

Can we distinguish modes of selective interactions using linkage disequilibrium?

Aaron P. Ragsdale
aaronpeaceragsdale@gmail.com

Department of Human Genetics, McGill University, Montreal, Canada
National Laboratory of Genomics for Biodiversity, Irapuato, Mexico

March 24, 2021

Abstract

Selected mutations interfere and interact with evolutionary processes at nearby loci, distorting allele frequency trajectories and correlations between pairs of mutations. A number of recent studies have used patterns of linkage disequilibrium between selected variants to test for selective interference and epistatic interactions, with some disagreement over the interpretation of observations from data. Interpreting this data is hindered by the relative lack of analytic or even numerical expectations for patterns of variation between pairs of loci under the combined effects of selection, dominance, epistasis, and demography. Here, I develop a numerical approach to compute the expected two-locus sampling distribution under diploid selection with arbitrary epistasis and dominance, recombination, and variable population size. I use this to explore how epistasis and dominance affect expected signed LD, including for non-steady-state demography relevant to human populations. Finally, I use whole-genome sequencing data from humans to assess how well we can differentiate modes of selective interactions in practice. I find that positive LD between missense mutations within genes is driven by strong positive allele-frequency correlations between pairs of mutations that fall within the same conserved domain, pointing to compensatory mutations or antagonistic epistasis as the prevailing mode of interaction within but not outside of conserved genic elements. The heterogeneous landscape of both mutational fitness effects and selective interactions within protein-coding genes calls for more refined inferences of the joint distribution of fitness and interactive effects, and the methods presented here should prove useful in that pursuit.

Introduction

Most new mutations that affect fitness are deleterious and tend to be eliminated from a population. The amount of time that a deleterious mutation segregates depends on the strength of selection against genomes that carry it, with very damaging mutations kept at low frequencies and purged relatively rapidly. In the time between mutation and fixation or loss, selected variants, both beneficial and damaging, can dramatically impact patterns of variation in nearby linked regions. This distortion away from neutral expectations is now empirically documented using sequencing data from an ever-growing set of study systems, but questions remain about the primary mode of interactions between multiple linked variants and their joint effects on genome-wide patterns of diversity.

In their foundational paper, [Felsenstein \(1974\)](#) recognized that linked selected variants reciprocally impede the efficacy of selection at each locus, a process known as selective interference. Linked selection reduces the fixation probability of advantageous mutations and increases that of deleterious mutations compared to expectations under single-locus models ([Felsenstein, 1974](#)). Allele frequency dynamics and correlations of linked selected variants are also predicted to deviate from expectations without interference. Under a multiplicative fitness model, where the fitness reduction of a genome carrying multiple deleterious variants is equal to the product of the fitness reduction of each mutation independently, we expect to see net linkage disequilibrium (LD) equal to zero for unlinked sites ([Felsenstein, 1974](#)). But for linked loci, those mutations are expected to segregate on different haplotypes more

often than together, leading to negative, or repulsion, LD, although the extent of LD depends non-trivially on the strength of selection and probability of recombination separating loci (??).

Non-additive effects, including dominance (i.e., interactions *within* loci) and epistasis (interactions *between* loci), further complicate our evolutionary models. A large fraction of nonsynonymous coding mutations are thought to be at least partially recessive (??), with average levels of dominance correlating with strength of selection (?), and dominance plays an important role in shaping expected equilibrium allele frequencies and the mutation load of strongly damaging disease mutations (?). On the other hand, epistasis differentially impacts the deleterious load for asexually and sexually reproducing organisms (??), has been invoked as an explanation for the evolutionary advantage of sex (???), and can drive incompatibilities that lead to postzygotic isolation during the process of speciation (?). Within populations, epistasis is known to cause signed LD to deviate dramatically from zero (??). However, despite appreciation of the effect of dominance on linked variation (??) and the evolutionary importance of epistatic interactions, we currently lack models for predicting correlation patterns between linked mutations under general selection models.

In this paper, I develop a numerical approach to solve for the two-locus sampling distribution under a general diploid selection model with variable recombination and single-population size history. I use this model to describe how epistasis and dominance shape expected patterns of signed LD, under both steady-state and non-equilibrium demography, that have been used to test for interference and epistasis in population genomic data. I then turn to human sequencing data and compare patterns of LD for synonymous, missense, and loss-of-function mutations within and between protein-coding genes and annotated domains. I show that while synonymous and missense variants display similar slightly positive average LD within genes, for missense mutations this signal is driven by correlations between pairs of mutations within, but not between, protein-coding domains, suggesting an importance for antagonistic epistasis or a prevalence of compensatory nonsynonymous mutations within conserved elements. Finally, I discuss how well we can expect to discriminate between modes of selective interactions using genome-wide patterns of LD.

Empirical observations

The most direct way to test for interactions between linked selected variants is through deep mutation scanning experiments, in which many distinct mutations are introduced within a target gene and then organismal fitness or some protein function is experimentally measured (????). For example, using the model system of the TEM-1 β -lactamase gene in *E. coli*, ? found evidence for synergistic epistasis, where multiple deleterious mutations had a greater effect on fitness than the multiplication of the observed effects of the individual mutations. The scale of mutation scanning experiments continues to improve dramatically, promising greater resolution of the fitness landscape in such model systems which can be compared to evolutionary theory (?).

Directed mutational studies are not possible in most natural populations, and we must turn to population genetic approaches to infer selective interactions between observed mutations. Motivated by theory that linked negatively selected mutations will display negative LD due to interference (?), and that epistasis will drive expected LD away from zero, a number of recent studies have used patterns of LD within classes of putatively selected variants to infer modes of selective interactions. In a notable study from ?, pairs of non-synonymous mutations were found to cluster more than expected along lineages in the *Drosophilid* species complex, and that those clustered mutations tended to preserve the charge of the protein and were in positive LD compared to pairs involving synonymous mutations, suggesting that compensatory nonsynonymous variants are tolerated and maintained. More recently, ? replicated this finding across a diverse set of genera, showing that such epistatic interactions are important for protein evolution.

? observed negative LD between loss-of-function variants in protein-coding genes (such as stop gains and losses, frameshifts, and other nonsense mutations) in both human and fruit fly populations. They interpreted this as evidence for widespread synergistic epistasis between these mutations, in which the fitness reduction of multiple mutations is greater than the product of that of each individual mutations independently. Within the past year or two, both ? and ? have reevaluated patterns of LD between coding variants in humans, fruit flies, and *Capsella grandiflora*, and suggested interference and dominance may instead be driving patterns of LD (?) or questioned whether LD between loss-of-function variants is significantly different from zero (?).

A number of factors impede our interpretation of patterns of signed LD between coding variants. First, for strongly deleterious or loss-of-function mutations, their rarity and low allele frequency mean that statistical measurement of LD and other diversity measures are very noisy. Second, comparisons are based on theory with limiting assumptions, such as steady-state demography, simple selection and interaction models, or unlinked loci. To generate predictions under more complex models, we rely on expensive forward simulations. Such simulations can help build intuition and test inference methods, but do not efficiently provide expectations for quantities of interest across the range of relevant parameters. Analytical and numerical methods for expected haplotype frequencies and LD under general selective interaction models are thus crucial for interpreting patterns of variation observed in data.

Results

Expected signed LD under steady-state demography

In the Methods, I expand on the moment system developed in ? to compute the expected sampling distribution Ψ_n of two-locus haplotypes under a general model of selective interactions. This sampling distribution stores the expected density or observed counts of pairs of biallelic loci with each possible haplotype frequency configuration in a sample of size n . In many cases it is simpler to visualize summaries (such as $\mathbb{E}[D]$) or slices of the distribution (Figures ?? and S1) instead of the full three-dimensional distribution. Here, I focus on low-order LD statistics, such as $\mathbb{E}[D]$ and $\mathbb{E}[D^2]$, and their decay with recombination distance as these are statistics that are commonly used to test for interactions between loci. Instead of unnormalized $\mathbb{E}[D^2]$ and $\mathbb{E}[D]$, I consider expectations for $\sigma_d^2 = \mathbb{E}[D^2]/\mathbb{E}[p(1-p)q(1-q)]$ and $\sigma_d^1 = \mathbb{E}[D]/\mathbb{E}[p(1-p)q(1-q)]$, where p and q are the derived allele frequencies at the left (A/a) and right (B/b) loci, respectively. These normalized statistics have two benefits: first, the mutation rate cancels so that expectations are robust to assumptions about the per-base mutation rate, and second, analytic expectations for these quantities are available under neutrality at constant population size for comparison (?).

I first consider the case of additive selection and epistasis, followed by dominance acting within loci but no epistatic effects, and then describe a general diploid selection model and consider simple gene-based dominance effects. In what follows I focus on parameters where the strengths of selection and dominance at each locus are equal, but note that the methods presented here allow for arbitrary and unequal selection and dominance at the two loci.

Additive selection and epistasis

For mutations under additive selection (i.e. $h = 1/2$) and no epistasis, we recover the well known Hill-Robertson interference result of negative LD between selected mutations, which is strongest for pairs of mutations that have selection coefficients $\gamma = O(1)$, or $s \approx 1/2N_e$ (Figure ??B). For strongly deleterious mutations, LD is close to zero even with tight linkage, as they almost always segregate at low enough frequencies that they are unlikely to interfere with each other (?).

With epistasis, mean signed LD is large for both weakly and strongly selected variants, with sign depending on the direction of epistatic interactions (Figure ??). Synergistic epistasis results in negative LD while antagonistic epistasis results in positive LD, and large positive LD can occur even when the epistatic effect is relatively weak. Epistasis-induced LD can extend over long distances, especially for strongly deleterious mutations. For example, in Figure ??F even moderately deleterious mutations with population-size-scaled selection coefficients of $\gamma = -10$ show large mean LD that extends well beyond 100 kilobases. More strongly deleterious interacting mutations are expected to show large signed LD over much larger recombination distances.

Dominance

The effects of epistasis between variants on LD have been widely studied. However, the combined effect of non-additive selection within a locus and interference has received less attention. ? used large-scale forward simulations to explore the effect of dominance on patterns of LD, showing that LD depends on both the

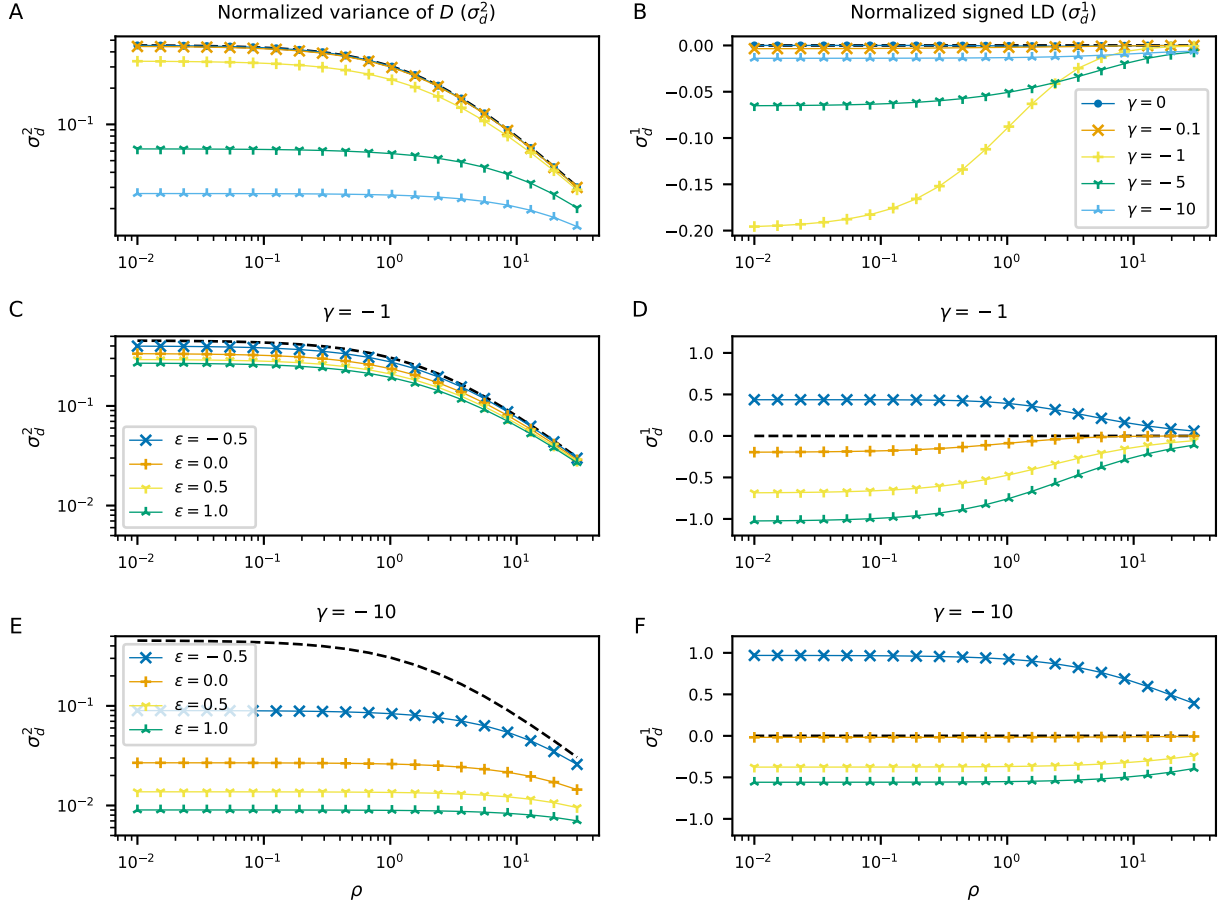


Figure 1: Additive selection and epistasis. Left panels show expectations for the decay of $\sigma_d^2 = \mathbb{E}[D^2]/\mathbb{E}[p(1-p)q(1-q)]$ with recombination distance, and right panels show expectations for $\sigma_d^1 = \mathbb{E}[D]/\mathbb{E}[p(1-p)q(1-q)]$. The scaled recombination rate is given by $\rho = 4N_e r$, where r is the probability of recombination between two loci. All expectations are at stationarity, and dashed lines are expectations under neutrality (Equation (??)). (A-B) Additive selection, with equal selection strength at both loci. Stronger selection reduces σ_d^2 (A), and in (B), we see that the negative LD effect of Hill-Robertson interference is strongest for intermediate selection ($\gamma \approx -1$ to -5 , where $\gamma = 4N_e s$). (C-F) Epistasis with additive selection. Synergistic epistasis ($\epsilon < 0$) decreases σ_d^2 while antagonistic epistasis ($\epsilon > 0$) increases it compared to expectations under a multiplicative fitness model ($\epsilon = 0$). Epistasis also pushes signed LD (σ_d^1) away from zero, with synergistic epistasis decreasing LD for negatively selected mutations and antagonistic epistasis increasing LD. The deviation of σ_d^1 from zero under epistasis can extend for long recombination distances, particularly for stronger selection.

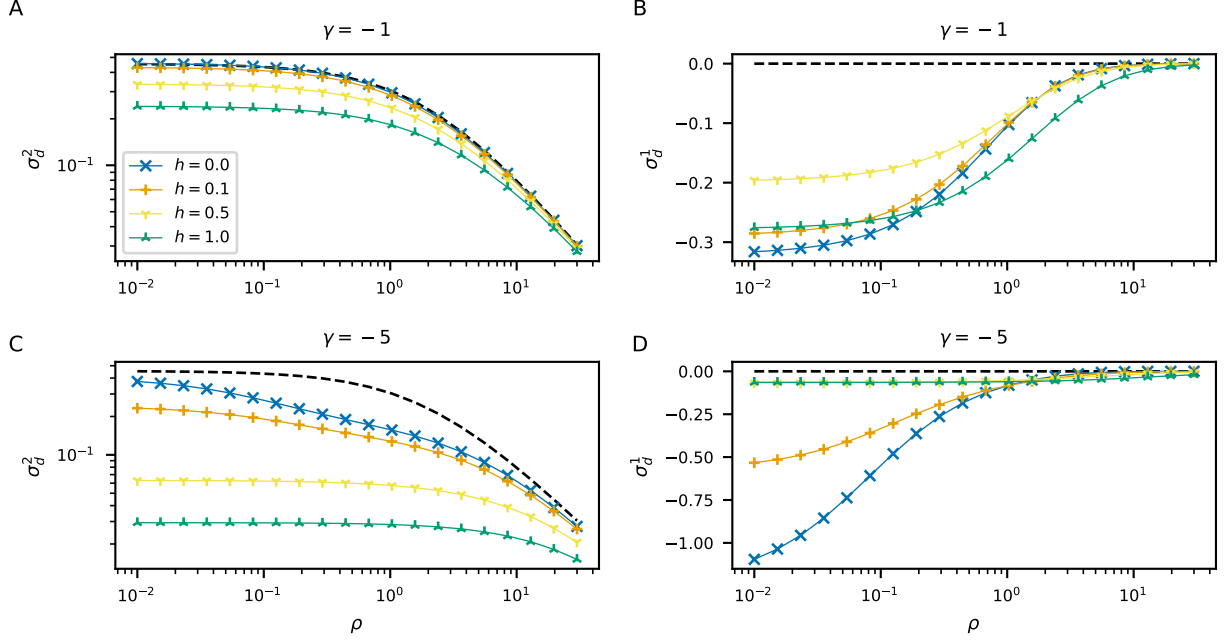


Figure 2: **Dominance effects.** (A and C) Pairs of recessive mutations ($h < 0.5$) have higher squared LD than additive mutations, while pairs of dominant mutations have lower expected D^2 . (B and D) Pairs of recessive deleterious mutations can have large negative LD, although this correlation decays more rapidly with larger recombination than LD under epistasis (Figure ??D and F). Notably, there is no simple relationship between strength of selection and dominance, particularly when s is $O(1/N)$. Note that this setup does not include epistatic effects (i.e., $\epsilon = 0$).

magnitude of the selection and dominance coefficients in a nonlinear way, but note that expensive large-scale simulations are needed to explore these effects as no closed form approach to computing those expectations exist. The method presented here provides such a closed form approach, so that expectations of signed LD and other two-locus statistics can be computed more efficiently and accurately.

The combined effect of the strength of selection and dominance on interference is indeed nontrivial, and some parameter regimes can cause strong negative LD between pairs of negatively selected variants. For strong selection, pairs of recessive mutations ($h < 1/2$) can have large negative signed LD (Figure ??D), although it decays rapidly with increasing distance between loci and is roughly zero for $\rho \gg 1$. For weakly selected mutations with $\gamma = -1$, we see that there is no monotonic effect of the level of dominance on negative LD, with both recessive and dominant pairs of mutations having more negative LD than the additive case.

General selection and gene-based dominance

While dominance and epistasis are well-characterized models of selection, a large family of selection models can be specified by assigning unique selection coefficients to each possible diploid pairs of haplotypes. If we assume the diploid genotype that is homozygous for the ancestral alleles (ab/ab) has fitness 1, then there are nine other possible diploid two-locus genotypes that could be given unique fitnesses (Table S1), noting that AB/ab and Ab/aB genotypes can have differing fitness effects.

The case with $s_{AB/ab} \neq s_{Ab/aB}$ can arise in a scenario where a mutation at either locus within a haplotype impacts some functional region or element, but a diploid individual carrying at least one copy that is free of mutations has minimal fitness loss. In this “gene-based dominance” scenario (e.g., ?), an AB/ab genotype has higher fitness than an Ab/aB type (a simple implementation of this model is given in the final column in Table S1). Such a gene-based fitness model gives similar expected positive signed LD to the model of

antagonistic epistasis (compare Figure ??E-F to G-H), although the interpretation of those two models can differ. With a highly parameterized space of possible general diploid selection models, multiple models with different biological interpretations can give similar patterns of expected signed LD.

The effect of population size changes on signed LD

The moment system for Ψ_n readily incorporates variable population size as well as mutation and recombination rates and selection parameters that change over time. Here I focus on non-equilibrium population size history and consider scenarios relevant to human demographic history, including bottlenecks and expansions. I explore two simple models (Figure ??A), one with an instantaneous expansion and another with a bottleneck followed by recovery. I also consider two demographic histories inferred using genome-wide genealogy reconstruction (?) applied to the ? dataset, and focus on size histories for the Yoruba from Ibadan, Nigeria (YRI) and Utahns of North and West European ancestry (CEU) (Figure ??B).

For each of the four histories, Figure ??C-H shows the dynamics of σ_d^1 for a given parameterization of two-locus selection, including synergistic and antagonistic epistasis, dominance within loci, and gene-based dominance. In general across each selection model, population size expansions are not expected to strongly affect σ_d^1 , whether that expansion occurs deeper in the past as in the simple expansion model or rapid expansion more recently, as in the YRI. On the other hand, population size reductions push signed LD to more extreme values and subsequent recoveries or expansion again reduce the magnitude of LD. Under no selection condition tested here do population size changes cause expected LD to change sign, showing that while the magnitude of deviation of LD from zero is sensitive to population size history, interpreting general patterns of the observed sign of LD in data should not be strongly affected by population size history.

Signed LD within protein-coding genes

In what follows I examine LD from human population sequencing data, focusing on mutations within protein-coding regions. I first look at average values of LD across genes for pairs of synonymous, missense, and loss-of-function mutations and then turn to patterns of LD partitioned by functional elements and distances separating mutations within genes.

Synonymous and missense mutations show similar levels of positive LD when considering pairs of mutations within the same gene. Loss-of-function mutations instead have more negative LD, suggesting differing modes of selective interactions for loss-of-function and missense mutations (Figure ??). Within each population, measurement noise gives 95% confidence intervals that overlap with zero across mutation class, although the pattern is remarkably consistent across African, European, and East Asian populations in the Thousand Genomes dataset. Comparing mean LD within genes in African and Eurasian populations, LD in Eurasian populations is somewhat larger on average, that is, more positive for missense mutations and more negative for loss-of-function mutations. This is in agreement with differences in expectations between populations that have or have not gone through a bottleneck in their recent past (Figure ??). For pairs of missense mutations in different protein-coding genes, signed LD decays to zero fairly rapidly and is lower than signed LD within genes even for mutations that are linked (distance less than 50 kb).

The similarity in signed LD observed between missense and synonymous mutations would suggest that interference between missense mutations is minimal. However, below I show that interactive effects differ dramatically between mutations that lie within annotated functional domains and mutations outside of functional domains, with these opposing effects effectively canceling out when averaging LD over entire genes. Due to the rarity of loss-of-function mutations, I focus only on comparisons between synonymous and missense mutations when looking at finer partitions of mutations within genes.

Large positive LD between pairs of missense mutations in conserved domains

To further explore the observation of positive LD between both synonymous and missense mutations within genes, I examine the role that annotated conserved domains in protein-coding genes play in driving this signal. Protein-coding domains are conserved elements of genes, often associated with some known functional or structural feature of a protein. As such, we would expect that purifying selection is stronger on average within

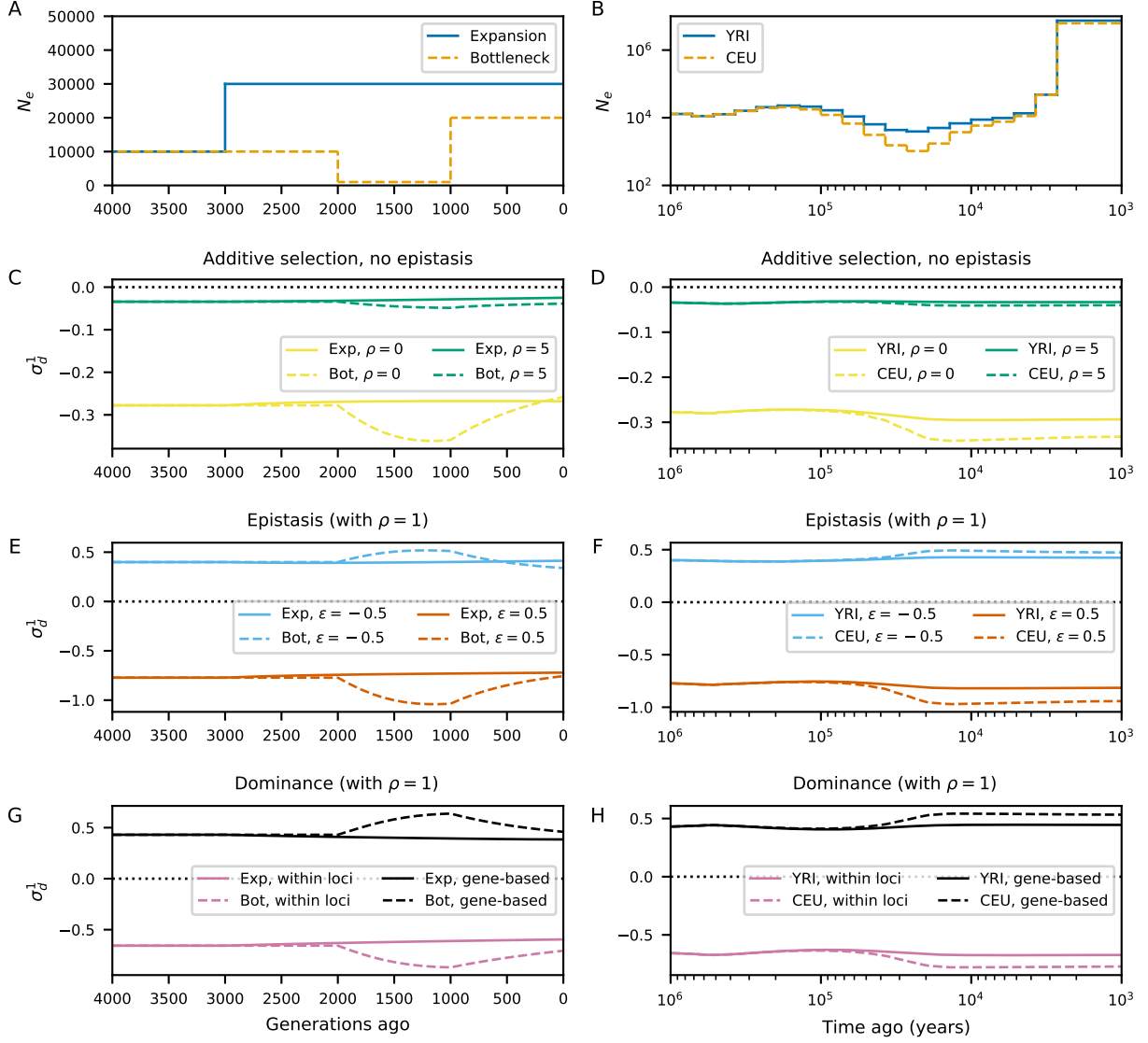


Figure 3: **The effects of demography on signed LD.** Left column: simulations under two toy models, one with an instantaneous expansion in the past, and the other with a 10-fold reduction in size followed by recovery. Right column: simulations under population size histories inferred using Relate (?) applied to populations from ?. In all simulations, γ is set to -2 at both loci. (A-B) Model illustrations. Note the differences in scale. (C-D) Additive selection with no epistasis, showing fully linked ($\rho = 0$) and more loosely linked ($\rho = 5$) pairs. Bottleneck events result in larger negative values of signed LD, but population size changes are not expected to change the sign of mean LD. (E-F) Synergistic ($\epsilon > 0$) and antagonistic ($\epsilon < 0$) epistasis. Again, bottleneck events result in larger signed LD. Here, $\epsilon = 1/2$ is a model with $\gamma_A = -2$, $\gamma_B = -2$ and $\gamma_{AB} = -6$, and $\epsilon = -1/2$ is a model with $\gamma_A = -2$, $\gamma_B = -2$ and $\gamma_{AB} = -2$. (G-H) Dominance within loci results in negative LD, while dominance acting at the gene level results in positive LD. Here, $h = 0.1$. Again, bottlenecks result in larger average LD in both cases.

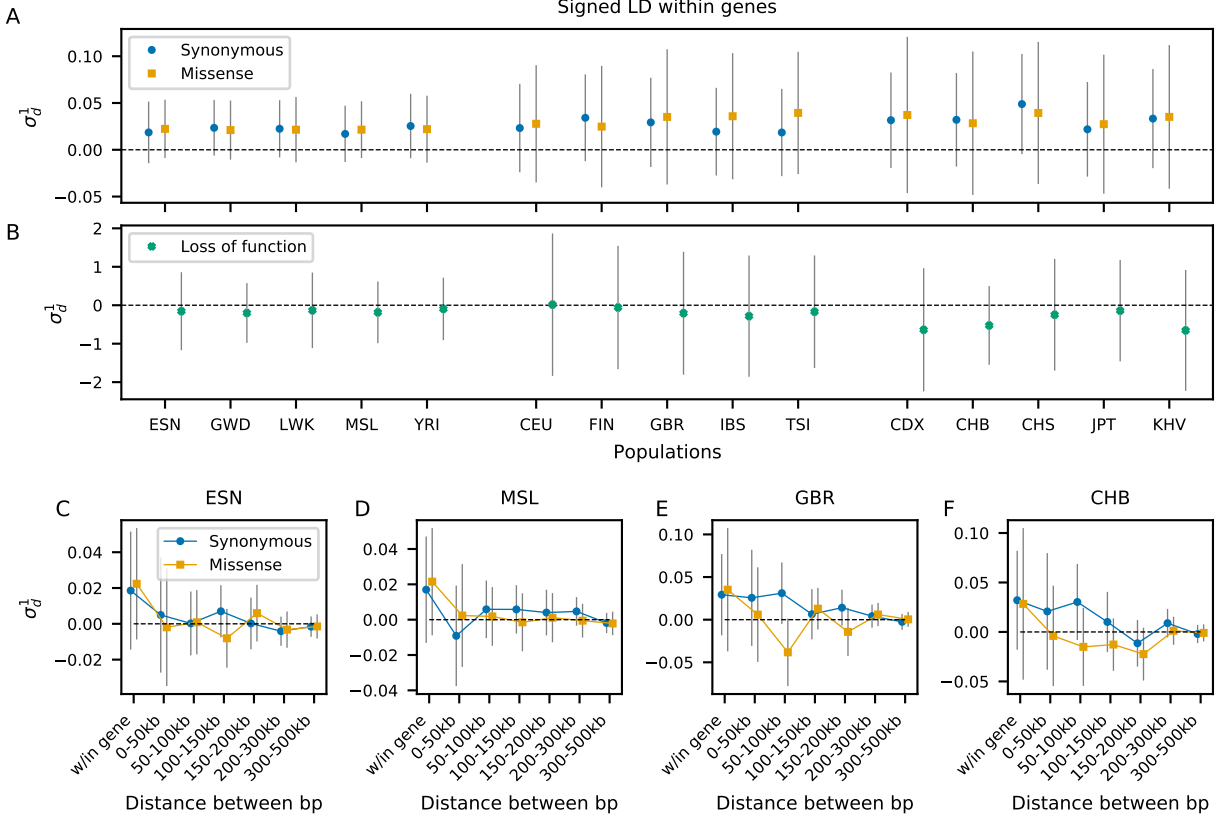


Figure 4: LD within and between protein-coding genes. (A-B) Pairs of variants in coding regions partitioned by classes of mutations. Loss-of-function mutations are defined as stop and start gains and losses and other nonsense mutations of high impact. In all populations, we observe positive mean signed LD for missense and synonymous variants, although the confidence interval of each measurement overlaps with zero. Similarly, we observe mean negative LD for pairs of loss-of-function variants within genes. There are only $\sim 200 - 300$ pairs of loss-of-function variants that fall within the same gene across the populations tested here, causing large uncertainty in each measurement. Gray lines indicate 2 standard errors of the measurement, computed by bootstrapping of sets of genes. (C-F) While pairs of synonymous and missense mutations within genes are consistently positive, pairs between genes are not consistent in their sign across populations. LD for all populations analyzed here is shown in Figure S3.

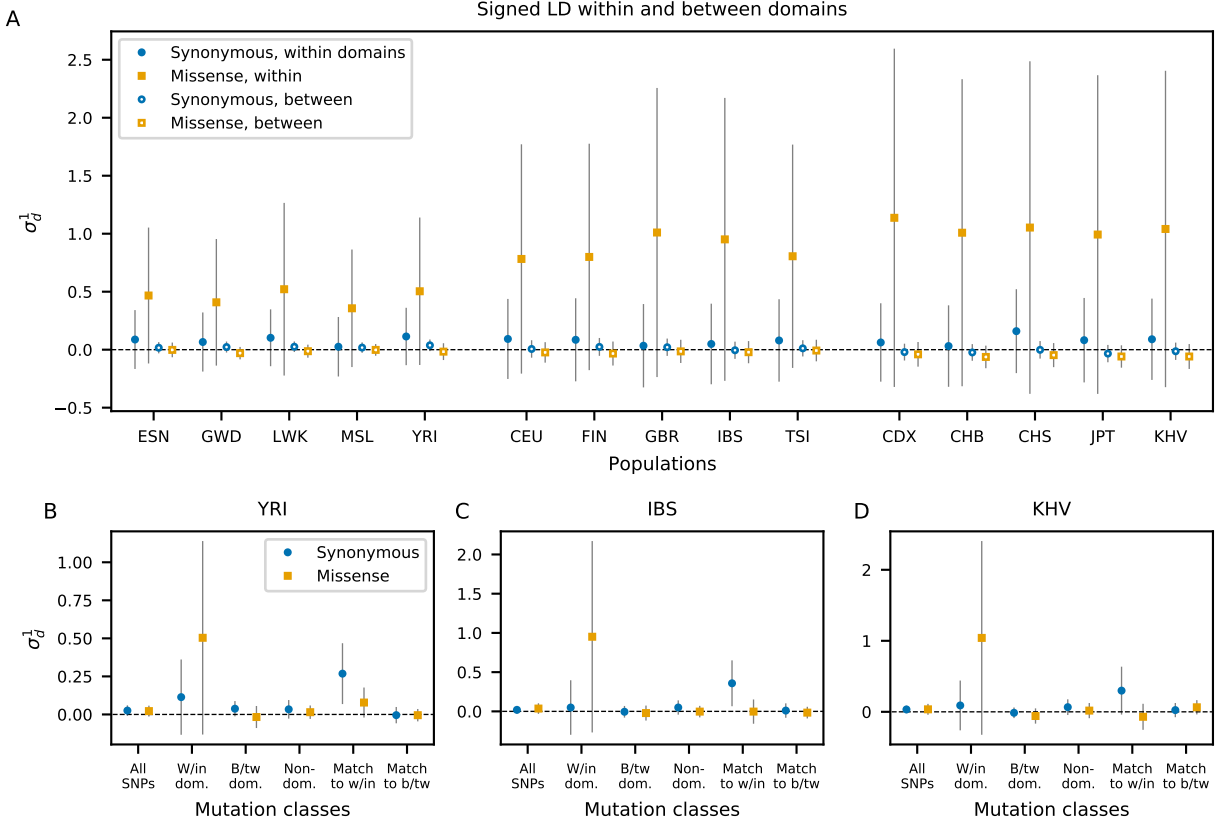


Figure 5: LD within and between annotated domains in protein-coding genes. (A) Considering only variants that fall within annotated domains in protein-coding genes, pairs of mutations that fall within the same gene display larger positive LD than pairs that lie in different domains. This pattern is most striking for missense mutations, for which pairs within domains have average signed LD that is consistently larger than for pairs that span two domains, or for pairs of synonymous mutations that are within the same domain. (B-D) The slightly positive LD that we observe for missense and synonymous mutations gene-wide (Figure ??A) is largely driven by the subset of pairs mutations that fall within the same annotated domain. LD for all populations analyzed here is shown in Figure S4.

domains than within the same gene but outside of a conserved domains. Indeed, the site-frequency spectrum (SFS) is skewed to lower frequencies for both missense and loss-of-function mutations within domains when compared to the same classes of mutations outside of domains, with much more negative values of Tajima's D within domains (Table S4). On the other hand, no difference is observed for synonymous mutations whether within or outside domains, suggesting roughly equivalent effects of selection (either direct or linked) on synonymous variation.

For pairs of missense mutations that both fall within the same functional domain, we observe large positive LD, roughly an order of magnitude larger than pairs of synonymous mutations that both fall within the same domain (Figure ??). This difference between pairs of missense and synonymous variants within domains is observed for both tightly linked ($< 10\text{bps}$) and more distant ($> 1\text{kb}$) pairs (Figure S5). Assuming most synonymous mutations are neutral, we would expect signed LD between missense mutations to be less than that between synonymous mutations under predictions from models of either Hill-Robertson interference, synergistic epistasis, or dominance (Figures ?? and ??). This observation is opposite to those expectations, suggesting a different prevailing interactive effect between nonsynonymous mutations within domains than those that are commonly tested or modeled.

The large positive signal of LD for missense mutations within the same domain does not extend to pairs of missense mutations that are in different domains. Pairs of missense and synonymous mutations show nearly equal levels of LD close to zero across domains, with missense mutations even slightly more negative than synonymous mutations (Figure ??). The interactive effect driving large LD within domains is therefore likely domain-specific. However, the average distance between pairs of mutations within domains is much smaller than between domains, so this observation may be primarily driven by the higher recombination distances between pairs of mutations across distinct domains.

Reduced LD between pairs of missense mutations outside of conserved domains

I also measure LD between pairs of mutations that both fall outside of annotated domains. For such mutations, I consider averages across all pairs of mutations outside of domains, as well as subsets of pairs matched to the same distances as pairs of mutation within and between domains as surveyed in the previous section (Figures ??B-D and S4). Averaging over all pairs of mutations within the same gene but outside domains, both synonymous and missense mutations have LD slightly above zero, with synonymous mutations observed to have slightly higher signed LD than missense mutations. This pattern is primarily driven by tightly linked variants. For pairs of mutations at distances matched to those within the domains, synonymous mutations have much larger positive LD than missense mutations, a signal that vanishes for pairs matched to distances between domains. For larger distances, both synonymous and missense mutations have LD approximately zero.

The role that tightly linked variants have in driving this signal can be clearly seen in the decay of signed LD with distance (Figure S6). Both pairs of synonymous and missense mutations with distances greater than a few hundred base pairs have average LD that fluctuates around zero. But for shorter distances, synonymous variants have large positive LD, while missense mutations have much lower LD except for pairs within the shortest distance bin (< 10 bp). This reduction in LD for missense mutations compared to synonymous mutations is consistent with Hill-Robertson interference between linked selected mutations instead of epistatic effects, as interference between either additive or recessive mutations but without epistasis results in signed LD that decays rapidly by recombination distances of $\rho \approx 1$.

Discussion

Previous theoretical and simulation studies have shown that interference and interactions between selected mutations reduce the efficacy of selection at linked loci, impacting substitution rates, the deleterious mutation load, and dynamics of segregating mutations (????). Interference between moderately deleterious mutations and synergistic epistasis are both expected to cause negative LD between selected mutations, which can be readily tested using population genetic data (???). Here, I first used a closed numerical approach to generate expectations for signed LD under a wide range of selective scenarios, and then compared patterns of LD in human populations between classes of coding mutations within genes using unbiased estimators for LD from unphased genotypes.

By taking broad genome- or gene-wide surveys of LD across functional classes of mutations, heterogeneous patterns of interactive effects that occur within functional sub-elements of genes can be missed. However, at the gene-wide level, we already find large differences in interactions between pairs of missense and loss-of-function mutations.

From gene-wide averages, LD between pairs of missense mutations does not appear to be different from pairs of synonymous variants, but LD between loss-of-function variants is substantially more negative. Hill-Robertson effects under additive selection are expected to be strongest for slightly to moderately deleterious variants (with $s \sim 1/N_e$), as strongly deleterious mutations are not expected to interfere with one another (?). However, loss-of-function mutations do not typically fall within this regime, and inference of the distribution of fitness effects (DFE) for new loss-of-function variants shows that a large majority are strongly deleterious (Supporting Information). Instead, negative synergistic epistasis between strongly deleterious mutations does produce large negative deviations of mean LD. Strongly deleterious *recessive* mutations can also produce this pattern, making it difficult to distinguish these effects with certainty. This difficulty is compounded by the

large measurement noise for $\mathbb{E}[D]$, especially for loss-of-function variants for which only a few hundred within-gene pair in human populations.

Non-uniform interactions between selected mutations within genes

Positive average LD between missense and synonymous mutations has previously been reported in humans and other species (??), and the similarity of their gene-wide LD might suggest that interference or interactions between missense mutations are minimal, or at least no stronger than those between synonymous mutations. However, averaging over all observed pairs of mutations within a gene masks element-specific interactive effects that drive LD in opposite directions. Nonsynonymous mutations found within protein domains, identified as conserved subunits of a gene with some structural or functional role, are more strongly selected against on average, but also have increased signed LD over synonymous mutations at the same distances within domains. Missense mutation outside of domains but at the same distances as those within domains have more negative LD, both compared to distance-matched synonymous mutations outside of domains and to mutations within domains. Neither dominance effects, synergistic epistasis, nor Hill-Robertson interference are expected to result in positive LD, so that some other interactive effect must be driving this positive LD signal within conserved domains.

There are a number of possible interaction scenarios that can result in positive signed LD between linked loci. One explanation is a prevalence of pairs of compensatory mutations that are tolerated to co-segregate at high frequencies within conserved domains (??). ? and ? have proposed such a mechanism to explain observed clusters of mutations in lineages of *Drosophila* and other species. Another possibility is a model of antagonistic, or diminishing returns epistasis, in which a single amino acid-changing mutation within a domain damages the functionality of that subunit, but additional mutations within that same domain reduce fitness by a factor less than the first mutation. A third possibility, related to the antagonistic epistasis case, is that selection acts on the functional domain as a unit instead of at loci within the domain individually (such as gene-based dominance). In this scenario, double heterozygotes have different fitnesses depending on whether the mutations are found on same haplotype or on different haplotypes.

We do not see an increase in LD between mutations that are found in different annotated domains within the same gene, which are on average at much larger recombination distances. Nor do we find increased LD outside of domains. Rather, for pairs of mutations outside of domains but at the same distance as those within the same domain, missense variants have considerably lower LD than synonymous variants. This difference between synonymous and missense pairs of variants largely disappears for SNPs separated by more than a few hundred or a thousand base pairs (Figure S6). This suggests that Hill-Robertson interference is the primary mode of interaction between mutations falling outside of domains, as epistasis is expected to impact LD over larger distances than what is observed, although the strength of epistasis is also likely to be a function of distance between mutations.

Taken together, selection on segregating variants within protein-coding genes is nonuniform, with both the overall strength of selection and interactions between variants differing between annotated elements. Missense and loss-of-function mutations within conserved domains are subject to stronger selection, skewing the SFS to lower frequencies. Typical approaches for inferring the distribution of fitness effects from population genomic data average over these differences by considering all nonsynonymous mutations together (??), but if loss-of-function variants are in fact recessive, even partially, this can impact inferences of the DFE (Supporting Information). It would be straightforward to adapt DFE-inference methods to infer more detailed and accurate representation of the heterogeneous effects of new mutations within genes by partitioning by missense and loss-of-function classes as well as by annotated domains. Similarly, the different modes of interaction between mutations across annotated domains mean that our standard models and simulation approaches may be too simple and will not capture the evolutionary trajectories and patterns of diversity that differ at finer scales.

Clustered mutations account for positive LD among synonymous mutations

In a single randomly mating population, neutral mutations are expected to have average signed LD of zero, but across all populations analyzed here, LD between synonymous mutations is positive. While selection on

some subset of synonymous variants is possible, it should be much weaker on average than between missense mutations, and any interference between selected synonymous variants should lead to negative LD.

One proposal is that spatial population structure may be responsible for the increase of LD observed between synonymous variants, as every natural population lives in some geography that will necessarily induce some structure. However, neutral evolution alone cannot create this effect. Under standard population genetic assumptions of constant mutation and recombination rates and independent mutation events, migration, admixture, and spatial structure do not result in nonzero $\mathbb{E}[D]$. Rather, for a set of mutations with *given* differences in allele frequencies between source populations, D can be nonzero in an admixed population between those particular mutations even when D is zero in the source populations (?). But this result relies on conditioning on directional differences in allele frequencies between populations. Expected mean allele frequency difference in the source populations is zero ($\mathbb{E}[(p_1 - p_2)(q_1 - q_2)] = 0$), as differences are equally likely to be positive or negative, and resulting $\mathbb{E}[D]$ is still zero when taking genome-wide averages. Instead, ? and ? used forward simulations under multi-population or explicitly spatial models and a multiplicative fitness function to explore the joint effect of structure, assortative mating, and interference, scenarios that can lead to nonzero LD. ? find positive LD among neutral mutations in their simulation, although it extends over much larger distances than observed in the data. Instead, positive LD in data between synonymous mutations decays to zero within just a few hundred bases (see Figures S6 and S5, and Figure 5 from ?).

Nonrandom mutational processes provide an alternative explanation for positive LD between neutral variants separated by short distances, in which clusters of mutations occur simultaneously in the same mutational event. Such multinucleotide mutations have been shown to affect patterns of LD over short distances, on the order of 10s to 100s of base pairs (?), and clustered mutational events may be common in humans (?). Indeed, a simple exponential model in which the fraction of pairs of mutations arising from a multinucleotide mutation event decays with distance fits the observed decay of σ_d^1 between synonymous variants (Supporting Information). We may therefore treat positive LD as the baseline expectation for tightly linked variants primarily due to clustered mutations, so that subsequent selective processes and interactions cause LD to deviate from that expectation, as seen for missense mutations across different sub-elements of genes. It may also be more appropriate to compare the negative LD observed between linked loss-of-function variants to that positive baseline expectation instead of zero, which would imply that they are more recessive or have stronger synergistic epistasis than from inferences assuming a neutral expectation of zero.

Challenges to distinguishing highly parameterized models of selective interactions

When partitioning measurements of LD by mutation classes or regions within genes, the decreasing number of pairwise comparisons leads to large estimated measurement noise. Within each population, confidence intervals of observed σ_d^1 often overlap with zero or overlap with that of other classes of mutations. While observed patterns are remarkably consistent across the 15 populations considered here, their joint evolutionary histories make formal testing of significance difficult due to shared variation, so that cannot be treated as independent measurements. Extensive forward simulations will likely be required to more thoroughly assess significance.

From a modeling perspective, in exploring or inferring multi-locus selective models, the number of plausible selection scenarios becomes quite large as we relax the strict assumptions of additivity and multiplicative fitness effects. The inclusion of dominance effects, epistasis, or other interactions leads to a rapid increase in the number of parameters to consider. This makes performing forward simulations that span the range of all such selective interaction scenarios burdensome.

Instead, closed, numerical approaches to compute expectations for two-locus statistics under a wide range of selection models, as presented here or in ? and ?, allow us to explore this highly parameterized space of models far more efficiently, and it opens the possibility for performing likelihood-based inference using signed LD or other two-locus summaries of the data. For example, inferring the joint distribution of dominance and selection is underpowered using the SFS alone, but because signed LD is sensitive to the levels of dominance (Figure ??), inferring the DFE with dominance may be feasible using the joint distribution of allele frequencies and LD.

Using a single low-order summary of signed LD, such as $\mathbb{E}[D]$ or σ_d^1 , is likely insufficient to confidently discriminate modes of selective interactions between linked mutations. Both dominance and epistasis can result in strongly negative LD, although the expected decay of LD under those scenarios can differ. Similarly, multiple interaction models can lead to positive signed LD. Among all interaction models, the extent of LD and rate of its decay also depends on the underlying distribution of selection coefficients among a class of mutations, which are unknown for a given pair of mutations, so that we must integrate over a distribution of fitness effects. This DFE, however, will have been inferred under a simple set of assumptions, such as additivity and interchangeability between sites within a gene, potentially biasing any inference using previously inferred DFEs to learn about patterns of interactions. Again, this underlies the need to jointly infer strengths and interactions of selected variants.

Finally, exploring additional summaries of the two-locus sampling distribution could provide additional power to distinguish between interaction models. The method developed here allows us to compute the full sampling distribution for a given sample size, and expectations for a large family of informative two-locus statistics can be computed directly from this distribution. These expectations can then be compared directly to empirical observations, which can be taken from either phased or unphased data (?). Additional work performing such an analysis is therefore warranted, which should provide a path forward for distinguishing between modes of selective interactions.

Methods

Existing theory and numerical methods

Many well-known properties of two-locus dynamics and equilibrium LD come from early work on the multi-locus diffusion approximation (????). This includes the result that genome-wide averages of signed LD are expected to be zero under neutrality. Under a two-locus biallelic model, where the left locus allows alleles A and a and the right locus allows alleles B and b , the standard covariance measure of LD is defined as $D = f_{AB} - f_A f_B$, where f_{AB} is the haplotype frequency of double-derived types carrying both A and B , and f_A and f_B are the marginal frequencies of the derived alleles at each locus. This covariance decays due to both drift and recombination at a rate proportional to the inverse of the effective population size and the distance separating loci (?):

$$\mathbb{E}[D]_{t+1} = \left(1 - \frac{1}{2N_e(t)} - r\right) \mathbb{E}[D]_t.$$

While $\mathbb{E}[D] = 0$, the variance of D is non-zero, and ? presented their groundbreaking result that the variance of D under neutrality and steady-state demography, normalized by the joint heterozygosity of the two loci, is

$$\sigma_d^2 = \frac{\mathbb{E}[D^2]}{\mathbb{E}[f_A(1-f_A)f_B(1-f_B)]} \approx \frac{5 + \rho/2}{11 + 13\rho/2 + \rho^2/2}, \quad (1)$$

where $\rho = 4N_e r$.

Analytic progress beyond these results has come haltingly. In the 1980s, recursions were developed to compute the two-locus sampling distribution under neutrality, that is, the probability of observing given counts of two-locus haplotypes in a sample of size n (?). This approach would continue to be developed and later form the foundation for the inference of local recombination rates from population genetic data (??).

To include selection, however, there have been relatively few advances beyond the Monte Carlo simulation approach taken by ?, albeit now with more powerful computational resources and sophisticated software for performing flexible forward simulation (e.g., ?, ?). Analytic results for two-locus distributions under selection are notoriously difficult, with a few notable flashes of progress. For example, ? considered the effect of a recent sweep on patterns of LD between neutral loci near the locus under selection, and in a recent paper, ? presented analytic solutions for patterns of LD between rare mutations under additive selection with epistasis. Nonetheless, such approaches are typically confined to steady-state demography and constrained selection models.

Numerical methods inhabit the space in between expensive discrete simulations and limited analytic solutions, providing a more efficient and practical method to compute expectations of two-locus diversity measures under a wider range of parameters and demographic scenarios. ? used a finite differencing approach to numerically solve the two-locus diffusion equation with additive selection at either locus, although that paper focused on the applicability of two-locus statistics to demographic inference. More recently, ? extended the ? system for $\mathbb{E}[D^2]$ to compute arbitrary moments of the distribution of D for any number of populations connected by migration and admixture. They also showed that such a moments-construction can be used to solve for the two-locus sampling distribution within a single population, though it requires a moment-closure approximation for nonzero recombination and selection. Below, I extend this approach to model arbitrary diploid selection, which encompasses dominance, epistasis, and other forms of selective interactions between two loci. In a concurrent study to this paper, ? developed numerical solutions to the same moments system, which they used to describe selected haplotype trajectories and the distortion of neutral diversity at loci variably linked to beneficial alleles that sweep to high frequencies under non-equilibrium demography.

The two-locus sampling distribution with arbitrary selection

The two-locus sampling distribution is the direct analog to the single-locus site frequency spectrum (SFS) of a given sample size. Instead of describing the density or number of mutations with a given allele count out of n samples, the two-locus distribution Ψ_n stores the density or number of pairs of loci with observed haplotype counts, so that $\Psi_n(i, j, k)$ is the number of pairs for which we observe i copies of the AB haplotype, j of type Ab , k of type aB , and $n - i - j - k$ of type ab . The size of Ψ_n grows rapidly, with $O(n^3)$ entries, which practically limits computational approaches to moderate sample sizes and a single population.

Under neutrality, a number of approaches exist to compute Ψ_n , including the recursion due to ? and ?, or more recent numerical approaches to the two-locus coalescent (?) or diffusion approximation (?). Selection is most easily included using the forward-in-time diffusion equation (??), where a standard approach is to first solve for the continuous distribution ψ of the density of two-locus haplotype configurations in the full population, and then integrate ψ against the multinomial sampling function to obtain Ψ_n .

Alternatively, ? showed that there exists a system of ordinary differential equations directly on the entries of Ψ_n , although that paper focused primarily on neutral dynamics and considered a model with selection at only one of the two loci. I briefly summarize this general approach below, but refer readers to that previous paper for detailed derivations of the drift, recombination, and mutation terms and the moment-closure approximation. Instead, here I focus on generalizing the selection operator to include epistasis, dominance, and other forms of two-locus interactions.

Moment equation for Ψ_n

The system of linear ordinary differential equations for the entries of Ψ_n takes the form

$$\Psi_n^{t+1}(i, j, k; t) = \mathcal{D}_{N(t)} \Psi_n^t + \mathcal{R}_r \Psi_{n+1}^t + \mathcal{U}_u \Psi_n^t + \mathcal{S}_{s_A, s_B, \dots, h_A, h_B, \dots} \Psi_{n+2}^t. \quad (2)$$

Here, $\mathcal{D}_{N(t)}$ is a sparse linear operator accounting for drift with population size $N(t)$, \mathcal{R} accounts for recombination with per-generation recombination probability r between the two loci, \mathcal{U} accounts for mutation, either under an infinite sites or biallelic reversible mutation model, and \mathcal{S} accounts for selection.

The moment system for Ψ_n can be derived directly from the diffusion approximation, or it can be found through a more intuitive process of tracking the dynamics of allelic states of a sample of size of n from the full population. We assume $n \ll N_e$, and r and s are $O(1/N_e)$ so that multiple coalescence, recombination, or selective events within the n lineages are rare in any given generation. For example, for drift events, we account for the probability that two lineages within the n tracked lineages at time $t + 1$ coalesced in the previous generation and, given a coalescence event, the probability that haplotype counts changed due to this coalescence event [see Supporting Information; ?, ?]. In typical diffusion approximation fashion, we multiple through by $2N_{ref}$ so that time is measured in $2N_{ref}$ generations, and we consider scaled parameters $\rho = 4Nr$, $\theta = 4Nu$, and $\gamma = 2Ns$.

Moment closure

In the absence of selection and for fully linked loci (i.e., $\rho = 0$), the system is closed and can be solved exactly. However, for nonzero recombination or selection, the entries of Ψ_n rely on the slightly sampling distributions with sample sizes $n + 1$ (for recombination and additive selection) or $n + 2$ (for non-additive selection). This is because if a recombination event occurs within one of n lineages being tracked by Ψ_n , we need to draw an additional lineage from the full population to recombine with that chosen lineage, thus requiring Ψ_{n+1}^t to find Ψ_n^{t+1} . Selection events similarly require extra lineages from the full population, which replace a chosen lineage that fails to reproduce with probability proportional to its selection coefficient, assuming negative selection.

This requirement of extra lineages for nonzero recombination and selection means that the system in (??) is not closed, so that we need a moment-closure approximation to solve for Ψ_n . As in ?, a jackknife approximation is used to estimate Ψ_{n+l} , for $l = 1$ or 2 , from Ψ_n (following the single-locus closure introduced in ?), so that $\hat{\Psi}_{n+l}(i, j, k) = \mathcal{J}_{n,l}\Psi_n$, although other accurate closure approximations are possible (?). This emits a closed approximate system,

$$\hat{\Psi}_n(i, j, k; t) = \mathcal{D}_{\nu(t)}\hat{\Psi}_n(t) + \mathcal{R}_\rho\mathcal{J}_{n,1}\hat{\Psi}_n(t) + \mathcal{U}_\theta\hat{\Psi}_n(t) + \mathcal{S}_{\gamma,h}\mathcal{J}_{n,2}\hat{\Psi}_n(t). \quad (3)$$

The jackknife approximation, which approximates an entry $\Psi_{n+l}(i, j, k)$ using nearby entries in Ψ_n , is more accurate for larger sample sizes (Figure S2), creating a tension between efficiency and accuracy: larger sample sizes result in more accurate solutions, as error in the jackknife is diminished, but computational complexity also grows rapidly in the number of entries of Ψ_n , which is $O(n^3)$. Accuracy and runtime are shown in Figure S2. In the results presented in this paper, sample sizes between $n = 30$ and $n = 80$ are used.

Derivations for the drift, recombination, and mutation operators and the jackknife moment-closure approximation can be found in section S1.3 of ?, and I repeat the main results in the Supporting Information of this paper.

Selection models with epistasis and dominance

To include selection, we consider a model where we draw lineages uniformly from the previous generation, but keep lineages with probability proportional to their fitness. In the absence of dominance, selection reduces to a haploid model, with acceptance and rejection probabilities depending on the fitnesses of each haploid copy, where haplotype Ab has fitness $1 + s_A$, aB has fitness $1 + s_B$, and AB has fitness $1 + s_{AB}$. We assume the doubly ancestral haplotype ab has fitness 1, so fitnesses are relative to that of ab haplotypes. The standard multiplicative fitness function assumes that $s_{AB} \approx s_A + s_B$, and a model for epistasis can be written as

$$s_{AB} = (s_A + s_B)(1 + \epsilon),$$

so that $\epsilon > 0$ implies synergistic epistasis, while $\epsilon < 0$ implies antagonistic epistasis.

To obtain the recursion equation under selection we consider drawing n lineages from generation t , which has an expected sampling distribution of haplotype counts given by Ψ_n^t . However, assuming $s \leq 0$ for each derived haplotype, each of those sampled lineages has probability of being rejected equal to the relative fitness of its haplotype state. If a lineage is rejected, a replacement is drawn from the full population. Under the assumption that $ns_{..} \ll 1$, the probability that more than one selection event occurs in any given generation is negligibly small, so that the case of multiple simultaneous rejections can be ignored. Then Psi_n^{t+1} relies only on Ψ_n^t and Ψ_{n+1}^t for additive selection. The full selection operator \mathcal{S} for additive selection is given in the Supporting Information.

To account for dominance within loci, or other general forms of two-locus selection, the selection operator no longer reduces to individual haplotypes, but instead we need to know the state of two-locus genotypes. For example, the fitness of an individual carrying an Ab haplotype depends on whether their second haplotype is ab , Ab , aB , or AB . We can therefore assign a selection coefficient to each possible diploid configuration, $s_{AB/AB}$, $s_{AB/Ab}$, and so on. Assuming that the doubly homozygous ancestral ab/ab genotype has relative fitness 1, this gives nine possible unique selection coefficients in the most general two-locus selection model.

Note that AB/ab and Ab/aB genotypes need not have the same selection coefficient, which allows for simulation under a gene-based model of dominance (Table S1).

The general selection operator follows the same approach as the haploid selection operator with epistasis described above. Now, in the case of a selection event rejecting a lineage within our tracked samples, we need to draw not only the replacement lineage from the full population but also a second haplotype from the full population to form the diploid genotype, as this determines the probability that we reject the focal haplotype. We again assume that $ns \ll 1$ for all genotype selection coefficients, so that we may assume at most a single selection event occurs in any given generation. This means that to find Ψ_n^{t+1} under a general two-locus selection model, which encompasses dominance within either locus, gene-based dominance, or a combination of dominance and epistatic effects, we need Ψ_{n+2}^t . Again, a full derivation and expressions for the general selection operator are given in the Supporting Information.

Low-order summaries of the sampling distribution

From Ψ_n , expectations for any two-locus statistic can be found by downsampling to the appropriate sample size. For example, to compute $\mathbb{E}[D]$, the sum is taken over all haplotype configurations $\mathbf{n} = (n_{AB}, n_{Ab}, n_{aB}, n_{ab})$, weighted by the density Ψ_n for that configuration:

$$\mathbb{E}[D] = \sum_{\mathbf{n}} \Psi_n(\mathbf{n}) \frac{n_{AB}n_{ab} - n_{Ab}n_{aB}}{n(n-1)}. \quad (4)$$

For large sample sizes, this is approximately equal to computing D by taking the maximum likelihood estimate for each allele frequency $f_i = n_i/n$, but the maximum likelihood estimate will be noticeably biased for small to moderate sample sizes. Other low-order two-locus statistics can be computed using the same approach, as implemented in *moments* following ?, which can be compared across sample sizes and between estimates from phased or unphased data. In this paper, I focus on $\sigma_d^2 = \mathbb{E}[D^2]/\mathbb{E}[p(1-p)q(1-q)]$ and $\sigma_d^1 = \mathbb{E}[D]/\mathbb{E}[p(1-p)q(1-q)]$, averaged over pairs of variants at all frequencies. However, allele-frequency conditioned statistics (such as keeping only loci below some frequency threshold as in ?) can be considered using this same approach.

Simulations of non-steady-state demography

I considered four variable population size histories, two simple toy models and two inferred from human populations in African and Europe. For each size history scenario, I tracked the evolution of $\Psi_n(t)$ for varying selection models, and Figure ?? plots the trajectory of σ_d^1 over time. The selection strength at both loci was fixed at $\gamma = 2$ for all scenarios, and I considered additive selection without epistasis for fully linked ($\rho = 0$) and more loosely linked ($\rho = 5$) loci. With $\rho = 1$, I also considered positive and negative epistasis ($\epsilon = 0.5$ and -0.5), recessive loci ($h = 0.1$), and a gene-based dominance model with $h = 0.1$.

The simple size change models both had ancestral $N_e = 10,000$, with one a 3-fold population expansion that occurs 3,000 generations ago, and the other a 10-fold reduction 2,000 generations ago followed by an expansion to double the ancestral size 1,000 generations ago. The size histories for YRI and CEU were inferred using Relate (?) applied to the phased autosomal data from ?, using default parameters as recommended in the Relate tutorial. Relate returns estimates of coalescence rates given time bins, and population sizes are estimated as their inverses. Estimates using Relate for population sizes in the very recent past ($< 3,000$ years, or ≈ 100 generations) diverged, so I truncated the history over this time period and assumed a constant size from the most recent non-diverged bin.

Analysis of human genomic data

Using the annotated variant call format (VCF) files from the phase 3 ? (Thousand Genomes) data release, I subset the genotype VCFs of the same release to autosomal variants that were annotated as either synonymous or nonsynonymous (including both missense and more damaging mutations such as stop gains, stop losses, and start losses). I further subset to samples within each non-admixed population in the African, European, and East Asian continental groups. Signed LD is sensitive to ancestral state identification, so I only kept

sites for which ancestral alleles were estimated with high confidence in both the VCF info field and the Thousand Genomes human ancestor reconstructed from a phylogeny of six primates.

In addition to ancestral state misidentification, measured LD is sensitive to phasing error, so I computed LD statistics using unphased genotypes following ?. This approach provides unbiased estimates for pairwise LD, under the assumption that individuals are not inbred. I considered pairs of mutations within the same mutation class (synonymous, missense, and loss-of-function) either within the same gene, in different genes within distance bins up to 500kb, and inside or outside of annotated domains within the same protein-coding genes. I used a dataset of annotated protein domains mapped to hg19 compiled by ? to determine if a given mutation falls within a domain or not.

Data and software availability

All data and software used in this paper are publicly available and open source. I downloaded the Thousand Genomes annotations and genotypes VCFs from the ftp server at <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>, and the Thousand Genomes human ancestor fasta file from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/. Protein domain information from ? was downloaded from <http://prot2hg.com/dbdownload.php>.

Implementation of moment equations to compute expectations for two-locus and linkage disequilibrium statistics are implemented in Python using Numpy (?) and sparse matrix solvers in Scipy (?). These methods are packaged within *moments*, and analyses here were performed using *moments* version 1.1.0, available from <https://bitbucket.org/simongravel/moments/src/master/> with extensive documentation at <https://moments.readthedocs.org>. Scripts to run all analyses, recreate figures, and compile this manuscript are available at https://github.com/apragdale/two_locus_selection.

Acknowledgements

I am grateful to Alex Diaz-Papkovich, Eric Friedlander, Simon Gravel, Mashaal Sohail, Matthias Steinrücken, and Kevin Thornton for helpful discussions and feedback on earlier versions of this manuscript that improved this work.