Supplementary Materials

Guillot et al.

The following sections provide additional information about aspects of the statistical analyses presented in the main text.

SUMMARY STATISTICS

The folded site frequency spectrum η is an array of frequencies showing the proportion $\eta_i = i/n$ of derived alleles at n sites relative to the sample size N. Many other summaries can be derived from the site frequency spectrum. Nei's homozygosity (Nei, 1978) is traditionally defined, with f_i being the frequency of each allele at the given site i, as

$$H = 1 - \frac{N\left(1 - \sum_{i=1}^{n-1}(1-f_i)^2 + f_i^2\right)}{N-1} \tag{1}$$

but can also be computed directly from the folded site frequency spectrum (Achaz, 2009) as

$$H = 1 - \frac{N\left(1 - \sum_{i=1}^{N-1}\eta_i((1-f_i^2) + f_i^2)\right)}{N-1} \tag{2}$$

The mean pairwise diversity θ_π can be computed from the folded site frequency spectrum (Achaz, 2009) as

$$\theta_\pi = \sum_{i=1}^{N-1}\eta_i i(N-i) \tag{3}$$

The summary statistics above were used to describe changes in genetic patterns due to marriage rules, but additional summary statistics (below) were used in the Approximate Bayesian Computation (ABC) analysis. Two of these summary statistics simply reflect the diversity of mtDNA, exactly as defined above

$$H^M, \theta_\pi^M$$

However, a series of new ζ summary statistics were developed as unbiased estimators of the relative genetic diversity on the autosomes and X chromosome. Due to the ascertainment bias of SNP chip data, the folded site frequency spectrum of biased observed data differs markedly from unbiased simulated data (Figure S1), so these spectra cannot be compared directly. We correct for this effect by instead comparing the difference between the site frequency spectra of autosomes and the X chromosome because this ratio carries information about population structure (Hedrick, 2007), as for instance, imposed by marriage rules. The ζ summary statistics have the generic form

$$\zeta_i^X = \frac{\eta_i^X - \eta_i^A}{\sum_{j=1}^{N} |\eta_j^X - \eta_j^A|} \tag{4}$$

where η_i and η_j are the frequencies of sites with i and j minor alleles on the autosomes or X chromosome for either the observed or simulated data. To chose summary statistics for ABC analysis, we used cross-validation to identify the set of summary statistics that yields the lowest prediction error. The set of summary statistics used in the ABC analysis for Rindi was

$$H^M, \theta_\pi^M, \zeta_1^X, ..., \zeta_6^X$$

The set of summary statistics used for testing Approximate Bayesian Computation on future unbiased genomic data was

$$\theta_\pi^M, \theta_\pi^A, \theta_\pi^X, \theta_\pi^Y, \zeta_1^X = \frac{\eta_1^X - \eta_1^A}{\sum_{j=1}^{N} |\eta_j^X - \eta_j^A|}, \zeta_2^X = \frac{\eta_2^X - \eta_2^A}{\sum_{j=1}^{N} |\eta_j^X - \eta_j^A|}, \zeta_1^Y = \frac{\eta_1^Y - \eta_1^A}{\sum_{j=1}^{N} |\eta_j^Y - \eta_j^A|}, \zeta_2^Y = \frac{\eta_2^Y - \eta_2^A}{\sum_{j=1}^{N} |\eta_j^Y - \eta_j^A|}$$

PARAMETERS

Table 1 summarizes the model parameters and values used in the simulation study (GAM regression) and empirical study of Rindi (ABC).

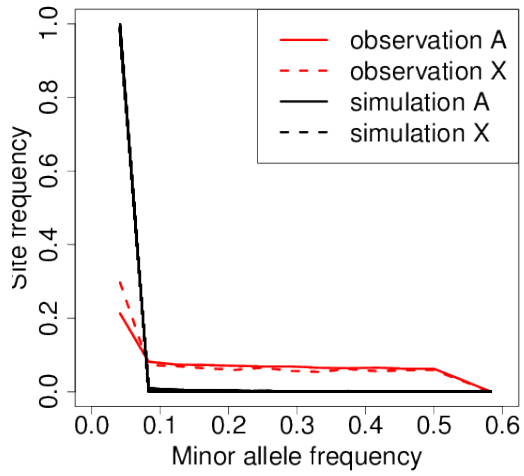

FIG. S1. The folded site frequency spectrum η for observed (red) and simulated (black) data on the autosomes (solid line) and X chromosome (dashed lines).

Parameter	Regression analysis	ABC analysis
Population size (N)	3,000	[5,000; 12,000]
Number of demes	20	20
Size of demes	150	[250; 600]
π_{mig}	[0; 1]	[0; 1]
π_{MBD}	[0; 1]	[0; 1]
μ_A	2.5×10^{-7} mut/site/gen	2.5×10^{-7} mut/site/gen
μ_X	2.5×10^{-7} mut/site/gen	2.5×10^{-7} mut/site/gen
μ_Y	2.5×10^{-7} mut/site/gen	2.5×10^{-7} mut/site/gen
μ_{mtDNA}	4.0×10^{-6} mut/site/gen	4.0×10^{-6} mut/site/gen
Generation time	25 years	25 years
Autosomal loci	32 bp; 200 loci	32 bp; 200 loci
X Chromosome loci	1,000 bp; 10 loci	1,000 bp; 10 loci
Y Chromosome locus	10,000 bp; 1 loci	10,000 bp; 1 loci
mtDNA locus	544 bp; 1 loci	544 bp; 1 loci
Sample size (mtDNA)	28	28
Sample size (nuclear loci)	24	24

Table S1. Model parameters and values used in the simulations for the regression and ABC studies.

LOCAL REGRESSION

Relationships between aspects of genetic diversity and marriage rule parameters were modeled using generalized additive models (GAM). Results for θ_π are shown in the main text. Results for homozygosity are shown below (Figure S2), where $H \to 1$ indicates low genetic diversity. Low mutation rates for nuclear DNA lead to a high frequency of non-segregating sites. Hence, H values are close to 1 for most nuclear regions (both observed and simulated data) (Hammer *et al.*, 2008), in contrast to the values that are more familiar from mtDNA studies.

APPROXIMATE BAYESIAN COMPUTATION

The accuracy of Approximate Bayesian Computation (ABC) was estimated using cross-validation tools (Beaumont *et al.*, 2010; Csillery *et al.*, 2012; Sunnåker *et al.*, 2013). ABC was used to estimate the value of the three (known) parameters N, π_{MBD} and π_{mig} for a randomly selected set of 1,000 simulations (Figure S3). These inferred values were then compared to the known values, measuring how far the 'predicted' (i.e., estimated) value is from the true value. This average distance prediction error over some parameter γ is computed over n simulations by:

$$E_{pred} = \frac{\sum_{i=1}^{n}((\gamma^* - \gamma)^2)}{n \cdot Var(\gamma)} \quad (5)$$

In an ideal inference setting, γ is expected to approach 0. ABC applied on the Rindi dataset has relatively high error, in part due to a loss of information from the ascertainment bias correction and the exclusion of Y chromosome data. ABC was used to test the potential power of this framework on an unbiased genomic dataset (i.e., full sequence data instead of SNPs) including the Y chromosome (Figure S4). These results are presented in the main text.

Computer Model

SMARTPOP simulates population genetic diversity forward-in-time under complex social constraints and structures. The code is freely available at:

`http://smartpop.sourceforge.net`

Figure S5 presents a flowchart showing the main elements of the Asymmetric Prescriptive Alliance model from the perspective of an individual. Note that a man's prescribed marriage partner is the Mother's Brother's Daughter (MBD). From the woman's perspective, the marriage partner is the Father's Sister's Son (FZS).

References

Achaz, G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1): 249–258.

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Knowles, L., Hickerson, M., and Scott, A. 2010. In defence of model-based inference in phylogeography. *Molecular Ecology*, 19: 436–446.

Csillery, K., Francois, O., and Blum, M. G. B. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3: 475–479.

Hammer, M. F., Mendez, F. L., Cox, M. P., Woerner, A. E., and Wall, J. D. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genetics*, 4(9): 8.

Hedrick, P. W. 2007. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution*, 61(12): 2750–2771.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3): 583–590.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. 2013. Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1): e1002803.

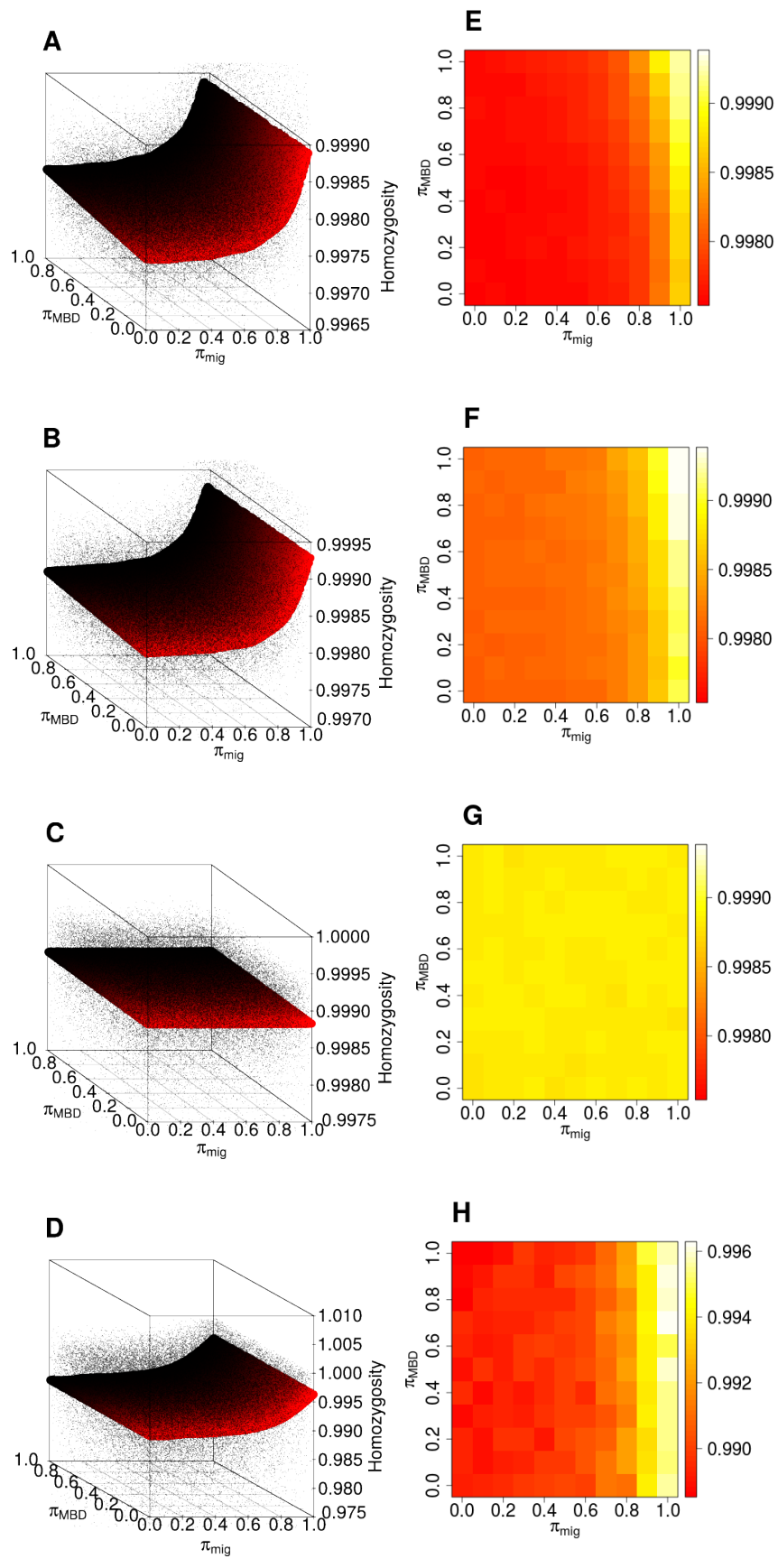

FIG. S2. Homozygosity (H) under Asymmetric Prescriptive Alliance across a grid of random values for π_{MBD} and π_{mig} for (A, E) the autosomes, (B, F) X chromosome, (C, G) Y chromosome and (D, H) mtDNA from 50,000 simulations (3,000 individuals, 20 demes). A-D show simulated values (black points) and fitted surfaces from the generalized additive models. E-H show average simulated homozygosity across the grid of parameters.

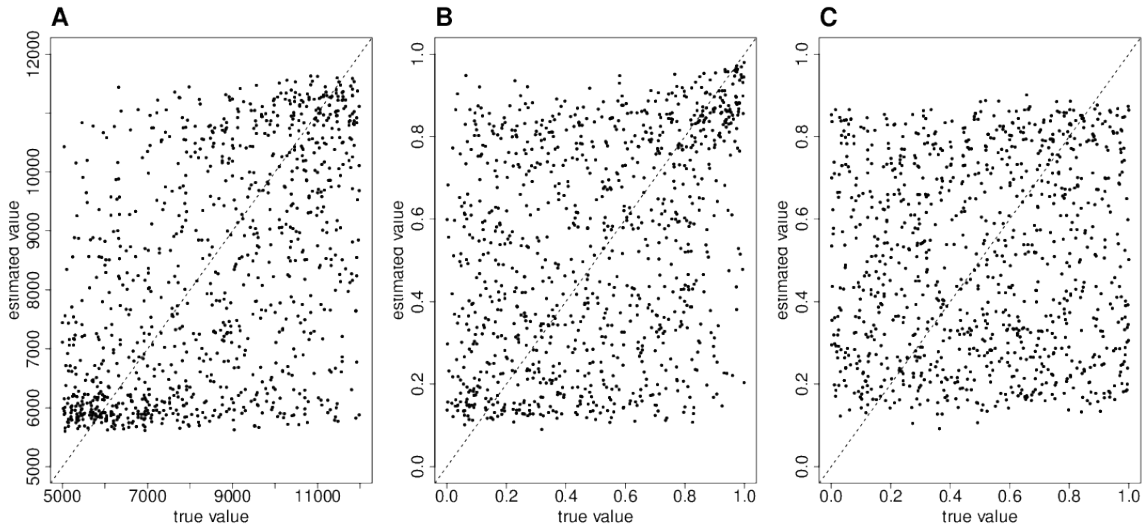

FIG. S3. Approximate Bayesian Computation cross-validation applied to the Rindi analysis. A, B and C represent estimated values of N, π_{mig} and π_{MBD}, respectively, against the true values of these parameters for a random set of 1,000 simulations.

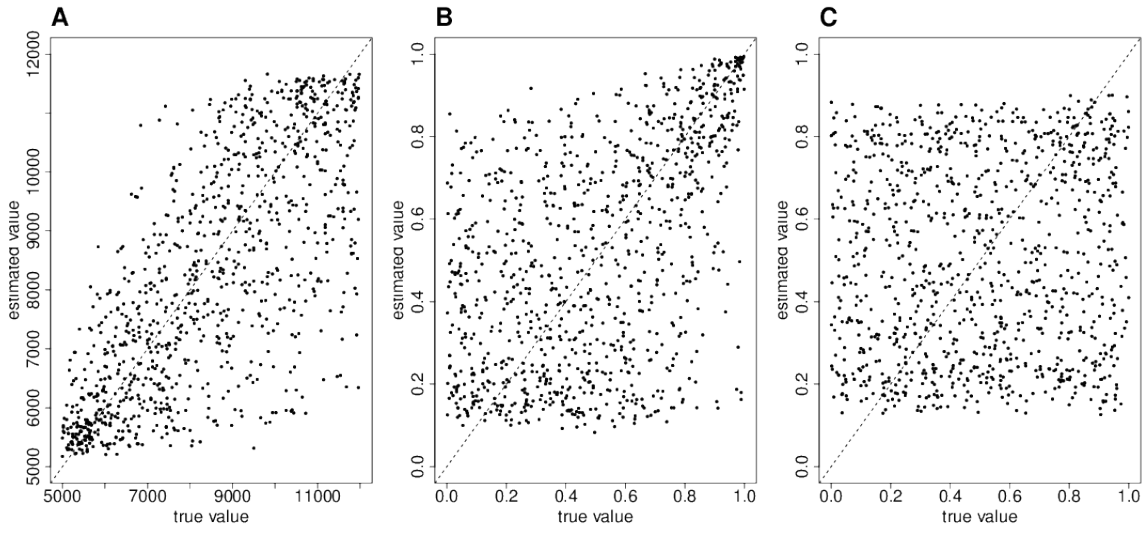

FIG. S4. Approximate Bayesian Computation cross-validation applied to simulated data without the SNP chip ascertainment bias. A, B and C represent estimated values of N, π_{mig} and π_{MBD}, respectively, against the true values of these parameters for a random set of 1,000 simulations.

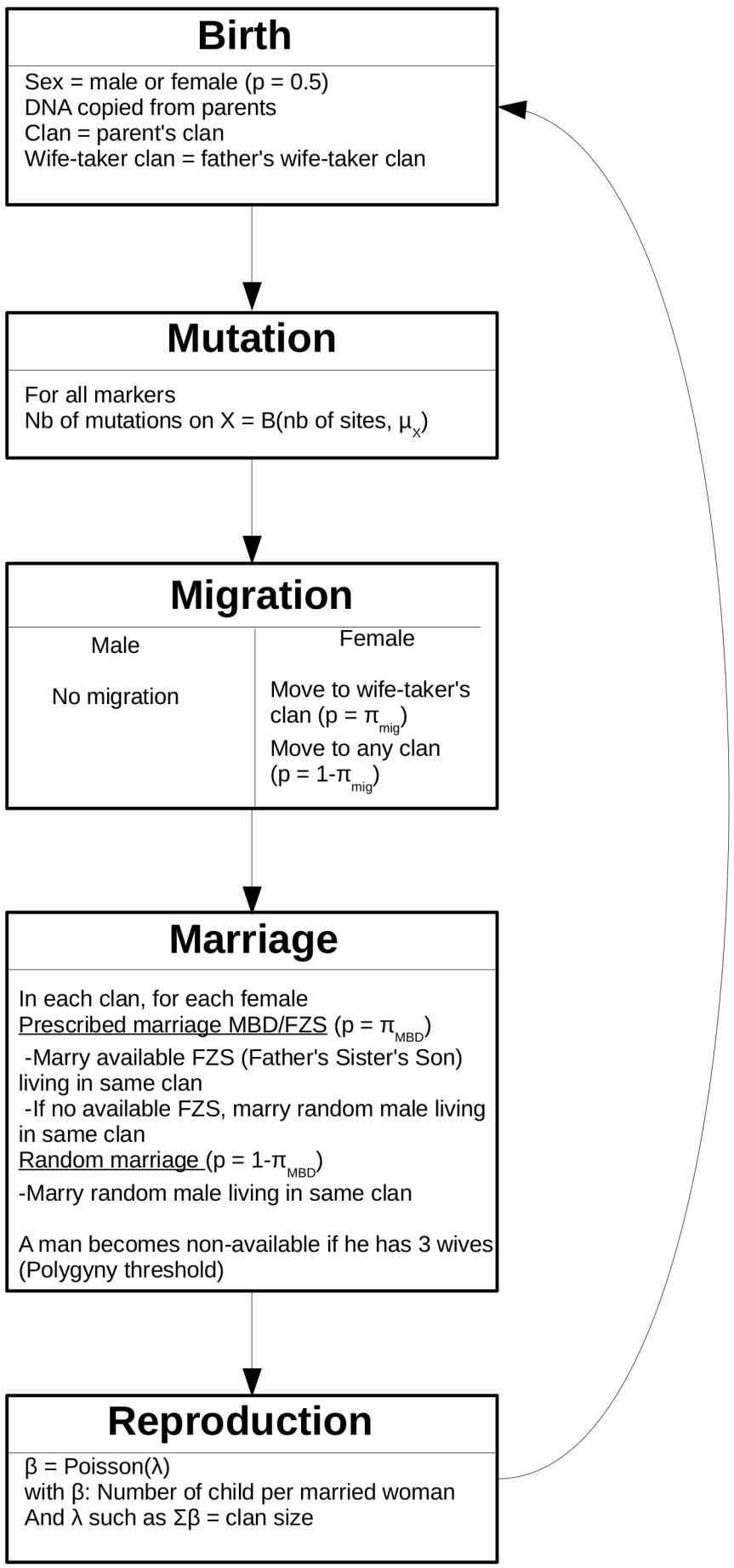

FIG. S5. Flowchart of the algorithm employed by SMARTPOP to model Asymmetric Prescriptive Alliance.