Genetic barriers to gene flow separate divergent substitution rates across a butterfly hybrid zone

Tianzhu Xiong[1,*], Xueyan Li[2], Masaya Yago[3], and James Mallet[1]

[1] *Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA*
[2] *Kunming Institute of Zoology, Chinese Academy of Science, Kunming 650223, China*
[3] *The University Museum, The University of Tokyo, Tokyo 113-0033, Japan*
[*] *Email:* txiong@g.harvard.edu

September 27, 2021

Abstract

Substitution rate defines the fundamental timescale of molecular evolution which often varies in a species-specific manner. However, it is unknown under what conditions lineage-specific rates can be preserved between natural populations with frequent hybridization. Here, we show in a hybrid zone between two butterflies *Papilio syfanius* and *Papilio maackii* that genome-wide barriers to gene flow can effectively separate different rates of molecular evolution in linked regions. The increased substitution rate in the lowland lineage can be largely explained by temperature-induced changes to the spontaneous mutation rate. A novel method based on entropy is developed to test for the existence of barrier loci using a minimal number of samples from the hybrid zone, a robust framework when system complexity far exceeds sample information. Overall, our results suggest that during the process of speciation, the separation of substitution rates can occur locally in the genome in parallel to the separation of gene pools.

I. INTRODUCTION

The rate of DNA sequence evolution is a critical parameter in evolutionary analysis. Both molecular phylogenetics and coalescent theory rely on observed mutations to reconstruct gene genealogies [1, 2], and so the rate of substitution is the predominant link from molecular data to information about the timing of past events [3]. Different species may have different rates of molecular evolution: generation time, the rate of spontaneous mutation, and the fixation probabilities of new mutations are three major factors determining the overall substitution rate in a species [4, 5]. Within a single species, a constant rate of mutation is often assumed across different populations, but recent data suggest that both mutation rates and mutation types can vary in a population-specific manner even within a single species [6].

On the other hand, differences in population-specific substitution rates may be difficult to detect, for gene flow and recombination between populations homogenize genomes and erode the signal. In an extreme scenario of a well-connected species, even if a particular population has an altered rate of evolution, its effect may quickly propagate elsewhere so that the entire species shares a single, average rate of molecular evolution. Between the regime of a panmictic species with a shared rate and the regime of two separated species with lineage-specific rates, lies a transitional regime of incipient species. Incipient species are lineages with significant levels of divergence, which are still capable of hybridization [7]. Notably, hybridization between incipient species often reveals an inhomogeneous landscape of genomic divergence due to reduced gene flow around loci resistant to hybridization ("barrier loci") [8]. What is the general picture of the rate of molecular evolution in this transitional regime? Under what conditions can lineage-specific rates be preserved despite hybridization? We propose a putative mechanism of the rate preservation, in which substitution rates will diverge only near linked regions of barrier loci, while they will mix in other parts of the genome.

The proposed mechanism of partial preservation of lineage-specific rates is tested on a hybrid zone between two recently diverged butterflies *Papilio syfanius* and *Papilio maackii.* The highland lineage *P. syfanius* forms a hybrid zone with the lowland lineage *P. maackii* throughout the eastern part of the Hengduan Mountains (China). Existing phylogenies based on a mitochondrial gene and a nuclear gene cannot distinguish the two lineages [9], but they are strongly diverged in multiple ecological traits [10] (Fig. S6-S7, Table S4). We first present a new method based on entropy to test for the existence of barrier loci which is valid even for a very small number of samples from the hybrid zone. Using this new method, we recovered a disproportionately high contribution of sex chromosomes in forming gene flow barriers (the large-Z effect) [11], and we also find significant evidence supporting the existence of barrier loci across the autosomes. We then show how different substitution rates between hybridizing lineages are preserved via genetic linkage to barrier loci across the genome. Finally, we discuss possible contributions of generation time and temperature to the divergent substitution rates in the butterfly system.

II. RESULTS

A correlation test for the existence of genetic barriers to gene flow

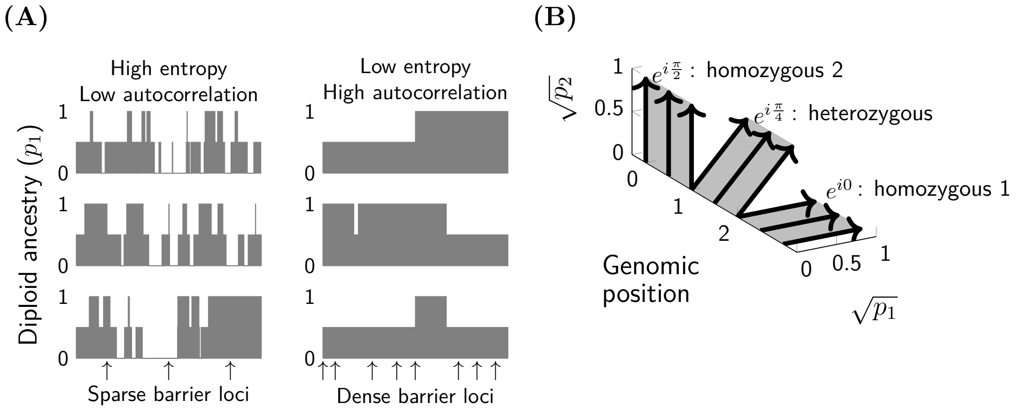

Figure 1: Analysis of entropy in hybrid genomes. **(A)** A demonstration of the relationship between the density of barrier loci and the randomness of ancestry configurations in three hybrid chromosomes. **(B)** The complex representation of diploid ancestry, where three ancestry states correspond to three phases of a unit complex phasor.

Most methods for detecting barrier loci require large sample sizes. There is a trade-off between the number of samples and the quality of genetic markers, so that only a small number of samples can be processed if the entire genome has to be sequenced with a dense distribution of markers. Moreover, there are many species, either endangered or difficult to collect, that cannot be captured in hundreds of individuals for analysis. We here develop a robust method to test for the existence of barrier loci under limited sampling, though at the expense of lacking the ability to identify them individually. The principle of the test is summarized below, and the mathematical detail is discussed in Supplementary Information Section 1.

Consider two diverging populations coming into secondary contact. Away from the hybrid zone, the reduction of gene flow due to barrier loci will preserve sequence divergence between populations, causing islands of genomic differentiation [12]. Near the hybrid zone, barrier loci tend to increase the length of so-called ancestry blocks (contiguous blocks of DNA with the same ancestry) [13]. The latter effect is attributed both to elevated linkage disequilibrium between barrier loci from the selection-migration balance [14, 15], and to the enrichment of a particular ancestry compared to the rest of the genome via linked selection [16, 17]. In other words, barrier loci decrease the randomness of the ancestry configuration in linked regions. The logic of the test is that if the genomic pattern of divergence is caused by barrier loci, the randomness of ancestry in hybrid individuals should co-vary with divergence between parental populations across the genome. If elevated divergence between parental populations is not caused by barrier loci, such correlation will disappear.

To directly measure the randomness of ancestries on a set of hybrid individuals in a given genomic region, we borrow the concept of entropy from information theory and signal processing. Entropy is a natural measure of the spread of a distribution over its all possible configurations. With recombination, parental sequences will mix randomly in the hybrid zone and create all kinds of new sequences of mixed ancestry. In other words, hybrid individuals can be found at many points in the space of all possible sequences of distinct ancestries, which is associated with high entropy [18]. With barriers to gene flow or any other structural restrictions to recombination, only a fraction of the entire sequence space is accessible, and the entropy will be low (Fig. 1A). We provide two complementary entropy measures, S_w and S_b, which capture the within-sample and between-sample randomness of ancestry from a complex-valued representation of ancestry signals (Fig. 1B). Overall, this method can be applied to a small number of unphased or phased hybrid individuals with estimated local ancestry blocks, along with samples from parental populations, to test for the existence of barrier loci. As a conservative test, it only responds to a sufficiently large number of barrier loci, and simulation shows that its false-positive rate is low (Fig. S2).

The prevalence of barrier loci between *P. syfanius* and *P. maackii*

We sampled 11 individuals from a transect covering both parental and hybrid populations between *P. syfanius* and *P. maackii* (Fig. 2A), and re-sequenced their genomes to an average coverage of 25×. The phylogeny of the assembled mitochondria re-affirms the previous finding that they are indistinguishable at the mitochondrial level (Fig. 2B). Genomic reads were mapped to the chromosomes of a previously assembled genome from a closely related species *Papilio bianor* [19], and unphased SNPs were used in subsequent analyses.

The entire sex (Z) chromosome stands out as a strong, integrated barrier to gene flow. Between the two parental populations (XY, KM), the relative divergence on the Z chromosome is strong (mean F_{ST}: 0.2∼0.4 on autosomes, and 0.78 on the Z chromosome, Fig. 2C), but the absolute divergence is weak (∼1% on autosomes and ∼1.5% on the Z chromosome, Fig. S14). Thus, the extreme F_{ST} on the Z chromosome is primarily caused by the chromosome's negligible genetic diversity and elevated sequence divergence between parental populations. In two hybrid populations, Z chromosomes are either fixed for the ancestry of either lineage, or show the signal of very recent hybridization (Fig. 2D). Together, the Z chromosome behaves as an integrated unit resistant to gene flow across the hybrid zone.

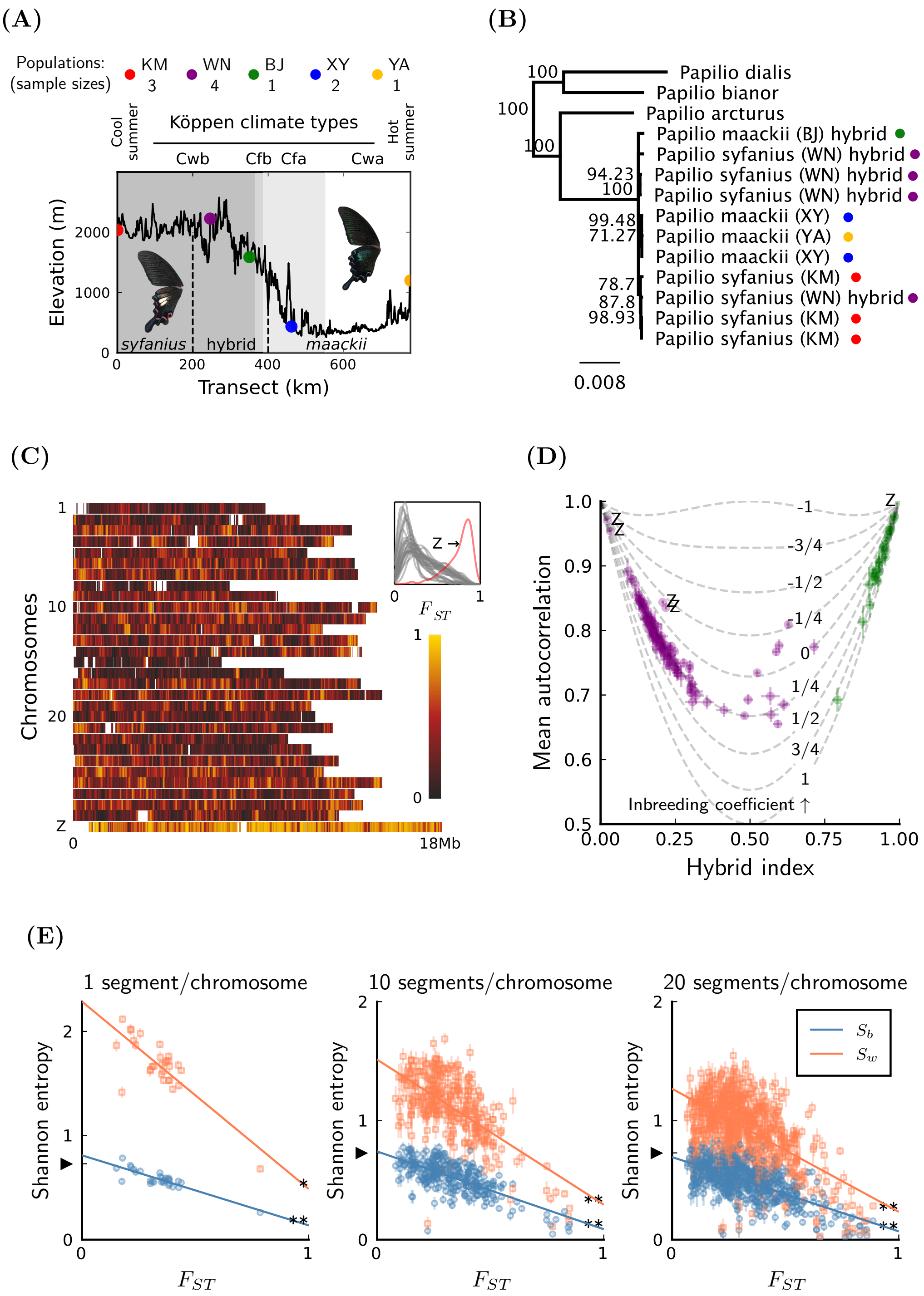

Figure 2: Structure of the hybrid zone between *P. syfanius* and *P. maackii.* **(A)** The transect covers five populations and a range of elevations with different climates. **(B)** The two lineages are indistinguishable in the mitochondrial tree. **(C)** F_{ST} distributions along 29 autosomes and the Z chromosome between populations KM and XY, calculated on 50kb windows. The inset shows the distribution density of F_{ST}. **(D)** The mean autocorrelation of each hybrid chromosome plotted against its hybrid index. Purple: population WN. Green: population BJ. Grey: isoclines of ancestry-inbreeding coefficients (Eq. 4). A positive ancestry-inbreeding coefficient implies excessive homozygosity in ancestry, which is the result of identity-by-descent from recent ancestors. A negative ancestry-inbreeding coefficient implies very recent hybridization, which produces excessive heterozygosity in ancestry. **(E)** The relationship between entropy in population WN and F_{ST} between populations (KM, XY). The significance of negative correlation is marked by asterisks. Two asterisks: Z-score> 5. One asterisk: $3 < Z$-score< 5. The black triangle on the vertical axis represents the theoretical maximum of S_b among 4 individuals.

Contrary to the Z chromosome, most autosomes in the hybrid populations show a strong signal of inbreeding. The highland hybrid population (WN) has an inbreeding coefficient ~ 0.25 for most autosomes, while some autosomes of the lowland hybrid population (BJ) have inbreeding coefficients larger than 0.5 (Fig. 2D). The lack of heterozygous ancestries within individuals indicates a very small breeding size in both localities. It is consistent with the field observation that the spatial distribution of this species is highly fragmented.

To test if major F_{ST} peaks across the genome are statistically associated with barrier loci, we divided each chromosome into 1, 10 or 20 segments, and computed both entropy measures S_w and S_b on the ancestry signal estimated for each segment in individuals from the highland hybrid population (WN). The presence of barrier loci is indicated by a significantly negative Pearson's correlation coefficient between (S_w, F_{ST}), or between (S_b, F_{ST}), where F_{ST} is calculated for the corresponding segments between the parental populations (XY, KM). We found that the correlation is always significantly negative for every choice of segment numbers (Fig. 2E), and is also significant if the Z chromosome is removed from the analysis, or if the local ancestry is estimated with different spatial resolutions (Fig. S18, Table S5-S6). Thus, there is strong evidence that the variation of F_{ST} is driven by the variation of barrier effect across the genome, and major F_{ST} peaks on the autosomes are indeed associated statistically with real barrier loci.

The preservation of lineage-specific substitution rates

Using three closely related species *P. bianor*, *P.dialis* and *P. arcturus* occupying different elevations as outgroups (Fig. 2B), we found that derived single-nucleotide variants accumulate more in the lowland lineage *P. maackii* across both synonymous sites and nonsynonymous sites. Using the three-population D_3 statistic [20], the bias is significant and invariant regardless of the outgroup species (Fig. 3A, Fig. S19). At least three hypotheses could explain the bias. Firstly, hybridization between all the outgroups and the highland *P. syfanius* might create asymmetric allele sharing and reduce the branch length leading to *P. syfanius*. Secondly, systematically greater gene copy number in the highland *P. syfanius* might suppress the call of a derived single-nucleotide variant in that lineage, because multiple copies of a gene decreases the genotype likelihood when they are all mapped to a single copy in the reference genome. Thirdly, the bias can simply be the result of increased substitution rates in the lowland *P. maackii*.

The four-population D_4 statistic (ABBA-BABA test) was used to infer the pattern of gene flow among the five species and test the first hypothesis [21]. A significantly non-zero D_4 implies the existence of gene flow. We found that a weak amount of gene flow most likely occurred between lineages from the same altitude: *P. syfanius* with *P. arcturus* (sympatric in highland), and/or *P. maackii* with *P. dialis* & *P. bianor* (sympatric in lowland) (Table S7, Fig. S20). Although we could not fit specific models of gene flow from D_4, gene flow between aforementioned lineages only produce non-negative D_3 when the outgroup is chosen between *P. bianor* and *P. dialis* (Supplementary Information Section 10), which contradicts observed negative D_3 statistics (Fig. 3A). The first hypothesis is thus disproved. To rule out the second hypothesis, note that we have already filtered out sites with excessive coverage or outside annotated genes prior to calculating the D_3 statistic, and the coverage for the rest of the sites showed no systematic increase of copy number in the highland species (Table S8-S9). Thus, we conclude that the bias of accumulating derived alleles is caused by an increased substitution rate in the lowland *P. maackii*.

Biased rates of substitution will not be observed, if gene flow completely homogenizes the divergence.

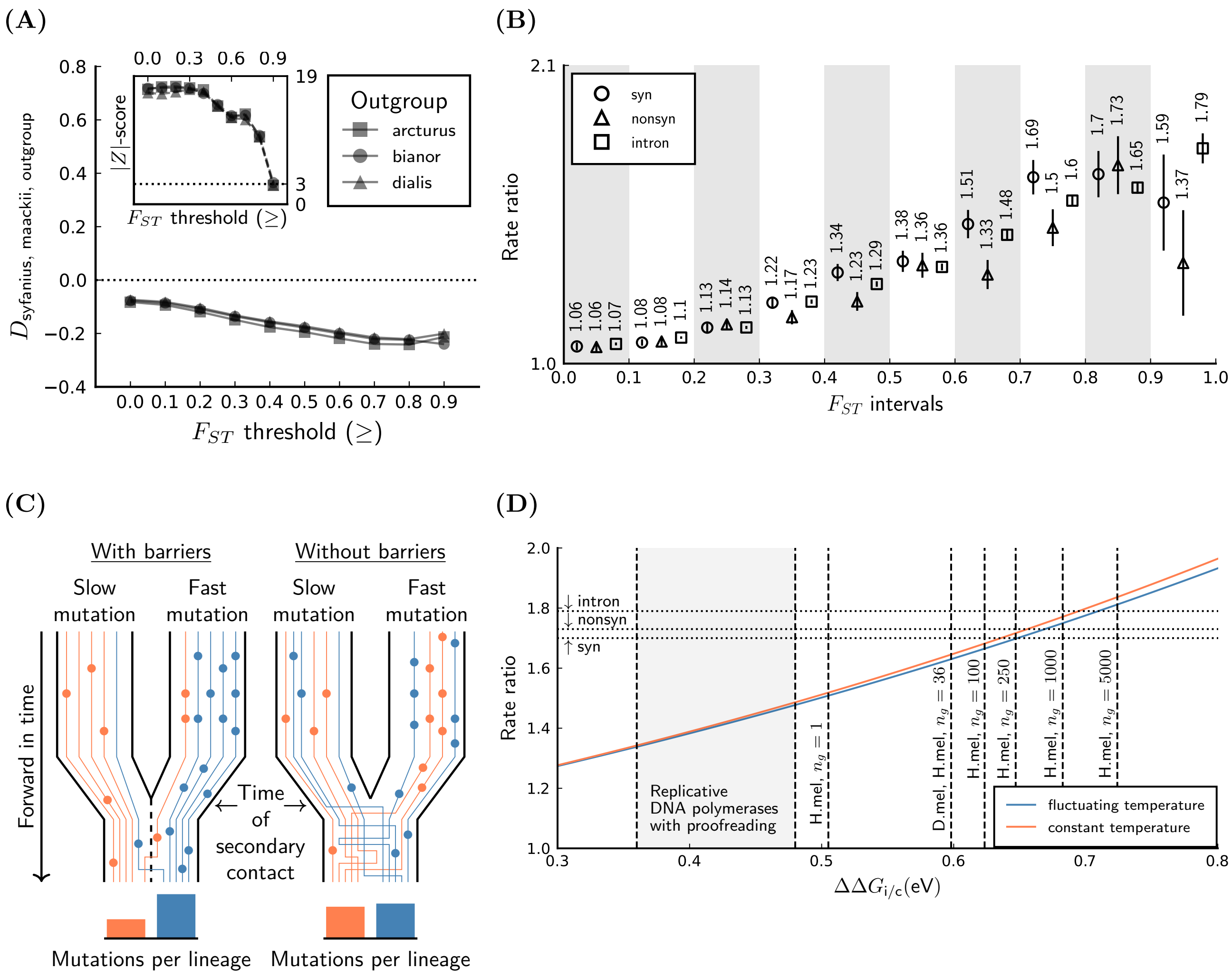

Figure 3: The separation of lineage-specific rates due to genetic linkage to barrier loci. **(A)** The three-population D-statistic on synonymous sites show that *P. syfanius* consistently shares a greater fraction of alleles with any outgroup. **(B)** The rate ratio between the lowland and the highland lineages increases with linked sequence divergence (F_{ST} on 50kb non-overlapping windows). **(C)** The conceptual model of how barrier loci separate the lineage-specific substitution rates during hybridization. **(D)** Predicted rate ratio of spontaneous mutations due to the difference of temperature between the lowland and the highland lineages. The per-generation mutation rate μ_g is taken from *Drosophila melanogaster* (D.mel) and *Heliconius melpomene* (H.mel). Each dashed vertical line represents a particular $\Delta\Delta G_{\text{i/c}}$ estimated from a pair of μ_g and n_g. Each dashed horizontal line is the maximum rate ratio observed for a particular type of sites.

The fact that a significant rate bias is present can only be the result of separating the spatial movement of parental lineages to different sides of the hybrid zone by gene flow barriers. In this conceptual model (Fig. 3C), it is predicted that the observed rate bias should monotonically increase with the effect of genetic barriers. Since we have established that many F_{ST} peaks are associated with barrier loci, the observed rate bias should also increase with local F_{ST} values in the genome. To quantitatively measure the bias of substitution rates, define rate ratio as the average substitution rate in the lowland species divided by the average substitution rate in the highland species. Therefore, a larger rate ratio is associated with higher bias. We estimated the rate ratio on synonymous sites, nonsynonymous sites, and introns between populations XY and KM, partitioned by the F_{ST} values on 50kb windows. We found that the monotonic relationship holds except for synonymous and nonsynonymous sites near extreme F_{ST} peaks (Fig. 3B), further confirming the proposed mechanism of rate preservation. All three types of sites show a maximum rate ratio within the range of 1.7 ∼ 1.8.

The effect of temperature and generation time

Multiple mechanisms could produce higher substitution rates in warmer conditions [22]. For most animals, warmer climate prolongs the breeding season and accelerates body development, such that the generation turn-over rate becomes faster. If the germline cell division number is a constant between two generations, more DNA duplication errors will accumulate with more generations in a year, which is known as the generation-time effect [4]. Temperature can also influence the spontaneous mutation rate in ectothermal species [23]. The high fidelity of DNA duplication is mainly attributed to the kinetic selection by DNA polymerases against mismatched base-pairings [24, 25]. Increasing temperature will increase the likelihood of overcoming the energy barrier of forming a mismatch, thus increasing the error rate in duplication. Finally, ecological selection might drive faster evolution in warm habitats, but this effect is not expected to affect coding and non-coding sites in the same way.

To assess the relative contribution of temperature and generation time to the biased substitution rate between *P. maackii* and *P. syfanius*, we gathered three lines of evidence: directly modeling the rate of spontaneous mutation under different temperatures, the mutation spectrum of single nucleotide variants, and the seasonal distribution of both butterflies.

Using museum specimens as the reference, we built the maximum entropy species range model [26], and the range model was combined with local temperature data [27] to calculate the ratio of spontaneous mutation rates due to temperature differences. Spontaneous mutation is modeled as a pseudo-first-order catalytic reaction [25]. If n_g is the number of germline cell divisions per generation, then the per-generation mutation rate μ_g is

$$\mu_g \approx n_g \exp\left(-\frac{\Delta\Delta G_{\mathrm{i/c}}}{k_B T}\right), \tag{1}$$

where $\Delta\Delta G_{\mathrm{i/c}}$ stands for the difference between the free energy barriers along the reaction paths leading to an incorrect v.s. a correct base-pairing. k_B is the Boltzmann constant and T is the thermodynamic temperature. As n_g is unknown in butterflies, we calculated the rate ratio under different values of n_g (Fig. 3D), including $n_g = 36$ for *Drosophila melanogastor* [28]. The result indicates that temperature difference is sufficient to induce a rate ratio of at least 1.5 ($n_g = 1$), or more than 1.6, if n_g is the same as *Drosophila*. This indicates that temperature-induced increase of spontaneous mutation rate is sufficient, in theory, to explain most of the increase in the rate of molecular evolution.

To find genomic evidence that temperature might have increased the spontaneous mutation rate between the two species, we argue that if only the generation-time effect causes the increase in substitution rates, the mutation spectrum will remain invariant, because mutation types should be independent of the number of DNA duplication cycles. However, different mutation types might respond differently to the change in temperature, for they undergo different chemical transformations. We mainly focused on the strong C:G mutational bias in our system (enriched single nucleotide mutations in the direction of C:G>*:*, see Fig. S22, Table S15-S20) [29, 30]. Using a kinetics argument, the C:G mutational bias suggests that increasing temperature might accelerate mutations on A:T sites more profoundly, as their mutational transition-states are less stable [31, 32, 33]. Thus, it is expected that the C:G mutational bias in the highland *P. syfanius* should be lower than that in the lowland *P. maackii*. To test this prediction, we selected single nucleotide mutations endemic to each population so that they are enriched for recent mutations, and we calculated the fraction of G:C>*:* mutations conditioning on the GC content. The maximum and the average temperatures from each population were used to calculate the free energy difference $\Delta\Delta G_{\mathrm{AT/GC}}$ between A:T>*:* and G:C>*:* mutations. If the prediction holds,

$\Delta\Delta G_{\mathrm{AT/GC}}$ should be positive. Despite the coarse approach, we recovered positive $\Delta\Delta G_{\mathrm{AT/GC}}$ for synonymous sites, nonsynoymous sites, and introns (Table S14, Fig. S23), but the estimated value is only marginally significant.

In terms of generation time, we used the temporal records from museum specimens and field observations to estimate the seasonal distribution of both species (Fig. S24). Annually, there are two main peaks for both species. However, the first peak in the highland *P. syfanius* is much higher than the second, while they are more even in the lowland *P. maackii*. We suspect that while the generation time is similar in both species, lowland populations might have an additional third brood in warm years [34]. The slight differences in generation time may account for the additional increase in substitution rates in the lowland species.

III. DISCUSSION

Hybridization between diverging lineages is often associated with varying levels of gene flow across the genome [35]. Functional loci under divergent selection, incompatibility genes, and structural rearrangements can all lead to the reduction of gene flow in linked genomic regions [15, 36, 37]. Barrier loci are important in the early stage of speciation as they prevent the homogenization of the entire genome, and in some cases even promote further divergence [38, 39]. When many barrier loci exist, it is difficult to model their joint dynamics analytically, and this militate against using a small sample size to test the predictions of each model. Unlike most previous approaches [40], our entropy-based method is essentially descriptive, and investigates the aggregate effect of barrier loci on hybrid ancestry. It produces robust and conservative results even when the sample size is unsuitable for locus-specific analysis.

The finding from the hybrid zone between *P. syfanius* and *P. maackii* indicates that barrier loci not only separate the genomic content in linked regions, but also allow for separation of substitution rates, a direct consequence of restricting the local gene genealogies to a specific genomic and environmental background. In fact, the D_3 statistic was initially conceived to detect gene flow [20], and is prone to mis-identification when substitution rates are different between lineages. We demonstrated that D_3 can thus be used to detect asymmetry in lineage-specific substitution rates, as long as gene flow could be rule out with additional information. To explain the large drop of observed rate ratio on nonsynonymous sites near extreme F_{ST} peaks, it is likely that these regions are enriched for derived mutations subject to strong selection, so that their substitution rates cannot be explained by temperature and generation time alone. Nonsynonymous sites also showed the least biased rate ratio in most F_{ST} partitions (except for $0.8 < F_{ST} < 0.9$), which is consistent with the prediction under the nearly neutral theory [41], because lowland species tend to have larger effective population sizes, which in turn suppress the fixation of nearly neutral nonsynonymous mutations.

Faster evolution in warmer climate has been observed in multiple systems [42, 43, 44]. Our result is consistent with a mechanism in which rising temperature directly accelerates the spontaneous mutation rates of these butterflies which lack intrinsic control of body temperature [45, 46, 23]. Although we cannot rule out other factors affecting mutation rates, we expect the intrinsic physiology to be similar between incipient species, as DNA replication is evolutionary conserved [47].

Together, an empirical model of speciation-with-gene-flow is revealed, where the divergence of local genomic content and the divergence of local rate of evolution co-occur between incipient species.

IV. MATERIALS AND METHODS

The complex-valued bi-ancestry signal

The space of all ancestry sequences is high-dimensional, and directly calculating the entropy in this space is not feasible with just a few samples. So we propose to measure only the pairwise correlation of ancestries among sites, which captures only the second-order randomness, but is sufficient for practical purposes. Consider a hybrid individual with two parental populations indexed by $k = 1, 2$. Assuming a continuous genome, let $p_k(l) = 0, \frac{1}{2}, 1$ be the diploid ancestry of locus l within genomic interval $[0, L]$. By definition, we have $p_1(l) + p_2(l) = 1$, i.e. the total ancestry is conserved everywhere in the genome. The bi-ancestry signal at locus l is defined as the following complex variable

$$z(l) = \sqrt{p_1(l)} + i\sqrt{p_2(l)} = e^{i \arccos \sqrt{p_1(l)}}, \tag{2}$$

where $i = \sqrt{-1}$ is the imaginary number. An advantage of using a complex representation for the bi-ancestry signal is that we can model different ancestries along the genome as different phases of a complex unit phasor ($e^{i\theta}$), such that the power of the signal at any given locus is simply the sum of both ancestries, which is conserved ($|z(l)|^2 = 1$). It ensures that we do not bias the analysis to any particular region or any particular individual when decomposing the signal into its spectral components. Note that the representation works for all kinds of ploidy by choosing $p_1 = 0, 1/n, 2/n, \cdots, 1$ for unphased n-ploid species, or $p_1 = 0, 1$ for fully phased data, which are equivalent to haploid genomes. For a comprehensive explanation we refer readers to Supplementary Material Section 1.3 .

Autocorrelation and hybrid index

The two-point autocorrelation function $A(l_1, l_2) = z(l_1)\overline{z(l_2)}$ measures the similarity between any two points along the ancestry signals, so that the mean autocorrelation, a, of a single signal, is

$$a = \frac{1}{L^2} \iint_{[0,L]^2} A(l_1, l_2)\, \mathrm{d}l_1\, \mathrm{d}l_2 = \left| \frac{1}{L} \int_0^L z(l)\, \mathrm{d}l \right|^2 \tag{3}$$

While hybrid index $h = \frac{1}{L}\int_0^L p_1(l)\, \mathrm{d}l$ is the average of the real-valued ancestry p_1 along the genome, averaging the complex ancestry gives us a measure of the overall similarity within a single individual. The (a, h) plane is a direct transformation of the widely used triangular plot of heterozygosity and hybrid index (Fig. S1). Our definition of ancestry-inbreeding coefficient is: conditioning on hybrid index, the deviation of heterozygosity in ancestry (H) from the random union of ancestry in a single chromosome.

$$F = 1 - \frac{H}{2h(1-h)} \tag{4}$$

It measures the balance between breeding within the hybrid population versus breeding with outside migrants, and defines a family of isoclines in the (a, h) plane (see Eq. S15, S16).

Within-sample spectral entropy

The mean autocorrelation is scale-independent as it does not distinguish long-range correlation from short-range correlation. To characterize the average autocorrelation at a given scale l, define the following scale-dependent autocorrelation function $B(l) = \frac{1}{L}\int_0^L z(\xi)\overline{z(\xi + l)}\, \mathrm{d}\xi$, where $z(l)$ is understood as a periodic function such that $z(\xi + l) = z(\xi + l - L)$ whenever the position goes outside of $[0, L]$.

The Wiener-Khinchin theorem guarantees that $z(l)$'s power spectrum $\zeta(f)$, which is discrete, and the autocorrelation function $B(l)$ form a Fourier-transform pair. It means that the power spectrum has exactly the same information as the autocorrelation function $B(l)$. Due to the uncertainty principle of Fourier transform, $B(l)$ that vanishes quickly at short distances (small-scale autocorrelation) will produce a wide $\zeta(f)$, and vice versa. So the entropy of $\zeta(f)$, which measures the spread of the total ancestry into each spectral component, also measures the scale of autocorrelation. In practice, $\zeta(f)$ is the square-modulus of the Fourier series coefficients of $z(l)$, and we fold the spectrum around $f = 0$ before calculating the within-sample entropy S_w. The formula used in the manuscript is

$$\begin{aligned} S_w &= -\sum_{n=0}^{+\infty} \zeta_n \ln \zeta_n \\ \zeta_n &= \begin{cases} |Z_n|^2 + |Z_{-n}|^2 & (n > 0) \\ |Z_0|^2 \quad (n = 0) \end{cases} \end{aligned} \tag{5}$$

where Z_n are the Fourier coefficients from the expansion $z(l) = \sum_{n=-\infty}^{+\infty} Z_n e^{i2\pi nl/L}$.

Between-sample spectral entropy

As ancestry configuration is far from random around barrier loci, it will also influence the correlation of ancestry between different individuals at the same locus. For a genomic region with dense barriers, two individuals could either be very similar in ancestry, or very different. This effect can be quantified by first calculating the cross-correlation $C_{j,j'}(l) = z_j(l)\overline{z_{j'}(l)}$ at position l between samples j and j', and then averaging across the genome: $c_{j,j'} = \frac{1}{L}\int_0^L C_{j,j'}(l)\,\mathrm{d}l$. The $J \times J$ dimensional matrix $\mathbf{C}$ with entries $c_{j,j'}$ describes the pairwise cross-correlation within the cohort of J samples. We also have $c_{j,j} \equiv 1$ as each sample is perfectly correlated with itself. The matrix $\mathbf{C}$ is Hermitian, so it has a real spectral decomposition with eigenvalues λ_j that satisfy $\sum_j \lambda_j/J = 1$. This process is very similar to performing a principal component analysis on the entire cohort of samples, and λ_j/J describes the fraction of the total ancestry that is projected onto the principal component j. If many loci co-vary in ancestry, the spectrum $\{\lambda_j\}$ will be concentrated near the first few components. Similarly, we use entropy to measure the spread of the spectrum, and hence the between-sample spectral entropy is defined as

$$S_b = -\sum_j \frac{\lambda_j}{J} \ln \frac{\lambda_j}{J} \tag{6}$$

Testing for asymmetric allele sharing

Given a species tree {{P$_1$,P$_2$},O}, where P$_1$ and P$_2$ are sister species and O is the outgroup, if mutation rate is constant and no gene flow with O, then on average the number of derived alleles within P$_1$ should equal the number of derived alleles within P$_2$. Let $\mathcal{S}$ be a collection of sites, f_s be the frequency of a particular site pattern at site $s \in \mathcal{S}$. "ABB" be the pattern where only P$_2$ and O share the same allele, and "BAB" be the pattern where only P$_1$ and O share the same allele, then the three-species D_3 statistic [20] is

$$D_{\mathrm{P_1,P_2,O}} = \frac{\sum_{s\in\mathcal{S}}(f_{s,\mathrm{ABB}} - f_{s,\mathrm{BAB}})}{\sum_{s\in\mathcal{S}}(f_{s,\mathrm{ABB}} + f_{s,\mathrm{BAB}})} \tag{7}$$

A significant deviation of $D_{\mathrm{P_1,P_2,O}}$ from 0 indicates that a process is breaking the symmetry in the system, where it could either be gene exchange with O or asymmetric mutation rate between P$_1$ and

P$_2$. To rule out the possibility of gene flow, we need the four-species D_4 statistic which considers the species tree {{{P$_1$,P$_2$},O$_1$},O$_2$} and site patterns ABBA versus BABA [21]:

$$D_{\mathrm{P}_1,\mathrm{P}_2,\mathrm{O}_1,\mathrm{O}_2} = \frac{\sum_{s\in\mathcal{S}}(f_{s,\mathrm{ABBA}} - f_{s,\mathrm{BABA}})}{\sum_{s\in\mathcal{S}}(f_{s,\mathrm{ABBA}} + f_{s,\mathrm{BABA}})} \tag{8}$$

As allele "B" is shared between an outgroup and a member of {P$_1$,P$_2$}, we can eliminate the influence of mutation rate, assuming no double-mutation. A significant deviation of $D_{\mathrm{P}_1,\mathrm{P}_2,\mathrm{O}_1,\mathrm{O}_2}$ from 0 indicates gene flow between the outgroups and {P$_1$,P$_2$}. The significance of both tests is computed using the block-jackknife with 1Mb blocks across the genome. Additionally, we estimated the rate ratio as follows. First we restrict to sites where all outgroups are fixed for the same allele, then rate ratio is computed as the ratio between the probability of observing a derived allele exclusive to P$_1$, and the probability of observing a derived allele exclusive to P$_2$. Let $I(\cdot)$ be the identity function, and f_s be the frequency of the derived allele, then:

$$\text{Rate ratio} = \frac{\sum_{s\in\mathcal{S}} f_{s,\mathrm{P}_1}(1-f_{s,\mathrm{P}_2})\Pi_{i\in\mathrm{outgroups}}I(f_{s,i}=0)}{\sum_{s\in\mathcal{S}}(1-f_{s,\mathrm{P}_1})f_{s,\mathrm{P}_2}\Pi_{i\in\mathrm{outgroups}}I(f_{s,i}=0)} \tag{9}$$

Museum specimens and climate data

Museum specimens with verifiable locality data of both *P. syfanius* and *P. maackii* were gathered from The University Museum (The University of Tokyo), Global Biodiversity Information Facility (see Supplementary Information Section 2), and individual collectors. Records of *P. maackii* from Japan, Korea and NE China were excluded from the analysis, so that most *P. maackii* individuals correspond to *ssp. shimogorii*, the subspecies that hybridizes with *P. syfanius*. Spatial principal component analysis was performed on elevation, maximum temperature of warmest month, minimum temperature of coldest month, and annual precipitation, all with 30s resolution from WorldClim-2 [27]. The first two PCAs, combined with tree covers [48], were used in MaxEnt-3.4.1 to produce species distribution models that use known localities to predict the occurrence probabilities across the entire landscape [26]. Outputs were trimmed near known boundaries of both species. Finally, the species distribution model was used as an integration kernel to calculate the geographic average of the maximum/minimum/mean temperatures for *P. maackii* and *P. syfanius* with climate data of 30s resolution from WorldClim-2 [27].

The reaction kinetics of spontaneous mutations

We assume that DNA duplication errors dominate the mutation process. Let μ_g be the per-generation mutation rate, and μ_d be the per-cell-division mutation rate, and n_g is the average number of germline cell divisions per generation. As $\mu_d \ll 1$, we can assume that no double-mutation occurs during a single generation, so $\mu_g \approx n_g\mu_d$. For a single cycle of germline cell division, let ΔG_i (or ΔG_c) be the sum of all free-energy barriers along the reaction path of duplicating a single site with an incorrect (or correct) pairing. μ_d is simply the conditional probability that a DNA duplication has occurred but the product is incorrect. Assuming the reaction is first-order in the concentration of DNA molecules, we have

$$\mu_d = \frac{\exp\left(-\frac{\Delta G_\mathrm{i}}{k_BT}\right)}{\exp\left(-\frac{\Delta G_\mathrm{i}}{k_BT}\right)+\exp\left(-\frac{\Delta G_\mathrm{c}}{k_BT}\right)} \approx \exp\left(-\frac{\Delta\Delta G_{\mathrm{i/c}}}{k_BT}\right), \tag{10}$$

where $\Delta\Delta G_{\mathrm{i/c}} = \Delta G_{\mathrm{i}} - \Delta G_{\mathrm{c}} > 0$ stands for the difference between energy barriers. So we have

$$\mu_g \approx n_g \exp\left(-\frac{\Delta\Delta G_{\mathrm{i/c}}}{k_B T}\right) \tag{11}$$

With the current estimate of per-generation mutation rate $\mu_g = 3 \times 10^{-9}/(\text{site} \cdot \text{generation})$ [49], and assuming T to be the room temperature 298.15K, we can calculate $\Delta\Delta G_{\mathrm{i/c}}$ under different values of n_g. In fact, $\Delta\Delta G_{\mathrm{i/c}}$ depends weakly on the temperature under consideration (273.15K to 313.15K), but strongly on n_g, so the exact temperature used in calculating $\Delta\Delta G_{\mathrm{i/c}}$ is not important.

For the C:G mutational bias, let $\Delta\Delta G_{\mathrm{AT/CG}}$ be the difference between the energy barriers associated with A:T>*:* mutations and C:G>*:* mutations. The fraction of A:T>*:* mutations, conditioning on a 50% GC-content, is

$$f_{\mathrm{AT}} = \left[1 + \exp\left(\frac{\Delta\Delta G_{\mathrm{AT/CG}}}{k_B T}\right)\right]^{-1}, \tag{12}$$

which is equivalent to

$$\ln\left(f_{\mathrm{AT}}^{-1} - 1\right) = (k_B T)^{-1}\Delta\Delta G_{\mathrm{AT/CG}}. \tag{13}$$

With multiple pairs of f_{AT} and T, we can perform linear regression to estimate $\Delta\Delta G_{\mathrm{AT/CG}}$.

Sampling, re-sequencing, and mitochondrial phylogeny

Eleven males of *P. syfanius* and *P. maackii*, with one male of *P. arcturus* and one male of *P. dialis* were collected from the field between July and August in 2018, and were stored in RNAlater at -20C prior to DNA extraction. E.Z.N.A Tissue DNA kit was used to extract genomic DNA, and KAPA DNA HyperPlus 1/4 was used for library preparation, with an insert size of 350bp and 2 PCR cycles. The library is sequenced on a Illumina NovaSeq machine with paired-end reads of 150bp. Adaptors were trimmed using Cutadapt-1.8.1, and subsequently the reads were mapped to the reference genome of *P. bianor* with BWA-0.7.15, then deduplicated and sorted via PicardTools-2.9.0. The realized coverage of 13 samples in repetitve and non-repetitive regions is summarized in Fig. S5, and the average coverage varies between 20× to 30×. Variants were called twice using BCFtools-1.9—one includes all samples, which was used in analyses involving outgroups, and the other one excludes *P. arcturus* and *P. dialis*, which was used in all other analyses. The following thresholds were used to filter variants: $10N <$DP$< 50N$, where N is the sample size; QUAL> 30; MQ> 40; MQ0F< 0.2. As a comparison, we also called variants with GATK4 and followed its best practises, and 93% of post-filtered SNPs called by GATK4 overlapped with those called by BCFtools. We used SNPs called by BCFtools throughout the analysis. Mitochondrial genomes were assembled from trimmed reads with NOVOPlasty-4.3.1 [50], using a published mitochondrial ND5 gene sequence of *P. maackii* as a bait (NCBI accession number: AB239823.1). The neighbor-joining mitochondrial phylogeny was built with Geneious Prime-2021.2.2 (genetic distance model: Tamura-Nei), and we used 10^4 replicates for bootstrapping.

Data availability

Relevant code is available at https://github.com/tzxiong/2021_Maackii_Syfanius_HybridZone.

AUTHOR CONTRIBUTIONS

T.X. and J.M. designed the project. X.L. provided the reference genome of *P. bianor* and facilitated the fieldwork. M.Y. provided most museum specimens used in the manuscript. T.X. collected and analyzed the samples. T.X. and J.M. wrote the manuscript.

ACKNOWLEDGEMENT

Butteflies were collected under the Lewis and Clark Fund for Exploration and Field Research (American Philosophical Society, 2017-2018) granted to Tianzhu Xiong. We thank Fernando Seixas for discussion on hybrid ancestry inference. We thank Kaifeng Bu for his comments on information-theoretic measures. Nathaniel Edelman, Neil Rosser, Miriam Miyagi, Yuttapong Thawornwattana, Sarah Dendy, John Wakeley, Robin Hopkins, and Naomi Pierce also contributed to the discussion at various stages of the project. We also thank Yuchen Zheng, Zhuoheng Jiang, Shaoji Hu, Feng Cao, and Shui Xu for their support in the field.

References

[1] Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**, 303–314 (2012).

[2] Wakeley, J. *Coalescent Theory: An Introduction* (Macmillan Learning, 2016).

[3] Bromham, L. & Penny, D. The modern molecular clock. *Nature Reviews Genetics* **4**, 216–224 (2003).

[4] Ohta, T. An examination of the generation-time effect on molecular evolution. *Proceedings of the National Academy of Sciences* **90**, 10676–10680 (1993).

[5] Lynch, M. Evolution of the mutation rate. *Trends in Genetics* **26**, 345–352 (2010).

[6] DeWitt, W. S., Harris, K. D., Ragsdale, A. P. & Harris, K. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences* **118** (2021).

[7] Darwin, C. *On the Origin of Species, 1859* (Routledge, 2004).

[8] Payseur, B. A. & Rieseberg, L. H. A genomic perspective on hybridization and speciation. *Molecular Ecology* **25**, 2337–2360 (2016).

[9] Condamine, F. L. *et al.* Fine-scale biogeographical and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. *Cladistics* **29**, 88–111 (2013).

[10] Kashiwabara, S. Why are *Papilio dehaanii* from Tokara IsIs. and Izu IsIs. beautiful? *Choken-Field* **6**, 6–16 (1991). URL https://ci.nii.ac.jp/naid/10027438692/en/.

[11] Presgraves, D. C. Evaluating genomic signatures of “the large X-effect” during complex speciation. *Molecular Ecology* **27**, 3822–3830 (2018).

[12] Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**, 375–402 (2009).

[13] Sedghifar, A., Brandvain, Y. & Ralph, P. Beyond clines: lineages and haplotype blocks in hybrid zones. *Molecular Ecology* **25**, 2559–2576 (2016).

[14] Slatkin, M. Gene flow and selection in a two-locus system. *Genetics* **81**, 787–802 (1975).

[15] Barton, N. H. Multilocus clines. *Evolution* 454–471 (1983).

[16] Szymura, J. M. & Barton, N. H. Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* **40**, 1141–1159 (1986).

[17] Martin, S. H., Davey, J. W., Salazar, C. & Jiggins, C. D. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology* **17**, e2006288 (2019).

[18] Caputo, P. & Sinclair, A. Entropy production in nonlinear recombination models. *Bernoulli* **24**, 3246–3282 (2018).

[19] Lu, S. *et al.* Chromosomal-level reference genome of chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience* **8**, giz128 (2019).

[20] Hahn, M. W. & Hibbins, M. S. A three-sample test for introgression. *Molecular Biology and Evolution* **36**, 2878–2882 (2019).

[21] Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**, 2239–2252 (2011).

[22] Rensch, B. *Evolution above the species level* (Columbia University Press, 1959).

[23] Waldvogel, A.-M. & Pfenninger, M. Temperature-dependence of spontaneous mutation rates. *Genome Research* gr–275168 (2021).

[24] Oertell, K. *et al.* Kinetic selection vs. free energy of DNA base pairing in control of polymerase fidelity. *Proceedings of the National Academy of Sciences* **113**, E2277–E2285 (2016).

[25] Wu, W.-J., Yang, W. & Tsai, M.-D. How DNA polymerases catalyse replication and repair with contrasting fidelity. *Nature Reviews Chemistry* **1**, 1–16 (2017).

[26] Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: An open-source release of Maxent. *Ecography* **40**, 887–893 (2017).

[27] Fick, S. E. & Hijmans, R. J. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315 (2017).

[28] Drost, J. B. & Lee, W. R. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among *Drosophila*, mouse, and human. *Environmental and Molecular Mutagenesis* **25**, 48–64 (1995).

[29] Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).

[30] Fu, L.-Y., Wang, G.-Z., Ma, B.-G. & Zhang, H.-Y. Exploring the common molecular basis for the universal DNA mutation bias: revival of Löwdin mutation model. *Biochemical and Biophysical Research Communications* **409**, 367–371 (2011).

[31] Cheng, K. C., Cahill, D. S., Kasai, H., Nishimura, S. & Loeb, L. A. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes GT and AC substitutions. *Journal of Biological Chemistry* **267**, 166–172 (1992).

[32] Oertell, K. *et al.* Transition state in DNA polymerase β catalysis: rate-limiting chemistry altered by base-pair configuration. *Biochemistry* **53**, 1842–1848 (2014).

[33] Gheorghiu, A., Coveney, P. & Arabi, A. The influence of base pair tautomerism on single point mutations in aqueous DNA. *Interface Focus* **10**, 20190120 (2020).

[34] Takasaki, H., Kawaguchi, N., Kuribayashi, T., Hasuo, R. & Kobayashi, S. Unusual successive occurrence of the Maackii Peacock (*Papilio maackii*; Papilionidae) at Okayama University of Science, southwestern Honshu lowland, in 2006 summer and autumn. *Naturalistae* 11–14 (2007).

[35] Via, S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 451–460 (2012).

[36] Schumer, M. *et al.* Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**, 656–660 (2018).

[37] Cheng, C. *et al.* Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* **190**, 1417–1432 (2012).

[38] Kirkpatrick, M. & Barton, N. H. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).

[39] Yeaman, S., Aeschbacher, S. & Bürger, R. The evolution of genomic islands by increased establishment probability of linked alleles. *Molecular Ecology* **25**, 2542–2558 (2016).

[40] Ravinet, M. *et al.* Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology* **30**, 1450–1477 (2017).

[41] Tomoko, O. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution* **40**, 56–63 (1995).

[42] Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences* **102**, 140–145 (2005).

[43] Wright, S., Keeling, J. & Gillman, L. The road from Santa Rosalia: a faster tempo of evolution in tropical climates. *Proceedings of the National Academy of Sciences* **103**, 7718–7722 (2006).

[44] Lin, G. *et al.* Evolutionary rates of bumblebee genomes are faster at lower elevations. *Molecular Biology and Evolution* **36**, 1215–1219 (2019).

[45] Chu, X.-L. *et al.* Temperature responses of mutation rate and mutational spectrum in an *Escherichia coli* strain and the correlation with metabolic rate. *BMC Evolutionary Biology* **18**, 1–8 (2018).

[46] Belfield, E. J. *et al.* Thermal stress accelerates *Arabidopsis thaliana* mutation rate. *Genome Research* **31**, 40–50 (2021).

[47] Miyabe, I., Kunkel, T. A. & Carr, A. M. The major roles of DNA polymerases epsilon and delta at the eukaryotic replication fork are evolutionarily conserved. *PLoS Genetics* **7**, e1002407 (2011).

[48] Hansen, M. C. *et al.* High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–853 (2013).

[49] Keightley, P. D. *et al.* Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution* **32**, 239–243 (2015).

[50] Dierckxsens, N., Mardulyn, P. & Smits, G. Novoplasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research* **45**, e18–e18 (2017).