Estimation of Cross-Species Introgression Rates using Genomic Data Despite Model Unidentifiability

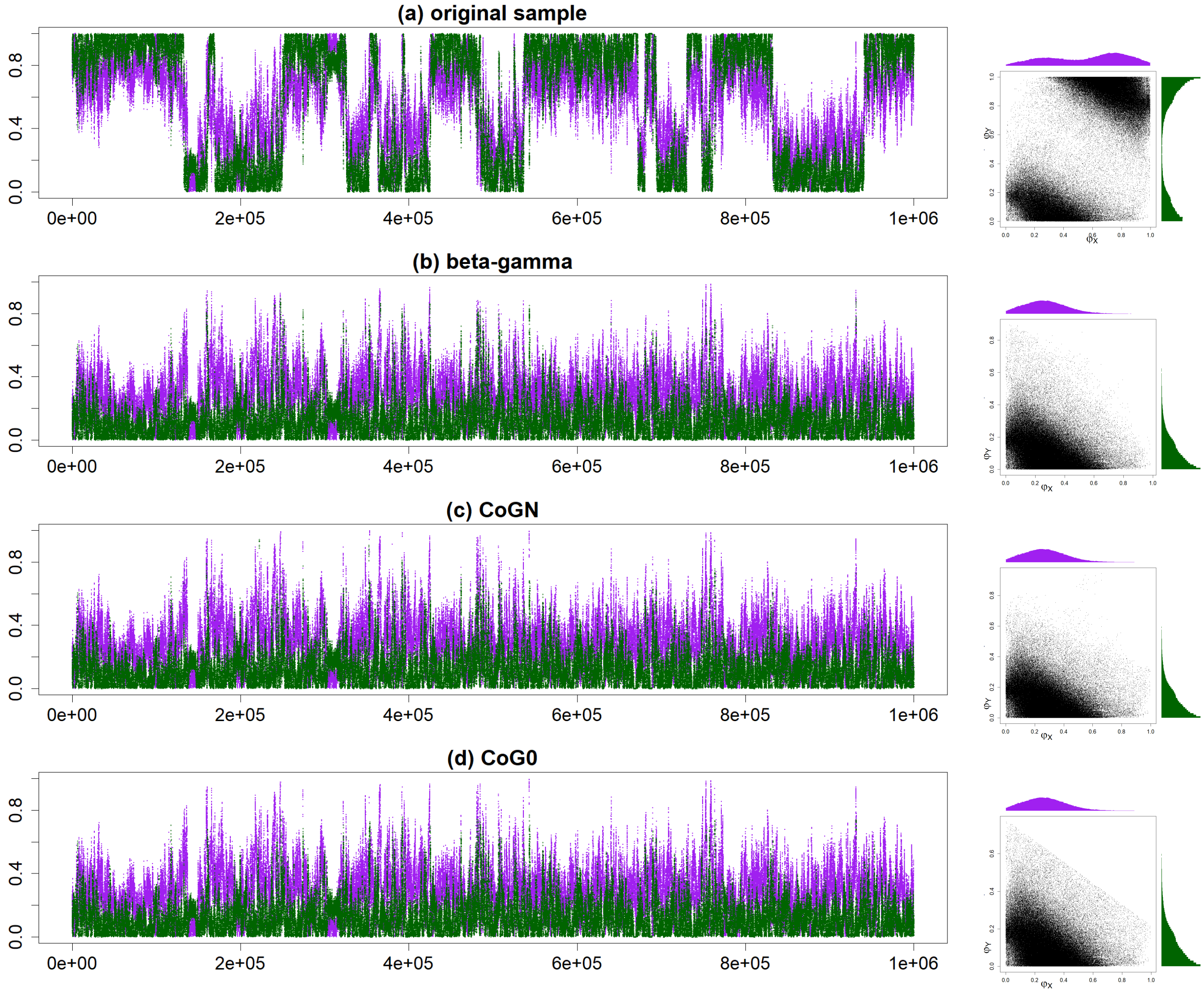
Ziheng Yang and Tomáš Flouri

Supplementary material

Table S1. Posterior means and 95% HPD CIs (in parenthees) for parameters in the MSci model of figure 1a from three simulated datasets

	truth (Θ)	mirror (Θ')	beta-gamma	CoG$_N$	CoG$_0$
$L = 500$ loci					
τ_R	0.01		0.0098 (0.0088, 0.0108)		
$\tau_X = \tau_Y$	0.005		0.0050 (0.0045, 0.0055)		
θ_A	0.002		0.0020 (0.0018, 0.0021)		
θ_B	0.01		0.0101 (0.0093, 0.0108)		
θ_R	0.002		0.0020 (0.0006, 0.0034)		
θ_X	0.002	0.01	0.0063 (0.0005, 0.0130)	0.0066 (0.0005, 0.0133)	0.0066 (0.0005, 0.0133)
θ_Y	0.01	0.002	0.0071 (0.0022, 0.0124)	0.0067 (0.0017, 0.0120)	0.0068 (0.0017, 0.0121)
φ_X	0.7	0.3	0.755 (0.472, 0.999)	0.764 (0.528, 0.999)	0.765 (0.530, 0.999)
φ_Y	0.2	0.8	0.447 (0.209, 0.670)	0.461 (0.214, 0.695)	0.462 (0.212, 0.695)
$L = 2000$ loci					
τ_R	0.01		0.0101 (0.0094, 0.0108)		
$\tau_X = \tau_Y$	0.005		0.0051 (0.0048, 0.0054)		
θ_A	0.002		0.0020 (0.0019, 0.0021)		
θ_B	0.01		0.0100 (0.0097, 0.0104)		
θ_R	0.002		0.0018 (0.0009, 0.0027)		
θ_X	0.002	0.01	0.0037 (0.0008, 0.0062)	0.0037 (0.0009, 0.0062)	0.0050 (0.0006, 0.0097)
θ_Y	0.01	0.002	0.0076 (0.0049, 0.0108)	0.0076 (0.0048, 0.0108)	0.0064 (0.0019, 0.0104)
φ_X	0.7	0.3	0.545 (0.178, 0.903)	0.545 (0.178, 0.903)	0.656 (0.449, 0.887)
φ_Y	0.2	0.8	0.398 (0.198, 0.598)	0.398 (0.198, 0.598)	0.450 (0.205, 0.684)
$L = 8000$ loci					
τ_R	0.01		0.0098 (0.0094, 0.0102)		
$\tau_X = \tau_Y$	0.005		0.0049 (0.0048, 0.0051)		
θ_A	0.002		0.0020 (0.0019, 0.0020)		
θ_B	0.01		0.0100 (0.0098, 0.0102)		
θ_R	0.002		0.0021 (0.0017, 0.0025)		
θ_X	0.002	0.01	0.0045 (0.0003, 0.0083)	0.0044 (0.0003, 0.0081)	0.0051 (0.0003, 0.0106)
θ_Y	0.01	0.002	0.0095 (0.0059, 0.0129)	0.0096 (0.0060, 0.0130)	0.0089 (0.0041, 0.0128)
φ_X	0.7	0.3	0.645 (0.442, 0.848)	0.645 (0.436, 0.848)	0.645 (0.436, 0.846)
φ_Y	0.2	0.8	0.334 (0.128, 0.632)	0.336 (0.129, 0.639)	0.310 (0.123, 0.539)

Note.— Empty values for Θ' mean the same values as for Θ. MCMC samples are processed using the three algorithms and then summarized. See figure S7 for the trace-scatter plots for the dataset of $L = 500$. The datasets, with each locus consisting of four sequences per species (or eight sequences per locus) and 500 sites per sequence, are simulated using the true parameter values (Θ).

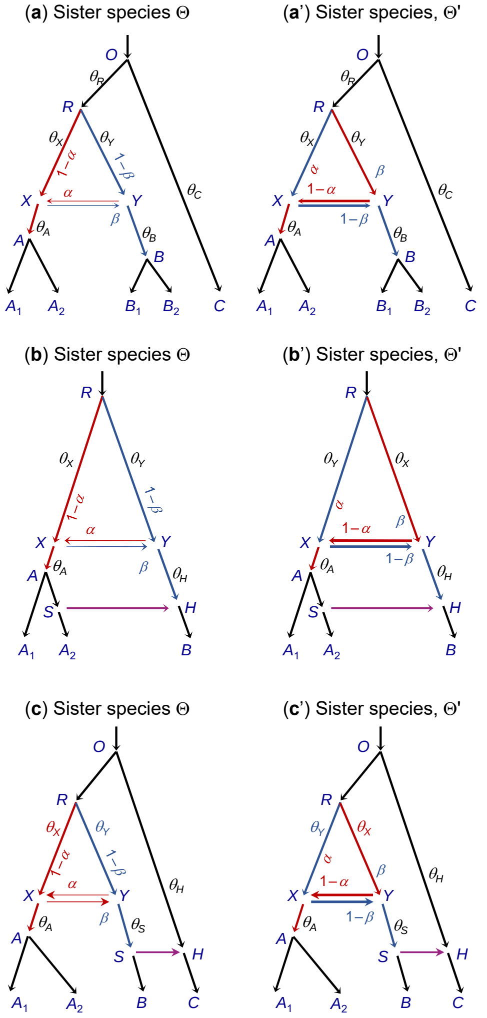

(a) original sample

(b) beta-gamma

(c) CoGN

(d) CoG0

Figure S1: Analysis of the first 500 exonic loci on chromosome 1 from the *Heliconius* data. See legend to figure 3.



(**a**) Sister species Θ

(**a**') Sister species, Θ'

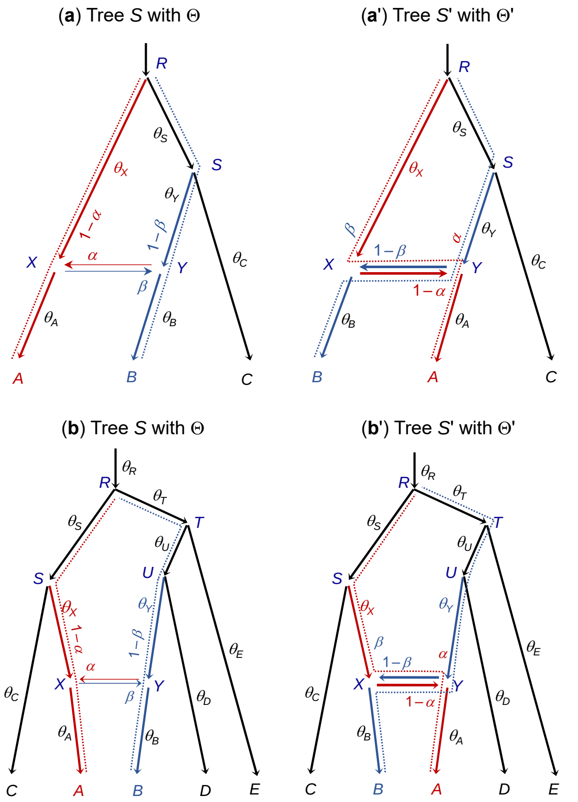
(**b**) Sister species Θ

(**b**') Sister species, Θ'

(**c**) Sister species Θ

(**c**') Sister species, Θ'

Figure S2: Three species trees (MSci models), each with a BDI event between sister species, exhibiting within-model unidentifiability. (**a** & **a**$'$) Subtrees are added to branches A, B, and R in the basic model of figure 1a. (**b** & **b**$'$) A BDI event between sister species X and Y with a unidirectional introgression involving descendant branches of X and Y. (**c** and **c**$'$) A BDI event between sister species X and Y with a unidirectional introgression involving one descendant branch and another branch that is not a descendant of X or Y. In all three cases, the parameter mapping is $\varphi'_X = 1 - \varphi_X$, $\varphi'_Y = 1 - \varphi_Y$, $\theta'_X = \theta_Y$, and $\theta'_Y = \theta_X$.



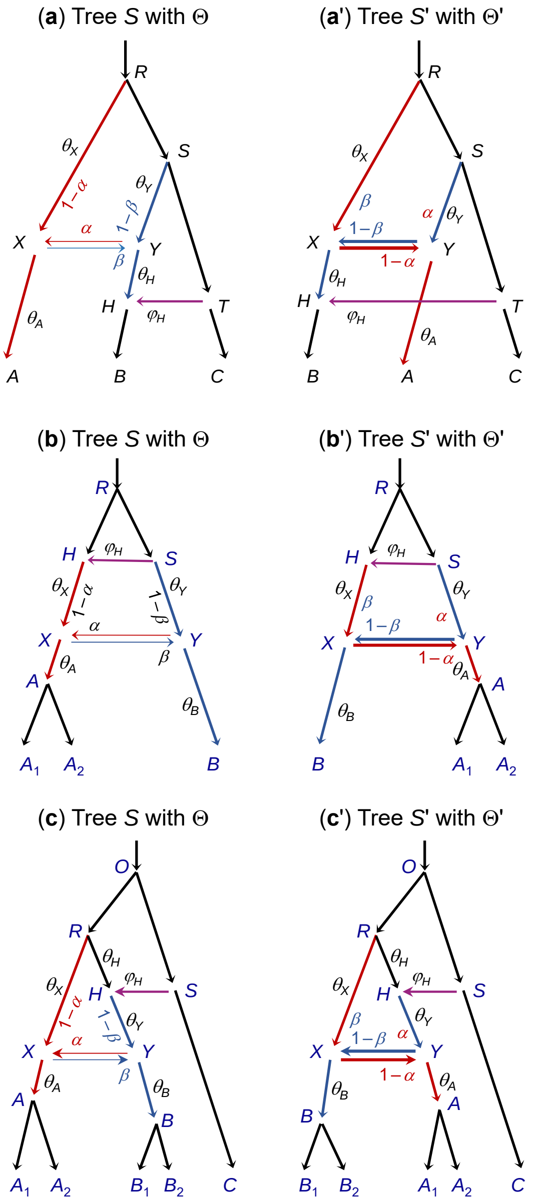
(**a**) Tree *S* with Θ

(**a'**) Tree *S*' with Θ'

(**b**) Tree *S* with Θ

(**b'**) Tree *S*' with Θ'

Figure S3: Two pairs of species trees or unidentifiable MSci models with a BDI event between non-sister species creating cross-model unidentiability. (**a** & **a**$'$) A pair of unidentifiable models with a BDI event between non-sister species. The dotted lines indicate the main routes taken by sequences sampled from species A and B, if the introgression probabilities α and β are $< \frac{1}{2}$. (**b** & **b**$'$) Another pair of unidentifiable models with a BDI event between non-sister species. The parameter mapping from Θ to Θ' in both cases is $\varphi'_X = 1-\varphi_Y$ and $\varphi'_Y = 1-\varphi_X$, with all other parameters (such as θ_X, θ_Y, θ_A, and θ_B) to be identical between Θ and Θ'.


(**a**) Tree *S* with Θ

(**a'**) Tree *S*' with Θ'

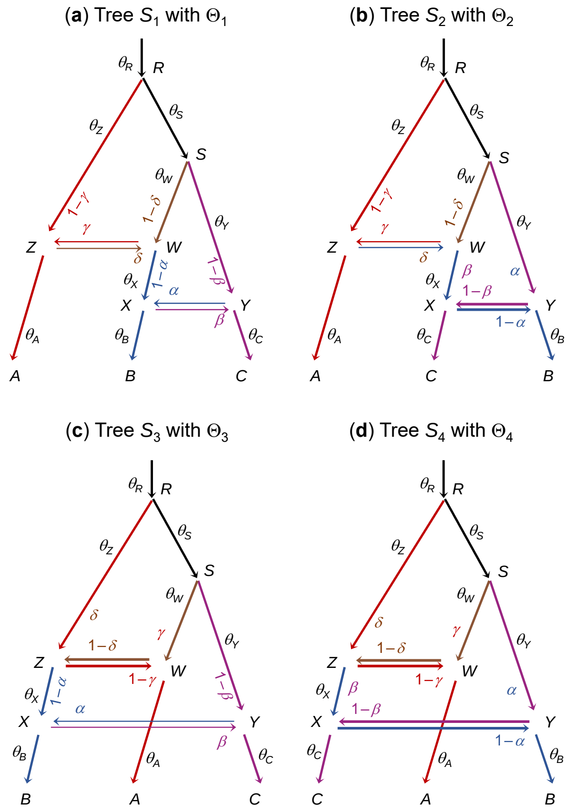
(**b**) Tree *S* with Θ

(**b'**) Tree *S*' with Θ'

(**c**) Tree *S* with Θ

(**c'**) Tree *S*' with Θ'

Figure S4: Three pairs of species trees (or unidentifiable MSci models) with one BDI event between non-sister species, illustrating the mapping of parameters (Θ and Θ′). In (**a**), *RXA* and *SYH* are non-sister species. In (**b**) & (**c**), nodes *X* and *Y* are non-sister species because of the unidirectional introgression event involving branches *RX* and/or *RY*. In each of the three cases, the mirror model (*S*′ with Θ′) is generated by pruning off branches *AX* at *X* and *BY* at *Y*, swapping places and reattaching, and applying the mapping $\varphi'_X = 1 - \varphi_Y$ and $\varphi'_Y = 1 - \varphi_X$.


(**a**) Tree S_1 with Θ_1

(**b**) Tree S_2 with Θ_2

(**c**) Tree S_3 with Θ_3

(**d**) Tree S_4 with Θ_4

Figure S5: Four species trees for species A, B, and C representing four unidentifiable models each with two BDI events between non-sister species. The cross-model parameter mappings concern only the introgression probabilities $\varphi_X \equiv \alpha$, $\varphi_Y \equiv \beta$, $\varphi_Z \equiv \gamma$, and $\varphi_W \equiv \delta$, while all other parameters are the same among the models. The colored lines indicate the main routes taken by sequences sampled from A (red), B (blue), and C (purple), if the introgression probabilities α, β, γ, and δ are all $< \frac{1}{2}$, from which the unidentifiability of the four models can be seen easily. Based on figure S9 of Finger et al. (2022).

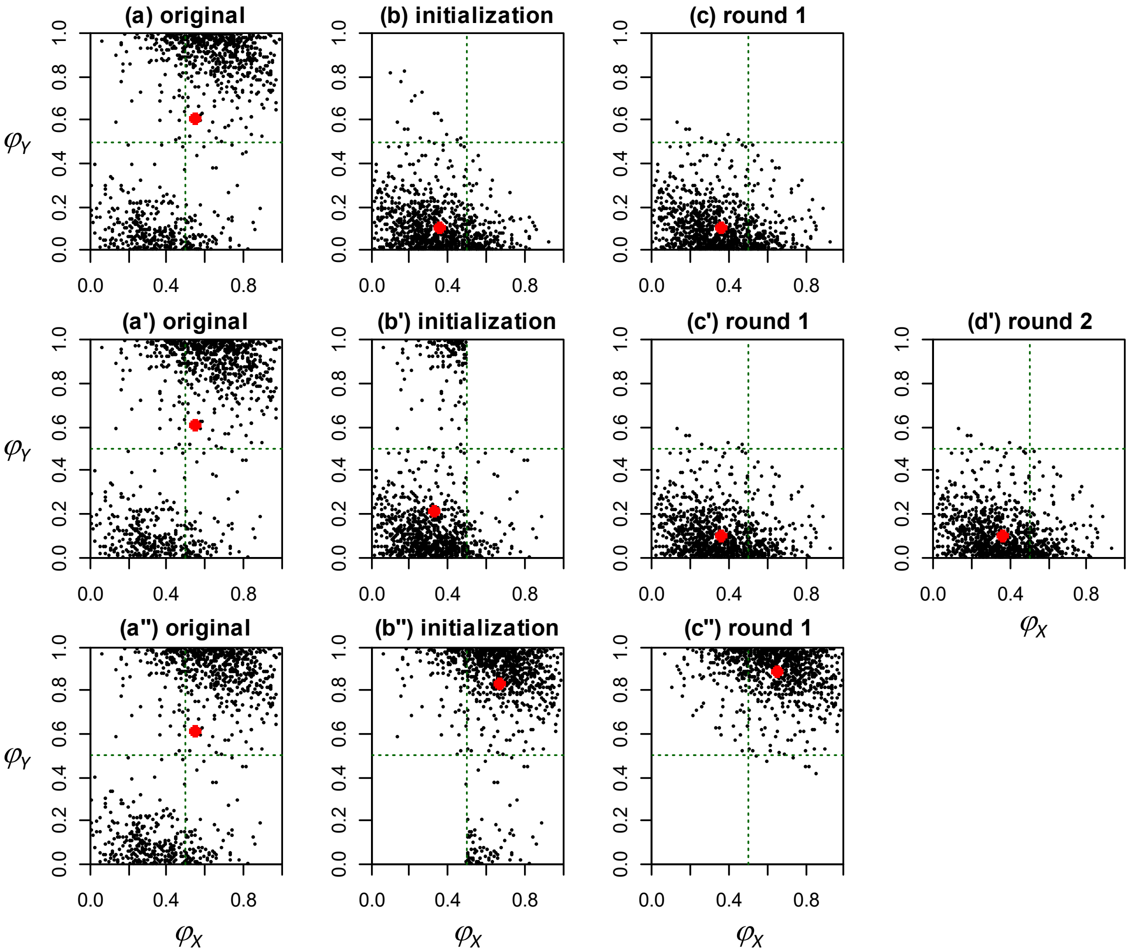


Figure S6: The CoG$_0$ algorithm moves sampled points to their mirror positions to be as close as possible to the center of gravity. Note that (φ_X, φ_Y) and its mirror position $(1-\varphi_X, 1-\varphi_Y)$ are mirror reflections of each other around the point $(\frac{1}{2}, \frac{1}{2})$. The 'original' sample consists of 1000 points, obtained from 'thinning' the MCMC sample from the BPP analysis of the 500 noncoding *Heliconius* loci of figure 3a. The mean (φ_X, φ_Y) = (0.544, 0.614) is indicated by the red dot in **a**, **a**′ & **a**″. The three rows illustrate three runs of the CoG$_0$ algorithm with different starting positions: (**a**-**c**) $\varphi_X + \varphi_Y < 1$, (**a**′-**d**′) $\varphi_X < \frac{1}{2}$ or $\varphi_Y < \frac{1}{2}$, and (**a**″-**c**″) $\varphi_X > \frac{1}{2}$ or $\varphi_Y > \frac{1}{2}$. In the first run, the initialization (under the condition $\varphi_X + \varphi_Y < 1$) moves 647 points above or right of the line $\varphi_X + \varphi_Y = 1$ to their mirror points below or left of the line, with the new mean (0.348, 0.111), indicated by the red dot (**b**). The algorithm then attempts to move points to their mirror positions to be closer to the red dot. Ten such points are moved, with the new mean (0.353, 0.107) (**c**). In the next iteration, no points move, so the algorithm terminates. In the second run (**a**′-**d**′), the initialization ($\varphi_X < \frac{1}{2}$ or $\varphi_Y < \frac{1}{2}$) moves 512 points from the upper right corner to their mirror points in the lower left, with the new mean (0.327, 0.216) (**b**′). Round 1 moves 136 points, with the new mean (0.353, 0.107) (**c**′). Round 2 moves one point, with the new mean (0.353, 0.107) (**d**′). In the third run, the initialization ($\varphi_X > \frac{1}{2}$ or $\varphi_Y > \frac{1}{2}$) moves 275 points from the lower left corner to their mirror points in the upper right corner, with the new mean (0.662, 0.832) (**b**″). Round 1 moves 75 points, with the new mean (0.647, 0.892), and the next round does not move any points, so the algorithm ends. The first two runs converge to the same mean (0.353, 0.107), while the third run converges to its mirror point (0.647, 0.892). If the original positions are taken as the initial positions (i.e., without initialization), the algorithm converges, after one iteration, to (0.647, 0.892), as in the second run. Note that the algorithm operates on four parameters $\Theta = (\varphi_X, \varphi_Y, \theta_X, \theta_Y)$ but only (φ_X, φ_Y) is shown here.

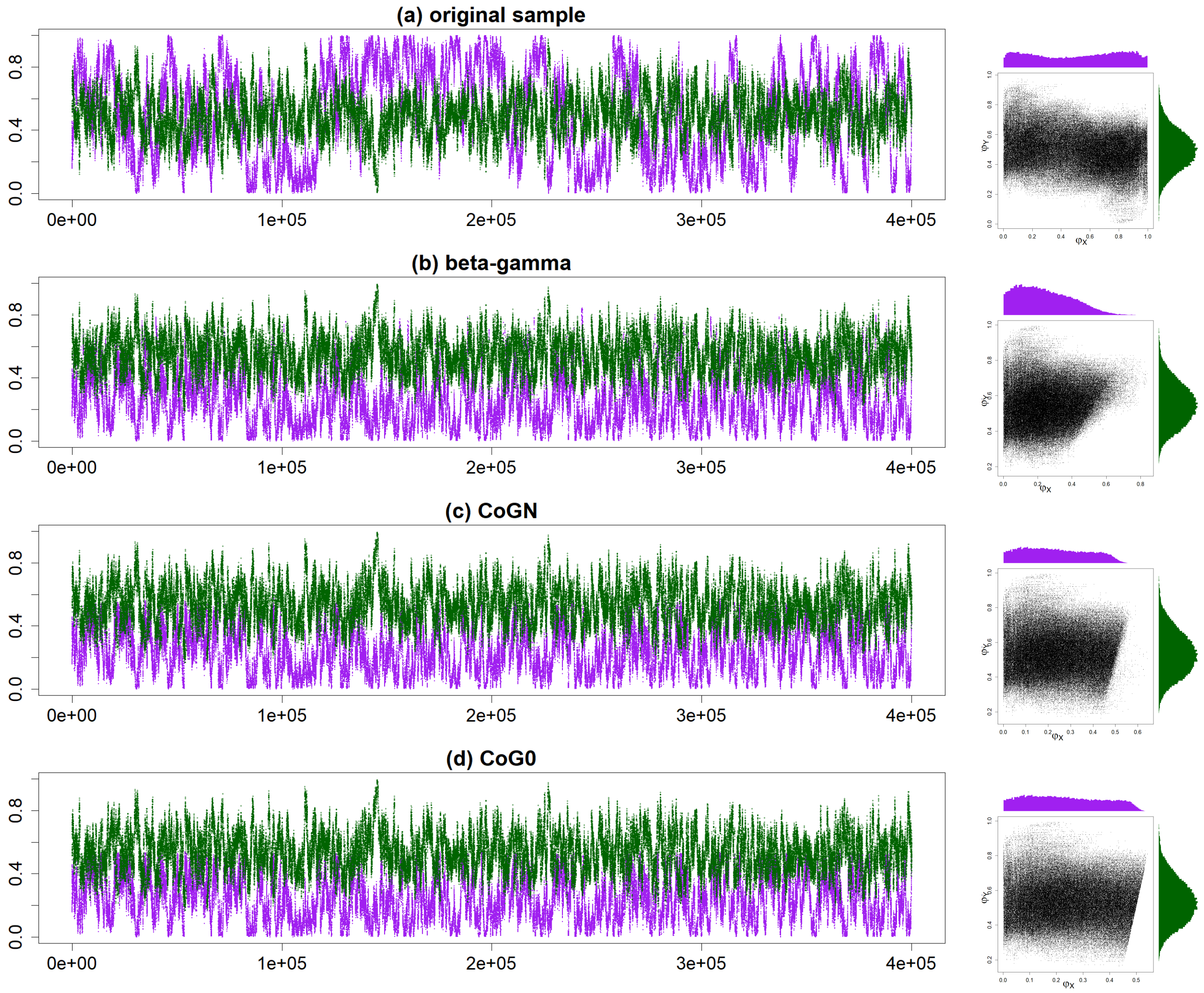


Figure S7: Trace plots of MCMC samples for φ_X (purple) and φ_Y (green) and 2-D scatter plots from BPP analysis of a dataset of $L = 500$ loci simulated under the BDI model of figure 1a. See table S1 for the true parameter values and posterior summaries. The plots are, from top to bottom, for (**a**) unprocessed sample and processed samples using (**b**) the β–γ, (**c**) the CoG$_N$, and (**d**) the CoG$_0$ algorithms. The true parameter values are $\Theta = (\varphi_X, \varphi_Y) = (0.7, 0.2)$, and the post-processing using all three algorithms mapped the samples to the mirror tower around $\Theta' = (0.3, 0.8)$.