Supplementary Materials

1 Details of the fitting procedure

The Gaussian kernel smoothing procedure described in the Methods section of the manuscript yields the approximate log likelihood value for each set of model parameters. This procedure does not eliminate noise entirely especially when the data-set contains high abundance haplotypes. High abundance haplotypes are less likely to occur in the simulations for higher haplotype discovery rates μ. Figure S1 illustrates this point.

Faced with the noise in the estimates of log likelihood we chose to extract the model parameters that maximize the likelihood by fitting a general parabolic surface to all of the estimated likelihoods which lie within 5% of the maximum. These parabolic fits yield the estimates of the model parameters, their confidence regions where the fitted log likelihood is reduced by three units and the accuracy of the estimated maximum likelihood as the root mean square deviation of the estimated log likelihoods from the fitted parabolic surface.

2 Example

To compare the fits of the models to the data, we computed the likelihoods of observing different haplotype datasets, i.e. sets of (abundance,homozygosity) pairs (c,h), given a particular model. We use Gaussian kernel smoothing to compute the log likelihood L. Smoothing is performed in logarithmic space where z is incremented by unity to assure that it remains positive. The kernel has a parameter λ which is the smoothing length scale in logarithmic space. We checked that the results are only weakly sensitive to λ which was accordingly fixed at λ=0.02.

The likelihood of the model is computed as:

$$Likelihood = \sum_i Ln(P_Sim(c_i, z_i))$$

over all observed size and homozygosity pairs. However, there are no simulated values for all pairs, so we replace

$$P_Sim(c_i, z_i) \approx P_{Interp}(c_i, z_i)$$

www.pnas.org/cgi/doi/10.1073/pnas.1714436116

where

$$P_{Interp}(c_i, z_i) = \frac{\sum_{j,k} P_{Sim}(j,k) * Ker(||c_i - j, z_i - k||)}{\sum_{j,k} Ker(||c_i - j, z_i - k||)},$$

and the sum runs on a lattice from

$$\begin{aligned} c_{min} &= max(1, e^{-log(c)-\lambda}) \\ c_{max} &= e^{-log(c)} \\ z_{min} &= max(1, e^{-log(h)-\lambda}) \\ z_{max} &= e^{-log(h)} \end{aligned}$$

and the kernel is a gaussian kernel normalized to one with a variance of $\sigma = 1/2 * (\lambda)^2$.

For example, if for a sepecific haplotype c=736 and z=14 then:

$$\begin{aligned} c_{min} &= 665 \\ c_{max} &= 813 \\ z_{min} &= 13 \\ z_{max} &= 16 \\ norm &= 37.2034 \\ sum &= 3.50644e - 10 \end{aligned}$$

And the likelihood of this haplotype is -25.3876.

3 Time Correlations

We compare the observed populations to the simulation distribution in equilbrium. In order to check that the simulations reach equilibrium and are not affected by the initial distribution, we sampled the family size distributions at different times $F(t)$, and computed the Spearmann correlation between the family size distribution vector at different times $C(\tau) = corr(F(t), F(t + \tau)$. When a family is present at time t and absent at $t + \tau$ or vice versa, we assigned it a size of 0. When all families were replaced (i.e. each family existing at time t does not exist at time $t + \tau$ we expect a correlation of -1, since each positive value at t will be 0 at $t + \tau$ and vice versa. One can clearly see in the following figure that correlations decrease to -1. We only sample the simulations when the correlations with the initial distribution decrease to -1.

4 Data format

The provided zip archive contains the haplotype data for the 23 sample populations labelled by the accepted abbreviations. Please see `https://www.haplostats.`

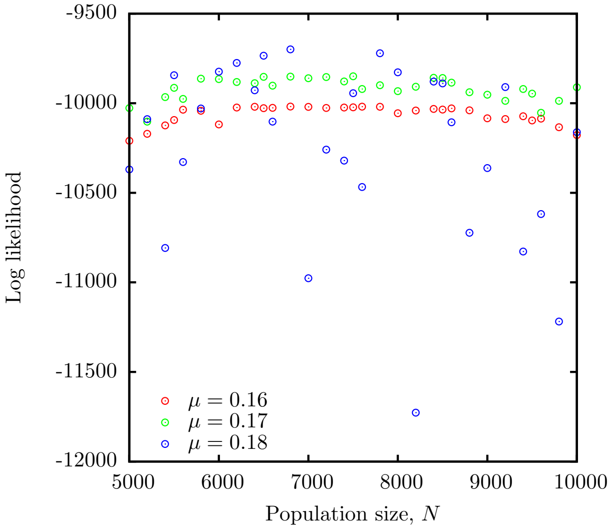


Figure S1: Approximate log likelihood of the HAWI sample given the model with multiplicative fitness and no additional selection mechanisms computed via the Gaussian kernel smoothing as a function of the model population size N for three values of haplotype discovery rate μ. For higher values of μ the empirically observed pairs are more likely to occur outside the cloud yielded by the simulation of the model, resulting in a higher level of noise.

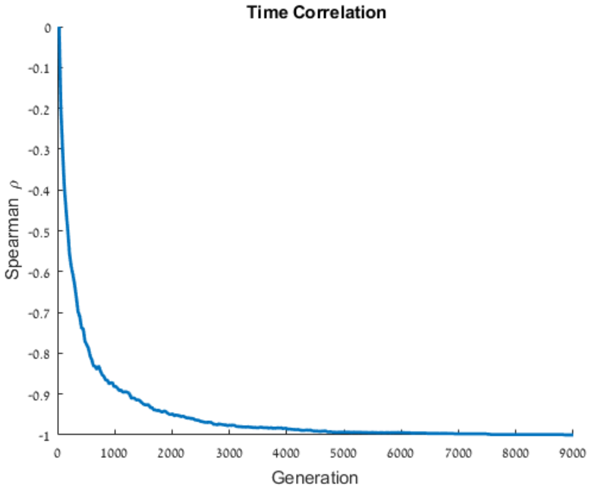


Figure S2: Time Correlations - Correlation between the Family size distribution as a function of the difference in generations.

`org/` for the definition of these abbreviations. Each file contains a space separated list of haplotype abundance,homozygocity pairs, one per line. Each line corresponds to a unique haplotype observed in the sample.

5 Population Sample Sizes

18 detailed race/ethnic sub-populations and 5 broad race/ethnic populations were studied. Each category was defined based on the registry donor self-identified race and ethnicity (SIRE). Broad categories are marked in bold and represent the sum of the detailed categories listed above each broad category in table. Some populations are merged populations, and we mark the merged population associated with each sub-population. EURCAU was never fitted because it is closely similar to CAU and often overlaps with it.

Symbol	Race Group	Sample Size	Global population
AAFA	African American	1184776	AFA
AFB	African	77984	AFA
CARB	Black Caribbean	90538	AFA
SCSEAI	South Asian	507364	API
FILII	Filipino	144044	API
HAWI	Hawaiian or other Pacific Islander	36252	API
JAPI	Japanese	90340	API
KORI	Korean	208560	API
NCHI	Chinese	290480	API
AINDI	Other Southeast Asian	110040	API
VIET	Vietnamese	113748	API
EURCAU	European Caucasian	3472992	CAU
MENAFC	MidEast/No. Coast of Africa	198752	CAU
MSWHIS	Mexican or Chicano	716492	HIS
SCAHIS	South/Cntrl Amer. Hisp.	425480	HIS
CARHIS	Caribbean Hispanic	332920	HIS
CARIBI	Caribbean Indian	42484	NAM
AMIND	North American Indian	109796	NAM
AISC	American Indian South or Central American	5926	NAM
SCAMB	Black South or Central America	4889	AFA
ALANAM	Alaska Native or Aleut	1376	NAM
AFA	African American*	1358187	
API	Asian and Pacific Islander*	993464	
CAU	Caucasian*	3671744	
HIS	Hipanic*	1141972	
NAM	Native American Indian*	159582	

6 Models Joint Probability Distribution

Model joint probability distribution P(c; z) measured via stochastic simulations for the model with the multiplicative fitness function and without additional selection mechanisms for different populations. Each circle corresponds to a pair (c,z) that occurred at least once among 105 uncorrelated snapshots of the population in steady state. The color represents ln P(c,z) (see color bar). Crosses are the (c,z) pairs observed in the specified population sample of the MHC haplotype data. Cross size is proportional to the logarithm of the number of pairs present in the sample.Black circle represents the approximate length scale λ of the Gaussian kernel smoothing.

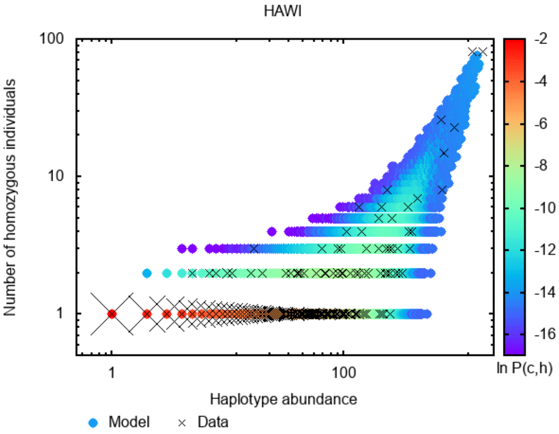


Figure S3: HAWI population

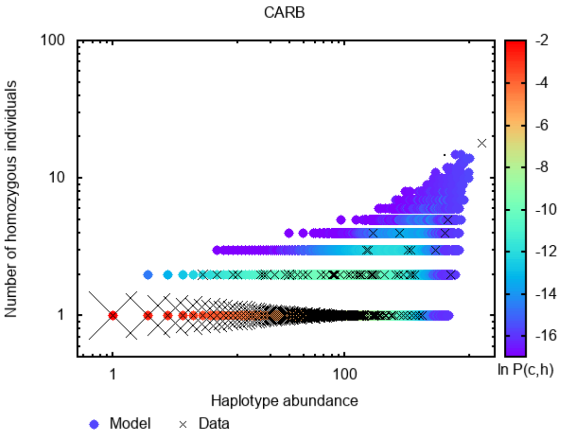


Figure S4: CARB population

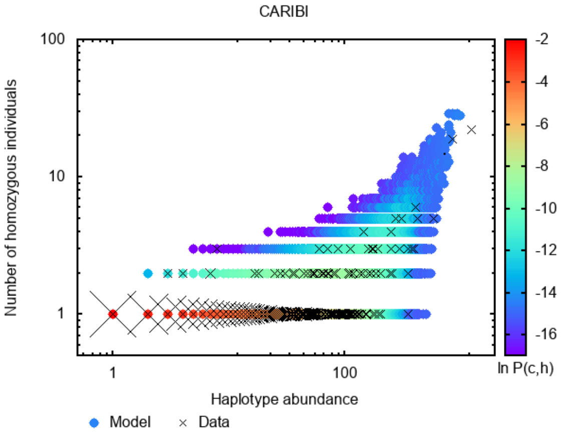


Figure S5: CARIBI population

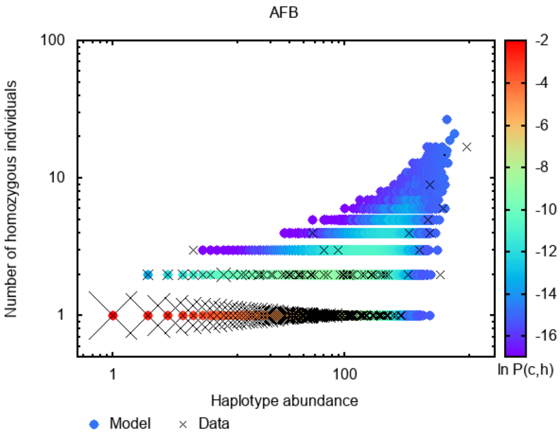


Figure S6: AFB population

7 Models And Populations

The above table shows all the populations and all the models that were fit for every population. The main limiting factor in the analysis of all populations was that too large populations, such as CAU took too long to run in complex model, and in some populations the optimization did not converge. $\surd$ represents runners that have managed to converge,× represents runs that failed to converge.

Populations & Models

Populations/Models	*Neutral*	*Multiplicative*	*Additive*	*Additive with decay*	*Additive with FDS*	*Additive with over dominance*	*Hybrid* $h_1 + h_2 - h_1h_2$	*Assortative*	*Hybrid with over dominance*	*Multiplicative with decay*	*Multiplicative with FDS*	*Population structure*
AAFA	✓	✗										
AFA	✓	✓										
AFB	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓
AINDI	✓	✓										
AMIND	✓	✓	✓		✓	✓	✓		✓	✓	✓	
API	✓	✓										
CARB	✓	✓	✓	✓	✓	✓	✓		✗	✓	✓	✓
CARHIS	✓	✓										
CARIBI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CAU	✗	✗										
FILII	✓	✓	✓			✓	✓			✓	✓	
HAWI	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓
HIS	✓	✓										
JAPI	✓	✓	✓			✓	✓	✗	✓	✓	✓	
KORI	✓	✓										

Populations & Models

Populations/Models	*Neutral*	*Multiplicative*	*Additive*	*Additive with decay*	*Additive with FDS*	*Additive with over dominance*	*Hybrid* $h_1 + h_2 - h_1 h_2$	*Assortative*	*Hybrid with over dominance*	*Multiplicative with decay*	*Multiplicative with FDS*	*Population structure*
MENAFC	✓	✓										
MSWHIS	✓	✗										
NAM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NCHI	✓	✓										
SCAHIS	✓	✓										
SCSEAI	✓	✓									✓	
VIET	✓	✓									✓	

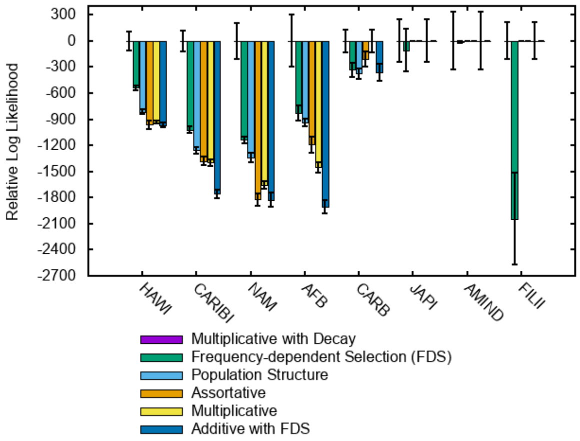
8 Comparison Of The Naive Models

The above table shows the reslults of the naive model on all populations.

	Neutral	Additive	Global Multiplicative
AAFA	-210980.55453		
AFA	-122870.69846		-67757.34845
AFB	-41016.84553	-29254.83215	-28437.11621
AINDI	-121068.41288		-104698.47258
AMIND	-39931.53231	-24272.87770	-23750.12645
API	-94493.79178		-53499.45159
CARB	-46689.07652	-31944.37530	-31098.67737
CARHIS	-91113.21129		-53355.79064
CARIBI	-21299.02252	-15805.37168	-15072.00191
FILII	-41499.23849	-25961.31774	-24685.86635
HAWI	-15413.79707	-10342.00502	-9864.63946
HIS	-231869.15870		-113854.11114
JAPI	-24611.46147	-14227.95389	-13842.83883
KORI	-46164.32001		-26305.48005
MENAFC	-96009.04724		-52657.94216
MSWHIS	-144799.26703		
NAM	-27099.33242	-17900.39211	-17070.70162
NCHI	-78076.73122		-38885.43224
SCAHIS	-135810.70427		-80124.00008
SCSEAI	-44843.46141		-30162.04784
VIET	-39196.56595		-20846.16900

9 Likelihood Comparison

Peak log likelihood for models with multiplicative fitness and additional mechanisms of selection. The values are compared with the multiplicative selection model with time decay. Thus, a value of 0 means similar likelihood to this model. Negative values represent a worse fit. Except for the additive with FDS, all models here are extensions of the multiplicative fitness model.


Relative Log Likelihood
300
0
-300
-600
-900
-1200
-1500
-1800
-2100
-2400
-2700
HAWI
CARIBI
NAM
AFB
CARB
JAPI
AMIND
FILII
Multiplicative with Decay
Frequency-dependent Selection (FDS)
Population Structure
Assortative
Multiplicative
Additive with FDS