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

February 21, 2021

Abstract

Selection acting on a mutation interferes and interacts with evolutionary processes at nearby loci, causing allele frequency and correlation patterns between pairs of selected mutations to deviate from expected single-locus dynamics. 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 interpretation of observations. 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, and variable recombination and population sizes. I use this to explore how different models of epistasis and dominance affect expected signed LD, including for non-steady-state demography relevant to human populations. Finally, I use whole-genome sequencing data to asses how well we can differentiate models of selective interference in practice, and find [[that while xxx, within classes/domains, we see...]].

Introduction

Most new mutations that affect fitness are deleterious and tend to be eliminated from a population. The average number of generations that a deleterious mutation segregates in a population depends on the strength of selection against genomes that carry it, with very damaging mutations kept at low frequencies in the population and purged relatively rapidly. But 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 has been well documented in practically every taxon and population for which we have genomic data. However, 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, Hill and Robertson (1966) recognized that linked negatively selected variants reciprocally impede the efficacy of population to remove those mutations. In general, linked selection reduces the fixation probability of advantageous mutations and increases that of deleterious mutations, compared to expectations under single-locus dynamics (Birky and Walsh, 1988). Allele frequency dynamics of linked selected variants are also predicted to deviate from expectations without interference. Under a multiplicative fitness model, where the fitness reduction of a haplotype carrying multiple deleterious variants is equal to the product of the fitness reduction of each mutation independently, we should expect to see those mutation segregate on different haplotypes more often than together, leading to negative, or repulsion, linkage disequilibrium (LD), although the extent of LD depends non-trivially on the strength of selection and probability of recombination separating loci (Hill and Robertson, 1966).

Non-additive effects further complicate our evolutionary models, including dominance (i.e., interactions *within* loci) and epistasis (interactions *between* loci), and both are thought to be widespread phenomena. A large fraction of nonsynonymous coding mutations are thought to be at least partially recessive (Huber

et al., 2018; Agrawal and Whitlock, 2011), with average levels of dominance correlating with strength of selection (Kacser and Burns, 1981)[1], and dominance plays an important role in shaping expected equilibrium allele frequencies and mutation load of strongly damaging disease mutations (Clark, 1998).[2] On the other hand, epistasis differentially impacts deleterious load for asexually vs sexually reproducing organisms (Kimura and Maruyama, 1966; Kondrashov, 1995) and has been invoked as an explanation for the evolutionary advantage of sex (Kondrashov, 1982; Charlesworth, 1990; Barton and Charlesworth, 1998), and epistatic interactions may also drive incompatibilities that lead to postzygotic isolation during the process of speciation (Turelli and Orr, 2000). Within populations, epistasis is known to cause signed LD to deviate dramatically from zero (Kondrashov, 1995; Charlesworth, 1990). However, despite patterns of dominance being appreciated as important factors shaping linked variation (Turelli and Orr, 2000; Zhao and Charlesworth, 2016), and the evolutionary importance of epistatic interactions, we currently lack models for predicting correlation patterns between linked mutations under general selection models.[3].

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 domains, and assess 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 induced within a target gene and fitness, or some protein function, is experimentally measured (Romero and Arnold, 2009; Steinberg and Ostermeier, 2016). For example, using the model system of the TEM-1 β-lactamase gene in *E. coli*, Bershtein *et al.* (2006) 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.

In most natural populations, directed mutation studies are not possible, 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 (Hill and Robertson, 1966), 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 Callahan *et al.* (2011), pairs of nonsynonymous 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 including synonymous mutations, suggesting that compensatory nonsynonymous variants were more generally tolerated.

More recently, Sohail *et al.* (2017) 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, from which they proposed that widespread synergistic epistasis between these mutations. Within the past year or two, both Sandler *et al.* (2020) and Garcia and Lohmueller (2020) 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 (Garcia and Lohmueller, 2020) or questioned whether LD between loss-of-function variants is significantly different from zero (Sandler *et al.*, 2020).

A number of factors impede our interpretation of patterns of signed LD between coding variants. First, for strongly deleterious or loss-of-function variants, their rarity and low frequency means that statistical

[1] correct citation?

[2] maybe also cite Haldane (1930)

[3] See McVean and Charlesworth (2000)

measurement of LD and other diversity measures are quite noisy. Second, comparisons are based on theory that are confined to limiting and simplistic assumptions, including steady-state demography, simple selection and interaction models, or unlinked loci. To generate predictions under more complex models, we rely on expensive forward simulations. These can be great for building intuition or testing inference methods, but do not efficiently provide expectations for quantities of interest across parameter regimes of interest. Theoretical and numerical studies of haplotype frequencies and LD under general selective interaction models would be of great benefit.

Existing theory and numerical methods

Many well-known properties of two-locus dynamics and equilibrium LD come from early work that centered on the multi-locus diffusion approximation (Kimura, 1955; Hill and Robertson, 1968; Ohta and Kimura, 1969, 1971). 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 has 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 Ohta and Kimura (1971) presented their famed result that the variance of D under neutrality and steady-state demography, normalized by the joint heterozygosity of the two loci, is

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

where $\rho = 4N_e r$.

Analytic progress beyond these well-known 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 (Golding, 1984). This approach would continue to be developed and later form the foundation for the inference of local recombination rates from population genetic data (Hudson, 2001; McVean *et al.*, 2004).

To include selection, however, there have not been many advances beyond the Monte Carlo simulation approach taken by Hill and Robertson (1966), albeit with somewhat more powerful computational resources and more sophisticated software for performing flexible forward simulation (e.g., Thornton (2014), Haller and Messer (2019)). Analytic results for two-locus distributions under selection are notoriously difficult, with a few notable flashes of progress. McVean (2007) considered the effect of a recent sweep on patterns of LD for loci nearby or spanning the locus under selection, and in an impressive recent paper, Good (2020) presented analytic solutions for patterns of LD between rare mutations under additive selection with epistasis. Nonetheless, such approaches are confined to steady-state demography and constrained selection models.

Numerical methods straddle the two approaches of expensive and noisy discrete simulation and 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.[4] Ragsdale and Gutenkunst (2017) solved 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, Ragsdale and Gravel (2019) extended the Hill and Robertson (1968) 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. This paper also showed that such a moments-construction can be used to solve for the two-locus sampling distribution within a single population, though requires a moment-closure approximation for nonzero recombination and selection. In the Methods, I extend this same approach to model arbitrary diploid selection, which encompasses dominance, epistasis, and other forms of selective interactions between two loci.

[4] cite other numerical approaches, and need to read: Hallgrímsdóttir and Yuster (2008)

Methods

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 where we see i AB haplotypes, j Ab haplotypes, and k aB haplotypes. Here, we use upper case to denote the derived allele, so the number of doubly ancestral haplotypes ab is $n - i - j - k$.

- Diffusion equation for $\psi(x, y, z)$
- Sampling to get Ψ_n
- Point to Ragsdale and Gravel (2019) supplement, with $\dot{\Psi}_n = \mathcal{D}\Psi_n$ ...
- Note that the selection operator

Drift and recombination

Selection models with epistasis and dominance

Moment closure and implementation

Low-order summaries of the sampling distribution

Expected LD under inferred human demographic history

Analysis of human genomic data

- Computing LD statistics commonly used for inference about selection models (Figure 1), using Ragsdale and Gravel (2020)
- Data analysis of Thousand Genomes Project data 1000 Genomes Project Consortium *et al.* (2015)
- Domain information from Stanek *et al.* (2020)

idea: LD within genes partitioned by whether two mutations fall within domains or are in different domains (distance as a proxy? or can we get information about whether they lie in the same gene or not?) - see Yeang and Haussler (2007), Ivankov *et al.* (2014), Taverner *et al.* (2020)

Data and software availability

- Thousand genomes VCFs and ancestral sequence (1000 Genomes Project Consortium *et al.*, 2015)
- Protein domain information (Stanek *et al.*, 2020)
- Implementation as `moments.TwoLocus` at https://bitbucket.org/simongravel/moments
- Scripts to recreate analyses, figures, and compile this manuscript at https://github.com/apragsdale/two_locus_selection

Results

Singed LD under steady-state demography

- Expected LD under models of epistasis and dominance at equilibrium (Figures 2 and 3)
  - σ_d^1, σ_d^2, r^2, for all frequencies
  - The same statistics conditioned on sample allele counts
  - Relationship between "Hudson slice" and allele count-conditioned LD

Additive selection and epistasis

Simple dominance within a locus

General selection and gene-based dominance

The effect of population size changes on signed LD

- Effect of demography on these same statistics
  - Bottleneck with recovery and expansion
  - Human-like size histories (using Speidel *et al.* (2019))
  - For each (in own column): 4
    - Top panel: demography
    - Next panel: additive selection and synergistic epistasis
    - Next panel: dominance with h = 0, 0.2
    - Next panel: gene-based dominance and compensatory mutations

Signed LD for mutations within and between genes

Large positive signed LD for missense mutations within annotated protein

domains

- Present data from loss-of-function (LOF), synonymous, and nonsynonymous variants from Thousand Genomes, both for pairs of loci within genes and for pairs of loci between nearby genes 5. Do we see any differences, and do the different patterns in African vs non-African populations tell us anything about the possible selection model for different categories of variants?

Some results:

- Within genes, both synonymous and missense mutations have LD slightly greater than zero, and are not significantly different from each other.
- On the other hand, LOF mutations have negative LD, although the small number and low frequency of LOF mutations in human populations means that the measurement is quite noisy.
- In fact, taking each population independently, all measurements have confidence intervals that overlap with zero
- But assuming independent measurements between populations (which is probably not correct, as populations have shared history and thus shared diversity patterns), can we say that synonymous and missense mutations have values significantly greater than zero and LOF mutations have values less than zero?

Within domains 6:

- Missense and synonymous mutations have positive LD when looking at pairs within the same annotated domain.
- Between domains, both are reduced to be roughly zero
- Within domains, missense mutations have signed LD much larger than synonymous mutations

Discussion

- What can we say about epistasis and models of dominance using LD statistics? Do we have power to distinguish models using low-order statistics of this kind?
- The role of numerical solutions to such models for population genetics analyses. They bridge the gap between theory limited by analytic intractability and expensive forward simulations such as `SLiM` and `fwdpy11`.
- Limitations and future directions:
  - multiple populations

  - strong selection leads to numerical instability

Acknowledgements

- Simon Gravel
- Kevin Thornton

Figures:

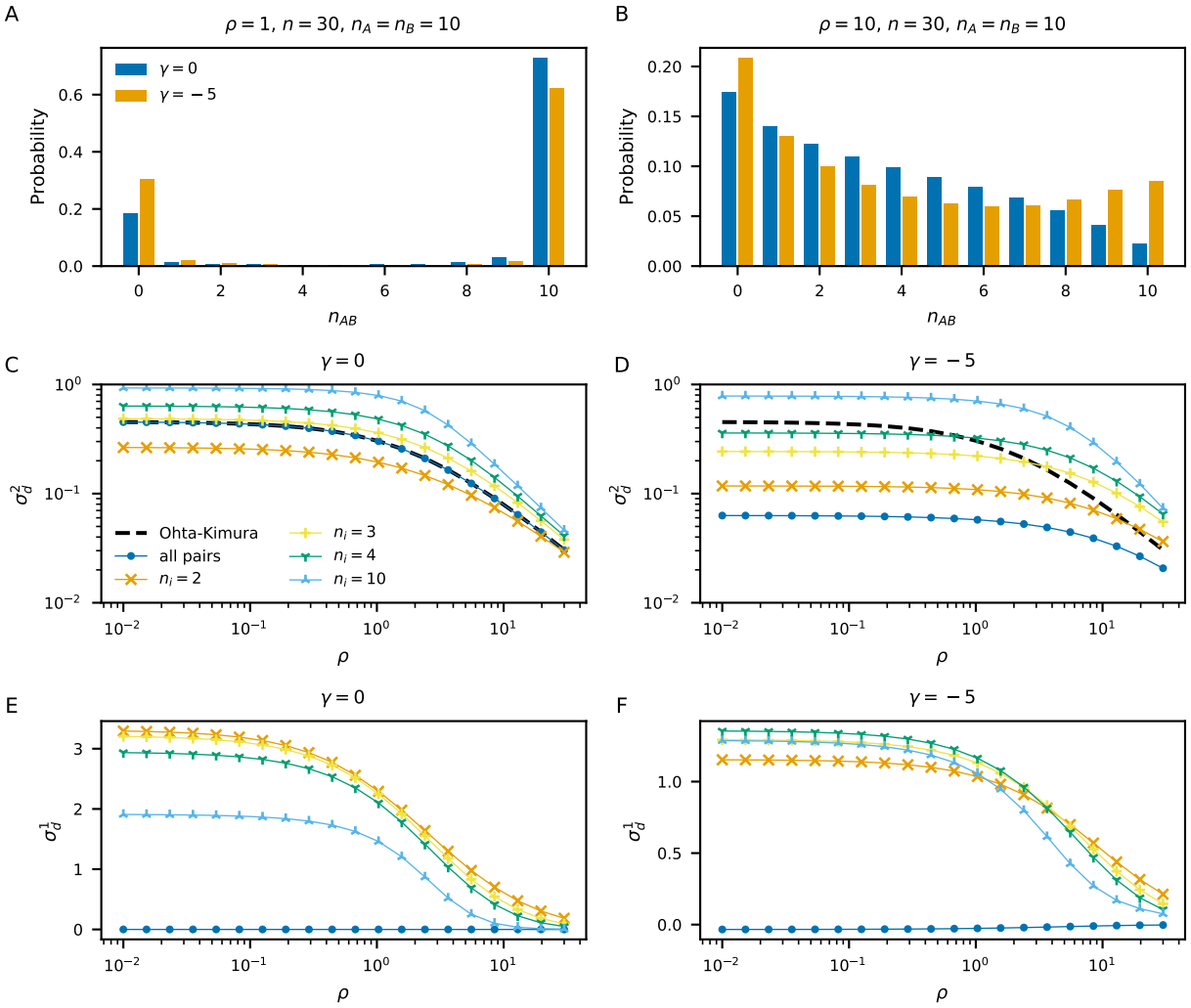

Figure 1: **Additive selection and allele count-conditioned LD.** (A and B) The distribution of AB haplotype counts in a sample size of 30, in which we observe 10 A alleles at the left locus, and 10 B alleles at the right locus. (C and D) The decay of normalized squared LD (σ_d^2) with scaled recombination distance for pairs of neutral and selected variants, respectively. The dashed line is the neutral expectation (Ohta and Kimura, 1971). (E and F) Similarly, the decay of $\sigma_d^1 = \mathbb{E}[D]/\mathbb{E}[p(1-p)q(1-q)]$ for neutral and selected variants, respectively. Note the difference in scale between the two panels.

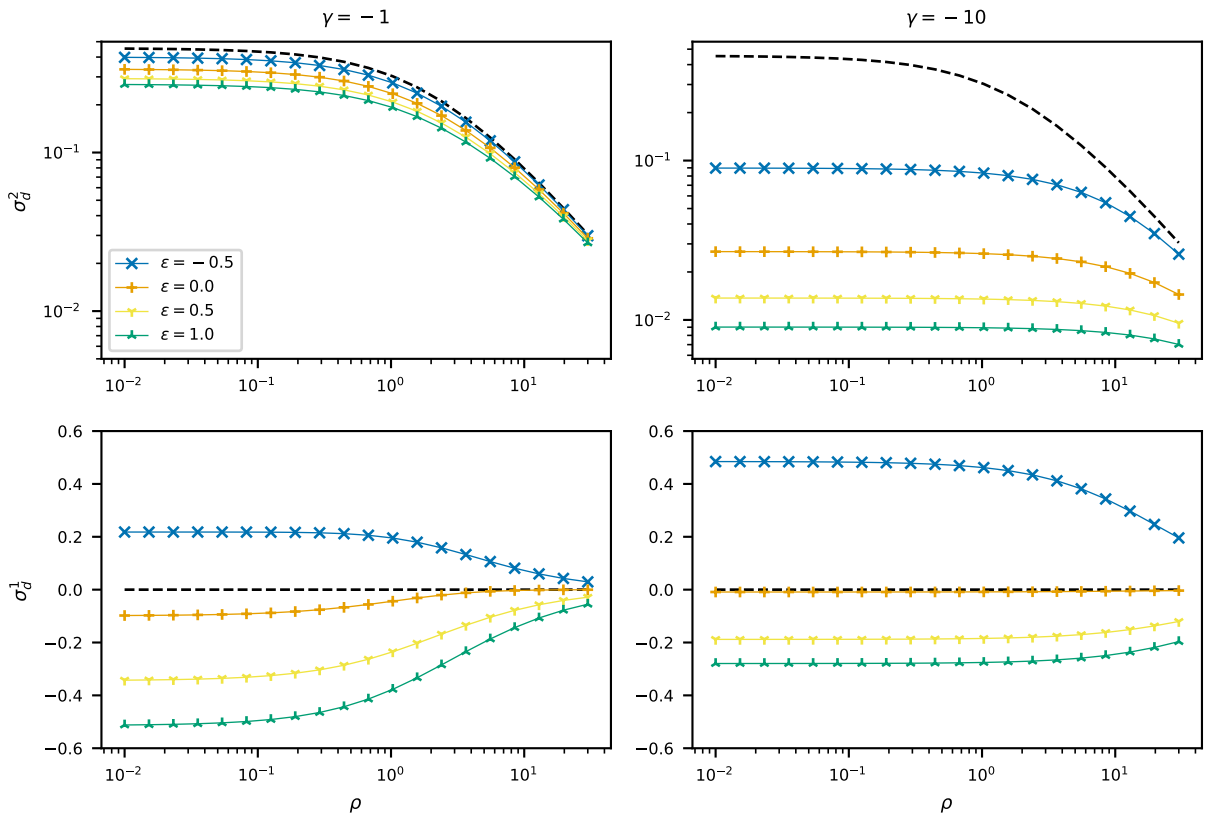

Figure 2: **Additive selection with epistasis.** (A) D^2, and (B) D.

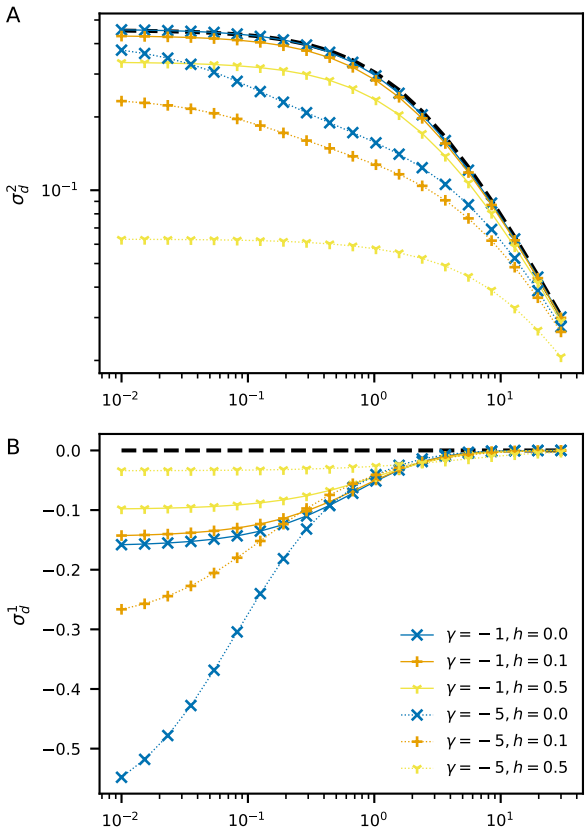

Figure 3: **Dominance effects.** (A) D^2, and (B) D.

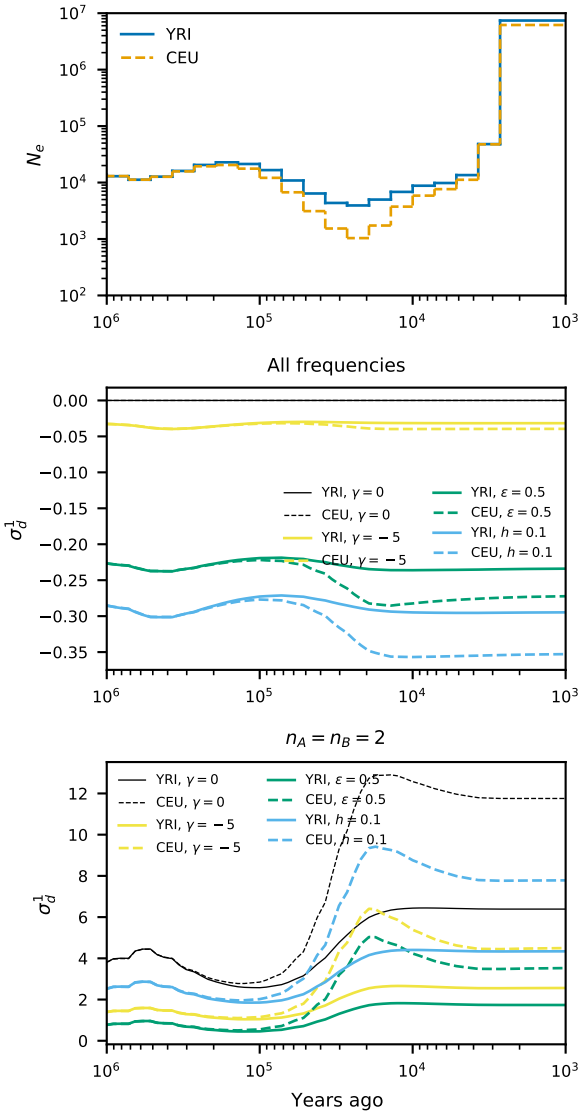

Figure 4: **The effects of demography on signed LD.**

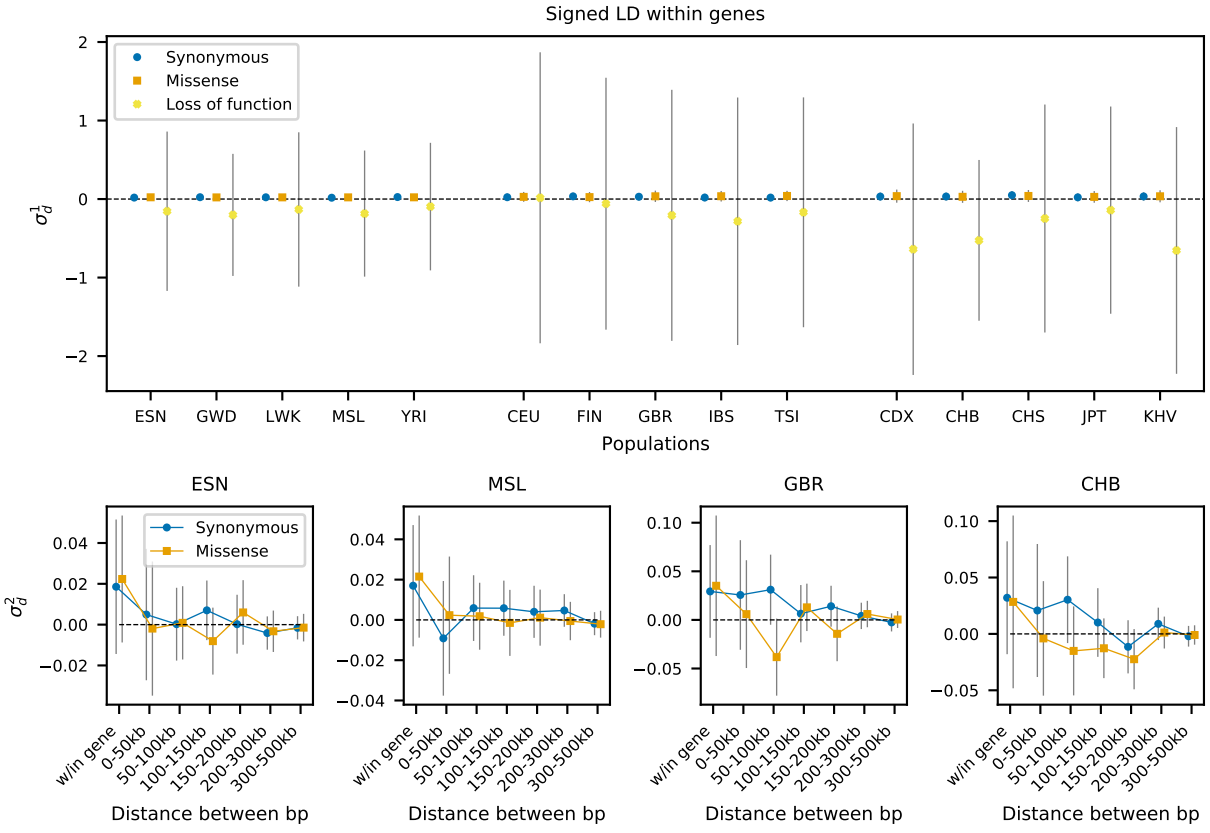

Figure 5: **LD within and between protein-coding genes.**

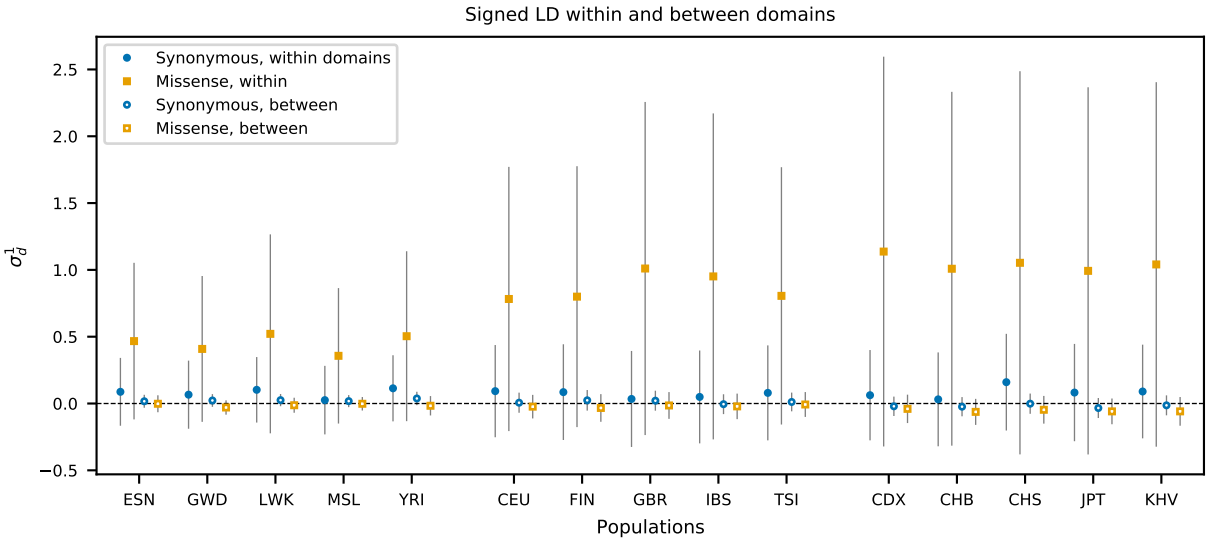

Figure 6: **LD within and between annotated domains in protein-coding genes.**

References

1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, *et al.*, 2015 A global reference for human genetic variation. Nature **526**: 68–74.

Agrawal, A. F., and M. C. Whitlock, 2011 Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics **187**: 553–566.

Barton, N. H., and B. Charlesworth, 1998 Why sex and recombination? Science **281**: 1986–1990.

Bershtein, S., M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik, 2006 Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature **444**: 929–932.

Birky, Jr, C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. U. S. A. **85**: 6414–6418.

Callahan, B., R. A. Neher, D. Bachtrog, P. Andolfatto, and B. I. Shraiman, 2011 Correlated evolution of nearby residues in drosophilid proteins. PLoS Genet. **7**: e1001315.

Charlesworth, B., 1990 Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet. Res. **55**: 199–221.

Clark, A. G., 1998 Mutation-selection balance with multiple alleles. Genetica **102-103**: 41–47.

Garcia, J. A., and K. E. Lohmueller, 2020 Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome.

Golding, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108**: 257–274.

Good, B. H., 2020 Linkage disequilibrium between rare mutations. BioRxiv **doi**: 10.1101/2020.12.10.420042.

Haldane, J. B. S., 1930 A note on fisher's theory of the origin of dominance, and on a correlation between dominance and linkage. Am. Nat. **64**: 87–90.

Haller, B. C., and P. W. Messer, 2019 SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol. Biol. Evol. **36**: 632–637.

Hallgrímsdóttir, I. B., and D. S. Yuster, 2008 A complete classification of epistatic two-locus models. BMC Genet. **9**: 17.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8**: 269–294.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**: 226–231.

Huber, C. D., A. Durvasula, A. M. Hancock, and K. E. Lohmueller, 2018 Gene expression drives the evolution of dominance. Nat. Commun. **9**: 2750.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics .

Ivankov, D. N., A. V. Finkelstein, and F. A. Kondrashov, 2014 A structural perspective of compensatory evolution. Curr. Opin. Struct. Biol. **26**: 104–112.

Kacser, H., and J. A. Burns, 1981 The molecular basis of dominance. Genetics **97**: 639–666.

Kimura, M., 1955 Random genetic drift in Multi-Allelic locus. Evolution **9**: 419–435.

Kimura, M., and T. Maruyama, 1966 The mutational load with epistatic gene interactions in fitness. Genetics **54**: 1337–1351.

Kondrashov, A. S., 1982 Selection against harmful mutations in large sexual and asexual populations.

Kondrashov, A. S., 1995 Dynamics of unconditionally deleterious mutations: Gaussian approximation and soft selection. Genet. Res. **65**: 113–121.

McVean, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175**: 1395–1406.

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155**: 929–944.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304**: 581–584.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63**: 229–238.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68**: 571–580.

Ragsdale, A. P., and S. Gravel, 2019 Models of archaic admixture and recent history from two-locus statistics. PLoS Genet. **15**: e1008204.

Ragsdale, A. P., and S. Gravel, 2020 Unbiased estimation of linkage disequilibrium from unphased data. Mol. Biol. Evol. **37**: 923–932.

Ragsdale, A. P., and R. N. Gutenkunst, 2017 Inferring demographic history using Two-Locus statistics. Genetics **206**: 1037–1048.

Romero, P. A., and F. H. Arnold, 2009 Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. **10**: 866–876.

Sandler, G., S. I. Wright, and A. F. Agrawal, 2020 Using patterns of signed linkage disequilibria to test for epistasis in flies and plants. BioRxiv **doi**: 10.1101/2020.11.25.399030.

Sohail, M., O. A. Vakhrusheva, J. H. Sul, S. L. Pulit, and others, 2017 Negative selection in humans and fruit flies involves synergistic epistasis .

Speidel, L., M. Forest, S. Shi, and S. R. Myers, 2019 A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. **51**: 1321–1329.

Stanek, D., D. M. Bis-Brewer, C. Saghira, M. C. Danzi, P. Seeman, *et al.*, 2020 Prot2HG: a database of protein domains mapped to the human genome. Database **2020**.

Steinberg, B., and M. Ostermeier, 2016 Shifting fitness and epistatic landscapes reflect trade-offs along an evolutionary pathway. J. Mol. Biol. **428**: 2730–2743.

Taverner, A. M., L. J. Blaine, and P. Andolfatto, 2020 Epistasis and physico-chemical constraints contribute to spatial clustering of amino acid substitutions in protein evolution.

Thornton, K. R., 2014 A c++ template library for efficient forward-time population genetic simulation of large populations. Genetics **198**: 157–166.

Turelli, M., and H. A. Orr, 2000 Dominance, epistasis and the genetics of postzygotic isolation. Genetics **154**: 1663–1679.

Yeang, C.-H., and D. Haussler, 2007 Detecting coevolution in and among protein domains. PLoS Comput. Biol. **3**: e211.

Zhao, L., and B. Charlesworth, 2016 Resolving the conflict between associative overdominance and background selection. Genetics **203**: 1315–1334.