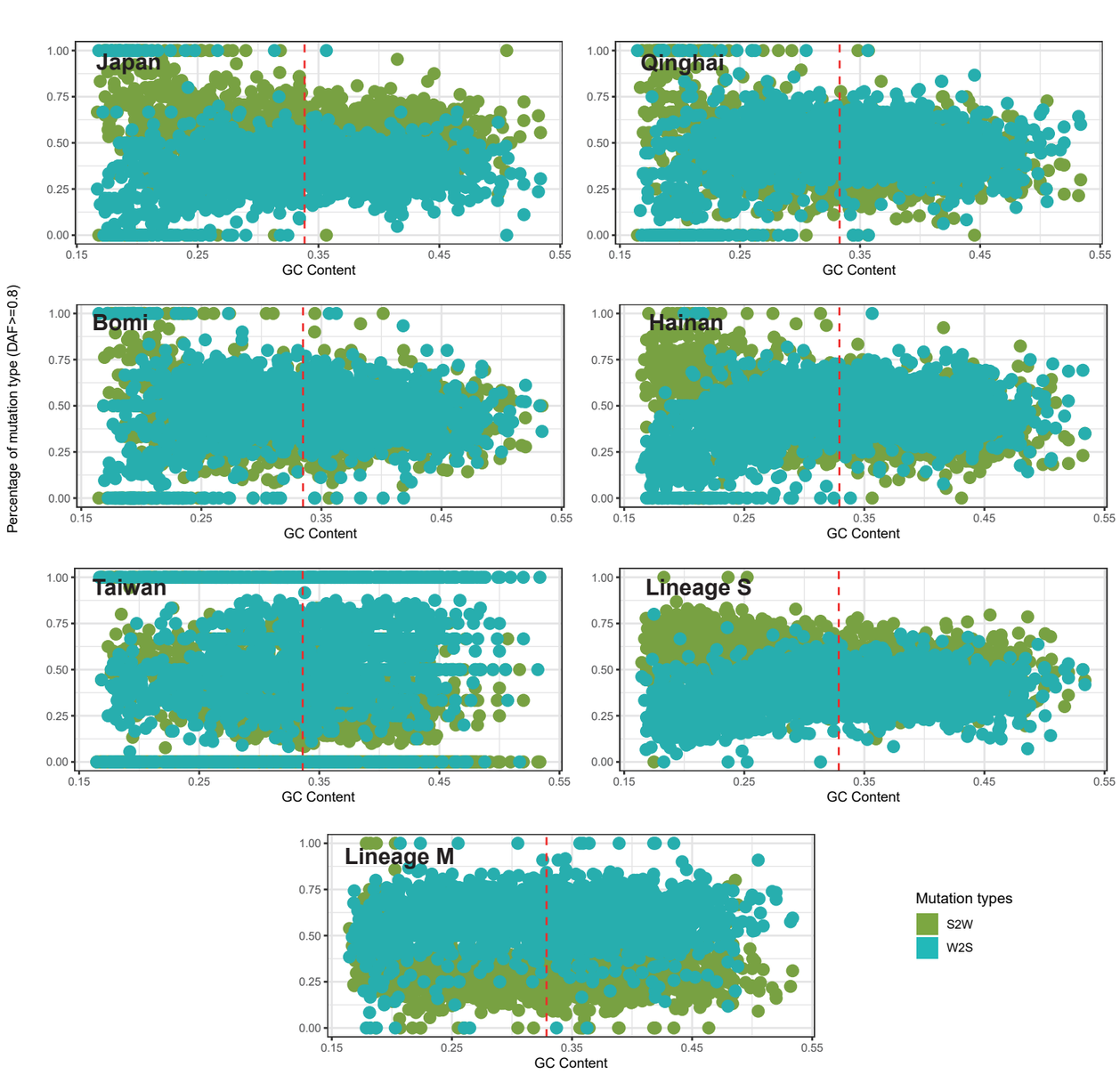

CDAF_001
3000
2000
1000
0
0.15
0.25
0.35
0.45
0.55
GC Content


Nucleotide diversity PC1
4
0
−4
−8
0.15
0.25
0.35
0.45
0.55
GC Content

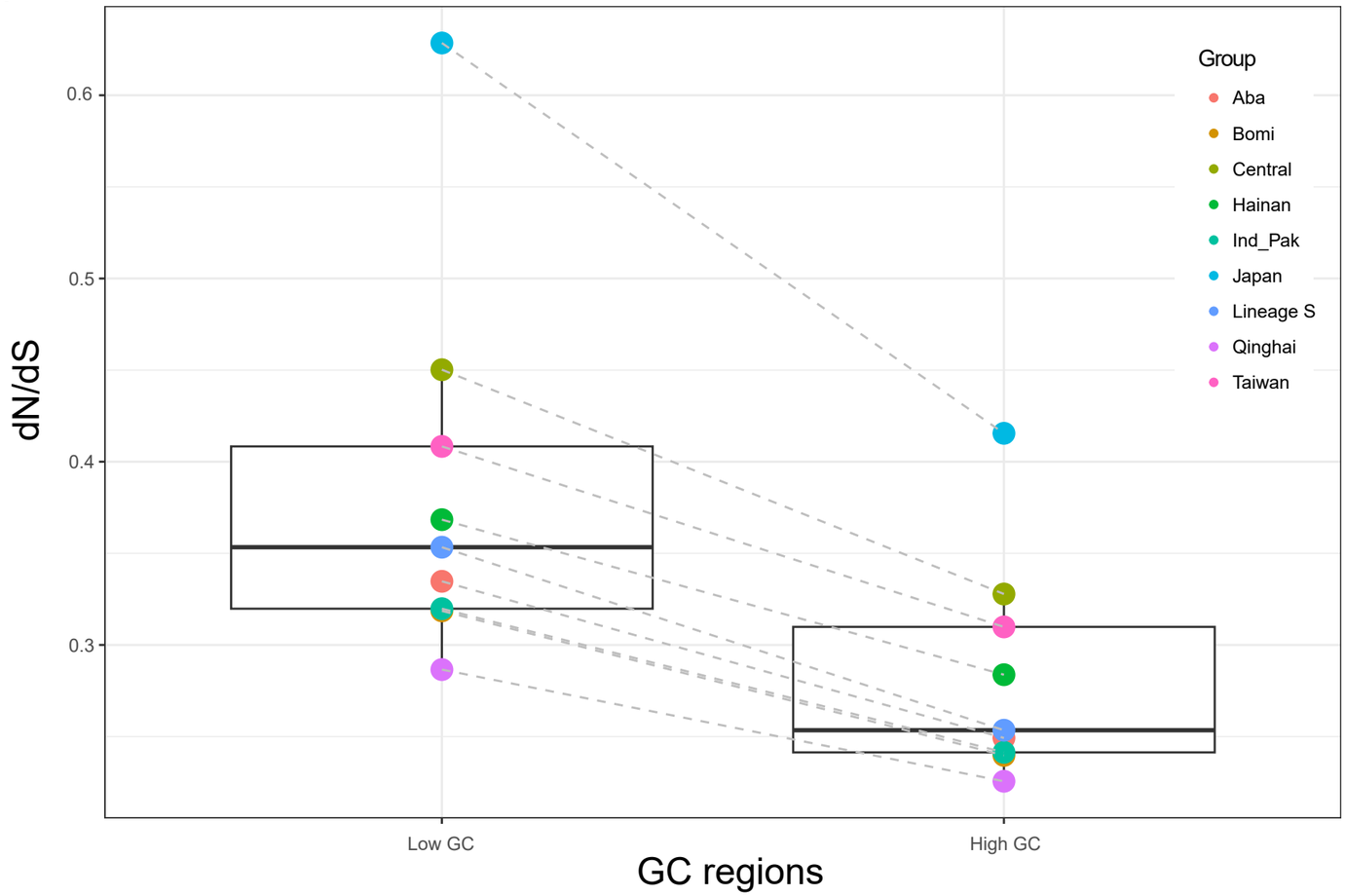

Recombination Rate PC1
2.5
0.0
−2.5
−5.0
0.15
0.25
0.35
0.45
0.55
GC Content


A
Low GC
High GC
Frequency of DAF
Mutation types
S2S
S2W
W2S
W2W

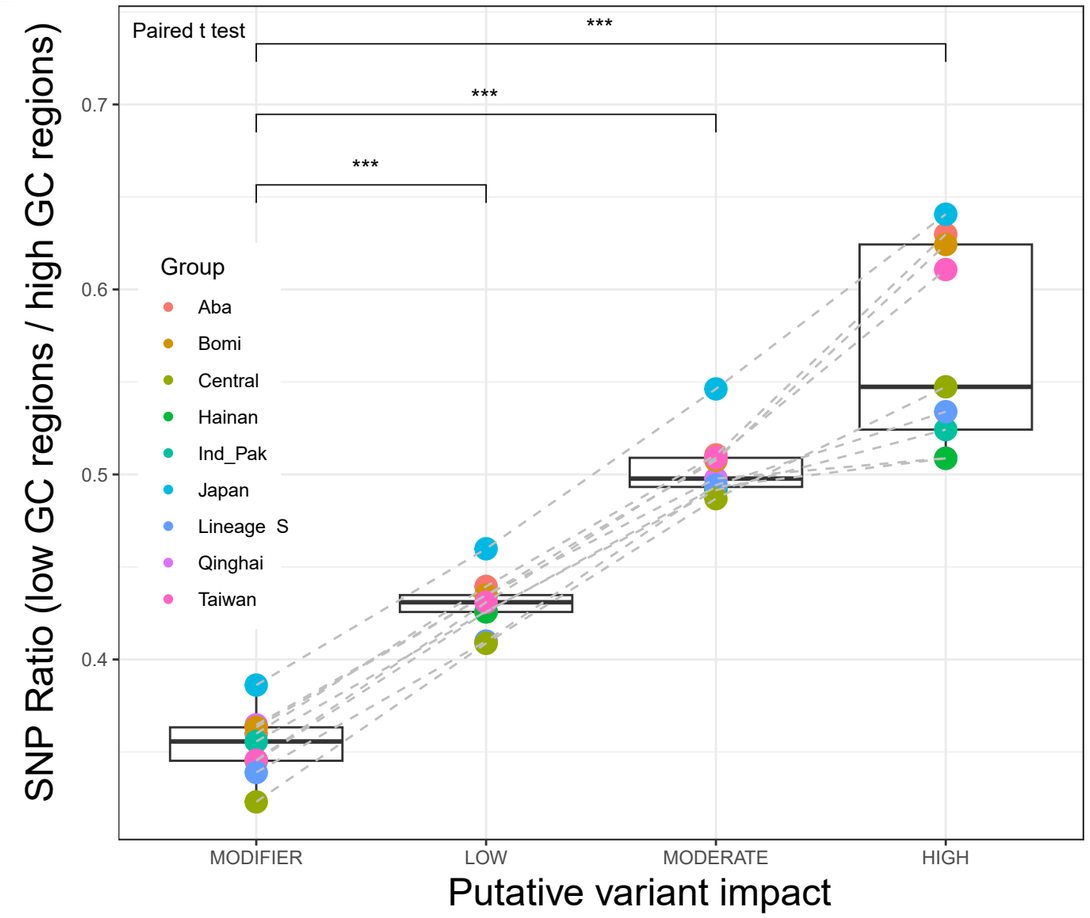

B
Low GC
High GC
Frequency of variants with DAF ≥ 0.8
Japan
Qinghai
Bomi
Hainan
Taiwan
Lineage S
Central
Lineage M
Mutation types
S2W
W2S


C
Japan
Qinghai
Bomi
Hainan
Taiwan
Lineage S
Lineage M
Percentage of mutation type (DAF>=0.8)
GC Content
Mutation types
S2W
W2S

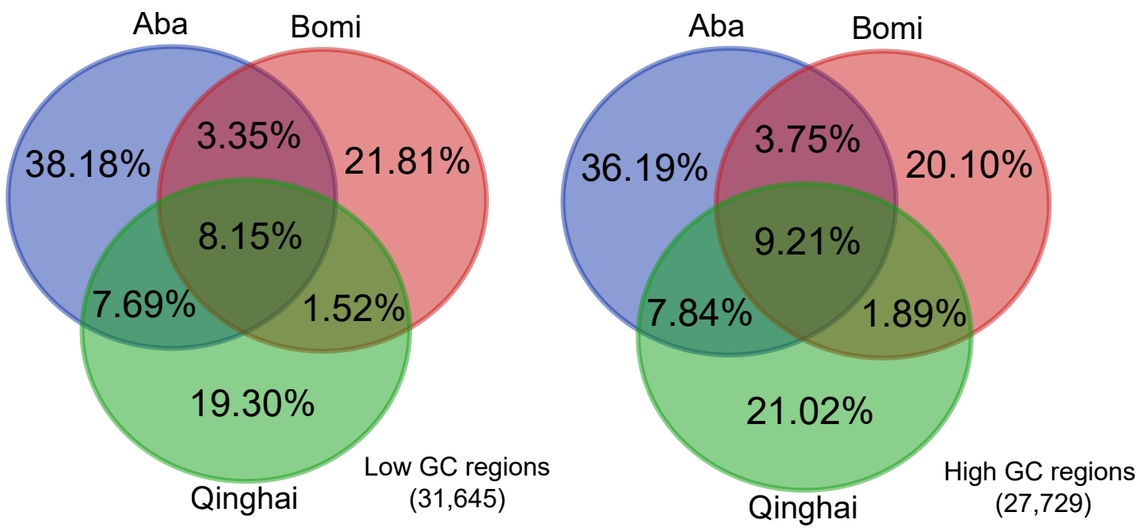
A



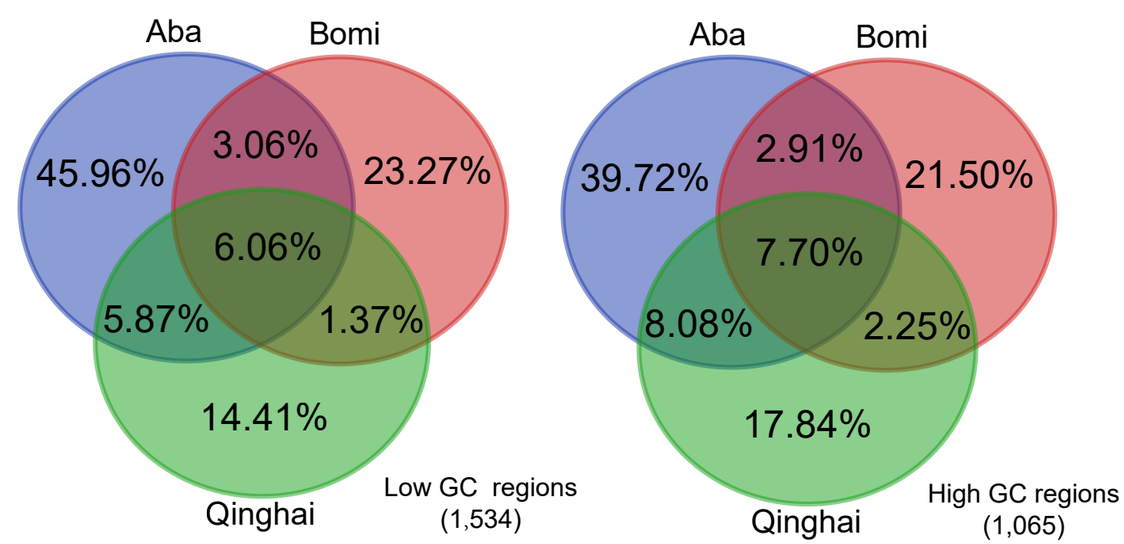
B



C

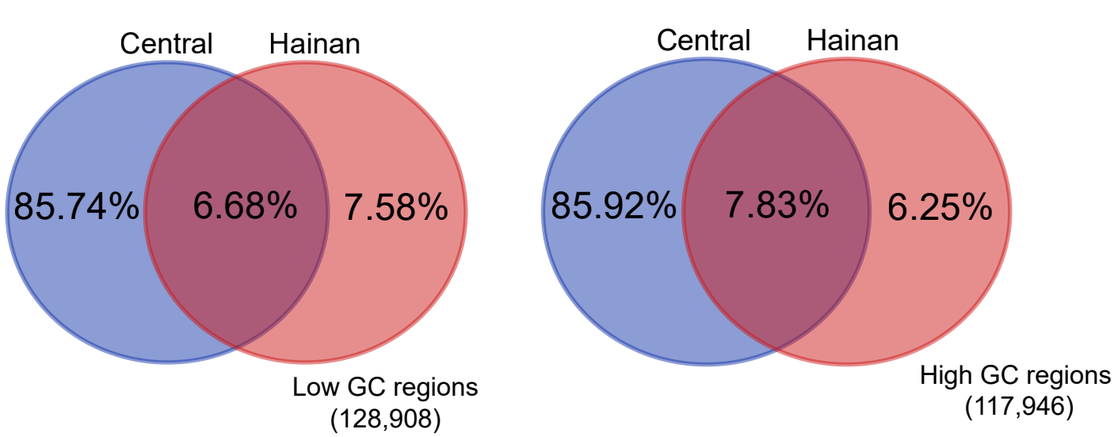


MODERATE impact variants

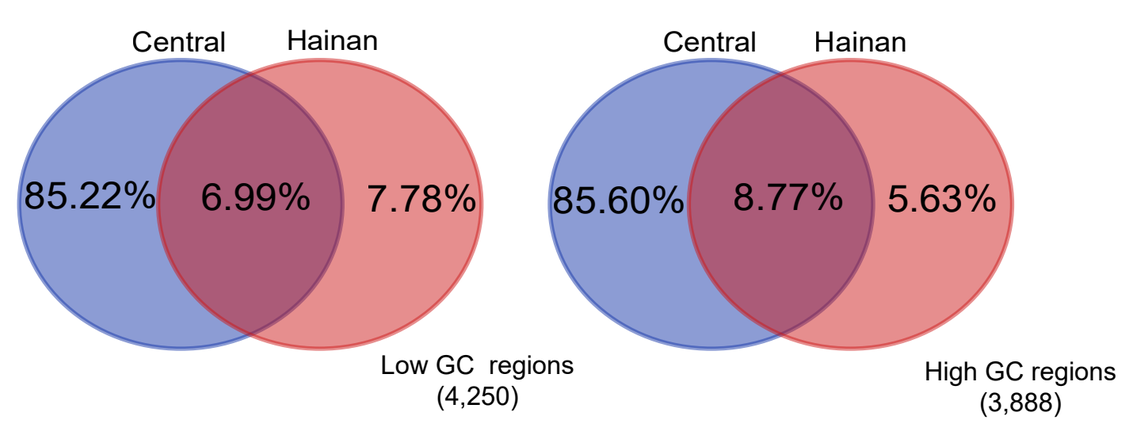


HIGH impact variants

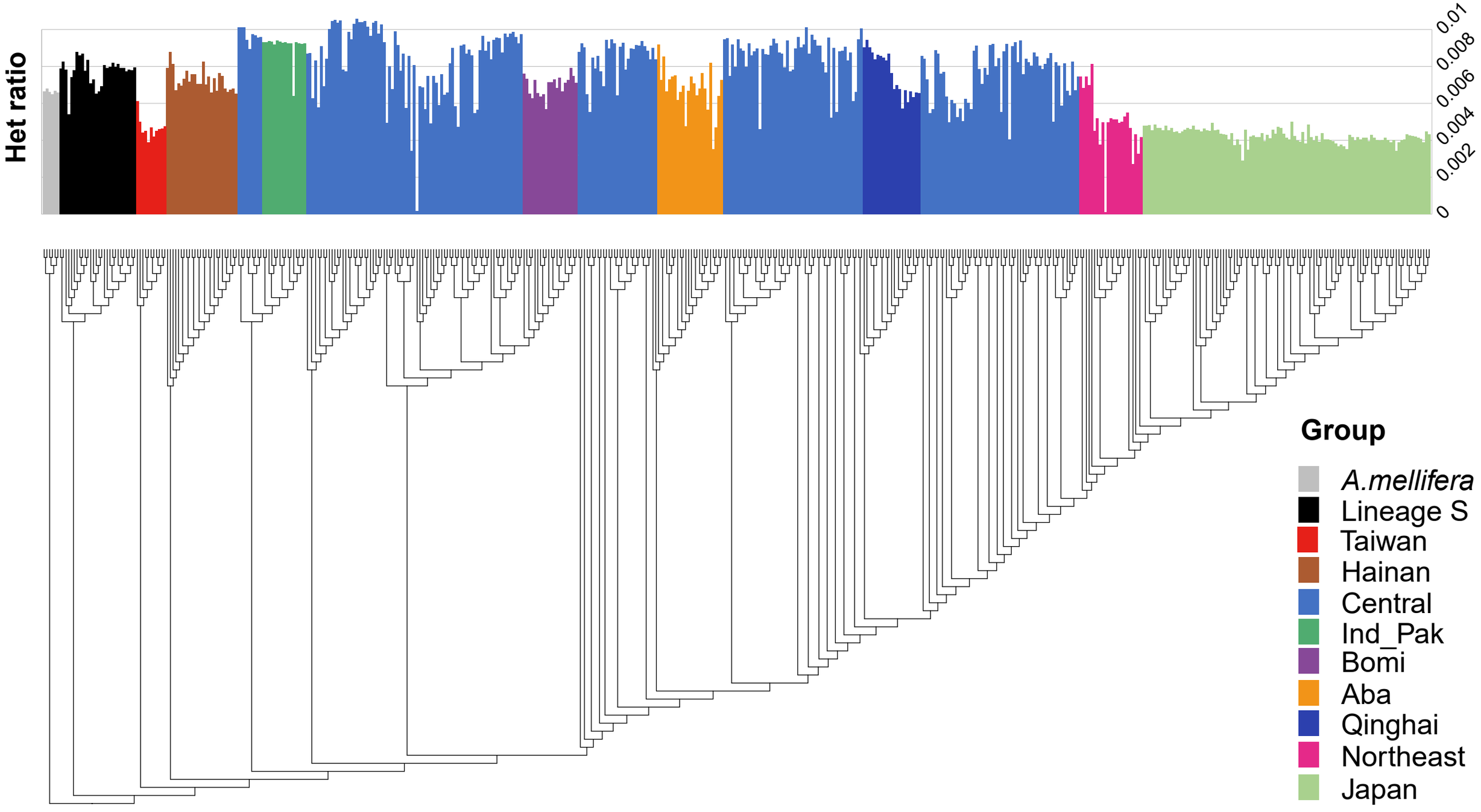
D

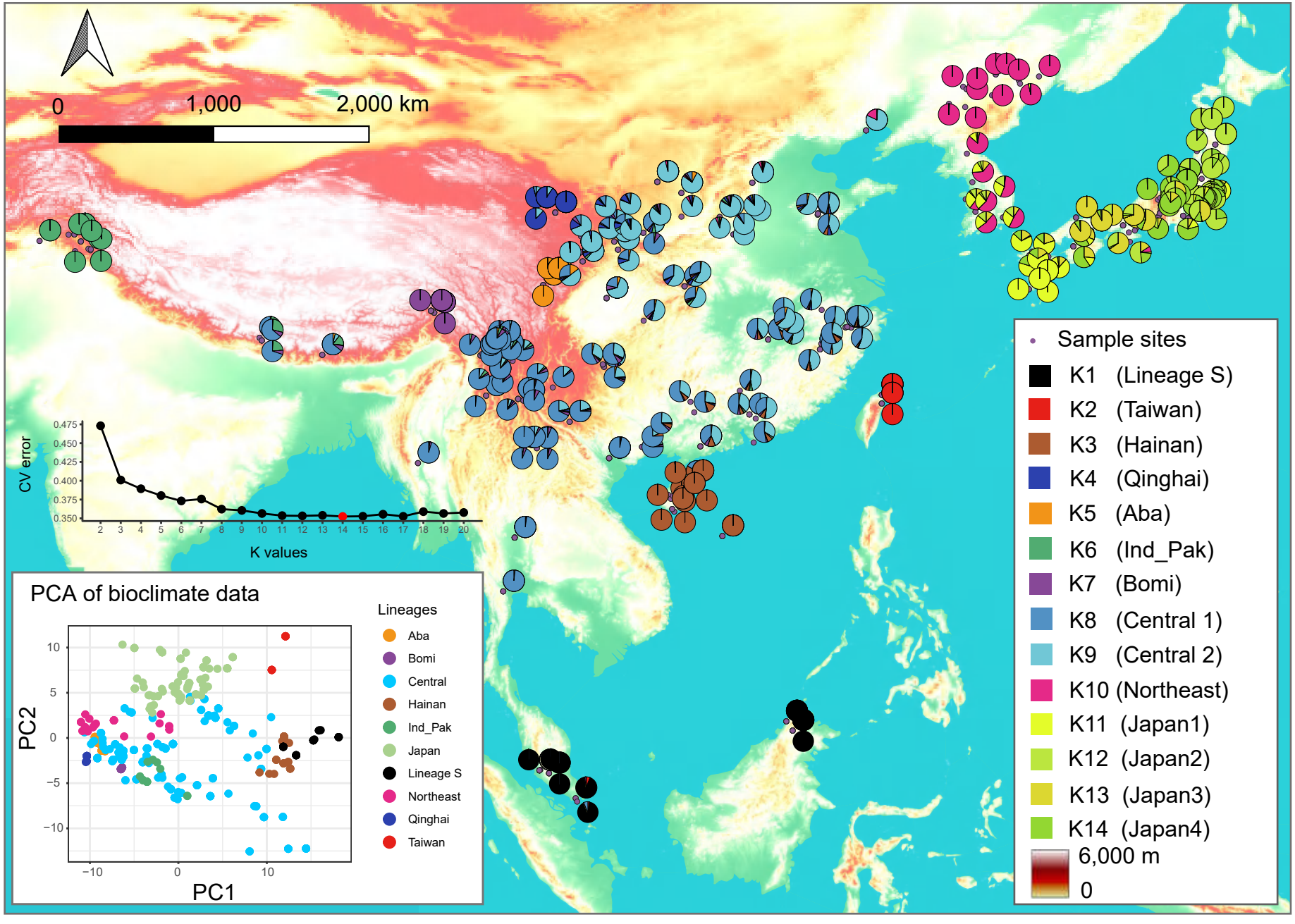


MODERATE impact variants



HIGH impact variants


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


0
1,000
2,000 km
CV error
0.475
0.450
0.425
0.400
0.375
0.350
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
K values
PCA of bioclimate data
PC2
10
5
0
−5
−10
−10
0
10
PC1
Lineages
Aba
Bomi
Central
Hainan
Ind_Pak
Japan
Lineage S
Northeast
Qinghai
Taiwan
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