Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements

David Murphy[1, 2, *], Eyal Elyashiv[1], Guy Amster[1, #] and Guy Sella[1, 3, #, *]

[1] Department of Biological Sciences, Columbia University, New York, NY 10027

[2] Genes and Human Disease Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104

[3] Program for Mathematical Genomics, Columbia University, New York, NY 10032

[#] These authors contributed equally to this work

[*] To whom correspondence should be addressed: david-murphy@omrf.org or gs2747@columbia.edu

Abstract

Analyses of genetic variation in many taxa have established that neutral genetic diversity is shaped by natural selection at linked sites. Whether the source of selection is primarily the fixation of strongly beneficial alleles (selective sweeps) or purifying selection on deleterious mutations (background selection) remains unknown, however. We address this question in humans by fitting a model of the joint effects of selective sweeps and background selection to autosomal polymorphism data from the 1000 Genomes Project. After controlling for variation in mutation rates along the genome, a model of background selection alone explains ~60% of the variance in diversity levels at the megabase scale. Adding the effects of selective sweeps driven by adaptive substitutions to the model does not improve the fit, and when both modes of selection are considered jointly, selective sweeps are estimated to have had little or no effect on linked neutral diversity. The regions under purifying selection are best predicted by phylogenetic conservation, with ~80% of the deleterious mutations affecting neutral diversity occurring in non-exonic regions. Thus, background selection is the dominant mode of linked selection in humans, with marked effects on diversity levels throughout autosomes.

Introduction

Selection at a given locus in the genome affects diversity levels at sites linked to it[1-27]. In particular, when a new, strongly beneficial mutation increases in frequency to fixation in the population, it carries with it the haplotype on which it arose, thus reducing levels of neutral diversity nearby, in what is sometimes called a "hard selective sweep"[2,3]. "Soft sweeps", particularly those in which an allele segregates at low frequency before becoming beneficial and sweeping to fixation, and "partial sweeps", in which a beneficial mutation rapidly increases to intermediate frequencies, also reduce neutral diversity levels near the selected sites[14-17,21,26]. Similarly, when deleterious mutations are eliminated from the population by selection, so are the haplotypes on which they lie. This process too reduces diversity levels near selected sites, in a phenomenon known as "background selection"[5,8,9,11-13,24,25,27]. Because the lengths of the haplotypes associated with selected alleles depend on the recombination rate, linked selection processes cause a greater reduction in levels of neutral genetic diversity in regions with lower rates of recombination or a greater density of selected sites. These predicted relationships have been observed in numerous taxa, including plants, Drosophila, rodents and primates, establishing the effects of linked selection in a wide range of species[4,18,19,23,28-35].

More recently, the advent of large genomic datasets and detailed functional annotations have made it possible to build maps of the effects of linked selection that predict levels of diversity along the genome[36,37]. The first effort predated the availability of genome-wide resequencing data, relying instead on information about incomplete lineage sorting among human, chimpanzee and gorilla, which reflects variation in diversity levels along the genome in the common ancestor of humans and chimpanzees[36]. This pioneering paper showed that a model of background selection fits variation in human-chimpanzee divergence levels along the genome remarkably well, with only a few parameters. Nonetheless, the estimate of the deleterious mutation rate was unrealistically high, much greater than estimates of the total mutation rate per site in humans[38,39] (SOM Section 5), raising the possibility that it was soaking up effects of other modes of selection, notably

those of selective sweeps. Subsequent work suggested that selective sweeps have had little effect on diversity levels in humans, with no more of a reduction of diversity around nonsynonymous substitutions than around synonymous ones[40,41], but the interpretation of these findings was contested. Notably, it was suggested that the effects of background selection are more pronounced around synonymous substitutions than nonsynonymous substitutions, thereby masking the effects of selective sweeps[42]. Thus, we still lack an understanding of the relative contribution of adaptation and purifying selection[10], as well as a map of their effects on human diversity levels.

Results

Model and inference. Here, we resolve these issues by considering the effects of background selection and selective sweeps on diversity levels jointly (Fig. 1 and SOM Section 1). To this end, we model the effects of background selection on neutral diversity levels as a function of genetic distance from regions that may be under purifying selection (following [8] and [43]), where the deleterious mutation rate per site and distribution of selection effects in a given type of region (e.g., exons) are parameters to be estimated (Fig. 1a). In turn, the effects of sweeps are modeled as a function of genetic distance from selected substitutions on the human lineage (following [44] and [45]), where the fraction of substitutions of a given type (e.g., nonsynonymous) that were beneficial and their distribution of selection effects are again parameters to be estimated (Fig. 1b). Importantly, our model should capture the effects of any kind of sweeps, be they hard, partial or soft, so long as they eventually resulted in a substitution and affected diversity levels nearby (see SOM Section D in [37]). Given the positions of different types of selected regions and substitutions, their corresponding selection parameters, and a fine-scale genetic map, the model allows us to calculate the marginal probability that any given neutral site in the genome is polymorphic in a sample (Fig. 1c). Provided measurements of polymorphism at neutral positions throughout the genome, we combine information across sites and samples to calculate the composite likelihood of selection parameters, and find the parameter values that maximize this likelihood (Fig. 1 and SOM Section 1). In addition to

parameter estimation, this approach yields a map of the expected neutral diversity levels along the genome (Fig. 1c).

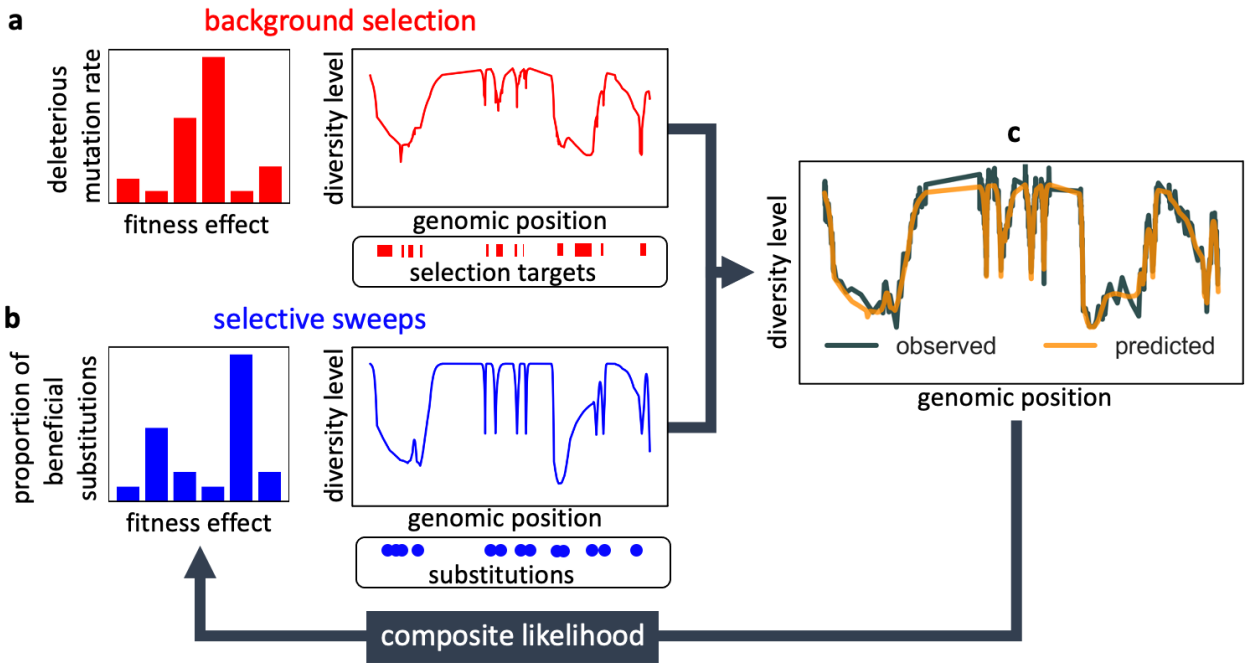

Figure 1. Modeling and inferring the effects of linked selection in humans. Given the targets of selection and corresponding selection parameters (a and b), we calculate the expected neutral diversity levels along the genome (c). We infer the selection parameters by maximizing their composite likelihood given observed diversity levels (c). Based on these parameter estimates, we calculate a map of the expected effects of selection on linked diversity levels.

To infer the effects of background selection and selective sweeps on human diversity levels, we analyze autosomal polymorphism data from 26 human populations, collected in Phase III of the 1000 Genomes Project[46]. Here, we focus on data from 108 genomes sampled from the Yoruba population (YRI), but in SOM Sections 7 and 9, we show similar results for other populations. To estimate diversity levels at neutral sites, we focus on non-genic autosomal sites that are the least conserved in a multiple sequence alignment of 25 supra-primates (see SOM Section 3.1). To control for variation in the mutation rates across neutral sites, we use estimates of the relative mutation rate for contiguous, non-overlapping blocks of 6000 putatively neutral sites, obtained from substitution rates in an eight-primate phylogeny (see SOM Section 3.3). To rely on a high-resolution genetic map built that is least

dependent on diversity levels, we use the map inferred from ancestry switches in African-Americans, which is highly correlated with other maps[47].

Background selection. We first focus on two of our best-fitting models of the effects of background selection (see below and SOM Section 4). In both cases, we take as targets of purifying selection the 6% of autosomal sites estimated as most likely to be under selection constraint. In one, we choose these sites using phastCons conservation scores obtained for a 99-vertebrate phylogeny that excludes humans[48]. In the other, we rely on Combined Annotation-Dependent Depletion (CADD) scores, which are based primarily on phylogenetic conservation (excluding humans) but also on information from functional genomic assays[49,50] (to avoid circularity, we use scores that were generated without using the McVicker et al. *B*-map[36] as input; see SOM Section 2.5). From these models, we obtain a map of predicted diversity levels (accounting for variation in mutation rates), which we can then compare to observed data (Figs. 2a and S24). As a measure of the precision of our

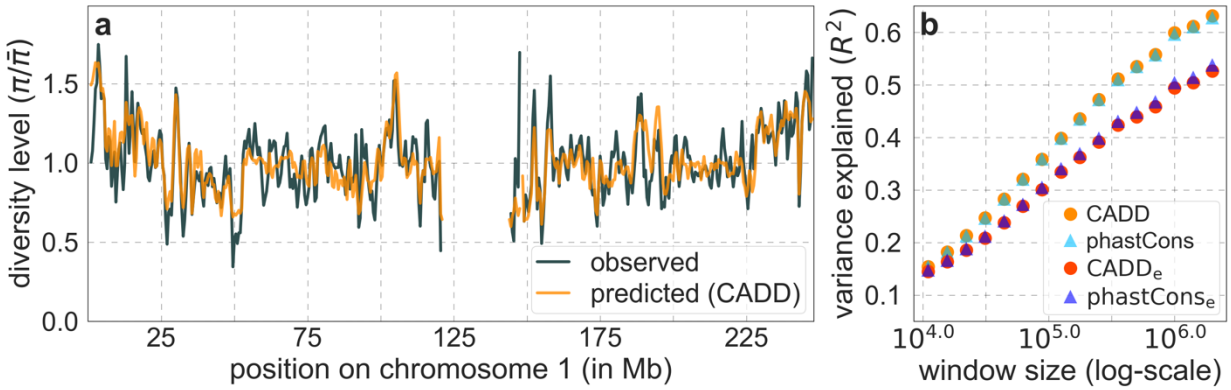

Figure 2. Comparison of diversity levels predicted by our best-fitting maps of background selection effects with observations. (a) Predicted and observed diversity levels along chromosome 1 in the YRI sample. Diversity levels are measured in 1 Mb windows, with a 0.5 Mb overlap, with the autosomal mean set to 1. (b) The proportion of variance in YRI diversity levels explained by background selection models on different spatial scales. Shown are the results for four choices of putative targets of selection: all sites with the highest 6% of CADD or phastCons scores (denoted CADD and phastCons, respectively) and the subset of these sites that are exonic (denoted CADD$_e$ and phastCons$_e$, respectively). See SOM Section 4 for similar graphs with other choices, and SOM Sections 7 and 9 for other populations. We show that our estimates are not inflated by over-fitting in SOM Section 6.

predictions, we consider the variance in diversity levels explained in non-overlapping autosomal windows (Fig. 2b). As can be seen, our predictions explain a large proportion of the variance across spatial scales: at the 1 Mb scale using CADD scores, they account for 59.9% of the variance in diversity levels, ~87% more than explained previously[36] (see SOM Section 4.6).

Selective sweeps. Next, we examine whether incorporating selective sweeps alongside background selection improves our predictions. Our inference should be able to tease apart the effects of selective sweeps, primarily because their effects, unlike those of background selection, should be centered around the locations of substitutions. Moreover, as noted, we expect to capture the effects of selective sweeps, be they hard, partial or soft[2,3,14-17,21], so long as they resulted in substitutions and substantially affected diversity levels (see SOM Section D in [37]). Indeed, previous work that applied a similar methodology to data from *Drosophila melanogaster* was able to identify distinct effects of background selection and sweeps[37]. To examine whether we can identify such effects in humans, we consider several choices of putatively selected substitutions along the human lineage, including any nonsynonymous substitutions or any nonsynonymous and non-coding substitutions in constrained regions, allowing each type to have its own selection parameters and considering different measures of constraint (see SOM Section 4.5). For no choice does incorporating sweeps improve our fit. In fact, in all cases our estimates of the proportion of substitutions resulting in sweeps with discernable effects on neutral diversity is approximately 0.

Moreover, in contrast to previous attempts[36,41], our model of background selection alone provides good quantitative fits to the diversity levels observed around different genomic features and in particular around nonsynonymous and synonymous substitutions (Figs. 3 and S48). Together, these results refute the hypothesis that reduced diversity levels around nonsynonymous substitutions in humans reflect "masked" effects of selective sweeps[42]; more generally, they indicate that selective sweeps resulting in substitutions had little effect on diversity levels in contemporary humans.

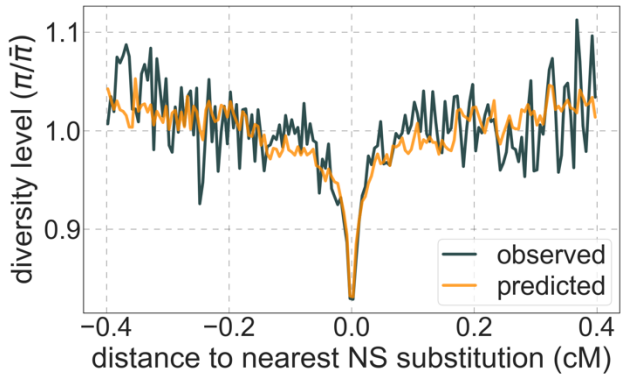

Figure 3. A background selection model predicts neutral diversity levels observed around human-specific nonsynonymous (NS) substitutions. Shown are the results for putatively neutral sites as a function of their genetic distance to the nearest nonsynonymous substitution (in 160 bins, each spanning 0.005cM). For observed values, we average diversity levels within each bin. For predicted values, we average predicted diversity levels for the best-fitting CADD-based model (using the mean mutation rate) within each bin, and then multiply by an estimate of relative mutation rates in that bin compared to the average (obtained from substitution data; see SOM Section 3.3). Both observed and predicted diversity levels are plotted relative to the autosomal mean. See Figs S48 and S50 for similar graphs for other genomic features and using data from other populations.

The lack of a footprint of sweeps in neutral diversity data does not imply that adaptation was rare in recent human evolution. Instead, much of it may have been driven by selection on genetically complex traits[40,41,51-53], i.e., traits with heritable variation arising from many segregating loci. The response to a shift in selection pressures on complex traits is expected to be highly polygenic, with short-term, rapid but tiny changes to allele frequencies at many loci that segregated before the shift, and a longer term, small excess in the fixation of alleles that change the trait in the direction favored by selection[54]. Importantly, polygenic adaptation is unlikely to result in sweeps and is likely to introduce only minor perturbations to the allele trajectories expected when selection pressures on traits remain constant[54], implying that its effects on neutral diversity levels should be minor[55]. In contrast, ongoing stabilizing selection on complex traits, i.e., selection acting to maintain traits near an optimal value, is thought to be common[56-59] and could have a substantial effect on neutral diversity levels[54]. Stabilizing selection induces purifying selection against minor alleles that affect complex traits[57,60,61], and purifying selection on these alleles could be a major source of background selection[54]. In other words, if much of the selection in

humans is driven by ongoing and changing selection pressures on complex traits, we may expect background selection to be the dominant mode of linked selection, as our results indicate.

The source of background selection. Focusing then on models of background selection alone, we ask which genomic annotations appear to be the sources of purifying selection. Previous work found selection on non-exonic regions to contribute little, to the extent that removing conserved non-exonic sites from a model of background selection had little effect on predicted diversity levels[36]. In contrast, when we include only conserved exonic regions in our inference, our predictive ability is considerably diminished (Fig. 2b).

Moreover, in models that include separate selection parameters for conserved exonic and non-exonic regions, purifying selection on non-exonic regions accounts for most of the reduction in linked neutral diversity (SOM Section 4.3). Our estimates suggest that ~80% of deleterious mutations affecting neutral diversity occur in non-exonic regions (e.g., in the model with the top 6% of phastCons scores, ~84% of selected sites and ~76% of mutations are non-exonic; with the top 6% of CADD scores, ~83% of selected sites and ~85% of deleterious mutations are non-exonic; see SOM Sections 4.3 and 4.6). Our estimates of the average strength of selection differ between exonic and non-exonic regions, but because the total reduction in diversity levels caused by background selection is fairly insensitive to the strength of selection (with the reduction being more localized for weakly selected mutations than for strong ones), the proportions of deleterious mutations that occur in these regions approximate their relative effects on neutral diversity levels[7] (see SOM Sections 4.3, 4.4 and 4.6). Thus, our estimates suggest that purifying selection on non-exonic regions accounts for ~80% of the reduction in linked neutral diversity. Moreover, including separate selection parameters for conserved exonic and non-exonic regions does not appear to improve our predictions (SOM Section 4.3 and Fig. S19).

Incorporating additional functional genomic information also does little to improve our predictions (SOM Sections 4.2 and 4.4). Notably, including separate selection parameters for coding regions and for each of the Encyclopedia of DNA Elements (ENCODE) classes of

candidate cis-regulatory elements (cCRE)[62], considerably diminishes our predictive ability (SOM Section 4.4). Indeed, using CADD scores[49,50], which augment information on phylogenetic conservation with functional genomic information, offers little improvement over relying on conservation alone (e.g., explaining 59.9% compared to 59.7% of the variance in diversity levels in 1Mb windows, a difference that is not statistically significant; SOM Section 6). Thus, at present, functional annotations that do not incorporate phylogenetic conservation appear to provide poorer predictions of the effects of linked selection and those that do offer little improvement over conservation alone (see SOM Sections 4.1- 4).

In turn, our predictions are fairly insensitive to the phylogenetic depth of the alignments used to infer conservation levels, although we appear to do slightly better using a 99-vertebrate alignment (excluding humans) compared to its monophyletic subsets (e.g., Figs. S14c and S33a and SOM Section 6.2). Our best-fitting models based on a variety of metrics are obtained using 5-7% of sites with the top CADD or phastCons scores as selection targets (Fig. S16 and S26), a percentage in good accordance with more direct estimates of the proportion of the human genome subject to functional constraint[63-66].

The deleterious mutation rate. Reassuringly, the deleterious mutation rates that we estimate for our best-fitting models are plausible (Fig. 4). Current estimates of the average mutation rate per site per generation in humans, including point mutations[38,39], indels[38], mobile element insertions[67], and structural mutations[68,69] lie in the range of $1.29 \times 10^{-8} - 1.38 \times 10^{-8}$ per base pair per generation (SOM Section 5). Further accounting for the length of deletions[38] (as a deletion that starts at a neutral site and includes selected sites should be included but a deletion that affects several selected sites should only be counted once), we find that the upper bound on the rate of deleterious mutations falls in the range of $1.29 \times 10^{-8} - 1.51 \times 10^{-8}$ per base pair per generation (SOM Section 5). The estimates for all of our best-fitting models fall well below this bound (Fig. 4). This is what we would expect, because not every mutation at putatively selected sites will be deleterious: some sites are misclassified as constrained and some mutations at selected sites are selectively neutral.

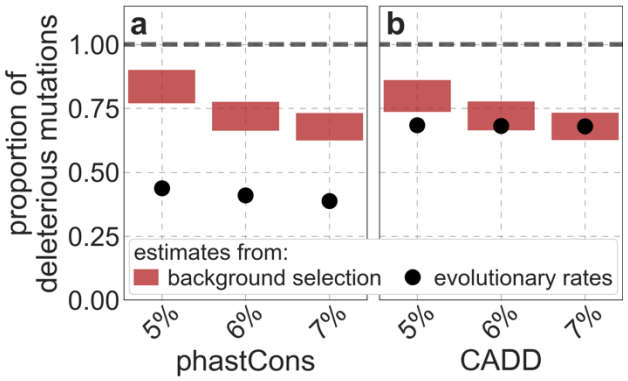

Figure 4. Estimates of the proportion of deleterious mutations at selected sites for our best-fitting background selection models. Shown are the results for models using 5-7% of sites with the highest phastCons scores (a) and CADD scores (b) as selection targets. For estimates based on fitting the models, we divide our estimates of the deleterious mutation rate per selected site by the estimate of the total mutation rate per site, where the ranges correspond to the range of estimates of the total rate, i.e., $1.29 \times 10^{-8} - 1.51 \times 10^{-8}$ per site per generation (SOM Section 5.1). For estimates based on evolutionary rates, we divide the estimated rates per site on the human lineage (from the common ancestor of humans and chimpanzees) at selection targets by the rates at matched sets of putatively neutral sites (see text and SOM Section 5.2 for details).

To test whether the estimated proportion of deleterious mutations is plausible, we compare it with an independent estimate based on the relative reduction in evolutionary rates at constrained vs. putatively neutral sites along the human lineage (these sets of sites were identified from an alignment that excludes humans; SOM Sections 3.1, 4.1 and 4.4). Because deleterious mutations at selected sites rarely fix in the population whereas neutral mutations fix with a much higher probability (which is the same at conserved and neutral sites), this relative reduction allows us to estimate the proportion of deleterious mutations[70]. To this end, we were also mindful of potential differences in point mutation rates between conserved and neutral sites, calculating rates separately for sites inferred to be GC versus AT in the human-chimpanzee ancestor and checking that other aspects known to affect mutation or substitution rates (i.e., the triplet context, CpG mutations and biased gene conversion) have little effect on our estimates (SOM Section 5). By this approach, our estimates are closer to (and even overlap) those obtained from fitting models of background selection based on CADD scores compared to those based on phastCons scores (Fig. 4); this is expected, given that CADD scores are much better than phastCons scores at

identifying constraint on a single site resolution[49,50], which markedly influences evolutionary rates at putatively selected sites (but not the predictions of background selection effects). We expect the two estimates to be similar but not identical, both because weak selection has a larger effect on evolutionary rates than on linked diversity levels[11-13,22,25] and because estimates based on the effects of background selection may absorb the deleterious mutation rate at selected sites that were not included in our sets but are closely linked to sites in them (SOM Section 5). In summary, given the fit to data and plausible estimates of the deleterious rates, it is natural to interpret our maps as reflecting the effects of background selection, i.e., as maps of B (defined as the ratio of expected diversity levels with background selection, π, and in its absence, π_0)[5].

Background selection in autosomes. Our maps are also well calibrated (Fig. 5). When we stratify diversity levels at putatively neutral sites by our predictions, predicted and observed diversity levels are similar throughout nearly the entire range of predicted values (e.g., $R^2 = 0.96$ when sites are in predicted percentile bins). One exception is for ~4% of sites in which background selection is predicted to be the strongest (i.e., with the lowest B), where our predictions are imprecise. This behavior is due to a technical approximation we employ in fitting the models (see SOM Section 1.5). The other exception is for ~2% of sites in which background selection is predicted to be the weakest (i.e., with B near 1), where observed diversity levels are markedly greater than expected. We observe similar behavior in all the human populations examined (Fig. S51), and we cannot fully explain it by known mutational and recombination effects (e.g., of base composition and biased gene conversion; SOM Section 8). This behavior could reflect ancient introgression of archaic human DNA into ancestors of contemporary humans (SOM Section 8.3), indicated also in other population genetic signatures[71-81]. Such introgressed regions are expected to increase genetic diversity and persist the longest in regions with low functional density and high recombination, corresponding to weak background selection effects[82-84].

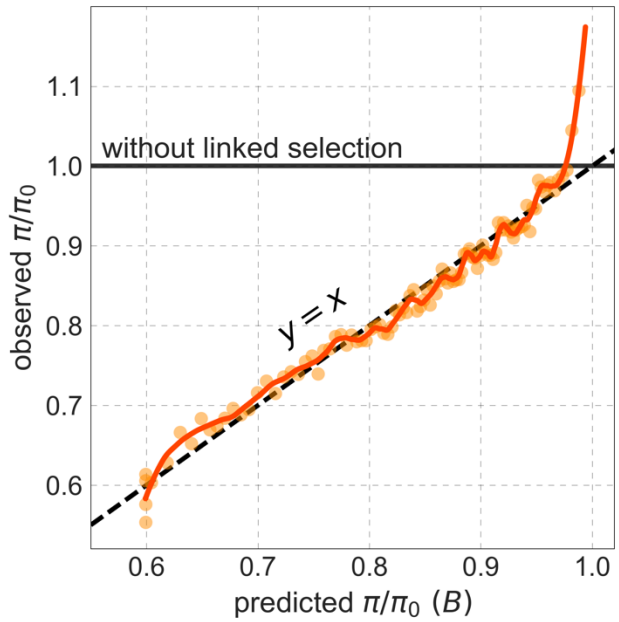

Figure 5. Observed vs. predicted neutral diversity levels across the autosomes. Shown are the results for the best-fitting CADD-based model. Light orange scatter plot: we divide putatively neutral sites into 100 equally sized bins based on the predicted *B*. For predicted values (x-axis), we average the predicted *B* in each bin. For observed values (y-axis), we divide the average diversity level by the estimate of the average relative mutation rate (obtained from substitution data; see SOM Section 3.3) in each bin, and normalize by the autosomal average of π_0 (estimated from fitting the model; see SOM Section 1.1). Dark orange curve: the LOESS fit for a similarly defined scatter plot but with 2000 rather than 100 bins (with span=0.1). For similar graphs corresponding to other models and using data from other populations, see SOM Sections 4 and 9, respectively.

Setting these outlier regions aside, we can use the maps to characterize the distribution of background selection effects in human autosomes. We note that background selection effects that are not captured by our models would cause us to underestimate the range and extent of background selection effects[37]. We find that diversity levels throughout almost all of the autosomes are affected by background selection, with a ~37% reduction in the 10% most affected sites, a non-zero (~2.1%) reduction even in the 10% least affected (after excluding outliers in the top 2% of bins; see Fig. 5), and a mean reduction of ~17%. These conclusions are robust across our best-fitting maps and populations (SOM Section 4 and Figs. S35 and S51). An important implication is that our maps of the effects of background selection provide a more accurate null model than currently used for other population

genetic inferences that rely on diversity levels, notably inferences about demographic history[85-89].

Conclusion. Our results indicate that background selection is the dominant mode of linked selection in human autosomes and the major determinant of neutral diversity levels on the Mb scale. They further reveal that background selection effects arise primarily from purifying selection at non-coding regions of the genome. Non-coding regions are known to exhibit substantial functional turnover on evolutionary timescales[63-66], and yet we find phylogenetic conservation to be the best predictor of selected regions, and that, at least at present, augmenting measures of conservation with functional genomic information in humans offers little improvement. It therefore remains unclear how much our maps can still be improved. Even without these potential refinements, our findings demonstrate that a simple model of background selection, conceived nearly three decades ago[5], is able to provide a reliable quantitative prediction of genetic diversity levels throughout human autosomes.

Acknowledgements. We thank M. Przeworski for helpful discussions throughout this work. We also thank Ipsita Agarwal, Eduardo Amorim, Peter Andolfatto, Ian Mathieson, Priya Moorjani, Itsik Pe'er, Joe Pickrell and Jonathan Pritchard for helpful discussions. We thank Yun Song for sharing unpublished results, and Martin Kircher, Lusiné Nazaretyan, Philipp Rentzsch and Max Schubach from the Kircher lab for generating CADD scores that were tailored for our purposes. We thank Ipsita Agarwal and Molly Przeworski for comments on the manuscript. This work was funded by NIH grant GM115889 to GS and NIH training grant T32GM008798 to DAM.

References

1. Hill, W.G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269-94 (1966).
2. Maynard Smith, J.M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23-35 (1974).
3. Kaplan, N.L., Hudson, R.R. & Langley, C.H. The "hitchhiking effect" revisited. *Genetics* **123**, 887-99 (1989).

4. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**, 519-20 (1992).
5. Charlesworth, B., Morgan, M.T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-303 (1993).
6. Wiehe, T.H. & Stephan, W. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. *Mol. Biol. Evol.* **10**, 842-54 (1993).
7. Hudson, R.R. How can the low levels of DNA sequence variation in regions of the drosophila genome with low recombination rates be explained? *Proc. Natl. Acad. Sci. USA* **91**, 6815-8 (1994).
8. Hudson, R.R. & Kaplan, N.L. Deleterious background selection with recombination. *Genetics* **141**, 1605-17 (1995).
9. Nordborg, M., Charlesworth, B. & Charlesworth, D. The effect of recombination on background selection. *Genet. Res.* **67**, 159-74 (1996).
10. Kim, Y. & Stephan, W. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **155**, 1415-27 (2000).
11. McVean, G.A. & Charlesworth, B. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**, 929-44 (2000).
12. Comeron, J.M. & Kreitman, M. Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**, 389-410 (2002).
13. Gordo, I., Navarro, A. & Charlesworth, B. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161**, 835-48 (2002).
14. Hermisson, J. & Pennings, P.S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335-52 (2005).
15. Przeworski, M., Coop, G. & Wall, J.D. The signature of positive selection on standing genetic variation. *Evolution Int. J. Org. Evolution* **59**, 2312-23 (2005).
16. Pennings, P.S. & Hermisson, J. Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* **23**, 1076-84 (2006).
17. Pennings, P.S. & Hermisson, J. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.* **2**, e186 (2006).
18. Wright, S.I. & Andolfatto, P. The impact of natural selection on the genome: emerging patterns in Drosophila and Arabidopsis. *Annu. Rev. Ecol. Syst.* **39**, 193-213 (2008).
19. Sella, G., Petrov, D.A., Przeworski, M. & Andolfatto, P. Pervasive natural selection in the Drosophila genome? *PLoS Genet.* **5**, e1000495 (2009).
20. Stephan, W. Genetic hitchhiking versus background selection: the controversy and its implications. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365**, 1245-53 (2010).
21. Coop, G. & Ralph, P. Patterns of neutral diversity under general models of selective sweeps. *Genetics* **192**, 205-24 (2012).
22. Charlesworth, B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. *J. Hered.* **104**, 161-71 (2013).
23. Cutter, A.D. & Payseur, B.A. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**, 262-74 (2013).
24. Nicolaisen, L.E. & Desai, M.M. Distortions in Genealogies due to Purifying Selection and Recombination. *Genetics* **195**, 221-30 (2013).

25. Good, B.H., Walczak, A.M., Neher, R.A. & Desai, M.M. Genetic diversity in the interference selection limit. *PLoS Genet.* **10**, e1004222 (2014).
26. Berg, J.J. & Coop, G. A Coalescent Model for a Sweep of a Unique Standing Variant. *Genetics* **201**, 707-25 (2015).
27. Cvijović, I., Good, B.H. & Desai, M.M. The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics* **209**, 1235-78 (2018).
28. Nachman, M.W. Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics* **147**, 1303-16 (1997).
29. Wright, S.I. *et al.* Testing for effects of recombination rate on nucleotide diversity in natural populations of Arabidopsis lyrata. *Genetics* **174**, 1421-30 (2006).
30. Payseur, B.A. & Nachman, M.W. Gene Density and Human Nucleotide Polymorphism. *Mol. Biol. Evol.* **19**, 336-40 (2002).
31. Nordborg, M. *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.* **3**, e196 (2005).
32. Andolfatto, P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. *Genome Res.* **17**, 1755-62 (2007).
33. Begun, D.J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.* **5**, e310 (2007).
34. Macpherson, J.M., Sella, G., Davis, J.C. & Petrov, D.A. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. *Genetics* **177**, 2083-99 (2007).
35. Cai, J.J., Macpherson, J.M., Sella, G. & Petrov, D.A. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* **5**, e1000336 (2009).
36. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
37. Elyashiv, E. *et al.* A Genomic Map of the Effects of Linked Selection in Drosophila. *PLoS Genet.* **12**, e1006130 (2016).
38. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLoS Genet.* **12**, e1006315 (2016).
39. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
40. Coop, G. *et al.* The Role of Geography in Human Adaptation. *Plos Genetics* **5**(2009).
41. Hernandez, R.D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920-4 (2011).
42. Enard, D., Messer, P.W. & Petrov, D.A. Genome-wide signals of positive selection in human evolution. *Genome Res.* **24**, 885-95 (2014).
43. Nordborg, M., Charlesworth, B. & Charlesworth, D. The effect of recombination on background selection. *Genet. Res.* **67**, 159-74 (1996).
44. Barton, N.H. The effect of hitch-hiking on neutral genealogies. *Genet. Res.* **72**, 123-33 (1998).
45. Gillespie, J.H. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155**, 909-19 (2000).
46. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
47. Hinch, A.G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-5 (2011).

48. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-50 (2005).
49. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-94 (2019).
50. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310-5 (2014).
51. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**, 1411-23 (2010).
52. Pritchard, J.K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nat. Rev. Genet.* **11**, 665-7 (2010).
53. Sella, G. & Barton, N.H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu. Rev. Genomics. Hum. Genet.* (2019).
54. Hayward, L.K. & Sella, G. Polygenic adaptation after a sudden change in environment. Preprint at https://doi.org/10.1101/792952 (2021)
55. Barton, N.H. Genetic hitchhiking. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **355**, 1553-62 (2000).
56. Wright, S. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. *J. Genet.* **30**, 243-56 (1935).
57. Robertson, A. A mathematical model of the culling process in dairy cattle. *Animal Science* **8**, 95-108 (1966).
58. Sella, G. & Barton, N.H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu. Rev. Genomics. Hum. Genet.* **20**, 461-93 (2019).
59. Walsh, B. & Lynch, M. *Evolution and Selection of Quantitative Traits*, (Oxford University Presss, 2018).
60. Simons, Y.B., Bullaughey, K., Hudson, R.R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, e2002985 (2018).
61. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97-159 (1931).
62. Consortium, E.P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
63. Ward, L.D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675-8 (2012).
64. Rands, C.M., Meader, S., Ponting, C.P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet.* **10**, e1004525 (2014).
65. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. USA* **111**, 6131-8 (2014).
66. Ponting, C.P. & Hardison, R.C. What fraction of the human genome is functional? *Genome Res.* **21**, 1769-76 (2011).
67. Gardner, E.J. *et al.* Contribution of retrotransposition to developmental disorders. *Nat. Commun.* **10**, 4630 (2019).
68. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).

69. Belyeu, J.R. *et al.* De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597-607 (2021).
70. Kimura, M. & Crow, J.F. The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics* **49**, 725-38 (1964).
71. Vernot, B. & Akey, J.M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017-21 (2014).
72. Steinrucken, M., Spence, J.P., Kamm, J.A., Wieczorek, E. & Song, Y.S. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Mol. Ecol.* **27**, 3873-88 (2018).
73. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14**, e1007641 (2018).
74. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-7 (2014).
75. Durvasula, A. & Sankararaman, S. Recovering signals of ghost archaic introgression in African populations. *Sci. Adv.* **6**, eaax5097 (2020).
76. Wall, J.D. & Hammer, M.F. Archaic admixture in the human genome. *Curr. Opin. Genet. Dev.* **16**, 606-10 (2006).
77. Wall, J.D., Lohmueller, K.E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823-7 (2009).
78. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-60 (2010).
79. Plagnol, V. & Wall, J.D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
80. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-22 (2010).
81. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359-71 (2015).
82. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection against Neanderthal Introgression. *PLoS Genet.* **12**, e1006340 (2016).
83. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881-91 (2016).
84. Schumer, M. *et al.* Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**, 656-60 (2018).
85. Terhorst, J., Kamm, J.A. & Song, Y.S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303-9 (2017).
86. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-6 (2011).
87. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919-25 (2014).
88. Steinrücken, M., Kamm, J., Spence, J.P. & Song, Y.S. Inference of complex population histories using whole-genome sequences from multiple populations. *Proc. Natl. Acad. Sci. USA* **116**, 17115-20 (2019).

89. Spence, J.P., Steinrücken, M., Terhorst, J. & Song, Y.S. Inference of population history using coalescent HMMs: review and outlook. *Curr. Opin. Genet. Dev.* **53**, 70-76 (2018).