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 19, 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,

One (!!) paragraph on epistasis, dominance, etc and overall importance. Thus, there is a lot of interest in disentangling and distinguishing between these processes.

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 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

In addition to impacting surrounding neutral variation, selected variants can interact and interfere with selection acting for or against other mutations. In their foundational paper, Hill and Robertson (1966) recognized that linked negatively selected variants impede the efficacy of a population to remove those mutations. Generally, 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; Barton, 1995). The joint allele frequency dynamics of linked variants under selection are also predicted to deviate from expectations without interference. Since haplotypes that carry multiple deleterious mutations have lower fitness than haplotypes that carry just one or the other under a multiplicative fitness model, 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. Nonetheless, this provides a simple, testable hypothesis under the standard selection model of additive effects within loci and multiplicative fitness between: we should observe average signed LD to be negative between pairs of selected variants.

Aside from selective interference under a multiplicative fitness model, where the relative fitness conferred by a haplotype carrying two deleterious mutations is equal to the product of the fitnesses of haplotypes carrying individual mutations, considerable attention has been given to non-multiplicative deviations in

the fitness function, known as epistasis. Early work focused primarily on mutations of strong effect and explored models of both synergistic and diminishing returns (or antagonistic or threshold) epistasis under sexual and asexual reproduction, often in the limiting case of free recombination between deleterious variants. Generally, for sexually reproducing organisms, deleterious load is decreased under simple models of synergistic epistasis compared to multiplicative fitness (Kimura and Maruyama, 1966; Kondrashov, 1995). Pervasive epistasis has been invoked as an explanation for the evolutionary advantage of sex (Kondrashov, 1982; Charlesworth, 1990; Barton and Charlesworth, 1998), as well as driving 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, with negative synergistic epistasis leading to an excess of negative LD (Kondrashov, 1982; Charlesworth, 1990).

While epistasis is the most commonly considered nonlinear effect influencing interactions between selected variants, non-additive effects *within* a locus may also cause deviations from our baseline models, either with or without epistasis. Dominance and multi-locus interactions are jointly important in shaping the expected equilibrium allele frequencies and mutation load due to strongly damaging disease mutations (Clark, 1998). However, despite dominance and partial recessiveness being appreciated as important factors shaping linked variation (Turelli and Orr, 2000; Zhao and Charlesworth, 2016), dominance has received less attention than epistasis in multi-locus selection models, and its impact on the joint segregation of negatively selected variants is poorly understood. There is ample evidence that a large fraction of selected mutations are at least partially recessive in humans (**?**), flies (**?**), and other species (**?**), and that average levels of dominance vary with the strength of selection (**?**).

- Patterns of dominance
  - Huber *et al.* (2018), Haldane (1930), Kacser and Burns (1981)

Ohta and Kimura (1969) and Ohta and Kimura (1971)

- Diffusion models for the neutral two-locus decay of LD σ_d^2

Hallgrímsdóttir and Yuster (2008)

-

Good (2020)

-
- Neutrality, additive, dominance, epistasis, general diploid selection models
- Analytic results: Older stuff, like some basic results from HR, through more recent results in Good (2020)
  - recursions under neutrality (Golding, 1984; Hudson, 2001)
- Recent numerical approaches (Ragsdale and Gutenkunst, 2017; Ragsdale and Gravel, 2019) and others that aren't me?
- Resorting to discrete forward simulations (a few examples)

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 regions, and assess how well we can expect to discriminate between modes of selective interactions from genome-wide patterns of LD.

Methods

The two-locus sampling distribution with arbitrary selection

Drift and recombination

Selection models with epistasis and dominance

Low-order summaries of the sampling distribution

Numerical solution

Moment closure approach

Implementation

- Numerical approach for solving the two-locus model with recombination, single-population demography, and general selection models, following Jouganous *et al.* (2017) and Ragsdale and Gravel (2019).

Validation

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?)

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

Population size changes [[do what]]

- Effect of demography on these same statistics
  - Bottleneck with recovery

  - Human-like size histories (using Speidel *et al.* (2019))

Data... main result in title

- 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. 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?

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:

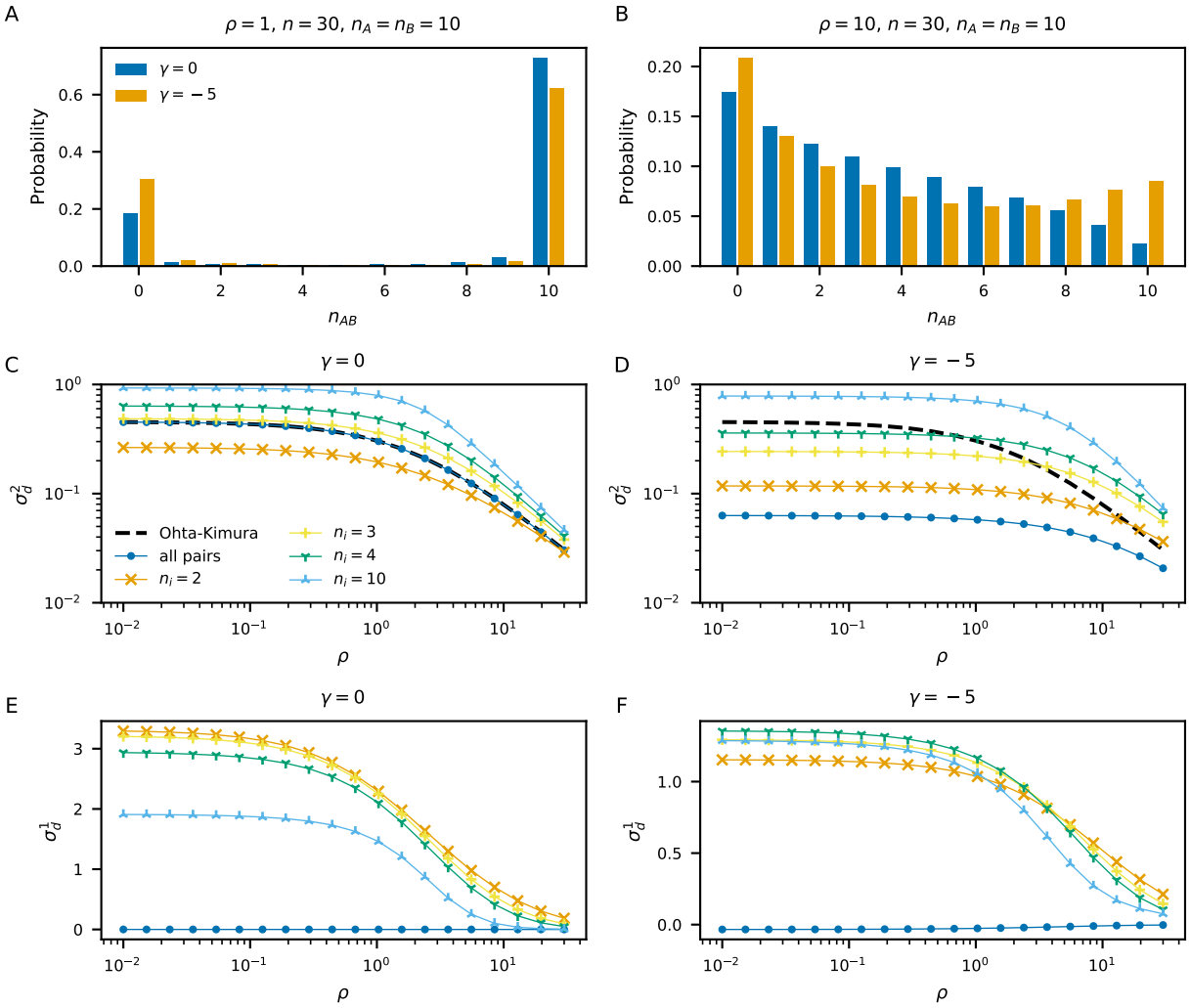

Figure 1: **Additive selection and allele count-conditioned LD.** (A and B) The distribution of AB haplotype counts in a sample size of 30, in which we observe 10 A alleles at the left locus, and 10 B alleles at the right locus. (C and D) The decay of normalized squared LD (σ_d^2) with scaled recombination 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.

4. Effects of demography on LD (bottleneck and recovery vs expansion, following Tennessen model)
5. LOF, missense, and synonymous variants, within and between genes.
6. Missense variants in genes partitioned by conservation? Tolerance to nonsynonymous changes?

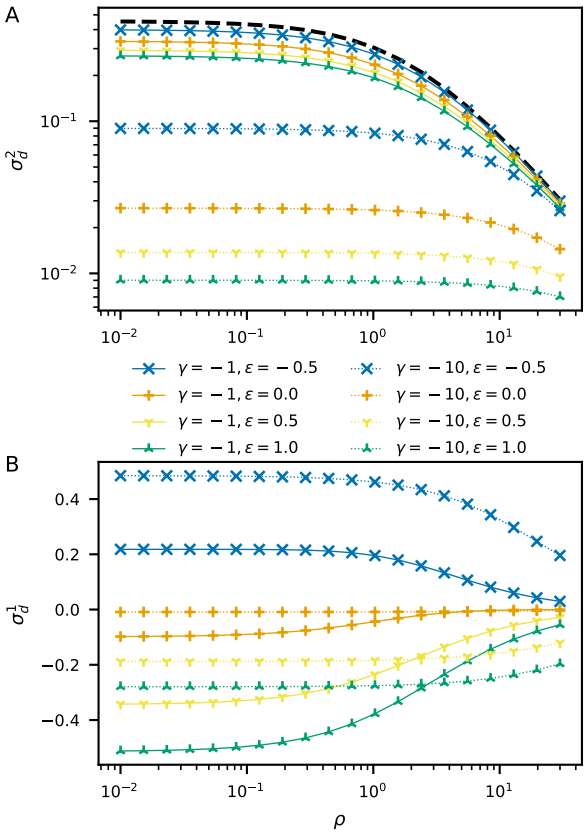

Figure 2: **Additive selection with epistasis.** (A) D^2, and (B) D.

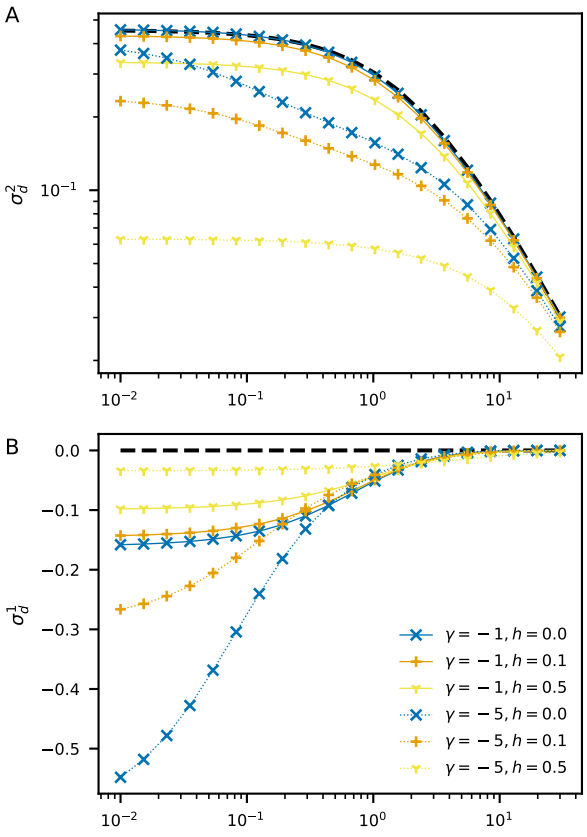

Figure 3: **Dominance.** (A) D^2, and (B) D.

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.

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics **140**: 821–841.

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.

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.

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 .

Jouganous, J., W. Long, A. P. Ragsdale, and S. Gravel, 2017 Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. Genetics **206**: 1549–1567.

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

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.

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.

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

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