Supporting Information for

Aaron P. Ragsdale*

March 1, 2021

Supporting Information

The diffusion equation and moment system for the two-locus sampling distribution

The two-locus diffusion equation with additive selection was first described by Kimura (1955) studied extensively in the 1960s and 70s, including by Hill and Robertson (1966) and Ohta and Kimura (1969). The continuous distribution $\psi(x_1, x_2, x_3)$ of haplotype frequencies in a population, where x_1 is the frequency of AB, x_2 of Ab, and x_3 of aB, is governed by the multi-dimensional Fokker-Planck equation:

$$\begin{aligned}\frac{\partial \psi}{\partial \tau} =& \frac{1}{2}\sum_{1\le i,j\le 3}\frac{\partial^2}{\partial x_i \partial x_j}\left[\frac{x_i(\delta_{i=j}-x_j)\psi}{\nu(\tau)}\right] \\ &- \frac{\rho}{2}\left(-\frac{\partial}{\partial x_1}D\psi + \frac{\partial}{\partial x_2}D\psi + \frac{\partial}{\partial x_3}D\psi\right) \\ &- \frac{\gamma_A}{2}\left[\frac{\partial}{\partial x_1}x_1(1-x_1-x_2)\psi + \frac{\partial}{\partial x_2}x_2(1-x_1-x_2)\psi - \frac{\partial}{\partial x_3}x_3(x_1+x_2)\psi\right] \\ &- \frac{\gamma_B}{2}\left[\frac{\partial}{\partial x_1}x_1(1-x_1-x_3)\psi - \frac{\partial}{\partial x_2}x_2(x_1+x_3)\psi + \frac{\partial}{\partial x_3}x_3(1-x_1-x_3)\psi\right].\end{aligned} \tag{S1}$$

D is the standard covariance measure of linkage disequilibrium,

$$D = x_1 - (x_1+x_2)(x_1+x_3) = x_1x_4 - x_2x_3,$$

γ_A and γ_B are the scaled selection coefficients at the left and right locus, and ρ is the scaled recombination rate between the two loci. Time τ is measured in $2N_e$ generations.

Given a function ψ that solves Equation S1, the two-locus sampling distribution for a sample size of n haploids can be found by integrating Ψ against the multinomial sampling function, so that

$$\Psi_n(i,j,k) = \binom{n}{i,j,k,n-i-j-k}\int\limits_{\substack{x_1,x_2,x_3\ge 0\\ x_1+x_2+x_3\le 1}}\int\int \psi(x_1,x_2,x_3)x_1^i x_2^j x_3^k(1-x_1-x_2-x_3)^{n-i-j-k}dx_1dx_2dx_3. \tag{S2}$$

- instead of (numerically) solving for ψ and using @ref(eq:multinomial}, we can apply the multinomial sampling formula directly to the PDE
- what we get instead is a system of differential equations on the entries of Ψ_n
- we can then solve for these noncanonical moments of ψ, which are the objects of interest in any case, and avoid having to solve this numerical PDE with unpleasant boundary conditions
- the same system can be found by deriving recursion from considering tracking n lineages within the population, in the vein of (**Wright1931-xx?**)

*aaronpeaceragsdale@gmail.com

Drift, mutation, recombination, and moment closure

- point to supplement of Ragsdale and Gravel (2019)

Selection

Data analysis

DFE for missense and LOF variants

Loss-of-function (LOF) variants show a dramatic skew toward low-frequency variants across all human populations (Table S4). Here, using the folded SFS for synonymous, missense, and LOF mutations across all autosomal genes, I infer DFEs for missense and LOF mutations independently. I consider a few different dominance coefficients to explore the effect of the assumed recessivity of the two classes of mutations.

The standard SFS approach to fitting the DFE involves first inferring a demographic history for the population using putatively neutral variants (here, synonymous mutations), and then fixing that demography and fitting a parameterized function for the distribution of selection coefficients for new mutations for the selected classes. DFE inference also requires an estimate for the total mutation rate of the different mutation classes, as much of the signal for strongly selected mutations comes from observing fewer mutations than expected given a known mutation rate (with the assumption that selection purges some fraction of strongly deleterious mutations which are unseen in the sample). Here, I fit demography and DFEs to the SFS from the Mende in Sierra Leone (MSL) using *moments* (Jouganous et al. 2017).

I used the mutation model from Karczewski et al. (2020) to estimate the total mutation rate across autosomal genes (uL, where u is the per-base mutation rate, and L is the total length of the coding genome). These values were $(0.1442, 0.3426, 0.0256)$ for synonymous, missense, and LOF mutations, respectively. Roughly two thirds of new mutations in coding regions are expected to be missense mutations, while only 5% of new mutations are LOF. I fit a demographic model to the synonymous variants, which included a population expansion in the deeper past and exponential growth in the recent past (Figure **??**A). Using the inferred optimal scaled mutation rate, $\theta = 4N_e uL$, I estimated $Ne \approx 12,300$, and assuming an average generation time of 29 years I converted the inferred genetic units to physical units. The best-fit model had a roughly two-fold expansion 400 thousand years ago, and then exponential growth over the past 20-30 thousand years, with a current effective size of $\sim 63,000$.

Under this demographic model, I fit a gamma distribution for the distribution of fitness effects to missense and LOF mutations (Table S5). For each fit, I fixed the scaled mutation rate for each mutation class, so that $\theta_{mis} = \frac{u_{mis}}{u_{syn}}\hat{\theta}_{syn}$ and $\theta_{lof} = \frac{u_{lof}}{u_{syn}}\hat{\theta}_{syn}$, where values of u were found using the GNOMAD mutation model (Karczewski et al. 2020). I tested three values for the dominance coefficient h, 0, 0.2 and 0.5. For missense mutations, $h = 0$ gave a poor fit to the data, and $h = 0.5$ fit best among the three tested dominance coefficients. For LOF variants, $h = 0$ also fit poorly, but $h = 0.2$ and $h = 0.5$ gave similar likelihoods, highlighting that inferring dominance using the SFS is poorly constrained. Regardless of the dominance coefficient assumed, however, the vast majority of LOF variants were inferred to be strongly deleterious, with only $\sim 10\%$ of new mutations having selection coefficients on the order $1/N_e$ or less.

Multinucleotide mutations and positive LD between linked synonymous variants

Multinucleotide mutations (MNMs) are complex mutational events that result in multiple mutations occurring on the same haplotype background in a single generation. Because MNMs fall on the same haplotype, those mutations will be in positive LD, and LD between those pairs that are very tightly linked will not be broken down all that rapidly. MNMs are expected to occur over relatively short distances, on the order of 10s or 100s of base pairs, making them a likely culprit of the observed positive LD among synonymous mutations at short distances.

Multinucleotide mutations can be easily incorporated into the moment system with a simple adjustment to the mutation operator. Instead of all mutations occurring independently in haplotypes with mutations

already segregating at the other locus, some fraction of new mutations could instead occur spontaneously and create a new pair of mutations with initial counts $n_{AB} = 1$ and $n_{ab} = n - 1$.

Here, I fit a simple exponential model for the fraction of new mutations at a given distance that arose through a MNM event, so that $P(MNM|d) = Ae^{-\lambda d}$, where d is the distance separating pairs of mutations. I considered all synonymous mutations within genes in the MSL population and used the same population size history model as inferred in the DFE section above for a demographic control. This left two parameters to be fit, A and λ, which I fit to the binned decay curve of σ_d^1. I needed to assume an average per-base recombination rate r across gene regions, and tested a number of values between 10^{-9} and 2×10^{-8}. The optimization was insensitive to the chosen value or r, because the decay of positive LD occurs rapidly. For any plausible value of r, the σ_d^1 decays to zero well before distances between pairs have scaled recombination rates $\rho = 4N_e rd$, and expected statistics for $\rho \ll 1$ vary only negligibly.

In fitting the LD decay of σ_d^1, the best fit parameters were $A = 0.132$, and $\lambda = 0.0103$. An exponential scaling of 0.01 implies that the vast majority of new mutations do not occur via MNMs for distances greater than 200 bp, though a substantial fraction $(10 - 15\%)$ occur via MNMs for very tightly linked loci with distances on the order $0 - 50$ base pairs. It is important to note that this does not mean that $10 - 15\%$ of new mutations occur via MNMs, since this fraction is contitioned on two mutations occurring at short distances.

Grouping Thousand Genomes populations based on clustering

Testing grouping populations that cluster together as a single population.

- Does not reduce CIs

Citation for Alex's work: https://www.abstractsonline.com/pp8/#!/9070/presentation/2384

Supplementary Tables

Table S1: General selection model for diploids and dominance models.

Diploid genotype	General model	Simple dominance	Gene-based dominance
AB / AB	$1 + s_{AB/AB}$	$1 + 2s_A + 2s_B$	$1 + 2s$
AB / Ab	$1 + s_{AB/Ab}$	$1 + 2s_A + 2s_B h_B$	$1 + 2s$
AB / aB	$1 + s_{AB/aB}$	$1 + 2s_A h_A + 2s_B$	$1 + 2s$
AB / ab	$1 + s_{AB/ab}$	$1 + 2s_A h_A + 2s_B h_B$	$1 + 2sh$
Ab / Ab	$1 + s_{Ab/Ab}$	$1 + 2s_A$	$1 + 2s$
Ab / aB	$1 + s_{Ab/aB}$	$1 + 2s_A h_A + 2s_B h_B$	$1 + 2s$
Ab / ab	$1 + s_{Ab/ab}$	$1 + 2s_A h_A$	$1 + 2sh$
aB / aB	$1 + s_{aB/aB}$	$1 + 2s_B$	$1 + 2s$
aB / ab	$1 + s_{aB/ab}$	$1 + 2s_B h_B$	$1 + 2sh$
ab / ab	1	1	1

Table S2: Haploid epistasis model.

Haplotype	Fitness
AB	$(1 + s_A + s_B)(1 + \epsilon)$
Ab	$1 + s_A$
aB	$1 + s_B$
ab	1

Table S3: Thousand Genomes Project population descriptions for populations used in this study.

Code	Description	Region
ESN	Esan in Nigeria	Africa
GWD	Gambian in Western Divisions in the Gambia	Africa
LWK	Luhya in Webuye, Kenya	Africa
MSL	Mende in Sierra Leone	Africa
YRI	Yoruba in Ibadan, Nigeria	Africa
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	Europe
GBR	British in England and Scotland	Europe
FIN	Finnish in Finland	Europe
IBS	Iberian Population in Spain	Europe
TSI	Toscani in Italia	Europe
CDX	Chinese Dai in Xishuangbanna, China	East Asia
CHB	Han Chinese in Beijing, China	East Asia
CHS	Southern Han Chinese	East Asia
JPT	Japanese in Tokyo, Japan	East Asia
KHV	Kinh in Ho Chi Minh City, Vietnam	East Asia

Table S4: Tamija's D for classes of coding mutations, both within annotated domains and outside of domains.

Population	Mutation type	Region	Tajima's D
ESN	Synonymous	All	-0.882
		In domain	-0.854
		Not in domain	-0.921
	Missense	All	-1.414
		In domain	-1.535
		Not in domain	-1.293
	Loss of function	All	-1.483
		In domain	-2.156
		Not in domain	-1.282
GWD	Synonymous	All	-1.011
		In domain	-0.981
		Not in domain	-1.052
	Missense	All	-1.566
		In domain	-1.678
		Not in domain	-1.452
	Loss of function	All	-1.697
		In domain	-2.328
		Not in domain	-1.501
LWK	Synonymous	All	-1.109
		In domain	-1.088
		Not in domain	-1.139
	Missense	All	-1.589
		In domain	-1.700
		Not in domain	-1.477
	Loss of function	All	-1.666
		In domain	-2.278
		Not in domain	-1.477
MSL	Synonymous	All	-0.983
		In domain	-0.959
		Not in domain	-1.017
	Missense	All	-1.501
		In domain	-1.603
		Not in domain	-1.400
	Loss of function	All	-1.559
		In domain	-2.303
		Not in domain	-1.332
YRI	Synonymous	All	-0.928
		In domain	-0.898
		Not in domain	-0.971
	Missense	All	-1.467
		In domain	-1.586
		Not in domain	-1.348
	Loss of function	All	-1.624
		In domain	-2.237
		Not in domain	-1.424
CEU	Synonymous	All	-0.417
		In domain	-0.392

Table S4: Tamija's D for classes of coding mutations, both within annotated domains and outside of domains. *(continued)*

Population	Mutation type	Region	Tajima's D
		Not in domain	-0.452
	Missense	All	-1.248
		In domain	-1.404
		Not in domain	-1.082
	Loss of function	All	-1.501
		In domain	-2.196
		Not in domain	-1.280
FIN	Synonymous	All	-0.058
		In domain	-0.047
		Not in domain	-0.075
	Missense	All	-0.883
		In domain	-1.048
		Not in domain	-0.710
	Loss of function	All	-1.200
		In domain	-2.034
		Not in domain	-0.906
GBR	Synonymous	All	-0.319
		In domain	-0.300
		Not in domain	-0.345
	Missense	All	-1.120
		In domain	-1.276
		Not in domain	-0.954
	Loss of function	All	-1.313
		In domain	-2.178
		Not in domain	-0.997
IBS	Synonymous	All	-0.689
		In domain	-0.664
		Not in domain	-0.724
	Missense	All	-1.424
		In domain	-1.560
		Not in domain	-1.279
	Loss of function	All	-1.636
		In domain	-2.349
		Not in domain	-1.378
TSI	Synonymous	All	-0.650
		In domain	-0.625
		Not in domain	-0.685
	Missense	All	-1.422
		In domain	-1.568
		Not in domain	-1.266
	Loss of function	All	-1.655
		In domain	-2.349
		Not in domain	-1.397
CDX	Synonymous	All	-0.374
		In domain	-0.366
		Not in domain	-0.385
	Missense	All	-1.179

Table S4: Tamija's D for classes of coding mutations, both within annotated domains and outside of domains. *(continued)*

Population	Mutation type	Region	Tajima's D
		In domain	-1.323
		Not in domain	-1.026
	Loss of function	All	-1.360
		In domain	-2.194
		Not in domain	-1.062
CHB	Synonymous	All	-0.598
		In domain	-0.593
		Not in domain	-0.606
	Missense	All	-1.389
		In domain	-1.528
		Not in domain	-1.239
	Loss of function	All	-1.586
		In domain	-2.344
		Not in domain	-1.298
CHS	Synonymous	All	-0.544
		In domain	-0.545
		Not in domain	-0.544
	Missense	All	-1.334
		In domain	-1.499
		Not in domain	-1.150
	Loss of function	All	-1.559
		In domain	-2.290
		Not in domain	-1.292
JPT	Synonymous	All	-0.371
		In domain	-0.368
		Not in domain	-0.376
	Missense	All	-1.194
		In domain	-1.355
		Not in domain	-1.019
	Loss of function	All	-1.410
		In domain	-2.272
		Not in domain	-1.086
KHV	Synonymous	All	-0.576
		In domain	-0.562
		Not in domain	-0.596
	Missense	All	-1.346
		In domain	-1.473
		Not in domain	-1.210
	Loss of function	All	-1.535
		In domain	-2.294
		Not in domain	-1.269

Table S5: DFEs inferred for missense and loss-of-function variants in MSL for varying values of h. General patterns are consistent across different chosen values of h, although for $h = 0$ results in poorer fits for both missense and LOF variants. Columns to the right of the log-likelihood (LL) column show proportions of new mutations with $|s|$ in each given bin.

Class	h	shape	scale	LL	$[0, 10^{-5})$	$[10^{-5}, 10^{-4})$	$[10^{-4}, 10^{-3})$	$[10^{-3}, 10^{-2})$	$[10^{-2}, \infty)$
Missense	0.0	0.093	768505	-678.2	0.260	0.062	0.077	0.096	0.505
	0.2	0.138	6660	-416.7	0.260	0.098	0.134	0.182	0.327
	0.5	0.147	2117	-392.0	0.282	0.114	0.159	0.214	0.231
LOF	0.0	0.132	99999054	-248.3	0.077	0.028	0.037	0.051	0.807
	0.2	0.177	477994	-226.7	0.083	0.042	0.063	0.095	0.717
	0.5	0.188	121419	-224.2	0.092	0.050	0.077	0.119	0.662

Supplementary Figures

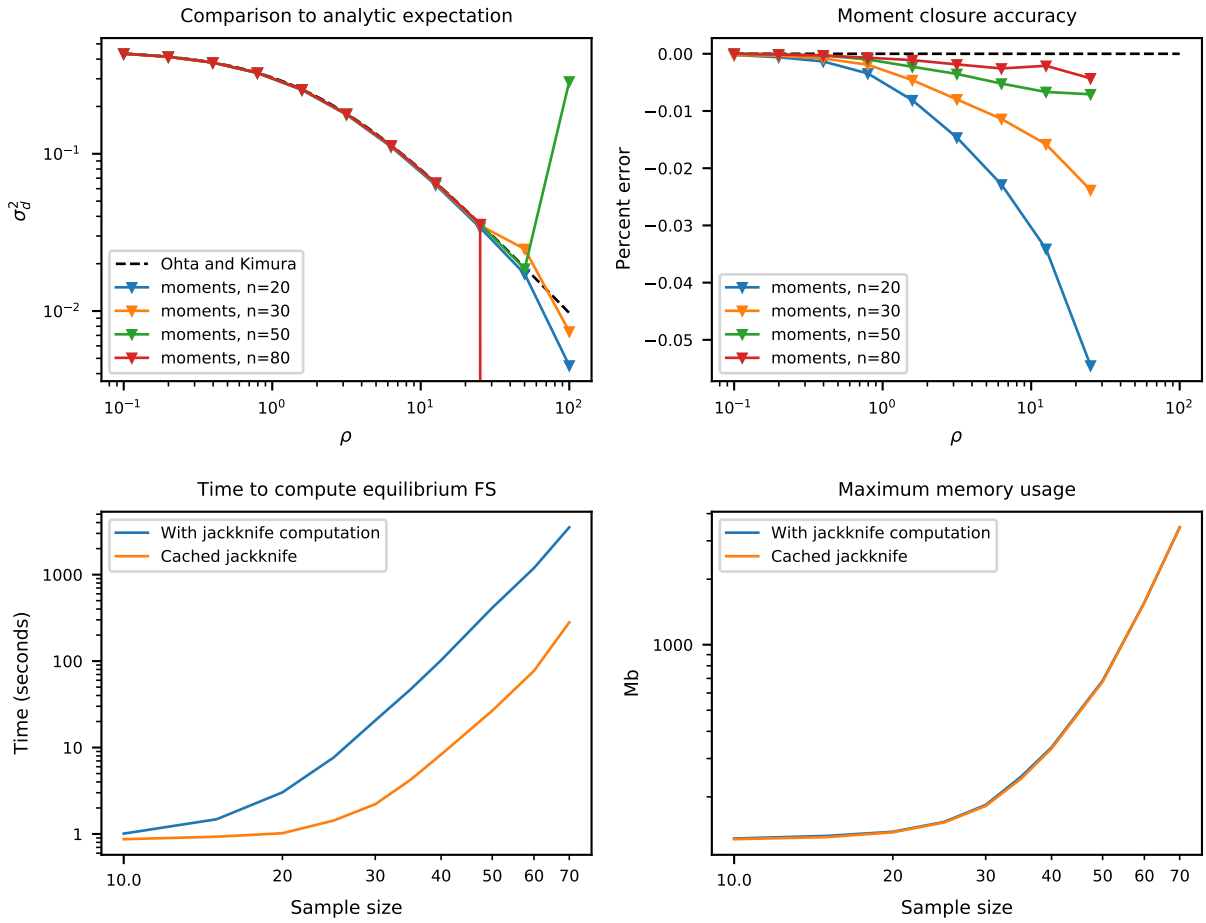

Figure S1: **Accuracy of the jackknife approximation and runtime** Small sample sizes can lead to large error in the closure approximation for larger recombination distances or selection coefficients. Generally, the jackknife approximation breaks down for recombination rates greater than $\rho \approx 30$. While increasing sample size leads to more accurate solutions, it comes at the cost of both increased runtime and memory usage. Most analyses performed in this paper used n between 40 and 70.

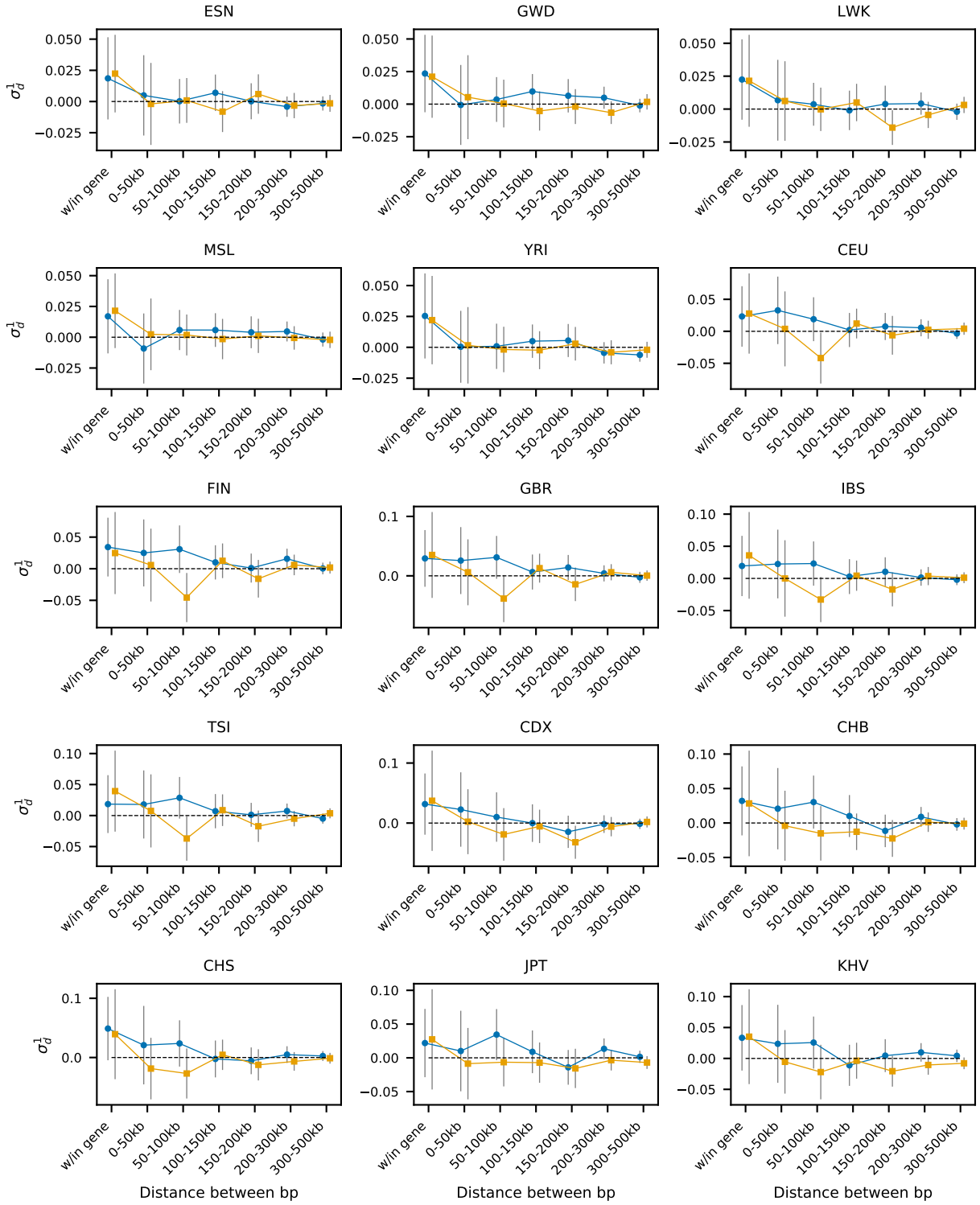

Figure S2: **LD within and between protein-coding genes.**

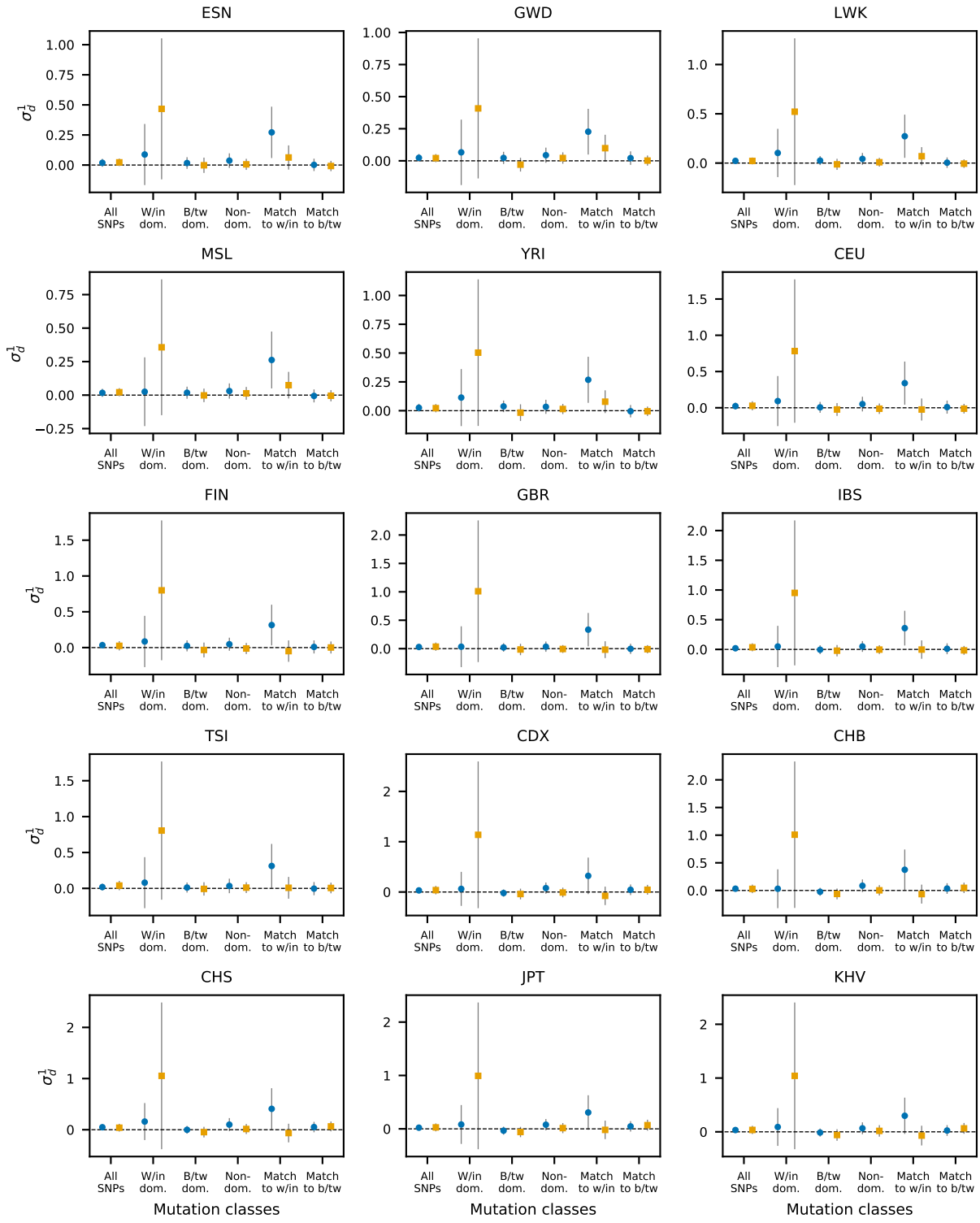

Figure S3: **LD within and between coding domains and pairs outside domains at matched distances.**

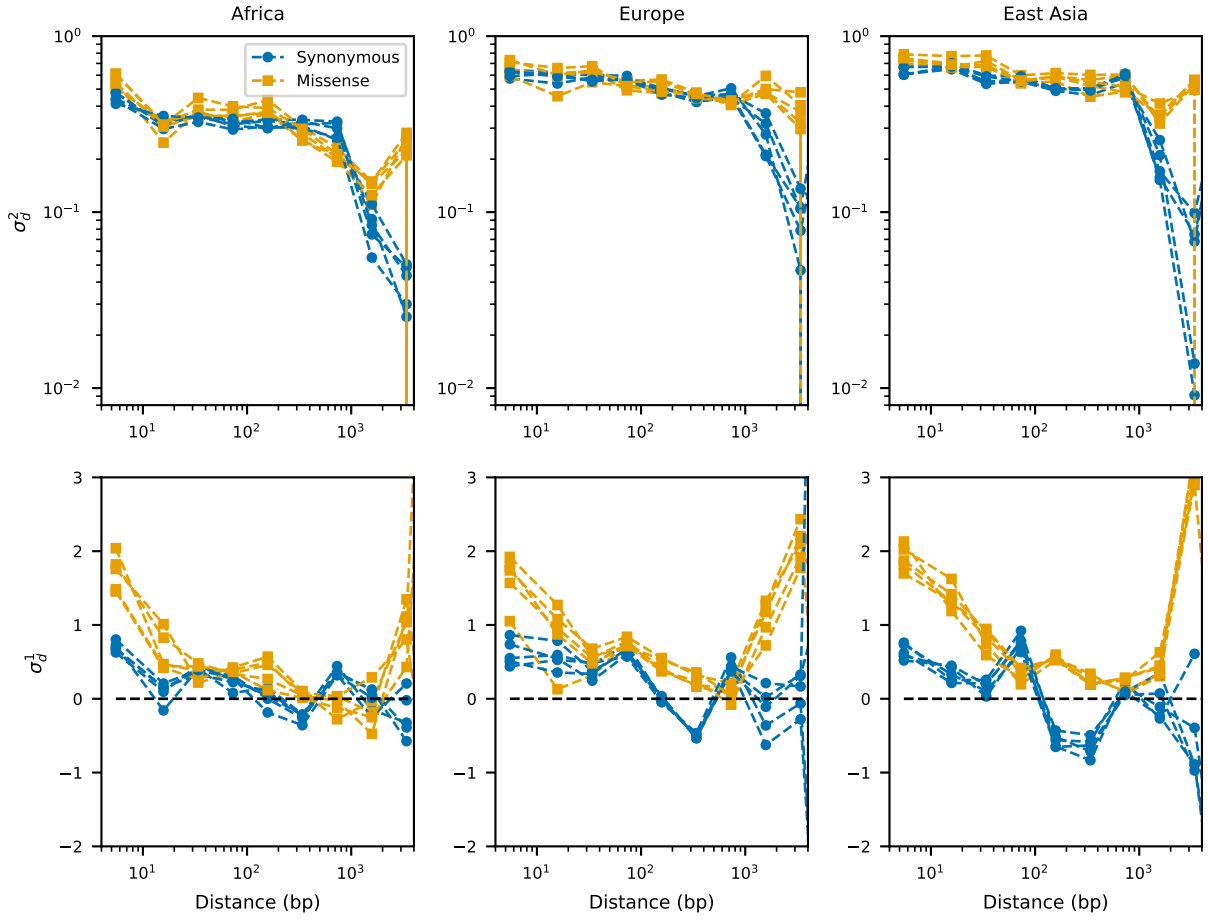

Figure S4: **LD decay for synonymous and missense mutations for pairs of mutations that fall inside the same domains.**

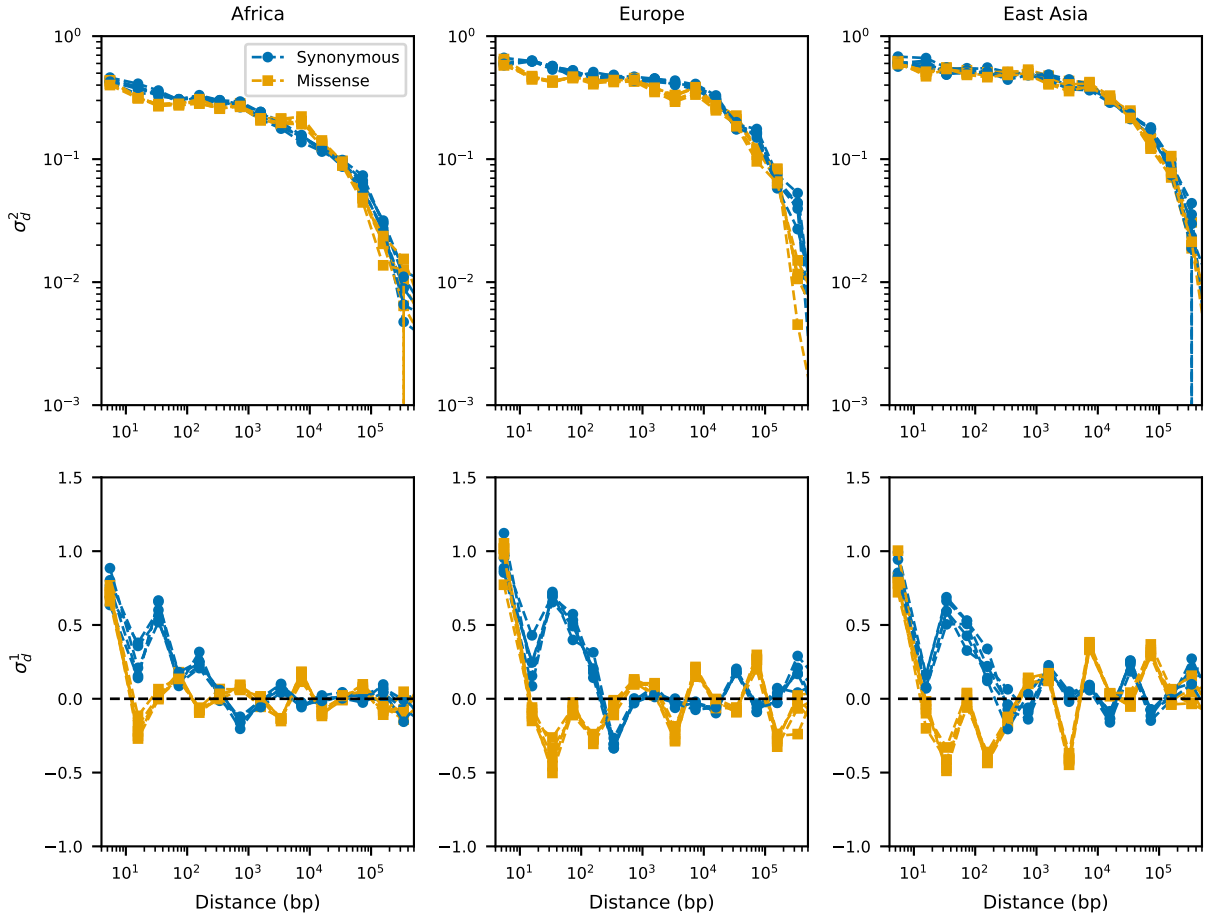

Figure S5: **LD decay for synonymous and missense mutations for pairs of mutations that fall outside of domains.**

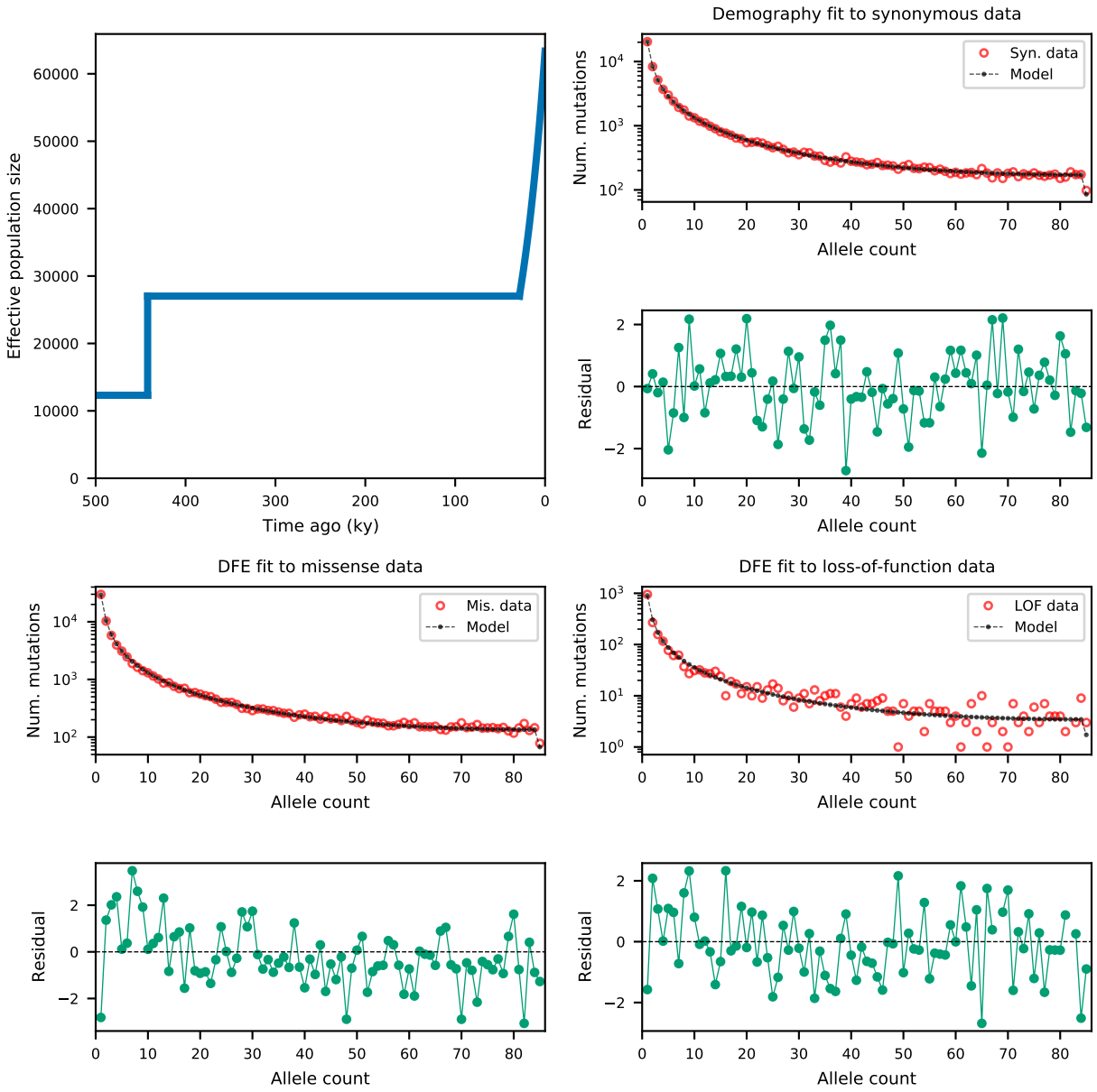

Figure S6: **Demography and DFE for MSL.** A demographic model was fit to the folded synonymous SFS, and DFEs were fit to missense and loss-of-function SFS. Shown here are DFEs fit with $h = 0.5$.

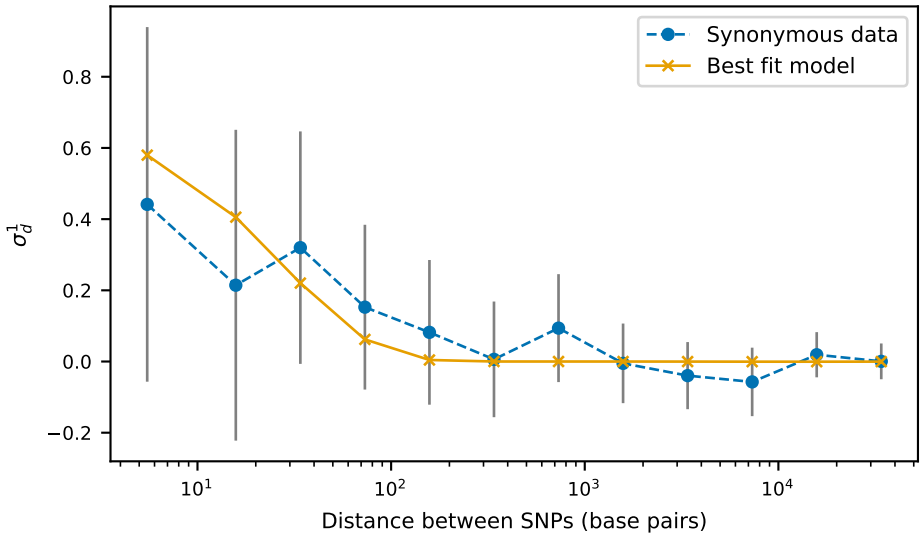

Figure S7: **Optimization of fraction of new mutations arising via multinucleotide mutations by distance.** A simple exponential function was fit to describe the probability that a pair of mutations arose through a MSM event at a given distance d, as $Ae^{-\lambda d}$. Across all recombination rates tested, the best fit parameters were $A = 0.13$ and $\lambda = 0.010$.

Supporting References

Hill, W G, and A Robertson. 1966. "The Effect of Linkage on Limits to Artificial Selection." *Genet. Res.* 8 (3): 269–94.

Jouganous, Julien, Will Long, Aaron P Ragsdale, and Simon Gravel. 2017. "Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation." *Genetics* 206 (3): 1549–67.

Karczewski, Konrad J, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581: 434–43.

Kimura, Motoo. 1955. "Random Genetic Drift in Multi-Allelic Locus." *Evolution* 9 (4): 419–35.

Ohta, T, and M Kimura. 1969. "Linkage Disequilibrium at Steady State Determined by Random Genetic Drift and Recurrent Mutation." *Genetics* 63 (1): 229–38.

Ragsdale, Aaron P, and Simon Gravel. 2019. "Models of Archaic Admixture and Recent History from Two-Locus Statistics." *PLoS Genet.* 15 (6): e1008204.