PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

The power of coalescent methods for inferring recent and ancient gene flow in endangered Bactrian camels

Tianqi Zhu[a], Zhen Wang[b,c], and Ziheng Yang[a,1]

Edited by Douglas Futuyma, Stony Brook University, Stony Brook, NY; received July 17, 2024; accepted January 9, 2025

Genomic sequence data harbor valuable information concerning the history of species divergence and interspecific gene flow and may offer important insights into conservation of endangered species. However, extracting such information from genomic data requires powerful statistical inference methods. A recent analysis of genomic sequence data found little evidence for gene flow from domestic Bactrian camels into the endangered wild Bactrian species. Nevertheless, the methods used to infer gene flow are based on data summaries and lack the power and precision to represent the complex phylogenetic history of the species with gene flow. Here, we apply Bayesian methods to genomic sequence data to test for both recent and ancient gene flow among the three species in the genus *Camelus* and to estimate the strength and timing of gene flow. We detect a strong signal of gene flow from domestic into wild Bactrian camels, confirming early evidence based on mitochondrial DNA and the Y chromosome. Overall gene flow appears to affect the autosomal genome uniformly, with similar effective rates of gene flow for exonic and noncoding regions. Estimation of species divergence times is seriously affected if gene flow is not accommodated in the analysis. Our results highlight the power of the coalescent model in analysis of genomic data and the utility of the coding as well as noncoding parts of the genome in elucidating the evolutionary history of modern species.

BPP | gene flow | genomics | introgression | camels

Significance

We analyzed genomic data from three species of camels to infer gene flow between species and to estimate species divergence times. In contrast to an earlier analysis, we found strong evidence for gene flow from the domestic two-humped Bactrian camel to the wild species, raising concerns about the impact of hybridization on the genetic integrity of the wild species. We detected gene flow between the domestic dromedaries and Bactrians. The results are consistent between coding and noncoding parts of the genome, and between models of gene flow. Our study highlights the power of the analysis of gene flow for assessing the persistence of modern species in fragmented habitats.

Since their domestications about 3,000 to 6,000 y ago, camels of the Old World (as well as the New World camels, llamas, and alpacas) have played critically important roles in multiple civilizations, transporting goods and people across continents, and providing milk, meat, wool, and draft (1, 2). The phylogenetic relationships of the camel species, their divergence times, history of domestications, and the genetic basis of their physiological adaptations have long been the focus of research (2–5). In today's world affected by climate change and desertification, camels have interested animal breeders and scientists as they provide sustainable milk and meat production (1). Studies of the origin and evolution of camels are thus not only important to our understanding of the past history of those iconic species but also to our future survival and well-being. There is an urgent need to conserve the critically endangered wild species, in addition to maintaining the genetic diversity in the two domestic species with different physiological adaptations to desert environments (1).

The ancestors of *Camelus* lived in the North American continent and split into New World (*Lamini*) and Old World (*Camelini*) camels ~16.3 ma (3, 6) (Fig. 1). Then the Camelini camels migrated via the Bering land bridge to the Old World, while the ancestors of llamas and alpacas spread to South America. There are two domestic species of camels in the Old World (*Camelus, Camelini*): the one-humped dromedary (*Camelus dromedarius*), found in the arid deserts of North Africa, East Africa, and the Arabian Peninsula, and the two-humped Bactrian camel (*Camelus bactrianus*), distributed in the cold desert areas of Northeast and Central Asia (7). The dromedaries and Bactrian camels split around 4.4 to 8 Ma (3, 6). There is also a wild Bactrian camel species (*Camelus ferus*) (8), with a split time from the domestic Bactrian species around 0.13 to 0.73Ma (9) (Fig. 1). This is long before any domestication event, suggesting that Bactrian camels were domesticated from a different wild population that is now extinct (9). Historically, the wild Bactrian camel was widely distributed throughout Asia, extending from the great bend of the Yellow River westward to central Kazakhstan (10), but today it is found only in the Mongolian Gobi desert, and Taklimakan and Lop Noor deserts in China. The number of *C. ferus* camels is estimated to be between a few hundred and 2,000 (11, 12). It is critically endangered due to habitat loss (13), and hybridization with escaped domestic camels (*C. bactrianus*) poses further threats to its genetic integrity. The wild dromedaries

Author affiliations: [a]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; [b]Shanghai Institute of Nutrition and Health, Key Laboratory of Computational Biology, Chinese Academy of Sciences, Shanghai 200031, China; [c]Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences; and [d]Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, United Kingdom

Author contributions: T.Z. performed research; T.Z. and Z.W. analyzed data; Z.W. contributed new reagents/analytic tools; Z.Y. designed research; Z.Y. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: z.yang@ucl.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2410949122/-/DCSupplemental.

Published XXXX.

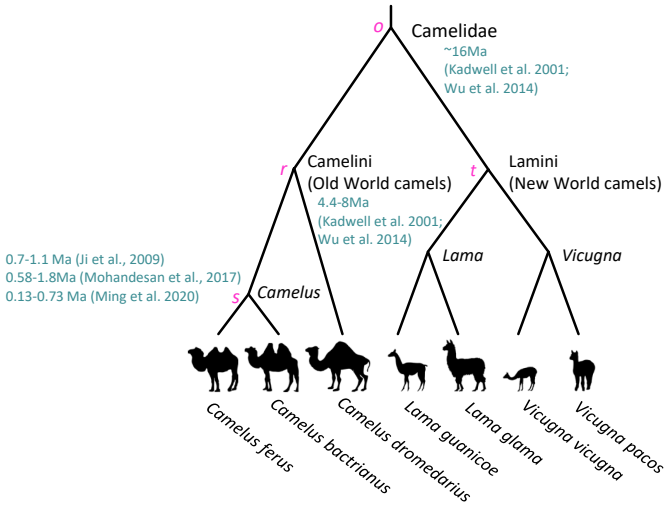

Fig. 1. The phylogeny of Camelidae, with rough estimates of split times. Note that molecular clock estimates of split times rely on assumed mutation rate and generation time, and involve large uncertainties (14). Branches are not drawn to scale. *C. ferus* and *C. bactrianus* Silhouette images are from Z.W., while other images are from https://www.phylopic.org/nodes/%54d3efe4-2dc0-4531-b641-1226cacb82a6/lama-glama-silhouettes .

became extinct <1,000 y ago, after the domestic species *C. dromedarius* appeared (3,000 to 4,000 y ago), contributing to its domestication with multiple introgressions (4).

In the New World, there are two domesticated species: llamas (*Lama glama*) and alpacas (*Vicugna pacos*), as well as their wild relatives: *Lama guanicoe* and *Vicugna vicugna* (Fig. 1). Far less genetic or genomic data have been generated for those species than for the Old-World camels (3).

Because of the paucity of fossils and the challenges to distinguish remains or bones from the wild and domestic Bactrian camels, and between the Bactrians and dromedaries (and their hybrids), inference of the demographic history of the camels has mostly relied on molecular data including ancient DNA. Indeed genomic data provide a rich source of information concerning the order and times of species splits and the presence and strength of gene flow between the extant and extinct camel species. Recently Ming et al. (9) generated genomic sequence data from the three species in the genus *Camelus*, and used the intergenic and intronic regions of the genome to infer the history of species divergence, domestication, and interspecific gene flow. A number of methods, including the *D*-statistic (15), ADMIXTURE (16), TREEMIX (17), and G-PHOCS (18) were used to infer gene flow. They found significant evidence for gene flow between the domestic dromedary and the Bactrian species, but surprisingly no unequivocal evidence for gene flow between the wild and domestic Bactrian camels. However, recent studies have demonstrated that the power to detect gene flow from genomic sequence data depends on the analytical methods used (19, 20). While G-PHOCS is a full-likelihood method, it had MCMC mixing issues and did not produce usable results from the dataset of ref. 9. ADMIXTURE and TREEMIX are not based on an explicit population genetic model of population divergence and gene flow, so that the parameter estimates from those methods may be hard to interpret.

While gene flow is often inferred using simple methods based on summaries of the genomic data (21), they lack the power and precision to estimate parameters that characterize the complex history of species divergence and hybridization/introgression. For example, most summary methods are unable to identify gene flow between sister lineages, or to infer the strength, direction, and timing of gene flow (20, 21). The past few years have seen significant progress in implementing and extending the multispecies coalescent (MSC) model (22) to accommodate gene flow (21, 23, 24). For example, two MSC models of gene flow have been implemented in the Bayesian Markov chain Monte Carlo (MCMC) program BPP (23, 24). The MSC-introgression (MSC-I) model assumes that gene flow occurs at a certain time point in the past (23, 25, 26), while the MSC-migration (MSC-M) model assumes continuous gene flow over extended time periods (24, 27). While the BPP program typically involves orders of magnitude more computation than summary methods, it has been successfully applied to datasets with over 10,000 loci albeit for a small number of species (19, 24).

Here, we reprocess the genomic data of ref. 9 to compile a multilocus dataset of noncoding regions and use BPP to analyze the data. We also compile and analyze an exonic dataset, to address the question whether gene flow has affected the coding and noncoding parts of the genome differently and to assess the utility of coding DNA as genetic markers for inferring the history of species divergence and gene flow.

Results

We processed the genomic data of ref. 9 to compile 10,000 noncoding segments (each of 1,000 base pairs) for four domestic Bactrian camels (*C. bactrianus*), two wild Bactrian camels (*C. ferus*), and two domestic dromedary camels (*C. dromedarius*) (Fig. 2*A* and *SI Appendix*, Table S1). We used the noncoding data to test for the presence or absence of gene flow among the three species and to construct a model of gene flow. We separated the 10,000 noncoding loci into four random subsets, each of 2,500 loci, and analyzed them as separate datasets using BPP, to reduce the computational load and to assess consistency among data subsets. Both the MSC-I and MSC-M models were used, assuming six, five, or four gene-flow events, with the five-rates and four-rates models excluding ancestral gene flow between the common ancestor of the two-humped camels (*C. bactrianus* and *C. ferus*) and the dromedaries (*C. dromedarius*) (Fig. 2*A*). Estimates of parameters including the rates of gene flow are summarized in Fig. 2*C* and *SI Appendix*, Table S2, while the results from Bayesian test of gene flow (20) are presented in *SI Appendix*, Table S3 and Table 1.

We then analyzed exonic data under the same models of gene flow (Fig. 2 *A* and *B*) to examine whether exonic data support the same gene-flow events as the noncoding parts of the genome and whether the two types of data produce similar estimates of parameters including the rates of gene flow.

We Detected Strong Evidence for Gene Flow from the Domestic Bactrian Camels to the Wild Species. The introgression probability was estimated to be $\varphi_{xy} = 12$ to 14% among the four subsets of noncoding data under the MSC-I model with six-, five-, or four-rates (Fig. 2 and *SI Appendix*, Table S2). Here, φ_{xy} is defined as the proportion of immigrants in the recipient species y from the donor species x. Under the continuous migration model (MSC-M), estimates of population migration rate M_{xy} were 0.100 to 0.151 under the six- and five-rates models (Fig. 2 and *SI Appendix*, Table S2), where $M_{xy} = m_{xy}N_y$ is measured in the expected number of migrants from species x to y per generation, with m_{xy} to be the proportion of immigrants in species y from x and N_y is the (effective) population size of species y. The results are consistent between runs and among the four data subsets (Fig. 2*C*). Bayesian test rejected the null model

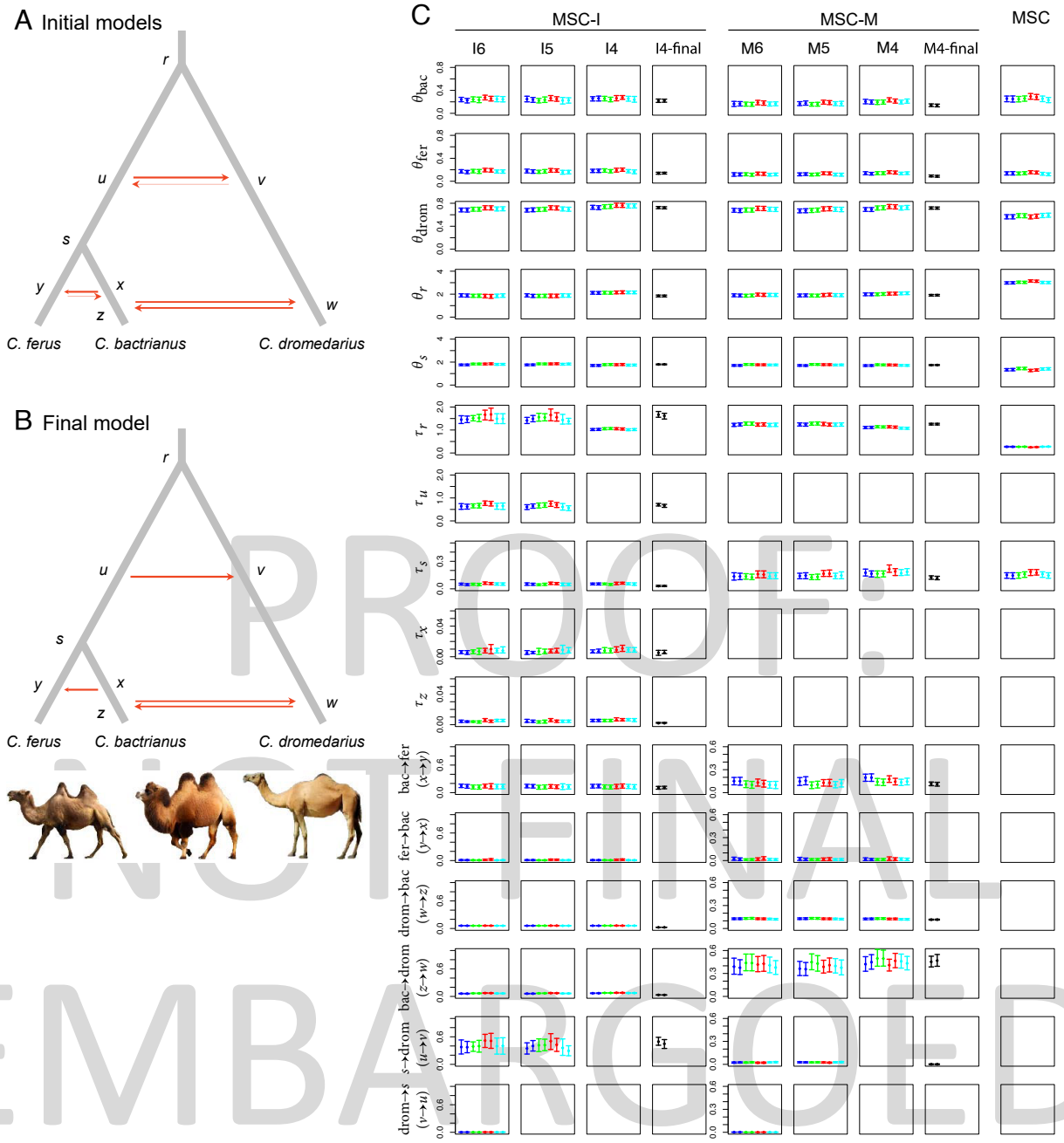

Fig. 2. (*A*) Species phylogeny with six gene-flow events for three species of camels in the genus *Camelus*. Gene-flow events are represented by horizontal lines and the two events not supported by the data are shown in thin lines. Gene flow is modeled using either a one-off major event in the MSC-I model or continuous migration in the MSC-M model. Five-rates and four-rates models are considered as well, in which the ancestral drom → *s* (*v* → *u*) gene-flow event or both the drom → *s* and *s* → drom events (*u* → *v*) are excluded. (*B*) Final model with four gene-flow events based on BPP analysis of the noncoding data. (*C*) Posterior means and 95% highest-probability-density (HPD) credibility intervals (CIs) for parameters under the MSC-I and MSC-M models of panels *A* and *B*, and under the MSC model with no gene flow in BPP analyses of the four subsets of noncoding data and of the full data of 10^4 noncoding loci. The columns represent different models: MSC-I with 6, 5, and 4 gene-flow events of panel *A* (I6, I5, I4), MSC-M with 6, 5, and 4 gene-flow events (M6, M5, M4), the final model of panel *B*, and MSC with no gene flow. The last six rows show the rates of gene flow (φ in MSC-I or $M = Nm$ in MSC-M). Each analysis is run twice, so that each of the I6–I4 and M6–M4 panels has eight sets of results (four data subsets and two runs). Parameters τ and θ are multiplied by 10^3. Note the different scales for divergence times (τ).

of no gene flow at the 1% level (with $2\ln \mathrm{BF}_{10} > 100$) in all four subsets under both the MSC-I and MSC-M models (Table 1 and *SI Appendix*, Table S3).

The result is consistent with previous analyses of mitochondrial DNA (28) and Y chromosome (29), which identified introgression from the domestic Bactrian species into the wild species, and considered it as a major threat to the genomic integrity and evolutionary independence of the wild species (1, 30). In the analysis of Ming et al. (9), TREEMIX detected no significant signal for migration between the wild and domestic Bactrian camels, but the residues from model fitting were indicative of some admixture between the wild and East Asian Bactrian populations. The ADMIXTURE analysis detected domestic Bactrian ancestry in several wild individuals with a proportion of 7 to 15%. However, such admixture proportions may not be easy to interpret as the apparent admixture may be due to ancestral polymorphism

Table 1. Summary of Bayesian tests of gene flow in BPP analyses of the noncoding data

	BPP									
Gene flow	MSC-I6	MSC-I5	MSC-I4	MSC-I4-final	MSC-M6	MSC-M5	MSC-M4	MSC-M4-final	ADMIXTURE†	TREEMIX†
bac → fer ($x \rightarrow y$)	++	++	++	++	++	++	++	++	? (7 to 15%)	–?
fer → bac ($y \rightarrow x$)	–	–	–		–	–	–		–	–
drom → bac ($w \rightarrow z$)	++	++	++	++	++	++	++	++	+ (1 to 10%)	+ (4 to 9%)
bac → drom ($z \rightarrow w$)	++	++	++	++	++	++	++	++	+	
s → drom ($u \rightarrow v$)	++	++		++	++	++		++		
drom → s ($v \rightarrow u$)	–				–					

Key to symbols:
"––": strong rejection of gene flow ($B_{10} < 0.01$);
"–": weak rejection of gene flow ($0.01 < B_{10} < 1$);
"+": weak support for gene flow ($1 < B_{10} < 100$);
"++": strong support for gene flow ($B_{10} > 100$);
"?": ambiguous result;
empty: no result.
†Proportions of ancestral contribution for ADMIXTURE and TREEMIX are from ref. 9.

(incomplete lineage sorting) as well as introgressive hybridization. Here, the MSC-I and MSC-M models implemented in BPP explicitly accommodate gene flow and test directly for evidence of gene flow while taking into account ancestral polymorphism.

Note that the introgression probability φ_{xy} in the MSC-I model is also the probability that any sequence entering species y is traced to species x (at the time of introgression) when one traces the history of the sampled sequences backward in time. In the MSC-M model, one can calculate a similar probability φ_0, defined as the probability that any sequence entering species y is traced to species x (irrespective of the time of migration) when one traces the history of the sampled sequences backward in time (31). In effect φ_0 measures the total amount of gene flow expected under the MSC-M model, comparable to φ under MSC-I. Let $M_{xy} = m_{xy}N_y$ be the population migration rate under MSC-M and $\Delta\tau = \Delta T\mu$ be the time duration of migration in mutational units while ΔT is the time duration in generations, with μ to be the mutation rate per site per generation. Then

$$\varphi_0 = 1 - e^{-m_{xy}\Delta T} = 1 - e^{-4M_{xy}\Delta\tau/\theta_y}, \quad [1]$$

where $\theta_y = 4N_y\mu$ is the mutation-scaled population size parameter for species y (equation 10 in ref. 31; see figure 1e in ref. 31 for example calculations).

The probability of introgression predicted under MSC-M (φ_0) was close to the estimated introgression probability φ under MSC-I (Fig. 3*A*), suggesting that the two models recovered about the same amount of gene flow from the domestic *C. bactrianus* to the wild *C. ferus* species (bac → fer or $x \rightarrow y$).

There Was no Evidence for Gene Flow from the Wild Bactrian Camels to the Domestic Species. Under the MSC-I model, the estimated introgression probability φ_{yx} were low, at ≈1% (Fig. 2*C* and *SI Appendix*, Table S2). The Bayes factor was in the range $0.01 < B_{10} < 1$ in all four data subsets (*SI Appendix*, Table S3), so that the data favored the model of no $y \rightarrow x$ introgression, even though the evidence was not significant at the 1% level.

Under the MSC-M model, the estimated migration rate M_{yx} were around 0.01 to 0.02 migrants per generation (*SI Appendix*, Table S2). The Bayes factor B_{10} was in general much smaller than 1, being <0.05 in some data subsets (*SI Appendix*, Table S3). There was even stronger rejection of gene flow than under MSC-I.

Table 2. Number (out of 48) of coding data subsets (each of 2,500 exons) in which the Bayes factor B_{10} rejects or supports gene flow under the final model of Fig. 2*B*

	MSC-I (I4-final)				MSC-M (M4-final)			
Gene flow	<0.01 ––	(0.01, 1) –	(1, 100) +	>100 ++	<0.01 ––	(0.01, 1) –	(1, 100) +	>100 ++
bac → fer ($x \rightarrow y$)	0	15	11	22	0	16	11	21
drom → bac ($w \rightarrow z$)	0	0	1	47	0	0	48	0
bac → drom ($z \rightarrow w$)	2	7	1	38	0	0	45	3
s → drom ($u \rightarrow v$)	0	6	14	28	5	30	5	8

Note: See legend to Table 1.

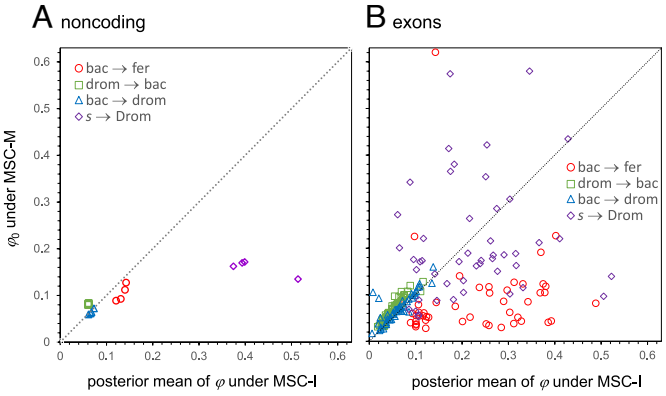

Fig. 3. Estimates of introgression probability (φ) under the final MSC-I model with four rates (Fig. 2*B*) and the predicted probability (φ_0) under the corresponding MSC-M model, calculated using Eq. **1** with the posterior means of parameters under MSC-M for (*A*) the noncoding and (*B*) the exonic data. In (*A*), there are four points for each introgression probability corresponding to four noncoding data subsets (each of 2,500 noncoding loci), while in (*B*) there are 48 exonic subsets.

This result is consistent with Ming et al. (9), who found using ADMIXTURE that the wild camels made nearly no contribution to the ancestry of domestic populations.

There Was Strong Evidence for Gene Flow between Domestic Bactrian and Dromedary Camels in Both Directions. The introgression probability under the MSC-I model was estimated to be ~ 6% from dromedaries to the Bactrians ($w \to z$) and ~7% in the opposite direction ($z \to w$) (Fig. 2*C* and *SI Appendix*, Table S2). Under the MSC-M model, the estimated migration rate (M) was 0.12 to 0.13 migrants per generation from dromedaries to the Bactrians and 0.36 to 0.50 in the opposite direction. The large differences in the estimates reflect the fact that dromedaries has a larger population size than the Bactrians ($\theta_{\text{drom}} > \theta_{\text{bac}}$, Fig. 2*C*) so that nearly equal proportions of migrants (φ) mean different numbers of migrants ($M = Nm$). Calculations using Eq. **1** suggest that the MSC-I and MSC-M models estimate similar amounts of gene flow between the two species (Fig. 3*A*).

The bac → drom (or $z \to w$) rates were slightly higher under the 4-rates model than under the 6-rates model (*SI Appendix*, Table S2). This may be due to the failure of the four-rates model to accommodate the ancient gene-flow event from the common ancestor of wild and domestic Bactrian camels to the dromedaries ($u \to v$, Fig. 2*A*) (see below).

Gene flow between the two domestic species of camels has been detected in previous studies, for example in the analysis of ref. 9 using ADMIXTURE and TREEMIX, with mixture proportions estimated to be in the range 1 to 10% (Table 1). The result is consistent with the well-known history of hybridization between the two species of camels in breeding practice, especially in Central Asia (1, 32).

There Was Strong Evidence for Gene Flow from the Bactrian Ancestor to the Dromedary, But No Evidence for Gene Flow in the Opposite Direction. The estimated introgression probability for the $s \to$ drom ($u \to v$) gene flow under the MSC-I model was in the range $\varphi_{uv} \approx 0.33$ to 0.51 (Fig. 2*C* and *SI Appendix*, Table S2). The migration rate under the MSC-M model (M_{uv}) was in the range 0.021 to 0.028 migrants per generation (*SI Appendix*, Table S2 and Fig. 2). While the rate may appear to be low, it has had a major impact on the genetic history of the species involved: because of the extended time period for migration, $\tau_r - \tau_s$ (Fig. 2*A*), the low rate per generation translates into a substantial proportion of the dromedary ancestry being traced back into the ancestral Bactrian lineage. Nevertheless, the MSC-I model recovered a much larger amount of gene flow than the MSC-M model, with $\hat{\varphi}^{(I)}_0 \gg \hat{\varphi}^{(M)}_0$ (Fig. 3*A*). One possible explanation for this discrepancy is that gene flow occurred at variable rates over time, and overall the MSC-I model achieved a better fit to the genomic data than the MSC-M model. Note that both MSC-I and MSC-M have the simple MSC model with no gene flow as a special case (with $\varphi_{uv} = 0$ for MSC-I and $M_{uv} = 0$ for MSC-M), so that a larger rate of gene flow also means better fit to the data.

The $s \to$ drom (or $u \to v$) gene flow involves two ancient sister lineages and does not appear to have been detected before. Note that most summary methods including the *D*-statistic are unable to detect gene flow between sister lineages (33).

In contrast to the strong evidence for gene flow from the Bactrian ancestor to the dromedaries ($s \to$ drom or $u \to v$), there was no evidence in support for gene flow in the opposite direction (drom $\to s$ or $v \to u$, Fig. 2*A*). The Bayes factor in support of gene flow B_{10} was much less than 1, at ≈0.09 under MSC-I and ≈0.018 under MSC-M, rejecting gene flow at the 10% and 5% levels, respectively (*SI Appendix*, Table S3). Note that the Bayesian test can lead to rejection of the alternative hypothesis, whereas in traditional hypothesis testing one may fail to support the alternative hypothesis but never reject it with great force. Consistently with the test, the estimated rates of drom $\to s$ ($v \to u$) gene flow were nearly zero under both the MSC-I and MSC-M models (*SI Appendix*, Table S2).

We Derived a Model of Gene Flow with Four Gene-Flow Events for the Three Camel Species. In summary, our analyses of the noncoding loci produced highly consistent results among the four data subsets, and between the MSC-I and MSC-M models although these make very different assumptions about the mode of gene flow. The consistency lends confidence in the inferred gene-flow events. There was also consistency among the six-, five-, and four-rates models of Fig. 2*A*, although the four-rates models produced slightly biased estimates of certain parameters due to its failure to accommodate the strong signal of gene flow from the Bactrian ancestor to the dromedaries ($s \to$ drom). Our analyses suggest that 2,500 noncoding loci may be informative enough to produce parameter estimates precise enough to draw useful biological conclusions.

Based on the parameter estimates from the noncoding data and the Bayesian test under the MSC-I and MSC-M models, we formulated a model of gene flow for the three species in the genus *Camelus*, with four gene-flow events (Fig. 2*B*). We applied this final model to the four subsets of noncoding data as well as the full dataset of 10,000 noncoding loci (Table 3 and Fig. 2*C*). The six-rates and five-rates models (Fig. 2*A* and *SI Appendix*, Table S2) include the final model with four rates as a special case and are thus overparameterized. The similarity in parameter estimates obtained for the data subsets under those models suggests that the cost of overparameterization was mostly computational while the biological results were essentially the same (34). The estimates under the final model from the full noncoding data (Table 3) were very similar to the estimates from the data subsets but with much narrower credibility intervals.

Table 3. Posterior means (with 95% HPD CIs in parentheses) of rates of gene flow (φ in MSC-I and $M = Nm$ in MSC-M) under the final model of Fig. 2*B* in BPP analyses of the noncoding and exons data

data	bac → fer ($x \to y$)	drom → bac ($w \to z$)	bac → drom ($z \to w$)	s → drom ($u \to v$)
MSC-I model (I4-final)				
Noncoding quarter 1	0.145 (0.108, 0.185)	0.061 (0.056, 0.065)	0.061 (0.054, 0.068)	0.344 (0.209, 0.490)
Noncoding quarter 2	0.122 (0.075, 0.171)	0.061 (0.056, 0.065)	0.068 (0.061, 0.075)	0.416 (0.289, 0.548)
Noncoding quarter 3	0.151 (0.099, 0.208)	0.061 (0.057, 0.066)	0.073 (0.066, 0.080)	0.515 (0.347, 0.659)
Noncoding quarter 4	0.134 (0.078, 0.192)	0.060 (0.055, 0.065)	0.065 (0.058, 0.072)	0.358 (0.216, 0.521)
Noncoding full data (10^4 loci)	0.150 (0.111, 0.164)	0.061 (0.058, 0.063)	0.067 (0.063, 0.070)	0.490 (0.390, 0.559)
Exons (averages)	0.227 (0.074, 0.399)	0.052 (0.044, 0.061)	0.062 (0.047, 0.078)	0.222 (0.047, 0.409)
MSC-M model (M4-final)				
Noncoding quarter 1	0.171 (0.119, 0.223)	0.130 (0.119, 0.141)	0.364 (0.275, 0.462)	0.024 (0.016, 0.032)
Noncoding quarter 2	0.113 (0.068, 0.160)	0.132 (0.121, 0.144)	0.446 (0.335, 0.559)	0.024 (0.017, 0.032)
Noncoding quarter 3	0.134 (0.089, 0.178)	0.127 (0.116, 0.138)	0.413 (0.322, 0.511)	0.023 (0.015, 0.032)
Noncoding quarter 4	0.114 (0.070, 0.159)	0.124 (0.114, 0.135)	0.384 (0.293, 0.487)	0.028 (0.020, 0.036)
Noncoding full data (10^4 loci)	0.126 (0.103, 0.149)	0.129 (0.124, 0.135)	0.449 (0.383, 0.519)	0.024 (0.020, 0.028)
Exons (averages)	0.095 (0.022, 0.176)	0.110 (0.091, 0.131)	0.191 (0.117, 0.274)	0.020 (0.007, 0.034)

Note: Estimates for exons are averages over the 48 subsets (each of 2,500 exons) of Fig. 4.

The Exonic Data Supported the Same Gene-Flow Events as the Noncoding DNA and Produced Similar Estimates of Rates of Gene Flow. We compiled 120,720 exons and separated them into 48 subsets, each of 2,500 loci. We merged the first two subsets to form a dataset of 5,000 loci and applied the 6-rates models of Fig. 2*A* (both MSC-I and MSC-M) to see whether the exonic data support the same gene-flow events as the noncoding data (*SI Appendix*, Table S4).

Under both MSC-I and MSC-M and for both exonic and noncoding data, there was weak rejection of the drom → s (or $v \to u$) and fer → bac (or $y \to x$) gene-flow events. Under MSC-I, the four rates in the final model of Fig. 2*B* were strongly supported by the exonic data (*SI Appendix*, Table S4). Under MSC-M, three of the four rates in the final model of Fig. 2*B* were strongly supported but the ancient gene-flow event from $s \to$ drom was weakly rejected, whereas it was strongly supported in the noncoding data. The estimated migration rate $M_{s\to \mathrm{Drom}}$ was small for both exonic data (0.017 with 95% HPD CI 0.005 to 0.029) and noncoding data (0.024, 0.020 to 0.028) (*SI Appendix*, Tables S2 and S4). Note that weak rejection and weak support are indecisive test results and may reflect low information content in the data. We conclude that the exonic and noncoding data were largely consistent concerning gene flow on the *Camelus* phylogeny.

We then fitted the final model of four gene-flow events of Fig. 2*B* to the 48 subsets of exonic data, each of 2,500 exons. The estimates are summarized in Fig. 4 and Table 3, with results for Bayesian testing summarized in Table 2.

The exonic data produced estimates of rates of gene flow between the two domestic species (drom → bac or $w \to z$, and bac → drom or $z \to w$, Table 3) similar to those from the noncoding data, with $\varphi_{wz} = 5\%$ and $\varphi_{zw} = 6\%$ under MSC-I for the exonic data, with the corresponding estimates to be ≈6.1% and 6 to 7% for noncoding data (Table 3).

The rate of gene flow from the domestic *C. bactrianus* to the wild *C. ferus* estimated from the exonic data varied considerably among the subsets or across the genome (Fig. 4), and furthermore, the MSC-M model in general predicted much less gene flow than the MSC-I model, with $\varphi_0 < \varphi$ (Fig. 3*B*). The CIs were much wider than for coding data, reflecting a lower information content in each subset of 2,500 exons. Overall, the exonic rate of gene flow for *C. bactrianus* → *C. ferus* was higher than for the noncoding data, with the average estimates at $\varphi_{xy} \approx 0.227$ for exons, in comparison with 0.12 to 0.15 for noncoding data (Table 3).

The reasons for this difference are unclear, but there is the intriguing possibility that the introgressed exonic alleles might be adaptive and their spread in the recipient *C. ferus* population might be accelerated by natural selection. Note that the rates of gene flow (both φ in MSC-I and M in MSC-M) estimated from genomic data are effective rates, reflecting the long-term effects of natural selection, drift, as well as introgressive hybridization (35). In contrast to the $x \to y$ rates, exonic data produced lower rates of ancestral gene flow ($s \to$ drom) than noncoding data, with $\varphi_{s\to \mathrm{drom}} = 0.222$ for exons vs. 0.34 to 0.52 for noncoding DNA (Table 3). Overall, the differences were small, and the exonic data produced similar estimates of rates of gene flow under the MSC-I and MSC-M models to those obtained from the noncoding data.

Finally, we note that the Bayes factor for gene flow was greater under MSC-I than under MSC-M, with $B_{10}^{(i)} > B_{10}^{(m)}$. Because $B_{10}^{(i)} = M_1^{(i)}/M_0$ and $B_{10}^{(m)} = M_1^{(m)}/M_0$, where M_0 and M_1 are the marginal likelihood values under the null model of no gene flow (H_0) and under the alternative model of gene flow (H_1), and the null model in the two tests is the same, we have

Fig. 4. Posterior means and 95% HPD CIs for parameters under the final MSC-I and MSC-M models with four gene-flow events (Fig. 2*B*) obtained from BPP analyses of 48 exonic data subsets (each of 2,500 exons). Each analysis is run twice, shown in the same color. The last four rows show the rates of gene flow (φ in MSC-I or $M = Nm$ in MSC-M). Parameters τ and θ are multiplied by 10^3. The same scales for the *y*-axis are used as in Fig. 2*C*.

$M_1^{(i)} > M_1^{(m)}$, and the MSC-I model fitted the exonic data better than the MSC-M model and was more powerful in inferring gene flow (Table 2). The same pattern was observed in the analyses of the noncoding data.

Estimation of Species Split Times May Be Seriously Biased if Gene Flow Is Not Accommodated. Estimates of species split times and population sizes measured in mutations (τs and θs) may be converted to absolute geological times and absolute population sizes by assuming a mutation rate and a generation time (9, 30). This is not pursued here as the uncertainties in mutation rate and generation time are hard to assess. Instead, we focus on estimation of the mutation-scaled split times and population sizes.

First, estimates of species split times (τ_r, τ_s) and population sizes (θ_{bac}, θ_{fer}, θ_{drom}, θ_r, θ_s) for the noncoding data tended to be larger than those from the exonic data, and to be more precise with narrower intervals (Figs. 2*C*, I4-final & M4-final, and 4). The larger estimates for the noncoding data reflect the higher mutation rate than for exonic data (Fig. 5). The smaller uncertainties in the estimates for the noncoding data reflect higher information content: as the species are closely related and the sequences are very similar, a high mutation rate means high information content so that an average noncoding locus is more informative than an average exonic locus.

For the noncoding data, the estimated age of the root (τ_r) was greater under the MSC-I model than under MSC-M (Fig. 2*C*). For the exonic data, the estimates were similar between the two models (Fig. 4). We suggest that such variations reflect the strong correlation between the estimated amount of gene flow ($u \to v$) and the species split time (τ_r). In particular, the MSC model assuming no gene flow produced much smaller estimates of τ_r, misinterpreting gene-flow events between the sister lineages as recent species divergence. This pattern is consistent with previous studies which found that ignoring gene flow leads to serious underestimation of species divergence times (36, 37).

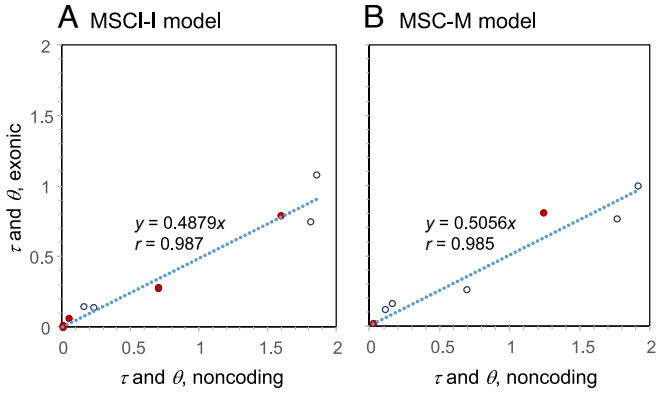

Fig. 5. Average estimates of species divergence times (τ, $\times 10^3$, red dots) and population sizes (θ, $\times 10^3$, black circles) obtained in BPP analyses of the exonic data plotted against estimates from the noncoding data under (*A*) the MSC-I and (*B*) the MSC-M models (Fig. 2*B*). (*A*) Under the MSC-I model, there are five θs and five τs, while (*B*) under MSC-M, there are five θs and two τs. Estimates for noncoding data are averages of posterior means over the four subsets (each of 2,500 noncoding loci) (Fig. 2*C*). Estimates for exonic data are averages of posterior means over the 48 subsets (each of 2,500 exons) (Fig. 4).

For the noncoding data, the estimated split time between *C. bactrianus* and *C. ferus* (τ_s) was smaller under the MSC-I model than under MSC-M (Fig. 2*C*), while for the exonic data, the opposite was true (Fig. 4). It is likely that gene flow between those two species has been ongoing over extended time periods so that the MSC-M model provides a better fit to the data than the MSC-I model. Note that τ_s does not represent the time of domestication, and instead represents the split time from the wild species of an extinct ancestral lineage that was later domesticated.

Effective population sizes vary greatly among species on the phylogeny (Figs. 2*C* and 4). The wild species *C. ferus* had the smallest size, while *C. dromedarius* had a larger size than *C. bactrianus*. Those extant species had much smaller populations than the ancestral lineages: $\theta_{bac}, \theta_{drom}, \theta_{fer} \ll \theta_r, \theta_s$. The results are consistent with genome-wide summaries published earlier (9). Note that the θ estimates for the extant species are smaller than the heterozygosity (π) or genetic diversity calculated based on single genomes (9, 30), because heterozygosity in the genomic data reflects the large population size in the distant past as well the small size in recent history.

The Mitochondrial Gene Tree Is Monophyletic for Each Species but the Data Are Not Informative About Ancestral Polymorphism. The maximum likelihood tree for mitochondrial genomes (*SI Appendix*, Fig. S2*A*) showed monophyly of sequences from each of the three extant species, reflecting their small population sizes. We estimated species split times on the species tree under the MSC model with no gene flow using BPP, treating the mitochondrial genome as one locus (*SI Appendix*, Fig. S2*B*). The 95% HPD intervals are wide, reflecting low information content in the data. Because the species split times on the species tree are drastically different (with $\tau_r \gg \tau_s$) and the extant species had small sizes, sequences from the same species coalesce very quickly, so that very likely only two sequences enter the root population in the species tree, and τ_o and θ_o are seriously confounded.

Discussion

We Constructed a Model of Species Divergence and Introgressive Hybridization for the Three Species of Old World Camels. Through testing for the presence of gene flow using the Bayesian test and estimating the rates of gene flow, we have constructed a model of species divergence and interspecific gene flow for the Old-World camels (Fig. 2*B*). Here, we review evidence for the inferred gene-flow events. Our model posits prevalent gene flow between the two domestic camel species (*C. dromedarius* ⇆ *C. bactrianus*). Human-aided hybridization between the two domestic camel species is practiced along multiple long-distance trade routes, to produce animals of great strength, with the robustness of the Bactrian, the endurance of the dromedary, tolerance of different climatic conditions, and improved meat and wool production (1).

Our model also posits strong gene flow from the domestic *C. bactrianus*) to the wild species (*C. ferus*). While the analysis of genomic data by Ming et al. (9) produced equivocal results concerning this gene flow, our result is consistent with previous analyses of mitochondrial and Y-chromosome data (28, 29). Interestingly, our analysis rejects gene flow in the opposite direction, from the wild *C. ferus* to domestic *C. bactrianus*. Hybrids in domestic Bactrians do occur but are rejected by farmers because of certain undesirable traits including ill temper, so that they do not contribute to the genetic makeup of the domestic Bactrians (38).

We inferred ancient gene flow from the common ancestor of the two-humped camels (*C. bactrianus* and *C. ferus*) to the dromedary species ($s \rightarrow$ drom), but no evidence of gene flow in the opposite direction. This gene-flow event does not appear to be suggested before, possibly because gene flow between sister lineages is unidentifiable by most summary methods.

We tested a model involving gene flow from the extinct wild dromedaries to the domestic dromedaries (*SI Appendix*, Fig. S1). However, with no sequence data from the extinct wild dromedary species, there is little information in the data to distinguish such a model of gene flow from a model of a large population size for the domestic dromedary species (large θ_{drom}). Inclusion of ancient genomes from the extinct dromedaries may shed light on this issue (4).

It has been suggested that spatial population structure generates signals in the genetic data that may be interpreted as gene flow between species (hybridization or admixture) by inferential methods ignoring population structure. Such methods may then be misled to infer gene flow between species when in reality the signal is due to population subdivision in the ancestral species (39, 40). We leave it to the future to assess the sensitivity of BPP inference of gene flow to ancestral population structure. Here, we discuss several lines of evidence suggesting that gene-flow events inferred in this paper (Fig. 2*B*) are plausible biologically and are unlikely to be artifacts due to ancestral population structure. First, there is abundant well-documented evidence for hybridization among the three species. For example, hybridization between the domesticated *C. bactrianus* and *C. dromedarius* is common as part of the breeding practice, and terms for the hybrids are used in many different languages (1). Gene flow from *C. bactrianus* into *C. ferus* may occur when escaped domestic camels hybridize with the wild species. Note that *C. bactrianus* and *C. ferus* are genetically very similar and can easily hybridize ($\tau_s \ll \tau_r$, Fig. 2).

Second, our analysis inferred gene flow from the domestic *C. bactrianus* to the wild *C. ferus* but rejected gene flow in the opposite direction. If the signal was due to population structure in the common ancestor of *C. bactrianus* and *C. ferus*, the signal should show up as gene flow in both directions. We suggest that the inferred gene-flow events are reliable, although our tests may miss certain gene-flow events.

It may also be noted that a model of ancestral population structure may be fitted to the genomic data using BPP and compared with models of gene flow using Bayes factors (see, e.g.,

figure 5a in ref. 24 and figure 4a in ref. 41 for examples of such models). Given that both the wild *C. ferus* and the domesticated *C. bactrianus* are known to have wide and overlapping geographic distributions in the recent past (10), a model of spatial population structure appears rather contrived.

Exons and Functional Genomes Are Useful Markers for Studying the Phylogenetic History of Extant Species with Gene Flow. There has been an emphasis of the fact that the coalescent model assumes neutral evolution of genetic markers, with mutations having no impact on the distribution of gene genealogies, and exons are often removed in coalescent-based analysis (e.g., refs. 9 and 42). Indeed a recent simulation study (43) found that several methods for detecting interspecific gene flow including $\delta a\delta i$ (44), FASTSIMCOAL2 (45), and BPP (23) produced excessive false positives when there was selection (selective sweeps, background selection, balancing selection, or adaptive introgression) but no gene flow. The results from that simulation appear to be erroneous because those methods had high false positive rates in data simulated under neutral evolution as well (43, figures 1, 3, and 5), but in such cases, there is no model violation and the methods are expected to work well. In the case of BPP, the authors did not use Bayes factors to test for gene flow but instead constructed a test based on the HPD CIs for φ, which has 100% false positive rate when the data are uninformative.

Here, in this paper, we addressed the suitability of coding DNA for coalescent-based analyses by analyzing both exonic and noncoding data. We found that exonic data produced highly similar parameter estimates to noncoding DNA. In particular species divergence times (τ) and population sizes (θ) are largely proportional between the two types of data, with estimates from the exons being about 0.5 times those from the noncoding regions: $\phi_C = 0.49\phi_{NC}$ ($r = 0.99$) under MSC-I and $\phi_C = 0.51\phi_{NC}$ ($r = 0.99$) under MSC-M, where ϕ_C and ϕ_{NC} are estimates of parameters (τ, θ) from the coding and noncoding data, respectively (Fig. 5). Similar linear relationships between parameter estimates from coding and noncoding data were observed in previous analyses of genomic data from the gibbons (46), *Anopheles* mosquitoes (47), and *Heliconius* butterflies (19).

While exons are expected to be under stronger selective constraints than noncoding regions, purifying selection removing deleterious nonsynonymous mutations in exons may be expected to have similar effects in closely related species, predominantly a reduction in neutral mutation rate. Species-specific directional selection responsible for morphological and behavioral adaptations of the species may affect coalescent-based analysis but such selection may be expected to be rare at the genome scale, unlikely to affect the analysis. Thus, we conclude that protein-coding genes, as well as other conserved elements in the genome, may be used as effective markers to infer the phylogenetic history of modern species with gene flow under the MSC models.

Materials and Methods

Sequence Data. Multilocus sequence alignments were compiled following the procedure of Ming et al. (9) for preparing data analyzed using the program G-PHOCS. See ref. 9 for detailed descriptions of the data. We expanded the dataset by including two dromedary and two wild Bactrian samples, with a total of eight samples used (*SI Appendix*, Table S1). The dataset consists of 10,000 noncoding genomic segments of length 1 kb (referred to as loci), separated by a gap of at least 30 kb between loci. Each locus consists of eight unphased diploid sequences. Heterozygous sites are represented by using ambiguity codes and the phase of multiple heterozygous sites in the same sequence were resolved computationally through likelihood calculation on the gene trees (18, 48, 49).

In addition, we compiled a dataset of exons, which were excluded in the analysis of Ming et al. (9). While exons are under selective constraints, they may be expected to reflect the same history of species divergence and gene flow as the noncoding data. We used the genome annotation of *C. ferus* (GCF_000311805.1) to identify known exons, and included all exons except short ones of less than 100bp long. A total of 120,720 exons (loci) were collected. Each exon was treated as an independent locus in the MSC model.

BPP Analysis of the Noncoding Data. The noncoding data were analyzed to test for the presence or absence of gene flow among the three species in the *Camelus* genus and to construct a model of gene flow. The 10^4 loci were split into four quarters, each of 2,500 loci, and analyzed separately using BPP. The JC mutation model (50) was used in calculation of the likelihood for the sequence alignment at each locus. A gamma prior was assigned to the age of the root, $\tau_r \sim G(2, 2{,}000)$, with the prior mean 0.001. Given τ_r, the ages of descendent nodes on the species tree had the uniform-Dirichlet prior distribution (51, equation 2). A gamma prior was assigned to population-size parameters on the species tree, $\theta \sim G(2, 2{,}000)$. Note that both θ and τ are measured in the expected number of mutations per site.

Gene flow was accommodated using either the MSC-I or the MSC-M models, with six gene-flow events assumed (Fig. 2*A*) (23, 24). We also considered a five-rates model and a four-rates model, excluding ancient gene-flow events involving the ancestor of domestic and wild Bactrian species. The MSC model with no gene flow was used for comparison as well. In the MSC-I model, the introgression probability was assigned a beta prior $\varphi \sim$ beta(1, 9), with the prior mean to be $1/(1 + 9) = 0.1$. In the MSC-M model, the migration rate was assigned a gamma prior $M = Nm \sim G(2, 20)$, with the prior mean 0.1. We used a burn-in of 40,000 MCMC iterations and took 2×10^5 samples, sampling every two iterations. Each analysis was run twice, with consistency between runs used to confirm MCMC convergence. Running time using 8 threads on a server was about 20 to 30 h under the six-rates models and shorter under simpler models (e.g., with 5, 4, or 0 rates of gene flow).

Bayesian Test of Gene Flow. Bayesian test of the null hypothesis $H_0 : \varphi = 0$ against the alternative hypothesis $H_1 : \varphi \neq 0$ was conducted according to ref. 20, using the Savage–Dickey density ratio to calculate the Bayes factor B_{10}. $B_{10} > 1$ (or < 1) means support for H_1 (or H_0), with $B_{10} > 100$ being significant support at the 1% level, $B_{10} < 0.01$ means significant support for H_0 at the 1% level, while $0.01 < B_{10} < 100$ means the test is not significant at the 1% level. We define a null interval, $\varphi < \epsilon$, in which the introgression probability φ is so small that it is of little significance. The Bayes factor representing the evidence for H_1 against H_0 contrasts the posterior odds for φ being inside the null interval against the prior odds:

$$B_{10} \approx \frac{1 - \mathbb{P}(\varphi < \epsilon|X)}{\mathbb{P}(\varphi < \epsilon|X)} \Big/ \frac{1 - \mathbb{P}(\varphi < \epsilon)}{\mathbb{P}(\varphi < \epsilon)}, \quad [2]$$

where $\mathbb{P}(\varphi < \epsilon)$ and $\mathbb{P}(\varphi < \epsilon|X)$ are the prior and posterior probabilities for $\varphi < \epsilon$, respectively (20, equation 7). Given the prior beta(1, 9) for φ, the prior probability $\mathbb{P}(\varphi < \epsilon) = 0.008964$ at $\epsilon = 0.001$ and 0.08648 at $\epsilon = 0.01$, given by the cumulative distribution function (CDF) for the beta distribution. The posterior probability is calculated by processing the MCMC sample under the model of gene flow (MSC-I) using shell scripts. We used $\epsilon = 0.001$ for the test and note that use of $\epsilon = 0.01$ produced similar results.

Similarly, we test whether the migration rate $M = Nm$ is significantly greater than zero by defining the null region $M < \epsilon$ with $\epsilon = 0.001$ or 0.01. Given the gamma prior G(2, 20), the prior probability $\mathbb{P}(M < \epsilon) = 0.001974$ for $\epsilon = 0.001$ and 0.01752 for $\epsilon = 0.01$, given by the CDF for the gamma distribution. Again the posterior probability $\mathbb{P}(M < \epsilon|X)$ was calculated by processing the MCMC sample under the MSC-M model (H_1).

Estimation of Species Divergence Times and Introgression Times. Bayesian estimation of parameters including the rates of gene flow (φ under MSI-I and M under MSC-M) and Bayesian test of gene flow allowed us to formulate a phylogenetic model for the three *camelus* species involving four gene-flow events (fig. 2*B*). We then used this model to analyze the four subsets

of noncoding data as well as the full data of 10^4 noncoding loci to estimate parameters including species divergence times and introgression times, and the rates of gene flow. Compared with phylogenetic methods, the coalescent-based time estimates account for ancestral polymorphism (52).

Analysis of the Exonic Data. The 120,720 exons were separated into 48 data subsets, each of 2,500 loci, which were analyzed separately, using the final model of gene flow constructed based on the noncoding data (Fig. 2*B*). The same settings were used as for the noncoding data to run BPP.

Analysis of Mitochondrial Genomic Data. We assembled an alignment of whole mitochondrial genomes for 16 Bactrians (*C. bactrianus*), 4 wild Bactrian species (*C. ferus*), and 11 dromedaries (*C. dromedarius*). We included two domestic South American species as well: one *V. pacos* and one *L. glama* (*SI Appendix*, Fig. S2). We inferred the maximum likelihood tree using RAXML-NG (53). In BPP analysis under the MSC, we treated the whole mitochondrial genome as one locus as all sites in the genome share the same genealogical tree. We used the gamma priors, $\theta \sim G(2, 400)$ with the mean 0.005, and $\tau \sim G(2, 20)$ with the mean 0.1. We used a burn-in of 2×10^5 iterations, and then took 2×10^5 samples, sampling every five iterations. The run took 15 min on one core.

Data, Materials, and Software Availability. Online supplemental information and data files (sequence alignments and bpp control files) have been deposited in Dryad (https://doi.org/10.5061/dryad.3xsj3txrk). All other data are included in the manuscript and/or SI Appendix.

ACKNOWLEDGMENTS. This study has been supported by China Natural Science Foundation Grants (T2122017 and 32070685) and China National Key R&D Program (2020YFA0712700) to T.Z., by China National Natural Science Foundation Grant (32070570) to Z.W., and by Biotechnology and Biological Sciences Research Council Grant (BB/X007553/1) and Natural Environment Research Council Grant (NSFDEB-NERC NE/X002071/1) to Z.Y.

1. P. A. Burger, E. Ciani, B. Faye, Old World camels in a modern world–A balancing act between conservation and genetic improvement. *Anim. Genet.* **50**, 598–612 (2019).
2. L. Ming, D. Siren, S. Hasi, T. Jambl, R. Ji, Review of genetic diversity in Bactrian camel (*Camelus bactrianus*). *Anim. Front.* **12**, 20–29 (2022).
3. H. Wu *et al.*, Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* **5**, 5188 (2014).
4. F. Almathen *et al.*, Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6707–6712 (2016).
5. E. Mohandesan *et al.*, Mitogenome sequencing in the genus *Camelus* reveals evidence for purifying selection and long-term divergence between wild and domestic Bactrian camels. *Sci. Rep.* **7**, 9970 (2017).
6. M. Kadwell *et al.*, Genetic analysis reveals the wild ancestors of the llama and the alpaca. *Proc. Biol. Sci.* **268**, 2575–2584 (2001).
7. B. Faye, G. Konuspayeva, "The encounter between Bactrian and dromedary camels in Central Asia" in *Camels in Asia and North Africa: Interdisciplinary Perspectives on their Past and Present Significance*, E. M. Knoll, P. Burger, Eds. (Austrian Academy of Sciences Press, Vienna, Austria, 2012), pp. 27–33.
8. R. Ji *et al.*, Monophyletic origin of domestic Bactrian camel (*Camelus bactrianus*) and its evolutionary relationship with the extant wild camel (*Camelus bactrianus ferus*). *Anim. Genet.* **40**, 377–382 (2009).
9. L. Ming *et al.*, Whole-genome sequencing of 128 camels across Asia reveals origin and migration of domestic Bactrian camels. *Commun Biol.* **3**, 1 (2020).
10. G. B. Schaller, *Wildlife of the Tibetan Steppe* (University of Chicago Press, Chicago, IL, 1998), pp. 151–162.
11. J. Hare, The wild Bactrian camel *Camelus bactrianus ferus* in China: The need for urgent action. *Oryx* **31**, 45–48 (1997).
12. R. P. Reading, H. Mix, B. Lhagvasuren, E. S. Blumer, Status of wild Bactrian camels and other large ungulates in South-Western Mongolia. *Oryx* **33**, 247–255 (1999).
13. J. Hare, *Camelus ferus*. The IUCN red list of threatened species 2008 (2008). 10.2305/IUCN.UK.2008.RLTS.T63543A12689285.en.
14. M. dos Reis, P. C. J. Donoghue, Z. Yang, Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**, 71–80 (2016).
15. R. E. Green *et al.*, A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
16. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
17. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
18. I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, A. Siepel, Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
19. Y. Thawornwattana, F. A. Seixas, J. Mallet, Z. Yang, Full-likelihood genomic analysis clarifies a complex history of species divergence and introgression: The example of the erato-sara group of Heliconius butterflies. *Syst. Biol.* **71**, 1159–1177 (2022).
20. J. Ji, D. J. Jackson, A. D. Leache, Z. Yang, Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipmunks. *Syst. Biol.* **72**, 446–465 (2023).
21. X. Jiao, T. Flouri, Z. Yang, Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.* **8**, nwab127 (2021).
22. B. Rannala, Z. Yang, Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656 (2003).
23. T. Flouri, X. Jiao, B. Rannala, Z. Yang, A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* **37**, 1211–1223 (2020).
24. T. Flouri, X. Jiao, J. Huang, B. Rannala, Z. Yang, Efficient Bayesian inference under the multispecies coalescent with migration. *Proc. Nat. Acad. Sci. U.S.A.* **120**, e2310708120 (2023).
25. D. Wen, L. Nakhleh, Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.* **67**, 439–457 (2018).
26. C. Zhang, H. A. Ogilvie, A. J. Drummond, T. Stadler, Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* **35**, 504–517 (2018).
27. J. Hey *et al.*, Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* **35**, 2805–2818 (2018).
28. K. Silbermayr *et al.*, High mitochondrial differentiation levels between wild and domestic Bactrian camels: A basis for rapid detection of maternal hybridization. *Anim. Genet.* **41**, 315–318 (2010).
29. S. Felkel *et al.*, A first Y-chromosomal haplotype network to investigate male-driven population dynamics in domestic and wild bactrian camels. *Front. Genet.* **10**, 423 (2019).
30. R. R. Fitak *et al.*, Genomic signatures of domestication in Old World camels. *Commun. Biol.* **3**, 316 (2020).
31. J. Huang, Y. Thawornwattana, T. Flouri, J. Mallet, Z. Yang, Inference of gene flow between species under misspecified models. *Mol. Biol. Evol.* **39**, msac237 (2022).
32. S. H. Akhmetsadykova, G. Konuspayeva, N. Akhmetsadykov, Camel breeding in Kazakhstan and future perspectives. *Anim. Front.* **12**, 71–77 (2022).
33. X. Jiao, Z. Yang, Defining species when there is gene flow. *Syst. Biol.* **70**, 108–119 (2021).
34. Y. Thawornwattana, J. Huang, T. Flouris, J. Mallet, Z. Yang, Inferring the direction of introgression using genomic sequence data. *Mol. Biol. Evol.* **40**, msad178 (2023).
35. S. H. Martin, C. D. Jiggins, Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.* **47**, 69–74 (2017).
36. A. D. Leaché, R. B. Harris, B. Rannala, Z. Yang, The influence of gene flow on species tree estimation: A simulation study. *Syst. Biol.* **63**, 17–30 (2014).
37. G. P. Tiley *et al.*, Estimation of species divergence times in presence of cross-species gene flow. *Syst. Biol.* **72**, 820–836 (2023).
38. A. Yadamsuren, E. Dulamtseren, R. P. Reading, *The Conservation Status and Management of Wild Camels in Mongolia* (Austrian Academy of Sciences, Vienna, Austria, 2012).
39. A. Eriksson, A. Manica, Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13956–13960 (2012).
40. R. Tournebize, L. Chikhi, Questioning Neanderthal admixture: On models, robustness and consensus in human evolution. *Nat. Ecol. Evol.* **9** (2024), 10.1038/s41559-024-02591-6.
41. D. Kornai, X. Jiao, J. Ji, T. Flouri, , Hierarchical heuristic species delimitation under the multispecies coalescent model with migration. *Syst. Biol.* **73**, 1015–1037 (2024).
42. N. Rosser *et al.*, Hybrid speciation driven by multilocus introgression of ecological traits. *Nature* **628**, 811–817 (2024).
43. M. L. Smith, M. W. Hahn, Selection leads to false inferences of introgression using popular methods. *Genetics* **227**, iyae089 (2024).
44. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
45. L. Excoffier, I. Dupanloup, E. Huerta-Sanchez, V. C. Sousa, M. Foll, Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
46. C. Shi, Z. Yang, Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.* **35**, 159–179 (2018).
47. Y. Thawornwattana, D. Dalquen, Z. Yang, Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol. Biol. Evol.* **35**, 2512–2527 (2018).
48. T. Flouri, X. Jiao, B. Rannala, Z. Yang, Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.* **35**, 2585–2593 (2018).
49. J. Huang, T. Flouri, Z. Yang, A simulation study to examine the information content in phylogenomic datasets under the multispecies coalescent model. *Mol. Biol. Evol.* **37**, 3211–3224 (2020).
50. T. H. Jukes, C. R. Cantor, *Evolution of Protein Molecules* (Academic Press, New York, NY, 1969), pp. 21–123.
51. Z. Yang, A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genom. Biol. Evol.* **2**, 200–211 (2010).
52. K. Angelis, M. dos Reis, The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Curr. Zool.* **61**, 874–885 (2015).
53. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

Author Query Form

Query reference	Query
Q1	Please review 1) the author affiliation and footnote symbols, 2) the order of the author names, and 3) the spelling of all author names, initials, and affiliations and confirm that they are correct as set.
Q2	Claims of priority or primacy are not allowed, per PNAS policy (https://www.pnas.org/authors/submitting-your-manuscript); therefore, the phrase “newly developed” has been deleted. If you have concerns with this course of action, please reword the sentence or explain why the deleted phrase should not be considered a priority claim and should be reinstated.
Q3	Certain compound terms are hyphenated when used as adjectives and unhyphenated when used as nouns. This style has been applied consistently throughout where (and if) applicable.
Q4	Please review 1) the author affiliation and footnote symbols, 2) the order of the author names, and 3) the spelling of all author names, initials, and affiliations and confirm that they are correct as set.
Q5	There is no provided division/section/unit for affiliations a–c. Please provide if this is available.
Q6	Per PNAS style city, country, and postcode are required in all affiliations. Please provide the same in affiliation c.
Q7	Please review the author contribution footnote carefully. Ensure that the information is correct and that the correct author initials are listed. Note that the order of author initials matches the order of the author line per journal style. You may add contributions to the list in the footnote; however, funding may not be an author’s only contribution to the work.
Q8	You have chosen to publish your PNAS article with the immediate open access option under a CC BY-NC-ND license. Your article will be freely accessible immediately upon publication; for additional details, please refer to the PNAS site: https://www.pnas.org/authors/fees-and-licenses. Please confirm this is correct.
Q9	If you have any changes to your Supporting Information (SI) file(s), please provide revised, ready-to publish replacement files without annotations.
Q10	As per PNAS style, italics for emphasis is not allowed in the text. Please check and remove in all instances.
Q11	Please confirm the edit made to the phrase “a burn-in.”
Q12	Note that all data deposited in a publicly accessible database (and therefore not directly available in the paper or SI) must be cited in the text with an entry in the reference list. References must include the following information: 1) author names, 2) data/page title, 3) database name, 4) a direct URL to the data, 5) the date on which the data were accessed or deposited (not the release date). For an example reference entry, visit https://www.pnas.org/author-center/submitting-your-manuscript#manuscript-formatting-guidelines. Please also indicate where the new reference citation should be added in the main text and/or data availability statement. Please add a reference for the following data: https://doi.org/10.5061/dryad.3xsj3txrk.
Q13	Abbreviations must be used at least twice in the acknowledgments section, as such BBSRC and NERC have been deleted.
Q14	Please provide accessed date (DD/MM/YYYY) in ref. 13.
Q15	Please provide page range in ref. 40.