

EVOLUTIONARY GENETICS

Negative selection in humans and fruit flies involves synergistic epistasis

Mashaal Sohail,^{1,2,3} Olga A. Vakhrusheva,⁴ Jae Hoon Sul,⁵ Sara L. Pulit,^{6,7} Laurent C. Francioli,⁷ Genome of the Netherlands Consortium,* Alzheimer's Disease Neuroimaging Initiative,† Leonard H. van den Berg,⁶ Jan H. Veldink,⁶ Paul I. W. de Bakker,^{7,8} Georgii A. Bazykin,^{4,9} Alexey S. Kondrashov,^{10,11‡} Shamil R. Sunyaev^{2,3‡}

Negative selection against deleterious alleles produced by mutation influences within-population variation as the most pervasive form of natural selection. However, it is not known whether deleterious alleles affect fitness independently, so that cumulative fitness loss depends exponentially on the number of deleterious alleles, or synergistically, so that each additional deleterious allele results in a larger decrease in relative fitness. Negative selection with synergistic epistasis should produce negative linkage disequilibrium between deleterious alleles and, therefore, an underdispersed distribution of the number of deleterious alleles in the genome. Indeed, we detected underdispersion of the number of rare loss-of-function alleles in eight independent data sets from human and fly populations. Thus, selection against rare protein-disrupting alleles is characterized by synergistic epistasis, which may explain how human and fly populations persist despite high genomic mutation rates.

Negative, or purifying, selection prevents the unlimited accumulation of deleterious mutations and establishes a mutation-selection equilibrium (1). The properties of negative selection are determined by the corresponding fitness landscape, the map that relates fitness to the “mutation burden” in an individual. Because of the difficulty of ascribing precise selection coefficients to different alleles, the mutation burden can be approximated by the total number of putatively deleterious mutations in an individual. Under the null hypothesis of no epistasis, selection acts on different mutations inde-

pendently, so that each additional mutation causes the same decline in relative fitness and fitness declines exponentially with their number.

By contrast, if synergistic, or narrowing (2), epistasis between deleterious alleles is present, each additional mutation causes a larger decrease in relative fitness. Synergistic epistasis reduces the mutation load under a given genomic rate of deleterious mutations (1, 3, 4) and can make sex and recombination advantageous (5). However, because neither the mutation burden nor fitness can be easily measured, data on fitness landscapes of negative selection remain inconclusive (6). Theory suggests that narrowing epistasis may emerge as a result of pervasive pleiotropy and the modular organization of biological networks (7). Some genome-wide investigations have found epistasis but no consistent directionality of effect (6, 8, 9).

We examined the distribution of the mutation burden in human and *Drosophila melanogaster* populations. In the absence of epistasis, alleles should contribute to the mutation burden independently (3), such that the variance of the mutation burden is equal to the sum of the variances at all loci or the additive variance (V_A) (10, 11) (Fig. 1). If mutant alleles are rare, the mutation burden follows a Poisson distribution with a variance (σ^2) equal to its mean (μ) (fig. S1).

In contrast, epistatic selection creates dependencies between deleterious alleles, so the total variance of the mutation burden is no longer equal to the additive variance (12). Selection with synergistic epistasis creates repulsion, or negative linkage disequilibrium (LD). As a result of this LD, the variance of the mutation burden is reduced by a factor of ρ (<1), which is determined by the strength of selection and the extent of epistasis, leading to an underdispersion ($\sigma^2 < V_A$) (12, 13)

(fig. S2). Antagonistic (diminishing returns) epistasis, instead, creates positive LD between deleterious alleles and increases the variance of the mutation burden leading to its overdispersion ($\sigma^2 > V_A$). Also, the difference between σ^2 and V_A is a genome-wide estimate of the net LD in fitness (11, 14). Using fully sequenced individual genomes from a population, we tested for synergistic epistasis without needing to measure fitness directly.

The ideal population for our test would be single-ancestry, outbred, nonadmixed, and randomly mating. We analyzed the Genome of the Netherlands (GoNL) Project (15), the Alzheimer's Disease Neuroimaging Initiative (ADNI), and Dutch controls from Project MinE, an amyotrophic lateral sclerosis study. For each of these, we obtained whole-genome sequences of unrelated individuals of European descent. We obtained similar data for Zambian flies from phase 3 of the *Drosophila* Population Genomics Project (DPGP3) (16). For each population, after applying stringent quality control filters (tables S8 to S12), we computed the mutation burden and corresponding σ^2 and V_A values, focusing on rare alleles for coding synonymous, missense, and loss-of-function (LoF) mutations (here defined as splice site disrupting or nonsense). For all of these data sets, the distribution of rare LoF alleles was underdispersed (Table 1).

On average, rare LoF alleles displayed variance (σ^2) reduced by a factor of ~ 0.95 , compared to additive variance (V_A). In contrast, rare synonymous and missense alleles were overdispersed. The GoNL project also provided a set of high-quality short insertions and deletions (indels), and in this data set, we observed an underdispersed distribution for the combined set of LoF alleles and frameshift indels (table S19). Overlaying the mutation burden distributions with Poisson distributions having identical means shows that the underdispersion is due to a depletion of individuals with a high number of deleterious alleles (figs. S12 to S17).

Even without epistasis, overdispersion in the mutation burden would be observed if genome-wide positive LD is present owing to population structure (Fig. 1) (17). If the population has a cline in average rare mutation burden (μ) due to, for example, a south-to-north expansion (15) followed by assortative mating, this may translate into an excess of σ^2 over V_A (figs. S3 and S4). Overdispersion may also be caused by DNA samples being sequenced or processed in different batches. A large proportion of the overdispersion in rare mutation burden computed on synonymous or missense alleles in the detailed GoNL samples could be attributed to geographic origin and sequencing batch (fig. S5 and tables S4 and S15). In contrast, LoF alleles were not significantly overdispersed by confounders (table S16). This is consistent with the results obtained for populations simulated under heterogeneous demography, which show that overdispersion in mutation burden decreases with the strength of negative selection (Fig. 2A) (11).

Given that overdispersion scales with selection strength, we constructed a “crucial” genome

¹Systems Biology PhD Program, Harvard Medical School, Boston, MA, USA. ²Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ³Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ⁴Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russia. ⁵Department of Psychiatry and Biobehavioral Sciences, University of California–Los Angeles, Los Angeles, CA, USA. ⁶Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, Netherlands. ⁷Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, Netherlands. ⁸Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands. ⁹Skolkovo Institute of Science and Technology, Skolkovo, Russia. ¹⁰Department of Bioengineering and Bioinformatics, M. V. Lomonosov Moscow State University, Moscow, Russia. ¹¹Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA.

*The members of the Genome of the Netherlands Consortium are listed in the supplementary materials.

†A complete listing of Alzheimer's Disease Neuroimaging Initiative (ADNI) investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

‡Corresponding author. Email: ssunyaev@rics.bwh.harvard.edu (S.R.S.); kondrashov@umich.edu (A.S.K.)

for humans, selecting only genes with an estimated selection coefficient against heterozygous protein-truncating variants exceeding 0.2 (11). An analogous essential genome was constructed for

D. melanogaster using the Database of Essential Genes (11). When only their crucial or essential genomes were considered, both humans (Fig. 2B and fig. S8) and *D. melanogaster* (fig. S9) showed an under-

dispersion in their missense mutation burden. In contrast, synonymous alleles remained overdispersed. Accordingly, we also observed that σ^2/V_A scales inversely with the strength of selection

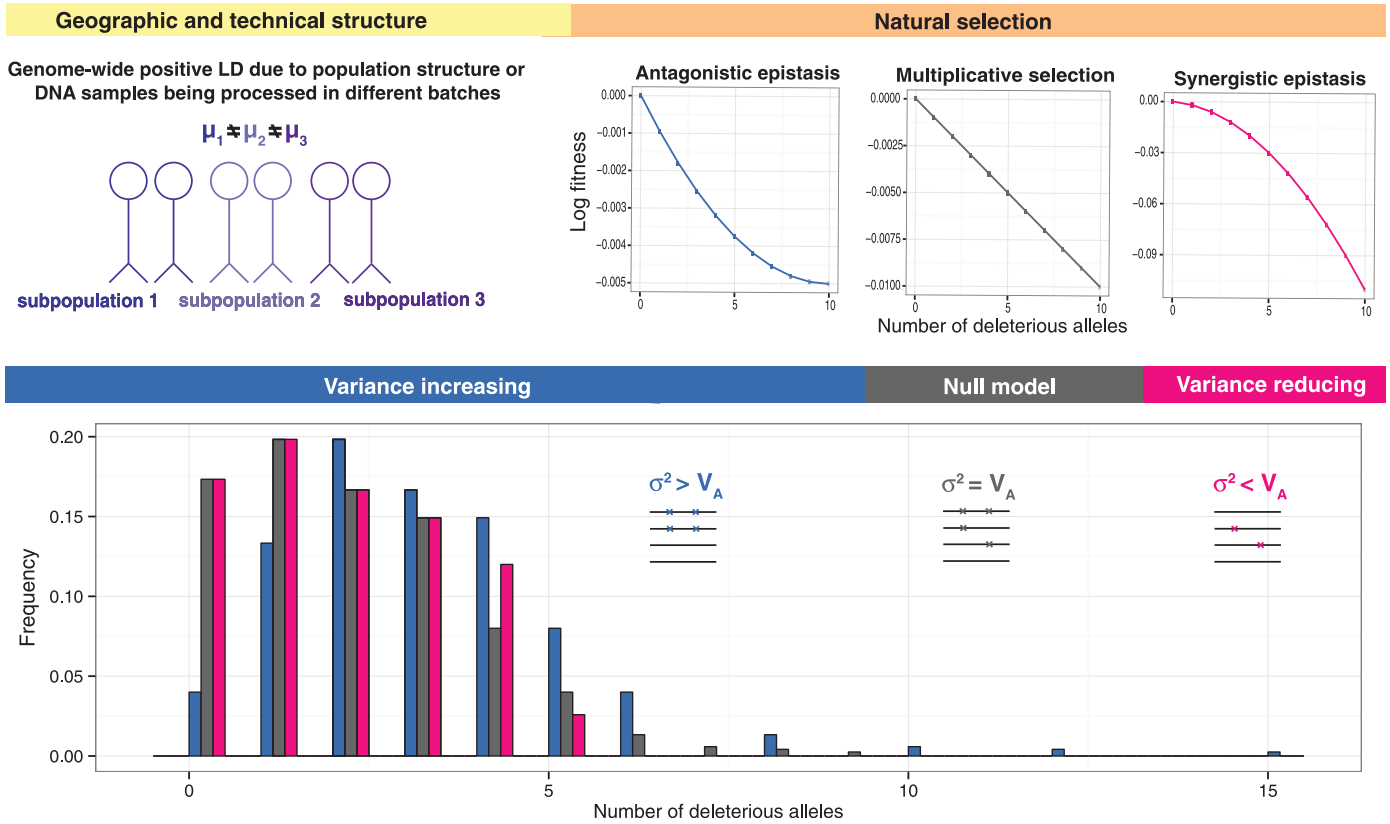


Fig. 1. Rare mutation burden under natural selection (orange, right) and population structure (yellow, left). The mutation burden (bottom panel) is shown under the null model (gray, the absence of epistasis and population

structure) and under variance-increasing (blue, antagonistic epistasis and population structure) and variance-reducing (pink, synergistic epistasis) models. μ_k is the mean of the mutation burden in subpopulation k within the population.

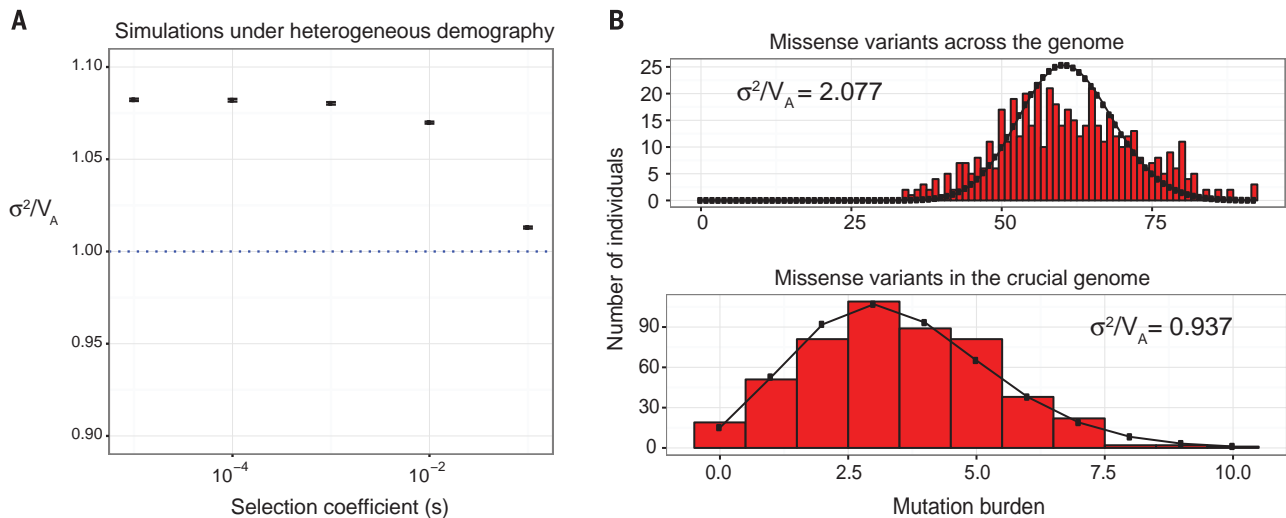


Fig. 2. Simulated and empirical distributions of rare missense mutation burden. (A) Simulations using SLiM 2.0 of unlinked sites under multiplicative selection in a finite population with heterogeneous demography (11). σ^2/V_A was calculated for the rare mutation burden computed on singletons at equilibrium, with the null expectation as shown (blue dotted line). Error bars show SEM (100 replicates). (B) Missense rare mutation

burden (red) computed on singletons across the genome ($\sigma^2/V_A = 2.077$) and only in the crucial genome ($\sigma^2/V_A = 0.937$) in the GoNL data set, overlaid with Poisson distributions (black) having identical means. The crucial genome for humans was constructed by selecting only genes with an estimated selection coefficient against heterozygous protein-truncating variants exceeding 0.2 (11).

acting on a gene for missense but not for synonymous alleles in the fly data sets (fig. S18) (11).

To investigate the significance of the underdispersion in rare LoF alleles, we generated an empirical null distribution for σ^2/V_A for each data set by resampling synonymous alleles at matched allele frequency as our test set of LoF alleles (Fig. 3) (11). We meta-analyzed the human data with three suitable (low inbreeding and admixture) non-European populations from phase I of the 1000 Genomes Project (18) (tables S1 and S2), and the fruit fly data with an American population from the *D. melanogaster* Genetic Reference Panel (DGRP) (19) (table S3). Meta-analysis across all data sets using Stouffer's method indicates that rare LoF alleles were significantly underdispersed in humans ($P = 0.0003$) and flies ($P = 9.43 \times 10^{-6}$) (11). Permuting functional consequences across variants, we confirmed the significance of our underdispersion signal in rare protein-altering mutations in humans (missense $P = 2.670 \times 10^{-4}$,

LoF $P = 0.002$) and *D. melanogaster* (missense $P = 9.43 \times 10^{-6}$, LoF $P = 0.0001$) (11). Furthermore, through regression analysis, resampling experiments, and simulations, we showed that the underdispersion signal persists after correcting for potential confounders and is not driven by outliers (tables S5, S17, and S18 and fig. S11) (11).

We also sought to determine the source of the observed negative LD and what it says about the shape of the fitness landscape. Directional selection with synergistic epistasis was proposed as a solution to the mutation load paradox (3, 4) and as a deterministic mechanism for the evolution of sex (5). However, as long as mutations are not unconditionally deleterious, they may be subject to stabilizing selection instead of directional selection, and this may also result in negative LD (20). Furthermore, in small populations, genetic drift in the presence of multiplicative selection may act as a random force to create negative LD, because mutations that

arise as unique events at different sites will be in repulsion (21, 22).

Although stabilizing selection is always narrowing and can thus be regarded as simply another way of generating synergy, a far lower mutational load is generated under stabilizing selection compared with purely directional selection (20). However, LoF alleles are likely to be unconditionally deleterious. With regard to the role of genetic drift, we validated with simulations of finite populations with realistic human demography that negative LD between unlinked sites is quantitatively negligible under a model of multiplicative selection (fig. S10). We also demonstrated that most of our signal in rare LoF alleles comes from net negative LD between completely unlinked alleles on different chromosomes (table S6) and very distant alleles on the same chromosome (figs. S6 and S7). If the source of negative LD is narrowing selection, then sexual reproduction has an evolutionary advantage for purely deterministic reasons. Our analysis cannot preclude the role of random chance or genetic drift in aiding this advantage by creating negative LD, as our signal, in part, comes from linked sites in the genome, although the majority does not.

Our empirical observations on properties of the fitness landscape for protein-disrupting variants have broader evolutionary implications, especially if the results extend to the broader class of mildly deleterious alleles. The question of how our species accommodates high deleterious mutation rates has long been pondered. Indeed, a newborn is estimated to have ~70 de novo mutations (23). The consensus for estimates for the fraction of the genome that is "functional" is that about 10% of the human genome sequence is selectively constrained (24). Thus, the average human should carry at least seven de novo deleterious mutations. If natural selection acts on each mutation independently, the resulting mutation load and loss in average fitness are inconsistent with the existence of the human population ($1 - e^{-7} > 0.99$). To resolve this paradox, it is sufficient to assume that the fitness landscape is flat only outside the zone where all the genotypes actually present are contained, so that selection within the population proceeds as if epistasis were absent (20, 25). However, our findings suggest that synergistic epistasis affects even the part of the fitness landscape that corresponds to genotypes that are actually present in the population.

Currently, although selection due to reproductive mortality in humans is deeply relaxed, there is still a substantial opportunity for selection (26, 27). Thus, our results suggest that even humans are experiencing ongoing narrowing negative selection.

Table 1. Negative linkage disequilibrium (LD) between rare LoF alleles in human and *D. melanogaster* genomes.

For humans, only singletons, and for flies, only alleles up to a minor allele count of 5, are included (see tables S2 and S3 for other frequency cut-offs). Net LD is normalized per pair of alleles and per pair of loci (11). A one-sided P value was obtained for σ^2/V_A by permutation, and a joint P value for all three human data sets shown (GoNL, ADNI, MinE) was computed by meta-analysis using Stouffer's method (11) (coding synonymous $P = 0.999$, missense $P = 5.155 \times 10^{-4}$, LoF $P = 0.002$). The number of samples is given in parentheses for each data set.

Variant type	Mean	σ^2/V_A	Net LD	
			Per pair of derived alleles	Per pair of loci
Humans				
<i>Genome of the Netherlands GoNL (495)</i>				
Synonymous	30.26	1.675	0.022	4.554×10^{-8}
Missense	60.88	2.077	0.018	3.609×10^{-8}
Nonsense	1.67	0.929	-0.039	-8.013×10^{-8}
Splice	0.90	0.953	-0.049	-1.008×10^{-7}
LoF	2.58	0.930	-0.029	-5.848×10^{-8}
<i>European ancestry ADNI (714)</i>				
Synonymous	38.99	2.077	0.028	2.709×10^{-8}
Missense	77.98	2.008	0.013	1.268×10^{-8}
Nonsense	2.10	0.933	-0.032	-3.126×10^{-8}
Splice	1.16	0.878	-0.104	-1.020×10^{-7}
LoF	3.26	0.930	-0.022	-2.126×10^{-8}
<i>Dutch MinE (601)</i>				
Synonymous	42.93	1.749	0.017	2.414×10^{-8}
Missense	79.34	1.960	0.012	1.675×10^{-8}
Nonsense	1.89	1.057	0.028	3.898×10^{-8}
Splice	0.95	0.972	-0.033	-4.641×10^{-8}
LoF	2.83	0.996	-0.001	-1.727×10^{-9}
<i>D. melanogaster</i>				
<i>Zambian DPGP3 (191)</i>				
Synonymous	3577.06	57.473	0.016	1.658×10^{-6}
Missense	2051.52	18.536	0.008	6.710×10^{-7}
Nonsense	10.21	0.928	-0.007	-4.139×10^{-7}
Splice	2.60	0.948	-0.020	-1.308×10^{-6}
LoF	12.81	0.929	-0.005	-3.298×10^{-7}

REFERENCES AND NOTES

- H. J. Muller, *Am. J. Hum. Genet.* **2**, 111–176 (1950).
- E. E. Shnol, A. S. Kondrashov, *Genetics* **134**, 995–996 (1993).
- M. Kimura, T. Maruyama, *Genetics* **54**, 1337–1351 (1966).

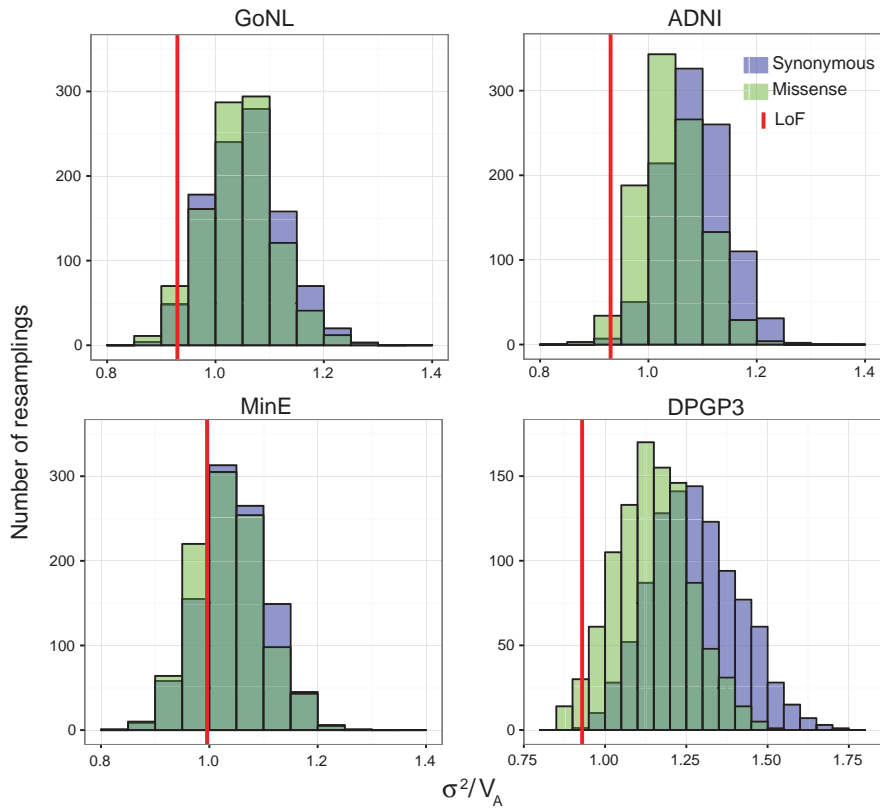


Fig. 3. Resampling distributions of σ^2/V_A for rare LoF mutation burden in humans and *D. melanogaster*. Synonymous (purple) and missense (green) alleles were resampled at the same allele frequency as LoF alleles to obtain empirical null distributions for σ^2/V_A in each data set. For humans, only singletons, and for flies, only alleles up to a minor allele count of 5, are included. A one-sided P value for σ^2/V_A of the rare LoF mutation burden (red) was obtained, and a joint P value for all three human data sets shown (GoNL, ADNI, MinE) was computed by meta-analysis using Stouffer's method (11) ($P = 0.0003$).

- J. F. Crow, M. Kimura, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 396–399 (1979).
- A. S. Kondrashov, *Nature* **336**, 435–440 (1988).
- J. A. G. M. de Visser, S. F. Elena, *Nat. Rev. Genet.* **8**, 139–149 (2007).
- H.-C. Chiu, C. J. Marx, D. Segrè, *Proc. Biol. Sci.* **279**, 4156–4164 (2012).

- C. Bank, R. T. Hietpas, J. D. Jensen, D. N. A. Bolon, *Mol. Biol. Evol.* **32**, 229–238 (2015).
- O. Puchta *et al.*, *Science* **352**, 840–844 (2016).
- M. Bulmer, *The Mathematical Theory of Quantitative Genetics* (Clarendon, 1980).
- Materials and methods are available as supporting materials.

- A. S. Kondrashov, *Genet. Res.* **65**, 113–121 (1995).
- B. Charlesworth, *Genet. Res.* **55**, 199–221 (1990).
- M. G. Bulmer, *Genet. Res.* **28**, 101–117 (1976).
- L. C. Francioli *et al.*, *Nat. Genet.* **46**, 818–825 (2014).
- J. B. Lack *et al.*, *Genetics* **199**, 1229–1241 (2015).
- R. C. Yang, *Genetics* **155**, 1449–1458 (2000).
- G. R. Abecasis *et al.*, *Nature* **491**, 56–65 (2012).
- T. F. C. Mackay *et al.*, *Nature* **482**, 173–178 (2012).
- B. Charlesworth, *Evolution* **67**, 3354–3361 (2013).
- J. Felsenstein, *Genetics* **78**, 737–756 (1974).
- N. H. Barton, S. P. Otto, *Genetics* **169**, 2353–2370 (2005).
- S. Besenbacher *et al.*, *Nat. Commun.* **6**, 5969 (2015).
- C. M. Rands, S. Meader, C. P. Ponting, G. Lunter, *PLOS Genet.* **10**, e1004525 (2014).
- Y. Leseqque, P. D. Keightley, A. Eyre-Walker, *Genetics* **191**, 1321–1330 (2012).
- S. C. Stearns, S. G. Byars, D. R. Govindaraju, D. Ewbank, *Nat. Rev. Genet.* **11**, 611–622 (2010).
- E. C. Larsen, O. B. Christiansen, A. M. Kolte, N. Macklon, *BMC Med.* **11**, 154 (2013).

ACKNOWLEDGMENTS

We are grateful to L. Mirny, G. McVean, and I. Adzhubey for scientific discussions; all members of the Sunyaev lab and two anonymous reviewers for comments that improved the manuscript; D. Jordan for providing SLiM 2.0 simulation runs; C. Cassa, D. Jordan, D. Weghorn, and D. Balick for providing genic selection estimates for humans; J. Fan for helping with analyses as a summer student; and J. Lack for help with *D. melanogaster* inversion data. This project was supported by NIH grants R01GM078598, R01GM105857, R01MH101244, and U01HG009088. Analysis of fruit fly data was performed at IITP RAS and supported by the Russian Science Foundation (grant no. 14-50-00150). Data sets used in this study can be accessed as follows: GoNL: www.nlggenome.nl/; ADNI: <http://adni.loni.usc.edu/>; Project MinE: www.projectmine.com; The 1000 Genomes Phase I Project: [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/); DGRP and DPGP3: www.johnpool.net/genomes.html.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/356/6337/539/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S22
Data file S1 to S3
Consortia
References (28–56)

8 July 2016; resubmitted 28 November 2016
Accepted 14 April 2017
10.1126/science.aah5238