Published in final edited form as:
Science. 2011 February 18; 331(6019): 920–924. doi:10.1126/science.1198878.

Classic selective sweeps were rare in recent human evolution

Ryan D. Hernandez[1,2], **Joanna L. Kelley**[1], **Eyal Elyashiv**[3], **S. Cord Melton**[1], **Adam Auton**[4], **Gil McVean**[4,5], **Guy Sella**[3,8], **Molly Przeworski**[1,6,7,8,9], and **1000 Genomes Project**

[1]Dept. of Human Genetics, University of Chicago, IL, USA

[3]Dept. of Ecology, Systematics and Evolution, Hebrew University, Israel

[4]Wellcome Trust Centre for Human Genetics, University of Oxford, UK

[5]Dept. of Statistics, University of Oxford, UK

[6]Dept. of Ecology and Evolution, University of Chicago, IL, USA

[7]Howard Hughes Medical Institute, University of Chicago, IL, USA

Abstract

Efforts to identify the genetic basis of human adaptations from polymorphism data have sought footprints of “classic selective sweeps”. Yet it remains unknown whether this form of natural selection was common in our evolution. We examined the evidence for classic sweeps in resequencing data from 179 human genomes. As expected under a recurrent sweep model, diversity levels decrease near exons and conserved non-coding regions. In contrast to expectation, however, the trough in diversity around human-specific amino acid substitutions is no more pronounced than around synonymous substitutions. Moreover, relative to the genome background, amino acid and putative regulatory sites are not significantly enriched for alleles that are highly differentiated between populations. These findings indicate that classic sweeps were not a dominant mode of adaptation over the past ~250,000 years.

Humans have experienced myriad adaptations since the common ancestor with chimpanzees and more recently have adapted to a wide range of environments. Efforts to infer the molecular basis of these adaptations from polymorphism data have largely been guided by the “classic selective sweep” model, in which a new, strongly beneficial mutation increases in frequency to fixation in the population (reviewed in (1, 2)). In this scenario, the allele ascends rapidly enough in frequency for there to be little opportunity for recombination to uncouple it from its genetic background, such that its rise sweeps out variation at linked sites, reducing linked neutral diversity in the population and distorting allele frequencies and patterns of linkage disequilibrium (3). In humans, the effects of sweeps are expected to persist for approximately 10,000 generations or about 250,000 years (4).

Identifying the footprint of a sweep against a noisy genomic background is challenging, because patterns of genetic variation reflect the effects of multiple modes of natural selection as well as of demographic history, mutation and recombination. To date, applications of statistical tests based on the sweep model have led to the identification of over 2000 genes as potential targets of positive selection in the human genome (2) and to the suggestion that diversity patterns in ~10% of the human genome have been affected by linkage to recent sweeps (e.g., (5)). The list of functionally characterized cases of genetic

[9]To whom correspondence should be addressed: mfp@uchicago.edu.
[2]Current affiliation: Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, CA, USA
[8]Joint senior authors

adaptations is short however, and the false discovery rate of selection scans is potentially high (6). Thus, it remains unknown if the well-documented cases are typical of human adaptations or if they represent rare instances where the genetic architecture of the adaptation was conducive to classic sweeps (7, 8), with most adaptations occurring by other modes (e.g., polygenic selection and selection on standing variation).

Two main lines of evidence have been advanced in support of the hypothesis that classic selective sweeps were common. First, regions of low recombination, in which a single sweep should have a larger span, exhibit lower diversity (after correcting for variation in mutation rates) as compared with regions of high recombination (9–11). Regions of low recombination also show greater differentiation between populations (12), as expected from local adaptation or, for some parameters, from the fixation of globally advantageous alleles (13). Second, under the sensible assumption that amino acid and conserved non-coding sites are enriched among targets of adaptation, one would expect that the signal of selection would be most clearly visible at or around such sites (e.g., (10, 14)). Consistent with this expectation, diversity levels decrease with the number of human-specific substitutions at amino acid or conserved non-coding sites (in 200–600 kb windows) (10) and genic regions show an enrichment of alleles that are highly differentiated between populations compared to non-genic regions (15, 16). These patterns are informative, but are only indirectly related to theoretical predictions. Moreover, some - possibly all - of these patterns may instead result from purifying selection acting on deleterious mutations at linked sites (“background selection”)(9–11, 16–18).

To evaluate the importance of classic sweeps in shaping human diversity, we analyzed resequencing data for 179 human genomes from four populations, collected as part of the low coverage pilot for the 1000 Genomes Project (19). These data overcome ascertainment biases arising in the study of genotyping data, with ~99% power to detect variants with a population frequency above 10% for 86% of the euchromatic genome(19).

We examined the extent to which selection impacts diversity levels at linked sites by calculating the average diversity as a function of genetic distance from the nearest exons, collating all exons across the genome (Fig. S1). To estimate neutral diversity levels, we focused only on non-conserved, non-coding and four-fold degenerate sites (11). We divided diversity by divergence to rhesus macaque (to which the contribution of ancestral polymorphism is minor (11)), in order to correct for systematic variation in the mutation rate. Our estimate of relative diversity appears little affected by the low fold coverage of individuals or variation in sequencing depth (Figs S2C–E). Scaled diversity levels are lowest near exons (Figs. 1 and S3), recovering half the drop by 0.03–0.04 cM, depending on the population and 80% by 0.07–0.1 cM (see (20)). Given that diversity is scaled by divergence, the trough in scaled diversity around exons does not reflect systematic variation in mutation rates as a function of the distance from exons, strong purifying selection on the sites themselves (which would decrease both diversity and divergence) or weak selection near exons (which should inflate, not decrease, diversity levels divided by divergence). Rather, the trough provides evidence for the effects of directional selection at linked sites, extending over a hundred kilobases.

This pattern is even more pronounced on the X chomosome, where is trough is deeper and wider, recovering diversity over twice the genetic distance (Figs. 1, S3). The greater footprint of linked selection on the X leads to a smaller ratio of X to autosome scaled diversity near exons than farther away, potentially confounding demographic analysis (21) (Fig. S4).

A similar effect is seen around conserved non-coding regions(CNCs), but the trough is more diffuse (Figs. 1, S3). Since CNCs tend to be linked to exons (the median distance of a CNC to the nearest exon is 0.08 cM), the trough around CNCs could be a byproduct of the effects of selection on exons (see below); alternatively, it could reflect less widespread selection on mutations in CNCs compared to exons (11, 22).

If the trough in scaled diversity results from classic sweeps at linked sites, it should be deepest around those changes most likely to have functional consequences, i.e., within exons, around amino acid substitutions. We tested this prediction by considering the average scaled diversity around human-specific amino acid fixations and, as a control for other evolutionary forces, around synonymous substitutions. Our rationale was as follows: human and chimpanzee species split approximately 5 Mya (e.g., (23)) so, assuming a constant rate of substitution, approximately 5% of human-specific substitutions could have left a detectable sweep in their wake (i.e., have occurred in the past 250,000 years). Thus, if a substantial fraction of amino acid changes are the result of classic sweeps, average diversity should be decreased around amino acid substitutions compared to synonymous substitutions. In the fly *Drosophila simulans,* diversity levels are indeed significantly lower and suggest that ~13% of amino acid substitutions involved classic sweeps (24). In contrast, human diversity levels around amino acid substitutions are not lower than around synonymous substitutions (Fig. 2; $p = 0.90$ for a window of size 0.02 cM around the focal substitution). This conclusion is robust to alternative approaches for inferring substitutions or estimating divergence and to the choice of genetic map (Fig. S5). The similar troughs indicate either that amino acid and synonymous mutations (including four-fold degenerate mutations; Fig. 2) experienced recurrent classic selective sweeps of similar intensities and rates, or more plausibly, that few amino acid substitutions resulted from classic sweeps.

Simulations suggest that even if only 10% of human-specific amino acid substitutions were strongly favored or if 25% of amino acid fixations were favored with weak effects, there should be a significant decrease in the diversity levels relative to what would be expected if all fixations were neutral (Fig. 3A; (20)). These simulations mimic the data structure, but do not fully capture the clustering of substitutions in the genome (Fig. S6). Because amino acid substitutions are more clustered with one another than with synonymous substitutions (Fig. S6A), this omission is conservative, leading to an *under*-estimate of the power to detect the effects of classic sweeps (Fig. S7, (20)). Thus, our finding strongly constrains the maximal fraction of protein changes that could have resulted from classic sweeps in the past 250,000 years.

The troughs in diversity around both synonymous and amino acid fixations could instead be due to strong purifying selection at linked sites. Indeed, we found that under a model of background selection (17), the expected troughs are of similar depths to the observed ones, with lower diversity predicted around four-fold degenerate substitutions than amino acid substitutions, as observed (Fig. 3B, (20)). Interestingly, even though the model assumes extremely weak purifying selection in CNCs, it predicts a trough around them as well (Fig. S8), indicating that the observed decrease in diversity around CNCs may primarily reflect selection on linked exons.

The prevalence of classic sweeps can also be evaluated by considering the genetic differentiation between the three population samples, whose ancestors occupied a range of environments. The two Eurasian populations (a population of European ancestry (CEU) and the combined Chinese and Japanese (CHB+JPT)) are thought to have split from the Yoruba (YRI) over 100 Kya, and CEU and CHB+JPT approximately 23 Kya (e.g., (25)). Given this time frame, local adaptation involving classic sweeps would have led to fixed differences or extreme differences in allele frequencies between YRI and CEU/CHB+JPT (and possibly

between CEU and CHB+JPT) at targets of selection (cf. (16)). Consistent with this expectation, there is an enrichment of highly differentiated single nucleotide polymorphisms (SNPs) between population pairs in genic compared to non-genic regions (Fig. 4A) (15, 16). However, the enrichments can also be explained by a 10–15% decrease in the effective population size near exons due to background selection, i.e., a trough similar to what is seen in Fig. 1 (16). In turn, CNCs are not significantly enriched for highly differentiated SNPs relative to non-conserved non-coding regions (Figs. 4B, S9B).

Although the enrichment of genic SNPs could in principle result from purifying selection alone, there are well-documented examples of adaptations among the most highly differentiated SNPs, notably in genes involved in pigmentation or infectious disease susceptibility (19)-in other words, there are at least a handful of loci that conform to the sweep model. To ask whether these cases represent the tip of the iceberg of sweeps yet to be discovered, we tested for an enrichment of highly differentiated alleles that cause amino acid changes or that lie in putative regulatory regions, at which *a priori* one would expect the highest fraction of changes to have phenotypic consequences and hence to be possible targets of sweeps. Since CNCs may be unusual regulatory elements, we also considered SNPs in UTRs or 1 kb upstream of the TSS, annotations in which over 10% of substitutions were estimated to be beneficial (22)and which are most strongly enriched for eQTLs in cell lines (26). These annotations all fall within genic regions, so a test of enrichment against the genomic background is confounded by background selection or other modes of selection that increase population differentiation at genic sites. Nonetheless, in comparisons between three human populations, enrichments of highly differentiated alleles at non-synonymous sites, 5′ and 3′ UTRs and within 1kb upstream of the TSS are either not significant when tested against the genomic background or only marginally significant (Figs. 4C–F, S10). This finding reflects the small numbers of cases of highly differentiated alleles (Fig. S11) and underscores how few local adaptations resulted from the extreme changes in allele frequencies between populations expected from classic sweeps. In particular, there are only four fixed amino acid differences between YRI and CEU, suggesting a rate of classic sweeps far below 10% since the two populations split (see (20)).

Moreover, extreme differentiation is also expected under a different model of sweeps, in which the beneficial allele was not a new mutation, but was already segregating in the population. While tests based on the frequency spectrum or the decay of linkage disequilibrium have low power to detect this mode of selection (e.g., (6, 27)), measures of differentiation should have substantial power so long as there was little or no gene flow between the populations and the allele was at low frequency when first favored (28) (as is likely to be the case for both neutral and previously deleterious alleles). Intriguingly, alleles with the largest differences in frequencies between populations, which should be most enriched for targets of selection(6), often segregate in both populations (Fig.S12) and tend to lie on a shorter haplotype than expected from a classic sweep (16), consistent with selection on standing variation rather than ongoing classic sweeps.

In summary, patterns of diversity around genic substitutions and of highly differentiated alleles are inconsistent with the expectation for frequent classic sweeps, but could result, at least in part, from background selection. Thus, while some substitutions in proteins and regulatory positions undoubtedly involved classic sweeps, they were too infrequent within the last 250,000 years to have had discernible effects on genomic diversity.

This conclusion does not imply that humans have experienced few phenotypic adaptations, or that adaptations have not shaped genomic patterns of diversity. Comparisons of diversity and divergence levels at putatively functional versus neutral sites, for example, suggest that 10–15% (and possibly as many as 40% (29)) of amino acid differences between humans and

chimpanzees were adaptive (e.g., (30)) as were 5% of substitutions in conserved non-coding regions (22, 29) and ~20% in UTRs (22). Given the paucity of classic sweeps revealed by our findings, an excess of functional divergence would point to the importance of other modes of adaptation. One way to categorize modes of adaptation is in terms of their effect on the allele frequencies at sites that affect the beneficial phenotype. In this view, classic sweeps bring new alleles to fixation; selection on standing variation or on multiple beneficial alleles brings rare or intermediate frequency alleles to fixation; and other forms of adaptation, such as selection on polygenic traits, increase or decrease allele frequencies to a lesser extent. Such changes in allele frequencies can decrease variation at closely linked sites - to a lesser extent than in a full sweep - and might therefore contribute to a reduction in diversity near functional elements (31), as well as to excess divergence. Alternatives to classic sweeps are likely for parameters applicable to human populations (7, 32); in particular, many phenotypes of interest are quantitative, and plausibly result from selection at many loci of small effect (8).

An important implication is that in the search for targets of human adaptation, a change in focus is warranted. To date, selection scans have relied almost entirely on the sweep model, either explicitly by considering strict neutrality as the null hypothesis and a classic sweep as the alternative or implicitly, by ranking regions by a statistic thought to be sensitive to classic sweeps and focusing on the tails of the empirical distribution. It appears that few adaptations in humans took the form that these approaches are designed to detect, suggesting that low hanging fruits accessible by existing approaches may be largely depleted. Conversely, the more common modes of adaptation likely remain undetected. Thus, in order to dissect the genetic basis of human adaptations and assess what fraction of the genome was affected by positive selection, we need new tests to detect other modes of selection, such as comparisons between closely related populations that have adapted to drastically different environments (e.g., (33)) or methods that consider loci that contribute to the same phenotype jointly (e.g., (34)). Moreover, if alleles that contribute to recent adaptations are often polymorphic within a population, genome-wide association studies should be highly informative.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank G. Coop, T. Long, G. McVicker, J. Pickrell, J. Pritchard and K. Thornton for helpful discussions, and G. Coop, A. Di Rienzo, J. Pritchard for comments. Supported by an NSF minority postdoctoral fellowship to RDH, NRSA postdoctoral fellowship GM087069 to JLK, WT086084MA to GM, ISF 1492/10 and NIH GM083228 to GS, as well as NIH GM20373 and GM72861to MP. M. P. is a Howard Hughes Medical Institute Early Career Scientist.

References

1. Sabeti PC, et al. Science. Jun 16.2006 312:1614. [PubMed: 16778047]
2. Akey JM. Genome Res. May.2009 19:711. [PubMed: 19411596]
3. Maynard Smith JM, Haigh J. Genet Res. Feb.1974 23:23. [PubMed: 4407212]
4. Przeworski M. Genetics. Mar.2002 160:1179. [PubMed: 11901132]
5. Williamson SH, et al. PLoS Genet. Jun.2007 3:e90. [PubMed: 17542651]
6. Teshima K, Coop G, Przeworski M. Genome Res. 2006; 16:702. [PubMed: 16687733]
7. Hermisson J, Pennings PS. Genetics. Apr.2005 169:2335. [PubMed: 15716498]
8. Pritchard JK, Pickrell JK, Coop G. Curr Biol. Feb 23.2010 20:R208. [PubMed: 20178769]

9. Hellmann I, et al. Genome Res. Jul.2008 18:1020. [PubMed: 18411405]
10. Cai JJ, Macpherson JM, Sella G, Petrov DA. PLoS Genet. Jan.2009 5:e1000336. [PubMed: 19148272]
11. McVicker G, Gordon D, Davis C, Green P. PLoS Genet. May.2009 5:e1000471. [PubMed: 19424416]
12. Keinan A, Reich D. PLoS Genet. 2010; 6:e1000886. [PubMed: 20361044]
13. Santiago E, Caballero A. Genetics. Jan.2005 169:475. [PubMed: 15489530]
14. Sabeti PC, et al. Nature. Oct 18.2007 449:913. [PubMed: 17943131]
15. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Nat Genet. Mar.2008 40:340. [PubMed: 18246066]
16. Coop G, et al. PLoS Genet. Jun.2009 5:e1000500. [PubMed: 19503611]
17. Charlesworth B, Morgan MT, Charlesworth D. Genetics. Aug.1993 134:1289. [PubMed: 8375663]
18. Charlesworth B, Nordborg M, Charlesworth D. Genet Res. Oct.1997 70:155. [PubMed: 9449192]
19. Durbin RM, et al. Nature. Oct 28.2010 467:1061. [PubMed: 20981092]
20. Supporting Online Material.
21. Hammer MF, et al. Nat Genet. Aug 29.2010
22. Torgerson DG, et al. PLoS Genet. Aug.2009 5:e1000592. [PubMed: 19662163]
23. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. Nature. Jun 29.2006 441:1103. [PubMed: 16710306]
24. Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. PLOS Genetics. 2010 In press.
25. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. PLoS Genet. Oct.2009 5:e1000695. [PubMed: 19851460]
26. Veyrieras JB, et al. PLoS Genet. Oct.2008 4:e1000214. [PubMed: 18846210]
27. Przeworski M, Coop G, Wall JD. Evolution Int J Org Evolution. Nov.2005 59:2312.
28. Innan H, Kim Y. Genetics. Jul.2008 179:1713. [PubMed: 18562650]
29. Eyre-Walker A, Keightley PD. Mol Biol Evol. Sep.2009 26:2097. [PubMed: 19535738]
30. Boyko AR, et al. PLoS Genet. 2008; 4:e1000083. [PubMed: 18516229]
31. Santiago E, Caballero A. Genetics. Feb.1995 139:1013. [PubMed: 7713405]
32. Ralph PL, Coop G. Genetics. Jul 26.2010
33. Perry GH, et al. Nat Genet. Oct.2007 39:1256. [PubMed: 17828263]
34. Orr HA. Genetics. Aug.1998 149:2099. [PubMed: 9691061]

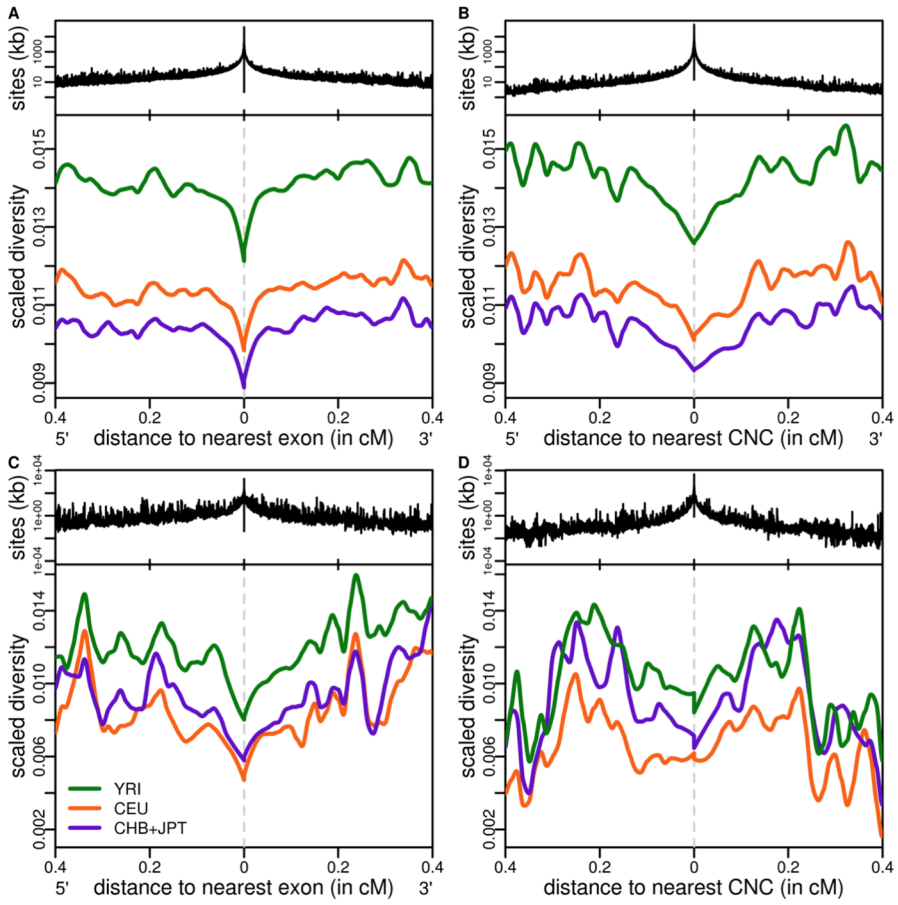

Figure 1.

Diversity levels divided by human-rhesus macaque divergence (at non-conserved, non-coding sites), as a function of genetic distance from exons and conserved non-coding regions. The top row (A–B) is for autosomes and the bottom row (C–D) for the X. Shown are LOESS curves obtained for a span of 0.1 and a bin size of 1.2×10^{-5} cM. Above each figure is a histogram of the number of kilobases in each bin (plotted on a log scale). See (20)for alternative versions.

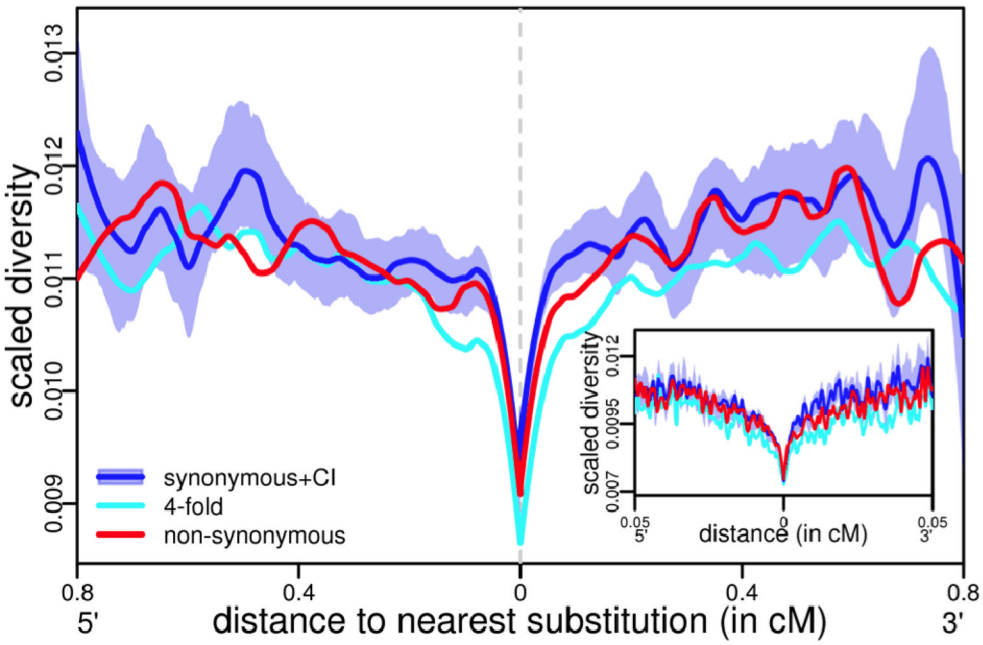

Figure 2.

Diversity levels divided by human-rhesus macaque divergence around human-specific substitutions across the autosomes. In the main plot, LOESS curves have a span of 0.2 and a bin size of 1.2×10^{-5} cM; the inset has a span of 0.05 to show added detail near the substitutions. The light blue shaded area represents the central 95%-tile of diversity estimates obtained from 100 bootstrap simulations. For alternative versions of this figure, including the same plot for YRI and CHB+JPT, as well as the X chromosome see Fig. S5.

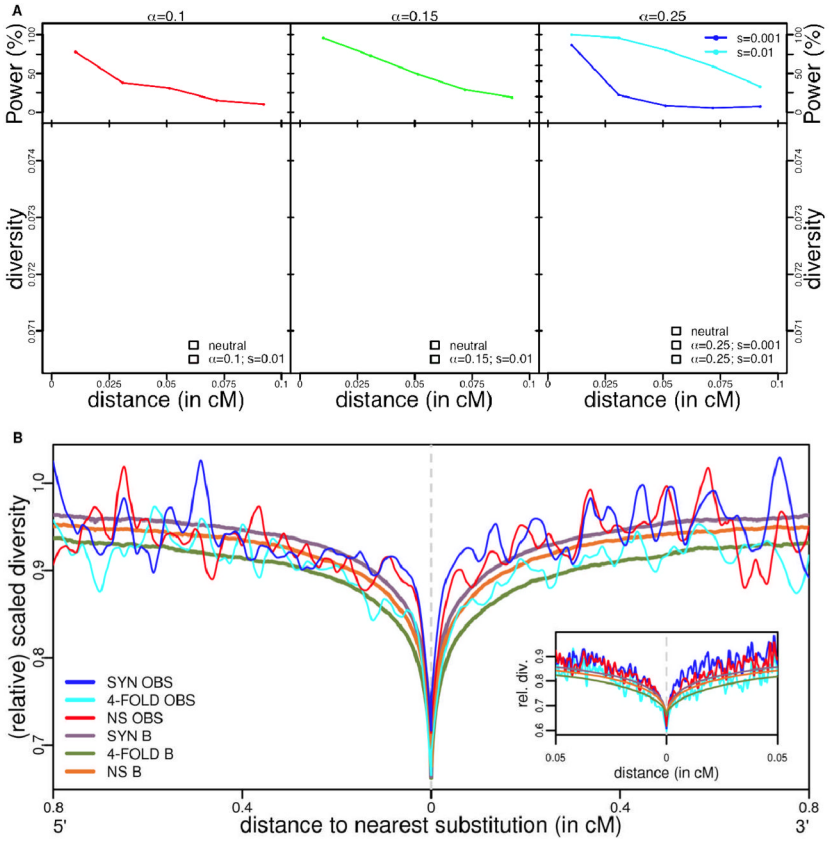

Figure 3.

A. The power to detect a decrease in diversity levels around amino acid substitutions due to classic sweeps. The top panel presents the power at a given genetic distance from the substitution, for the three sets of selection parameters (see (20)). The bottom panel shows the diversity patterns expected around amino acid substitutions for four sets of selection parameters, as well as for a model of purely neutral fixations, after LOESS smoothing (with a span of 0.2). For each set of parameters, the shaded area represents the central 95%-tile obtained from 100 bootstrap simulations. The depth of the trough reflects the fraction of substitutions that were beneficial and its width the typical strength of selection (24). **B.** Relative diversity levels around non-synonymous, synonymous and four fold degenerate synonymous substitutions predicted under a model of background selection(see (20)). B is the predicted diversity level relative to what is expected with no effects of background selection, i.e., under strict neutrality, taking into account variation in mutation rates (11). OBS is the observed value of average scaled diversity (i.e., diversity divided by divergence to rhesus macaque). For the expected diversity around exons and CNCs, as well as predictions for the X chromosome, see Fig. S8.

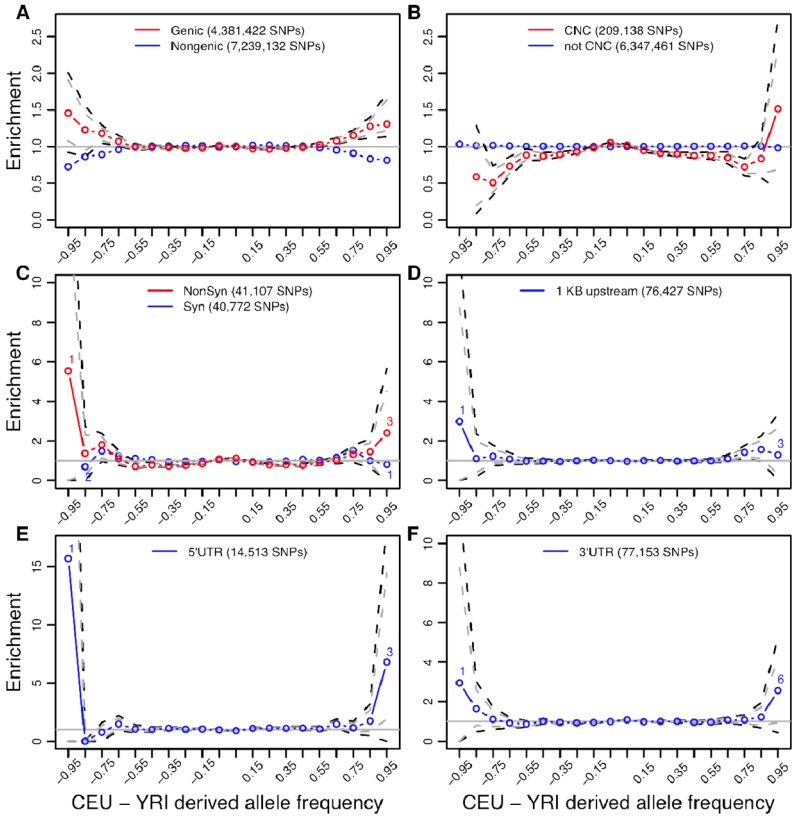

Figure 4.

A. Enrichment of highly differentiated SNPs in genic compared to non-genic sites. Shown is the CEU-YRI comparison (other population comparisons are in Fig. S9A). The total number of SNPs considered in each pairwise comparison is provided in the legend of each plot. Central 90% and 98% confidence intervals are shown with gray and black dashed lines, respectively; they were obtained by bootstrapping 200 kb regions 1000 times (16). **B**. Enrichment of highly differentiated SNPs in conserved non-coding compared to non-conserved, non-coding positions (20), for the CEU-YRI comparison. Other population comparisons are in Fig. S9B. **C–F.** Enrichment of specific genic annotations relative to the genomic background showing the central 90% and 98% confidence intervals for the CEU-YRI comparison with gray and black dashed lines, respectively. Note that an enrichment of 0 corresponds to no SNPs with that level of differentiation, so the confidence interval is not estimated in this case. Other population comparisons are shown in Fig. S10. For the numbers of each bin, see Fig. S11. Enrichments calculated on the folded frequency spectrum are shown in Fig. S13B. For a comparison of synonymous and non-synonymous SNPs in an alternative resequencing dataset, see Fig. S14.