

# Purifying selection on noncoding deletions of human regulatory elements detected using their cellular pleiotropy

David W. Radke<sup>a,b,c,d,1</sup>, Jae Hoon Sul<sup>e</sup>, Daniel J. Balick<sup>b,c,d</sup>, Sebastian Akle<sup>b,c,d,f</sup>, Alzheimer's Disease Neuroimaging Initiative<sup>\*</sup>, Robert C. Green<sup>c,d</sup>, and Shamil R. Sunyaev<sup>b,c,d,1</sup>

<sup>a</sup>Program in Genetics and Genomics, Biological and Biomedical Sciences PhD Program, Harvard Medical School, Boston, MA, 02115, USA; <sup>b</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA; <sup>c</sup>Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA; <sup>d</sup>Broad Institute of Harvard and MIT, Cambridge, MA, 02142 USA; <sup>e</sup>Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA, 90095, USA; <sup>f</sup>Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA, 02138, USA; <sup>\*</sup>Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [http://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf)

This manuscript was compiled on May 19, 2020

1 **Genomic deletions provide a powerful loss-of-function model in non-**  
2 **coding regions to assess the role of purifying selection on human**  
3 **noncoding genetic variation. Regulatory element function is char-**  
4 **acterized by non-uniform tissue/cell-type activity, necessarily linking**  
5 **the study of fitness consequences from regulatory variants to their**  
6 **corresponding cellular activity. We used deletions from the 1000**  
7 **Genomes Project (1000GP) and a callset we generated from genomes**  
8 **of participants in the Alzheimer's Disease Neuroimaging Initiative**  
9 **(ADNI) in order to examine whether purifying selection preserves**  
10 **noncoding sites of chromatin accessibility (DHS), histone modifica-**  
11 **tion (enhancer, transcribed, polycomb-repressed, heterochromatin),**  
12 **and topologically associated domain loops (TAD-loops). To exam-**  
13 **ine this in a cellular activity-aware manner, we developed a sta-**  
14 **tistical method, Pleiotropy Ratio Score (PlyRS), which calculates a**  
15 **correlation-adjusted count of "cellular pleiotropy" for each noncod-**  
16 **ing base-pair by analyzing shared regulatory annotations across**  
17 **tissues/cell-types. Comparing real deletion PlyRS values to simu-**  
18 **lations in a length-matched framework and using genomic covari-**  
19 **ates in analyses, we found that purifying selection acts to preserve**  
20 **both DHS and enhancer sites, as evident by both depletion of dele-**  
21 **tions overlapping these annotations and a shift in the allele fre-**  
22 **quency spectrum of overlapping deletions towards rare alleles. How-**  
23 **ever, we did not find evidence of purifying selection for transcribed,**  
24 **polycomb-repressed, or heterochromatin sites. Additionally, we**  
25 **found evidence that purifying selection is acting on TAD-loop bound-**  
26 **ary integrity by preserving co-localized CTCF binding sites. Notably,**  
27 **at regions of DHS, enhancer, and CTCF within TAD-loop boundaries**  
28 **we found evidence that both sites of tissue/cell-type-specific activity**  
29 **and sites of cellularly pleiotropic activity are preserved by selection.**

purifying selection | genomic deletions | noncoding regulatory elements  
| cellular pleiotropy

1 **L**arge-scale sequencing studies have provided tremendous  
2 insight into biological function and human disease, with  
3 statistical signatures of natural selection serving as a primary  
4 identifying feature. The classic example is the analysis of  
5 selective constraints on protein coding genes evident from the  
6 depletion of missense or nonsense genetic variants. These ad-  
7 vances, however, are not directly translatable to the analysis  
8 of noncoding DNA, which has increasingly become a focus  
9 of human genetics research. Functional genomic studies have  
10 revealed numerous regions of regulatory activity marked by  
11 chromatin accessibility and histone modification (1, 2). Associ-

ation signals for common human phenotypes are dramatically  
enriched in these regulatory regions of the genome (3), show-  
casing the importance of specialized cellular function. In  
contrast to protein-coding sequences, the function of regula-  
tory sequences is not determined by triplet codon structure  
thereby providing no obvious analog to protein-truncating  
single nucleotide variants (SNVs) to identify loss of function.  
This ambiguity of the mutational consequences of individual  
nucleotides within regulatory sequences complicates the ability  
to study their function through the lens of purifying natural  
selection. Previous work focusing on SNVs within noncoding  
regions developed sophisticated genetic models that relied on  
functional proxies such as transcription factor binding sites,  
nucleotide conservation across species, or machine learning (4-  
11). However, it is difficult to clearly interpret these findings  
in terms of selection against the loss of regulation. In contrast  
to SNVs, deletions are a class of variation that provide a direct  
loss of normal regulatory function at a locus by physically  
removing the sequence of a regulatory element in at least

## Significance Statement

We used natural genomic deletions as a loss-of-function model to assess the role of purifying selection in preserving human noncoding regulatory sites. We examined this in a cellular activity-aware manner through development of a statistical method, Pleiotropy Ratio Score (PlyRS), which calculates an adjusted count of "cellular pleiotropy" for each noncoding base-pair by analyzing correlations from shared regulatory annotations across tissues/cell-types. By comparing real deletion PlyRS values to simulations, we found that purifying selection acts to preserve both DHS and enhancer sites and TAD-loop boundary integrity by preserving co-localized CTCF binding sites. Notably, we found evidence at these regulatory regions that both sites of tissue/cell-type-specific activity and sites of cellularly pleiotropic activity are preserved by selection.

D.W.R., R.C.G., and S.R.S. designed research; D.W.R. performed research; D.W.R., J.H.S., D.J.B., and S.A. contributed new reagents/analytic tools; D.W.R. and S.R.S. analyzed data; and D.W.R. and S.R.S. wrote the paper.

R.C.G. receives compensation for advising AIA, Grail, Humanity, UnitedHeath, Verily and Wamberg, and is co-founder of Genome Medical. The remaining authors declare no conflict of interest.

<sup>1</sup>To whom correspondence should be addressed. E-mail: [davidradke@fas.harvard.edu](mailto:davidradke@fas.harvard.edu) or [ssunyaev@rics.bwh.harvard.edu](mailto:ssunyaev@rics.bwh.harvard.edu).

31 a heterozygous manner. This logic underlies experimental  
32 studies of regulatory function using CRISPR/Cas9 systems  
33 (12, 13). Yet, natural population genetic variation provides  
34 a more systematic and genome-wide view of the action of  
35 selection on deletions. Work done by sequencing consortia  
36 has demonstrated reduction of deletion variation in various  
37 categories of regulatory sequences (14–16).

38 The hallmark feature of human regulatory elements is their  
39 non-uniform activity across tissues and cell types. Here, we  
40 offer a population genetic analysis of natural deletions in light  
41 of variable regulatory activity across tissues. Deletions that  
42 remove sites of genomic regulation with pleiotropic cellular  
43 effects (what we term "cellular pleiotropy", i.e. the same regula-  
44 tory element locus is active in more than one tissue or cell-type)  
45 might be expected to be, on average, more deleterious (i.e.  
46 fitness-reducing) than deletions that remove cell-type-specific  
47 sites, since any changes at the DNA level to these regulators  
48 potentially affects multiple tissues/cell-types simultaneously.  
49 Another possibility is that since tissue/cell-type-specific reg-  
50 ulation is what enables widespread cellular diversity, these  
51 regulatory elements must be under strong selective constraint  
52 to preserve their specialized biological function. These two  
53 potential modes of selection preserving regulation of cellular  
54 activity are not mutually exclusive, as selection may be oper-  
55 ating to remove overlapping deletions to preserve the utility  
56 of both types of regulators. Prior work has provided sug-  
57 gestive evidence that tissue activity count is a contributor  
58 to selective constraint in regulatory sequences (10, 16, 17).  
59 Studying purifying selection on noncoding deletions is thus  
60 inherently tied to the cellular activity of corresponding deleted  
61 regulatory sequences. To address this, we have developed a  
62 statistical method, Pleiotropy Ratio Score (PlyRS), to quantify  
63 the amount of tissue/cell-type activity (i.e. cellular pleiotropy)  
64 for individual nucleotides in light of the hierarchical devel-  
65 opmental structure of human tissues and cell types, while  
66 controlling for their correlation rather than using a simple  
67 tissue/cell-type count. We then analyzed separately several  
68 diverse epigenomic features (open chromatin, histone modi-  
69 fications, and topologically associated domain loops) taking  
70 into account non-independence of these individual annotations  
71 across tissues and cell types using our PlyRS values. In this  
72 way, we assessed the effect of purifying selection on millions  
73 of nucleotide positions in the human genome by examining  
74 patterns of PlyRS values within naturally occurring deletion  
75 sequences.

76 Reduction of genetic variation and a shift in the allele fre-  
77 quency spectrum (AFS) towards rare variants are two key  
78 signatures of purifying selection. If selection is operating on  
79 the removal of deleterious deletions overlapping regulatory  
80 regions, we would expect to see both a reduction in deletion  
81 variation overlapping the important regulatory features and a  
82 shift in the AFS of remaining overlapping deletions towards  
83 rarer frequencies, relative to neutral expectations. These  
84 conditions on segregating deletions should be simultaneously  
85 present to conclude that purifying natural selection is acting  
86 to preserve a particular regulatory epigenomic feature(s), as  
87 either reduced deletion counts or a shift in the deletion AFS  
88 alone may indicate deletion calling artifacts or confounding  
89 genomic covariates. Both of these signatures are prone to  
90 various biological and technical confounders, particularly for  
91 structural variation. For example, the accuracy of deletion

calls is influenced by their length and allele frequency (AF) 92  
(18). Longer deletions have more prevalent missing coverage 93  
and common deletions are observed more often in the popula- 94  
tion, so these types of deletions are more likely to be correctly 95  
identified using current methods based on analyzing short-read 96  
sequencing data. Variant calling accuracy also depends on 97  
the mappability of the sequence (19). Another known issue is 98  
the observed negative correlation of deletion length and AF 99  
(14, 20). This could be due to underlying biology, deletion 100  
caller algorithm biases, or both. In addition to technical con- 101  
founders, biological factors unrelated to the direct pressure of 102  
selection may affect the degree of variation (i.e. the number of 103  
segregating mutations) and the AFS. For example, the degree 104  
of variation is linearly proportional to mutation rate; however, 105  
the deletion mutation rate at fine-scale is still unknown and 106  
could be influenced by sequence GC content and other local 107  
genomic properties. In contrast to the overall variation, the 108  
normalized AFS is not affected by mutation rate, at least 109  
for relatively small sample sizes, but together with degree 110  
of variation could be influenced by complex mechanisms like 111  
background selection. To address these complications, we 112  
simulated length-matched positions of each real deletion while 113  
keeping the original AF label, and took into account relevant 114  
genomic confounding variables co-occurring with the same 115  
deletion. Using this framework, we compared the observed 116  
diversity and AFS of real deletions to the expectations based 117  
on computer simulations using analyses of PlyRS values across 118  
their coordinates. 119

## 120 Results

**Pleiotropy Ratio Score (PlyRS).** To score deletions with respect 121  
to their effect on regulatory function, we considered both 122  
the number of removed elements and the activity of each 123  
element across cell and tissue types. In contrast to SNVs, a 124  
noncoding deletion can potentially remove regulatory function 125  
at a genomic locus along two distinct "axes" (see SI Appendix, 126  
Fig. S1 for a cartoon). One axis ("horizontal") corresponds 127  
to the amount of regulatory space removed by the deletion 128  
irrespective of its tissue/cell-type activity. The other axis 129  
("vertical") corresponds to the combined amount of regulatory 130  
activity across tissues and cell types of each base-pair (i.e. 131  
the cellular pleiotropy of a regulatory coordinate). Thus, for 132  
any deletion overlapping regulatory sequences, there will be a 133  
simultaneous removal at that locus along both axes, which we 134  
quantify by a counting score for each axis. 135

For the horizontal axis we count deleted base-pairs with a 136  
regulatory annotation from any tissue/cell-type (SI Appendix, 137  
Note S1a). We do not require removal of an entire regulatory 138  
element for this horizontal count, since deletion of even a 139  
partial regulatory element sequence can render it inoperable 140  
(21). Additionally, since regulatory element boundaries are 141  
not perfectly aligned between tissues, it could be the case that 142  
a partial deletion of an element observed in one tissue may 143  
correspond to a complete deletion of the element observed 144  
in another tissue. Consequently, for a deletion overlapping 145  
a regulatory element(s), the horizontal axis count score can 146  
range from as low as 1 (only a single regulatory base-pair 147  
deleted) to as high as the length of the deletion (all base-pairs 148  
along the deletion length overlap a regulatory element[s]). 149

A simple numerical count of the number of tissues/cell- 150  
types where a regulatory element locus has activity is not 151

152 sufficient for properly specifying cellular pleiotropy, because  
153 this count can be heavily influenced by the cellular diversity  
154 of the particular tissues/cell-types included in the analysis.  
155 For example, a count of 3 in an analysis performed with heart  
156 tissue, lung tissue, and ten blood cell-types would not have the  
157 same interpretation as a count of 3 in an analysis performed  
158 with heart tissue, lung tissue, and only one blood cell-type.  
159 In the former, it could be that the count of 3 comes from  
160 three highly-correlated blood cell-types, but in the latter, the  
161 count of 3 would have to come from the more developmentally  
162 diverse set of all three tissues/cell-types. Therefore, to enable  
163 proper "counting" of cellular pleiotropy, we developed a sta-  
164 tistical method, called Pleiotropy Ratio Score (PlyRS), which  
165 calculates a correlation-adjusted count of cellular pleiotropy  
166 for each base-pair in the noncoding genome (*SI Appendix, Note*  
167 *S1b*).

168 We use the PlyRS value at any given base-pair along the  
169 length of a deletion to provide the counting score along the ver-  
170 tical axis. At any base-pair coordinate within a deletion, the  
171 PlyRS value can range from 0-indicating no tissue/cell-type  
172 included in the analysis has annotated activity-to a maximum  
173 of 1-indicating that all tissues/cell-types analyzed have anno-  
174 tated activity. Between these extreme bounds, PlyRS does  
175 not simply calculate the fraction of tissues/cell-types where  
176 the base-pair exhibits regulatory activity, rather it weights  
177 this proportion relative to the overall correlation of regulatory  
178 activity across these tissues/cell-types along the genome. For  
179 example, a base-pair active in three highly related cell types  
180 would be assigned a lower PlyRS value than a base-pair active  
181 in three (or potentially less) unrelated cell types. Thus, as a  
182 consequence of the PlyRS method calculation, counts of ele-  
183 ments active in tissues/cell-types with common activity will be  
184 down-weighted while counts of elements active in tissues/cell-  
185 types with rare activity will be up-weighted. Similarly, for  
186 each base-pair that has only tissue/cell-type-specific activity,  
187 the PlyRS value will be different for that particular tissue/cell-  
188 type depending on how its activity covaries across the genome  
189 with the other regulatory tissues/cell-types being analyzed.  
190 *SI Appendix* Fig. S2 and Fig. S3, respectively, illustrate how  
191 PlyRS corresponds to the raw tissue/cell-type count and how  
192 PlyRS compares to tissue/cell-type-specific counts.

193 **Construction of Deletion and Regulatory Datasets.** To exam-  
194 ine potential selective constraints on deletions within regula-  
195 tory regions, we needed fine-resolution of genomic coordinates  
196 for both deletions and regulatory regions as well as high-  
197 confidence deletion allele frequencies from population data.  
198 For this, we compiled deletion data from two callsets and  
199 regulatory data from seven callsets, and applied additional  
200 filters relevant to our analysis. See Materials and Methods for  
201 additional criteria used to ensure high-quality datasets.

202 We used deletion data from the 1000 Genomes Project  
203 Consortium Phase 3 callset (1000GP) of breakpoint-resolved  
204 deletions for which deletions were genotyped in 2,504 individ-  
205 uals from 26 modern human populations (14) (*SI Appendix,*  
206 *Note S2a*). We additionally used deletions that we called and  
207 genotyped across 752 individuals sequenced as part of the  
208 Alzheimer's Disease Neuroimaging Initiative (ADNI) (22) (*SI*  
209 *Appendix, Note S2b* and *Note S3*), using the CNV algorithm  
210 GenomeSTRiP (23). We restricted our analysis to noncoding  
211 deletions. As expected, the bulk (>80%) of deletions in our  
212 datasets remaining after filtering were rare (below 1% AF).

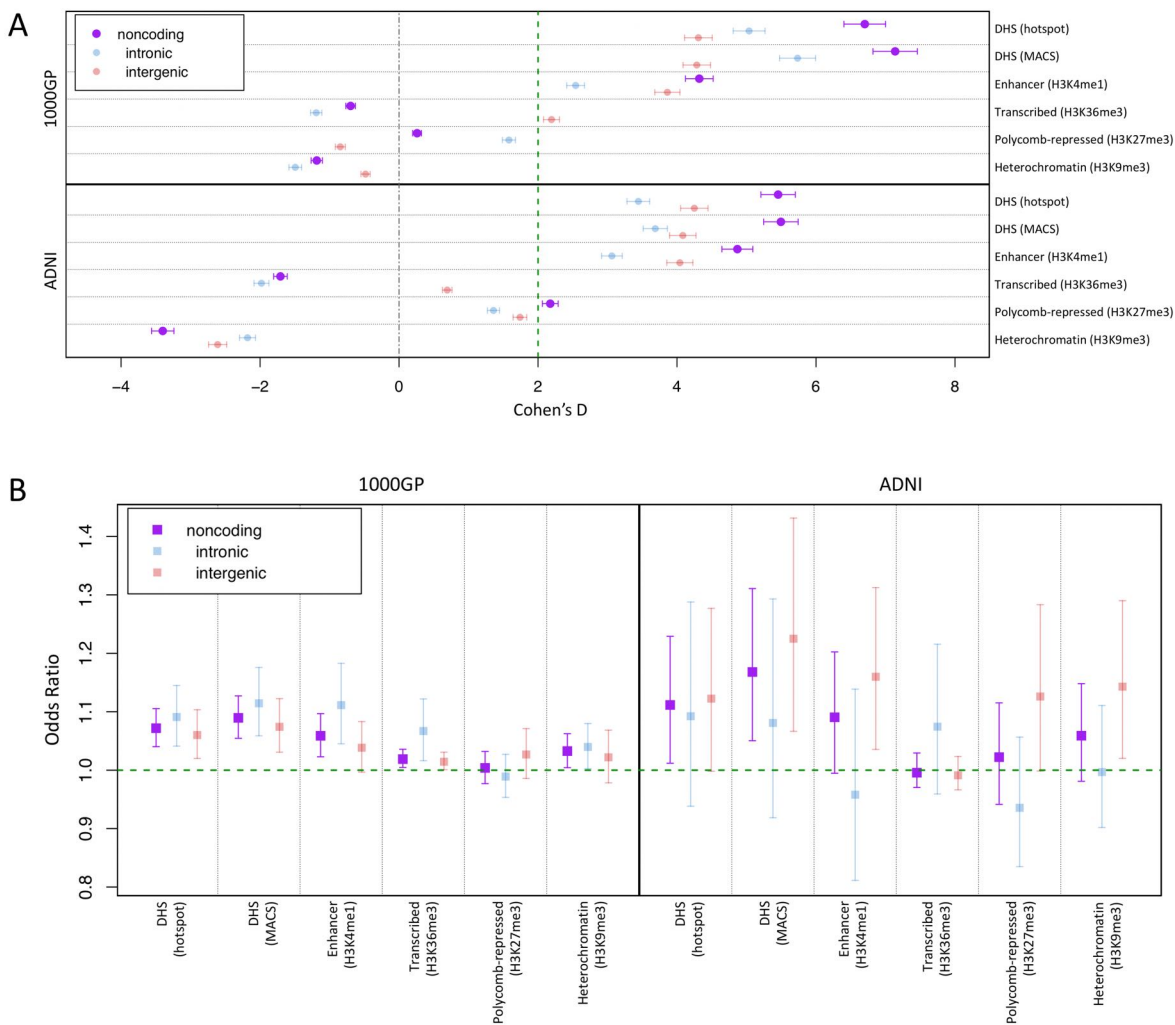
213 To analyze genomic deletions within regulatory regions, we  
214 used regulatory data from the NIH Roadmap Epigenomics  
215 Consortium (REC) (2). In particular, we used two callsets of  
216 chromatin accessibility data (DNase I hypersensitivity "DHS")  
217 and four callsets of histone modification data (H3K4me1  
218 "enhancer", H3K36me3 "transcribed", H3K27me3 "polycomb-  
219 repressed", and H3K9me3 "heterochromatin"). Two sets of  
220 DHS annotation (hotspot and MACS) were used to check for  
221 consistency in the analyses. DHS annotations are typically  
222 associated with sites of open chromatin allowing accessibility  
223 for regulator binding and histone annotations are typically  
224 associated with sites of specific regulatory activity, as noted.  
225 We additionally used regulatory data that demarcate topologi-  
226 cally associated domain loops (TAD-loops) (24), which are  
227 associated with local genomic regions of physically interacting  
228 regulatory activity.

229 **Depletion of Variation at DHS or Enhancer Sites.** We first  
230 tested whether there was evidence of depletion of noncod-  
231 ing deletion variation overlapping chromatin accessibility and  
232 histone modifications (*SI Appendix, Note S4a*). We corrected  
233 for the confounding effects of mappability, deletion length,  
234 and allele frequency using simulations. For each real dele-  
235 tion in both the 1000GP and ADNI datasets, we randomly  
236 simulated 1,000 deletions of the same length to occur on the  
237 same chromosome and same noncoding genomic compartment  
238 space (intronic or intergenic) using only uniquely mappable  
239 sequence coordinates (*SI Appendix, Note S2d*) for both real  
240 deletions and simulated deletions. For more detail on our  
241 length-matched simulations, see *SI Appendix, Notes S5a-S5c*.  
242 We summed the PlyRS values calculated per base-pair along  
243 the length of every deletion. This sum, denoted PlyRS<sub>sum</sub>  
244 (*SI Appendix, Note S1c*), corresponds to the total cellular  
245 pleiotropy (for a specific regulatory feature) of the deletion,  
246 encompassing both the horizontal and vertical "axes" along  
247 which purifying selection may be operating on the deletion (*SI*  
248 *Appendix, Note S1*). We compared PlyRS<sub>sum</sub> values for both  
249 real and simulated deletions and quantified depletion across  
250 regulatory features using Cohen's D statistic (*SI Appendix,*  
251 *Note S5b*).

252 Panel A of Fig. 1 shows PlyRS<sub>sum</sub> effect sizes from compar-  
253 ing real data to simulations and indicates significant depletion  
254 of deletions (Cohen's D>2, corresponding to 2 std. dev.)  
255 overlapping DHS or enhancer regions. We did not detect  
256 a significant depletion for deletions overlapping transcribed,  
257 polycomb-repressed, or heterochromatin epigenomic features.  
258 The depletion of deletions overlapping DHS or enhancer sites  
259 was significant not only in the full deletion sets, but also in  
260 both the intronic and intergenic genomic compartments. Addi-  
261 tionally, we found concordance between effect sizes in 1000GP  
262 and ADNI datasets for DHS or enhancer deletion depletions,  
263 suggesting reliable capture of biological information from dele-  
264 tion callsets with differing characteristics (*SI Appendix, Tables*  
265 *S1-S2*). These results suggest that purifying selection may be  
266 operating broadly on deletions to preserve DHS and enhancer  
267 epigenomic features. *SI Appendix, Table S5d1 (Note S5d)* lists  
268 the effect sizes found in the depletion simulations.

269 **Shift in Allele Frequency Spectrum at DHS or Enhancer Sites.**  
270 We next tested whether there was a shift in the allele frequency  
271 spectrum of noncoding deletions overlapping the chromatin ac-  
272 cessibility and histone modification epigenomic features. The





**Fig. 1.** Depletion of deletions and shift of deletion allele frequency spectrum overlapping regulatory sites. (A) We calculated  $\text{PlyRS}_{\text{sum}}$  (*SI Appendix, Note S1c*) for every deletion to quantify overlap with sites of chromatin accessibility or histone modification. We plot the degree of reduction in the  $\text{PlyRS}_{\text{sum}}$  for real deletions relative to simulation. This reduction is measured using Cohen's D, which is the effect size of a t-test on  $\text{PlyRS}_{\text{sum}}$  values (*SI Appendix, Notes S5a-S5b*) in units of standard deviation (plotted with 95% confidence intervals on the mean reduction). Two units of effect size (Cohen's = 2) approximately corresponds to the 95% confidence interval of significance in depletion. Higher values of Cohen's D indicate larger depletion within those sets compared to simulation. In presence of the true effect, there is sample size dependence on the underlying t-test, and the expected value of Cohen's D would be higher for larger datasets. (B) For each deletion we determined the magnitude of  $\text{PlyRS}_{\text{sum}}$  depletion, calculated as a ratio between its  $\text{PlyRS}_{\text{sum}}$  and the average  $\text{PlyRS}_{\text{sum}}$  of its length-matched simulated deletions (*SI Appendix, Note S6b*), for sites of chromatin accessibility or histone modification. We tested whether  $\text{PlyRS}_{\text{sum}}$  depletion magnitude depends on allele frequency (deletions categorized as rare [ $\text{AF} \leq 1\%$ ] or common), using multivariate logistic regression in the presence of genomic covariates (*SI Appendix, Note S6a*). We plot the regression odds ratio (OR) with 95% profile likelihood-based confidence intervals. Results above 1 indicate positive correlation of the magnitude of  $\text{PlyRS}_{\text{sum}}$  depletion with allele frequency. This corresponds to an excess of rare alleles overlapping the regulatory feature in the real dataset compared to simulation, which is the expected result for features being preserved by the action of purifying selection against overlapping deletions.

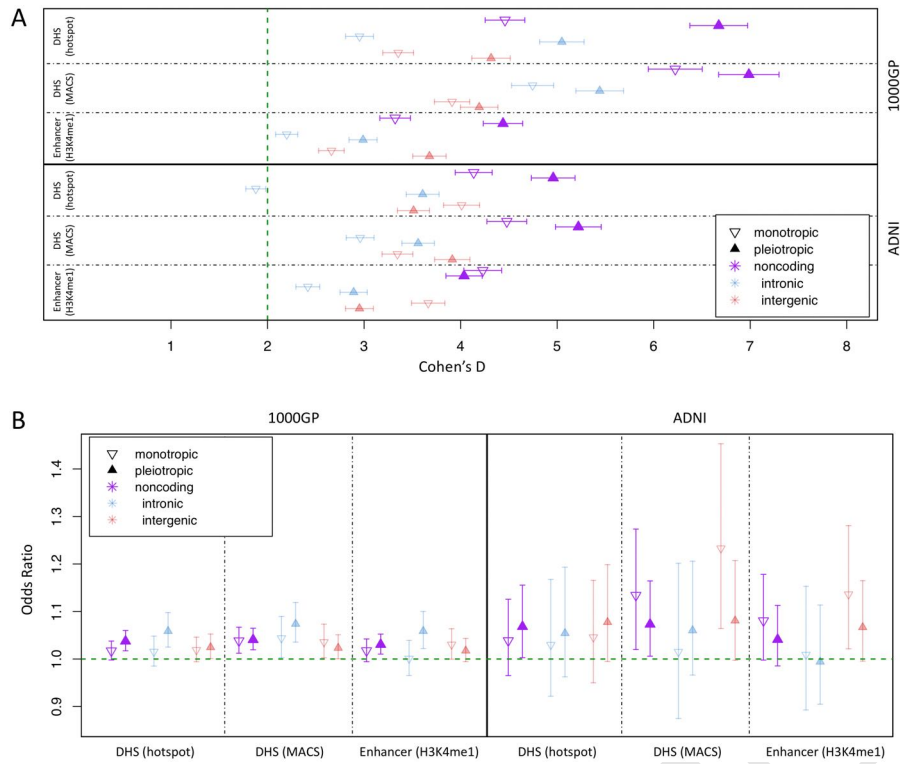
273 analysis of allele frequency distribution is important because  
274 the total degree of variation can be confounded by mutation  
275 rate (unlike SNVs, we do not have good models for mutation  
276 rate along the genome for deletions [(25)]). The allele frequency  
277 distribution, when normalized, does not depend on mutation  
278 rate for relatively small populations (within the limits of the  
279 infinite sites approximation), but due to the recent explosive  
280 growth of the human population, this assumption may break  
281 down for extremely large sample sizes at which point recent  
282 recurrent mutations become relevant. However, for the sample  
283 sizes analyzed here, the allele frequency distribution can be  
284 assumed to be independent of mutation rate, with the chance  
285 of recurrent mutations being small. This is especially true for  
286 deletions which would require recurrent mutations to occur at  
287 the same breakpoints (start and end coordinates being identical).  
288 Therefore, a shift in the allele frequency spectrum of  
289 real deletions in our datasets compared to simulated deletions  
290 would likely reflect the action of purifying selection. Still, the  
291 allele frequency distribution can be affected by a number of  
292 variables unrelated to selective pressure. To take into account  
293 the potential effect of background selection, we controlled  
294 for regional (50kb +/- deletion coordinates) SNV nucleotide  
295 diversity and recombination rate, as well as distance to the  
296 nearest transcription start site. We additionally controlled  
297 for regional GC content. Due to technical confounders, allele  
298 frequency is expected to be influenced by deletion length so  
299 we also controlled for length explicitly. We accounted for  
300 these genomic covariates using multivariate logistic regression,  
301 testing whether  $\text{PlyRS}_{\text{sum}}$  depletion magnitude depended on  
302 allele frequency (deletions categorized as rare [ $\text{AF} \leq 1\%$ ] or  
303 common; *SI Appendix, Note S6a*). To measure the magnitude  
304 of potential  $\text{PlyRS}_{\text{sum}}$  depletion for each deletion, we calculated  
305 a ratio between its  $\text{PlyRS}_{\text{sum}}$  and the average  $\text{PlyRS}_{\text{sum}}$   
306 of its length-matched simulated deletions (*SI Appendix, Note*  
307 *S6b*). If purifying selection is, in fact, acting against deletions  
308 overlapping regulatory features, we would expect the largest  
309  $\text{PlyRS}_{\text{sum}}$  depletions to be found in common deletions (in  
310 our test, an odds ratio [OR] above 1 which shows positive  
311 correlation with allele frequency).

312 Panel B of Fig. 1 shows that for deletions overlapping  
313 DHS sites the OR significantly (confidence interval [CI] 95%)  
314 exceeded 1 in both datasets, indicating the action of purifying  
315 selection. Additionally, for deletions overlapping enhancer  
316 sites the OR significantly exceeded 1 in the 1000GP dataset,  
317 while the lower CI boundary of the OR was nearly significant,  
318 at 0.995, in the ADNI dataset. All intronic and intergenic  
319 genomic compartment sets for DHS or enhancer features had  
320 mean odds ratios  $>1$  (except ADNI intronic enhancers at  
321 0.96). *SI Appendix, Table S6c1, (Note S6c)* lists the odds  
322 ratios found in the logistic regressions. These results suggest  
323 that purifying selection may be preserving DHS and enhancer  
324 epigenomic features by reducing allele frequencies of overlapping  
325 deletions. On the other hand, there is a lack of consistent  
326 allele frequency shift for genomic compartment sets for transcribed,  
327 polycomb-repressed, and heterochromatin features in  
328 both deletion datasets, with the mean OR sometimes falling  
329 below 1 and the OR CI often extending below 1. In light of  
330 the insufficient evidence across datasets for an excess of rare  
331 alleles for these features, combined with the lack of reduction  
332 in variation described above, we focused the analysis below  
333 on DHS and enhancer epigenomic features which showed sta-

tistical significance of both key signatures of broad selection  
against overlapping deletions.

**Differential Selection on Preserving Cellular Activity.** The results described above have indicated that purifying selection is acting against the total cellular pleiotropic burden ( $\text{PlyRS}_{\text{sum}}$ ) of noncoding deletions, preserving both DHS and enhancer regulatory sites. However, these analyses do not clarify if purifying selection preserves DHS or enhancer sites of both tissue/cell-type-specific activity and cellularly pleiotropic activity. One possibility is that deletions removing regulatory elements active in multiple tissues/cell-types incur a greater fitness cost. Another possibility is that since tissue/cell-type-specific elements are vital to organismal development, deletions removing them are subject to a stronger selective effect. It could also be the case that purifying selective pressure on deletions is acting to preserve both types of regulatory sites simultaneously. To distinguish between these scenarios, we calculated two additional PlyRS measures,  $\text{PlyRS}_{\text{sum-mono}}$  and  $\text{PlyRS}_{\text{sum-pleio}}$  (*SI Appendix, Note S1c*).  $\text{PlyRS}_{\text{sum-mono}}$  included the sum of PlyRS values of each deleted base-pair for which a base-pair is only associated with regulatory activity in one tissue/cell-type.  $\text{PlyRS}_{\text{sum-pleio}}$  included the sum of PlyRS values of each deleted base-pair for which that base-pair is associated with regulatory activity in more than one tissue/cell-type. The sum of these two components is the original measure of total cellular pleiotropic burden,  $\text{PlyRS}_{\text{sum}}$ . With these additional PlyRS measures, we performed the same analyses as above to examine both a potential reduction in variation and a shift in allele frequency, now applied separately to each component of  $\text{PlyRS}_{\text{sum}}$ . This allowed us to determine, within the same sets of real deletions, which scenario of regulatory activity preservation was contributing to the signal of depletion in variation and shift in the AFS as found above.

Fig. 2A shows a significant depletion of variation for DHS or enhancer sites corresponding to both tissue/cell-type-specific activity and for cellularly pleiotropic activity in both 1000GP and ADNI datasets. The effect size of this reduction in variation for  $\text{PlyRS}_{\text{sum-mono}}$  or  $\text{PlyRS}_{\text{sum-pleio}}$  was greater for  $\text{PlyRS}_{\text{sum-pleio}}$  for both noncoding regulatory features, except for enhancer sites in ADNI deletions where the effect size was comparable (error bars overlapping). *SI Appendix, Tables S5d2-S5d3 (Note S5d)* lists the effect sizes found in the depletion simulations, including those for intronic and intergenic compartments where depletion values did not consistently favor greater reduction of  $\text{PlyRS}_{\text{sum-pleio}}$ . Fig. 2B shows that the magnitude of deletion depletion overlapping DHS or enhancer sites leads to a significantly shifted AFS at both sites of tissue/cell-type-specific activity and cellularly pleiotropic activity. For DHS or enhancer sites in all genomic compartments, the mean odds ratios of the magnitude of depletion for  $\text{PlyRS}_{\text{sum-mono}}$  or  $\text{PlyRS}_{\text{sum-pleio}}$  in association to allele frequency were  $>1$  in both deletion datasets (except ADNI intronic enhancers), and were comparable between  $\text{PlyRS}_{\text{sum-mono}}$  and  $\text{PlyRS}_{\text{sum-pleio}}$ . *SI Appendix, Tables S6c2-S6c3 (Note S6c)* lists the odds ratios found in the logistic regressions. These results collectively indicate that purifying selection is acting to preserve DHS or enhancer sites of tissue/cell-type-specific activity as well as cellularly pleiotropic activity.



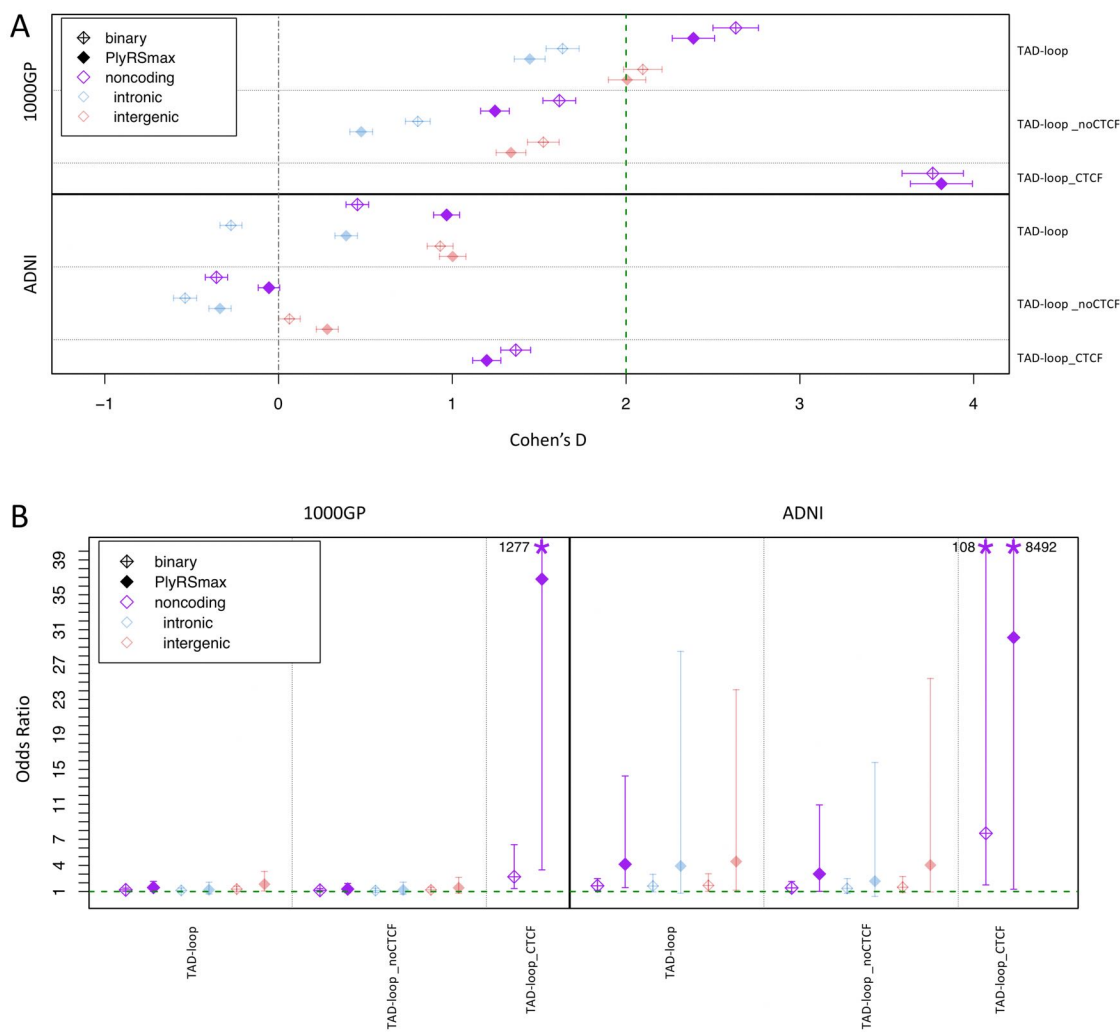
**Fig. 2.** Depletion of deletions and shift of allele frequency spectrum overlapping DHS or enhancer sites of variable cellular activity. (A) We calculated  $\text{PlyRS}_{\text{sum-mono}}$  (monotropic) and  $\text{PlyRS}_{\text{sum-pleio}}$  (pleiotropic) (*SI Appendix, Note S1c*) for every deletion to quantify overlap with DHS or enhancer sites. We plot the degree of reduction in  $\text{PlyRS}_{\text{sum-mono}}$  (or  $\text{PlyRS}_{\text{sum-pleio}}$ ) for real deletions relative to simulation measured using Cohen's D (with 95% confidence intervals on the mean reduction). (B) For each deletion we determined the magnitude of  $\text{PlyRS}_{\text{sum-mono}}$  (monotropic) (or  $\text{PlyRS}_{\text{sum-pleio}}$  [pleiotropic]) depletion, calculated as a ratio between its  $\text{PlyRS}_{\text{sum-mono}}$  (or  $\text{PlyRS}_{\text{sum-pleio}}$ ) and the average  $\text{PlyRS}_{\text{sum-mono}}$  (or  $\text{PlyRS}_{\text{sum-pleio}}$ ) of its length-matched simulated deletions (*SI Appendix, Note S6b*), for DHS or enhancer sites. We tested whether  $\text{PlyRS}_{\text{sum-mono}}$  (or  $\text{PlyRS}_{\text{sum-pleio}}$ ) depletion magnitude depends on allele frequency (deletions categorized as rare [ $\text{AF} \leq 1\%$ ] or common), using multivariate logistic regression in the presence of genomic covariates (*SI Appendix, Note S6a*). We plot the regression odds ratio with 95% profile likelihood-based confidence intervals.

393 **Purifying Selection on CTCF Sites within TAD-loops.** We also  
 394 investigated whether there was evidence of depletion of variation  
 395 and a shift in the AFS of deletions overlapping topologically  
 396 associated domain loops (TAD-loops). These large regions of  
 397 self-interacting DNA facilitate cis-regulatory effects at a  
 398 wider scale than that of individual regulators (26, 27) and  
 399 so deletions removing a TAD-loop boundary may be under  
 400 strong purifying natural selection to preserve the TAD-loop  
 401 integrity. The distance between TAD-loop boundaries is greater  
 402 than our longest deletions (25kb limit, [*SI Appendix, Note*  
 403 *S2c*]), and consequently deletions in our datasets can only  
 404 overlap with at most one TAD-loop boundary. Additionally,  
 405 the TAD-loop boundary data (*SI Appendix, Note S4b*) are  
 406 less precise than chromatin accessibility or histone modifica-  
 407 tion annotations, so the number of base-pairs of a deletion  
 408 overlapping a TAD-loop boundary may not reflect actual dele-  
 409 teriousness of the mutation but rather correspond to imprecise  
 410 annotations on the edges. These characteristics of TAD-loop  
 411 boundary annotation mean that using  $\text{PlyRS}_{\text{sum}}$  to define the  
 412 total cellular pleiotropy of overlapping deletions can propagate  
 413 a potential bias in the measure. To avoid this and still test  
 414 whether purifying selection may be operating on deletions  
 415 overlapping TAD-loop boundaries, we measured overlap both  
 416 as a binary variable and by calculating the maximal  $\text{PlyRS}$   
 417 value ( $\text{PlyRS}_{\text{max}}$ , *SI Appendix, Note S1c*) along the length  
 418 of an overlapping deletion. We performed the same analyses  
 419 as for the chromatin accessibility or histone modification  
 420 annotations (*SI Appendix, Notes S5a-S5c*).

421 Rao Huntley et al. (24) identified that a large majority  
 422 (86%) of TAD-loop loci had binding from the insulator pro-  
 423 tein CTCF, which ensures integrity of DNA loops, and conse-  
 424 quently, TAD-loop fidelity (28, 29). Given this critical function  
 425 of CTCF and its presence within most TAD-loop boundaries,  
 426 we suspected that deletions that overlap TAD-loop loci might

427 be under stronger purifying selection if a deletion also simulta-  
 428 neously overlaps a CTCF site within the TAD-loop boundary  
 429 (*SI Appendix, Note S4c*). To elucidate this, in addition to  
 430 identifying the full set of deletions overlapping TAD-loop  
 431 annotation (TAD-loop), we further refined deletions into two  
 432 subsets (*SI Appendix, Note S5c*): deletions overlapping TAD-  
 433 loop but not simultaneously overlapping a CTCF binding  
 434 site (TAD-loop<sub>noCTCF</sub>) and deletions overlapping TAD-loop  
 435 while simultaneously overlapping a CTCF binding site (TAD-  
 436 loop<sub>CTCF</sub>). Only about 1% of all deletions in our datasets  
 437 overlapped TAD-loop<sub>CTCF</sub>, so we ignored intronic and inter-  
 438 genic designations in the analysis (but maintained them in  
 439 simulations).

440 Fig. 3A shows the effect sizes of binary overlap or  $\text{PlyRS}_{\text{max}}$   
 441 overlap from comparing real deletions to simulations and in-  
 442 dicates that, with respect to the full set of TAD-loops being  
 443 overlapped (irrespective of whether CTCF sites are simulta-  
 444 neously overlapped), there was minimal depletion of deletion  
 445 variation, if any. However, as also seen in Fig. 3A, separation  
 446 into TAD-loop<sub>noCTCF</sub> and TAD-loop<sub>CTCF</sub> subsets revealed  
 447 that a signal of depletion was evident only for deletions over-  
 448 lapping TAD-loop<sub>CTCF</sub>. Deletions in the ADNI dataset ex-  
 449 hibited the same characteristic pattern of greater reduction  
 450 in variation in TAD-loop<sub>CTCF</sub> versus TAD-loop<sub>noCTCF</sub> as was  
 451 seen in the 1000GP dataset; however, the reduction seen in  
 452 ADNI deletions overlapping TAD-loop<sub>CTCF</sub> was not statisti-  
 453 cally significant. We did not find any difference between the  
 454 effect size of depletion for binary overlap compared to the  
 455  $\text{PlyRS}_{\text{max}}$  overlap measure, suggesting that there may not  
 456 be stronger selection against deletions overlapping the most  
 457 cellularly pleiotropic TAD-loop<sub>CTCF</sub>. *SI Appendix, Tables*  
 458 *S5d4-S5d5 (Note S5d)* lists the effect sizes found in the TAD-  
 459 loop depletion simulations. We also examined whether the  
 460 depletion magnitude of binary overlap or  $\text{PlyRS}_{\text{max}}$  overlap



**Fig. 3.** Depletion of deletions and shift of allele frequency spectrum overlapping TAD-loop regulatory sites. (A) We calculated a binary variable and  $\text{PlyRS}_{\max}$  (SI Appendix, Note S7c) for every deletion to quantify overlap with sites of TAD-loop. We plot the degree of reduction in the binary variable (or  $\text{PlyRS}_{\max}$ ) for real deletions relative to simulation measured using Cohen's D (with 95% confidence intervals on the mean reduction). (B) For each deletion we determined the magnitude of binary variable (or  $\text{PlyRS}_{\max}$ ) depletion, calculated as the difference between the binary variable (or  $\text{PlyRS}_{\max}$ ) and the average binary variable (or  $\text{PlyRS}_{\max}$ ) of its length-matched simulated deletions (SI Appendix, Note S6b), for sites of TAD-loop. We tested whether binary variable (or  $\text{PlyRS}_{\max}$ ) depletion magnitude depends on allele frequency (deletions categorized as rare [AF<=1%] or common), using multivariate logistic regression in the presence of genomic covariates (SI Appendix, Note S6a). We plot the regression odds ratio (OR) with 95% profile likelihood-based confidence intervals.

461 at TAD-loop loci exhibited dependence on allele frequency  
 462 using the same logistic regression framework as above with  
 463 chromatin accessibility and histone modification annotations.  
 464 Fig. 3B shows compelling evidence of a shift in the deletion  
 465 AFS based on the magnitude of depletion at  $\text{TAD-loop}_{\text{CTCF}}$ ,  
 466 for which the mean odds ratio estimate for binary overlap  
 467 in 1000GP was 2.70 (minimum [min] 95% CI: 1.35) and in  
 468 ADNI was 7.67 (min CI: 1.76). The mean odds ratio estimate  
 469 for  $\text{PlyRS}_{\max}$  overlap of  $\text{TAD-loop}_{\text{CTCF}}$  in 1000GP was 36.80  
 470 (min CI: 3.49) and in ADNI was 30.11 (min CI: 1.27). The  
 471 excess of rare alleles overlapping  $\text{TAD-loop}_{\text{CTCF}}$  dramatically  
 472 exceeded the shift for  $\text{TAD-loop}_{\text{noCTCF}}$ , which displayed only a  
 473 modest effect in the ADNI dataset (min CI: 1.02) and was not  
 474 significant in the 1000GP dataset. These results collectively  
 475 suggest that purifying selection may be acting to preserve  
 476 TAD-loop integrity by specifically preserving CTCF binding  
 477 motifs within TAD-loop boundaries. SI Appendix, Tables S6c4-

S6c5 (Note S6c) lists the odds ratios found in the TAD-loop  
 478 logistic regressions. 479

## 480 Discussion

481 Using the clarity of genomic deletions to identify loss of noncoding  
 482 regulatory function, we have examined whether purifying  
 483 selection is operating to preserve noncoding regulatory sites of  
 484 chromatin accessibility (DHS), histone modification (enhancer,  
 485 transcribed, polycomb-repressed, and heterochromatin), and  
 486 topologically associated domain loops (TAD-loops). Analysis  
 487 of selection in the noncoding genome is motivated by prior find-  
 488 ings in human genetics from genome-wide association studies  
 489 that conclude most of heritability is due to relatively common  
 490 noncoding alleles within regulatory annotations (3). Initially,  
 491 these findings appeared inconsistent with the expectation that  
 492 disease-associated alleles are under pressure from purifying  
 493 selection. However, recent studies demonstrated that complex



494 trait effect sizes are negatively correlated with allele frequency,  
495 hinting at the action of purifying selection (30–32). These  
496 observations put the question of the effect of noncoding regula-  
497 tory alleles on function and fitness at the forefront of genomic  
498 studies ranging from basic evolutionary genetics to the alle-  
499 lic architecture of common human traits. Since a principal  
500 characteristic of human regulatory element function is their  
501 non-uniform activity across tissues and cell types, interpreting  
502 fitness consequences from genetic variants in noncoding regions  
503 is thus inherently linked to corresponding regulatory element  
504 cellular activity. To incorporate this defining feature into the  
505 study of noncoding purifying selection, we have developed  
506 a statistical method, Pleiotropy Ratio Score (PlyRS), which  
507 quantifies the extent of abundance of cellularly pleiotropic  
508 activity for individual base-pairs.

509 Using our PlyRS method, our results indicate that purifying  
510 selection acts on both DHS and enhancer sites, as evident by  
511 both the depletion of deletions overlapping these annotations  
512 and a shift in the allele frequency spectrum of overlapping  
513 deletions towards rare alleles. Using simulated deletions in  
514 a length-matched framework and covariate-aware analyses,  
515 we notably found statistically significant evidence at DHS or  
516 enhancer regions that both sites of tissue/cell-type-specific  
517 activity and sites of cellularly pleiotropic activity are preserved  
518 by selection. We find some evidence that cellularly pleiotropic  
519 variants may be subject to a stronger reduction in variation  
520 than cell-type-specific variants. However, ambiguity between  
521 tissue/cell-type-specific and cellularly pleiotropic sites in terms  
522 of AFS shifts indicates that the strength of purifying selection  
523 across both types of regulatory site cellular activities may  
524 be roughly equivalent. Additional analysis on larger datasets  
525 would be needed to accurately quantify the relative contribu-  
526 tions of selection on sites of variable regulatory activity.

527 In contrast to the findings above, we did not find evidence  
528 of purifying selection acting on other epigenomic annotations  
529 such as transcribed, polycomb-repressed, or heterochromatin  
530 sites, consistent with previously reported findings (14, 16). In  
531 the absence of statistical confirmation, we can conclude that,  
532 notwithstanding any specific regulatory locus potentially being  
533 under selective constraint, these classes of epigenomic regu-  
534 lators as a whole are not selectively preserved in noncoding  
535 space. These results underscore the importance of DHS and  
536 enhancer annotations for specifying critical cellular regulation.  
537 Notably, our findings parallel the observation in human genet-  
538 ics that the largest fraction of heritability resides in regulatory  
539 space marked by DHS or enhancer features (3, 33). Our results  
540 additionally support the hypothesis that an aggregate selective  
541 burden may occur on long deletions that overlap multiple DHS  
542 or enhancer sites simultaneously (14, 34). We find suggestive  
543 evidence that this may be the case for deletions longer than  
544 the median length in our datasets, especially on those that  
545 overlap cellularly pleiotropic sites (*SI Appendix*, Fig. S4).

546 We have also presented evidence that purifying selection  
547 is operating to preserve TAD-loop boundary integrity by pre-  
548 serving co-localized CTCF binding sites. However, we did  
549 not find statistical evidence that selection is acting against  
550 deletions overlapping TAD-loop boundaries without simultane-  
551 ous removal of CTCF sites. We found conclusive statistically  
552 significant evidence for this preservation of TAD-loop<sub>CTCF</sub>  
553 sites in 1000GP but only a qualitative trend for this in ADNI.  
554 The difference in significance for these findings between dele-

tion datasets may simply be due to the difference in power  
to see this effect, as there are 4x the number of deletions in  
1000GP in comparison to ADNI. We did not find statistically  
significant evidence in either dataset that the sites of high-  
est cellular pleiotropy of TAD-loop<sub>CTCF</sub> provides additional  
signal for purifying selection beyond that for TAD-loop<sub>CTCF</sub>  
sites of any cellular pleiotropy. This equivalence may again be  
due to lack of power: either five primary tissues/cell-types in  
TAD-loop boundary analysis are not numerous enough to see a  
difference (compared to the 25 primary tissues/cell-types used  
in the analysis of chromatin accessibility and histone modifi-  
cation features), or deletions overlapping cellularly pleiotropic  
TAD-loops are already so few in number that power is limited  
(only 4% [1000GP] or 8% [ADNI] of all deletions in our  
datasets). As with the DHS and enhancer findings mentioned  
above, larger datasets may provide the power needed to clarify  
the relative contributions of selection on TAD-loop and CTCF  
sites of variable activity, as well as provide better resolution of  
TAD compartments versus TAD-loop boundaries which may  
improve analyses. The PlyRS method is flexible and easily  
allows for the addition of new and larger regulatory datasets  
as they become available.

## Materials and Methods

We used deletions from two datasets, the 1000 Genomes Project  
(1000GP, [(14)]) and the Alzheimer's Disease Neuroimaging Initia-  
tive (ADNI [(22)], *SI Appendix*, Note S3), to examine selective  
constraint within regulatory regions. The two deletion datasets  
have different callset properties, enabling robustness of the analy-  
sis. 1000GP consists of deletions derived from low-coverage whole  
genome sequencing (WGS) that span a wider length range and are  
genotyped from individuals of diverse demographic histories. ADNI  
consists of deletions derived from high-coverage WGS data that are  
on average longer and more rare, using genotypes from the subset of  
individuals that we determined were of European ancestry as iden-  
tified by principal components analysis. For both deletion datasets,  
we restricted our analyses to noncoding deletions by removing any  
deletion that overlapped any exon or UTR by one base-pair or more,  
as exonic deletions have been previously shown to be under strong  
purifying selection because of their protein-altering effects (35). We  
also examined only deletions occurring on autosomes because sex-  
chromosome functional elements may involve complex sex-biased  
regulation (36) which might be subject to unique selective prop-  
erties. To mitigate non-uniform (i.e. biased) deletion callability in the  
noncoding genome which might distort the AFS of the remaining  
set of deletions, we additionally excluded deletions overlapping any  
regions of low mappability, segmental duplications, centromeres,  
and reference assembly gaps. Additional details on the deletion  
datasets and filtering criteria are given in *SI Appendix*, Note S2.  
Specific characteristics of the deletion datasets are shown in *SI  
Appendix*, Table S1 (1000GP) and Table S2 (ADNI). An extended  
description of the ADNI dataset construction process is given in *SI  
Appendix*, Note S3. Information on obtaining ADNI data access,  
including files we deposited [in-progress] for this project, can be  
found at: <http://adni.loni.usc.edu/data-samples/access-data/>.

We used regulatory data from the NIH Roadmap Epigenomics  
Consortium (REC) for definition of regulatory breakpoints as well  
as uniform processing across multiple tissue/cell-types (2). We use  
annotation data for sites of chromatin accessibility (DNase I hyper-  
sensitivity "DHS") and histone modification (H3K4me1 "enhancer",  
H3K36me3 "transcribed", H3K27me3 "polycomb-repressed", and  
H3K9me3 "heterochromatin"). Two sets of DHS annotation (hotspot  
and MACS) were used to check for consistency. We used all 25  
primary tissues/cell-types (*SI Appendix*, Note S4a and Table S3)  
for which data were available across all six callsets for each tissue/cell-  
type. We additionally used TAD-loop boundary regulatory data  
consisting of a callset of 5 primary tissues/cell-types [(24)], *SI  
Appendix*, Note S4b). Additional details on the regulatory datasets  
are given in *SI Appendix*, Note S4. Identity of the tissues and



623 cell-types analyzed from REC is shown in *SI Appendix*, Table S3.  
624 For all analyses involving DNase hypersensitivity or histone modification  
625 regulatory features, we excluded deletions (and genomic space)  
626 overlapping TAD-loop boundaries (*SI Appendix*, Note S5c),  
627 as deletions disrupting TAD-loop integrity may already be under  
628 purifying selection owing to the potentially resulting cis-regulatory  
629 effects. In this way, we ensure reliable interpretation of selective  
630 effects on deletions disrupting chromatin accessibility or histone  
631 modification, without introducing potential confounding from selective  
632 pressure from co-localized TAD-loop disruption, which we  
633 analyzed separately.

634 To examine potential purifying selection against deletions to  
635 preserve regulatory features, we examined deletion overlap in the  
636 context of regulatory tissue activity. To properly "count" tissue  
637 activity removed by deletions overlapping regulatory features, we developed  
638 a statistical method called Pleiotropy Ratio Score (PlyRS),  
639 which calculates a correlation-adjusted count of cellular pleiotropy  
640 for each base-pair in the noncoding genome. A description of the  
641 calculation of PlyRS, and derived PlyRS measures calculated for  
642 deletions, is given in *SI Appendix*, Note S1. Source code of PlyRS  
643 calculation can be downloaded from the repository [in-progress] on  
644 Github: <https://github.com/davidwradke/PlyRS>.

645 To determine if the action of purifying selection is occurring  
646 against deletions overlapping regulatory sites, we required the identification  
647 of two key signatures: reduction of genetic variation overlapping  
648 the sites and a shift in the allele frequency spectrum (AFS)  
649 towards rare variants of the remaining alleles overlapping the sites.  
650 These signatures were assessed in light of results from deletion  
651 simulations. A description of the simulation procedure and significance  
652 calculation of reduction in variation is given in *SI Appendix*,  
653 Note S5. Descriptions of the procedure involving multivariate  
654 regression on deletion genomic covariates and significance  
655 calculation of shift in allele frequency spectrum are given in *SI  
656 Appendix*, Note S6.

657 **ACKNOWLEDGMENTS.** We thank members of the Sunyaev  
658 lab and Matt Maurano for helpful discussion on the topic. We  
659 thank Sheila Sutti Diamond for administrative help related to  
660 the Alzheimer's Disease Neuroimaging Initiative data used and  
661 generated in this project. We also thank Sung Chun, Nikolaos  
662 Patsopoulos, Deb Farlow, and Ruth McCole for helpful discussion  
663 related to the ADNI work. We thank the ADNI individuals for their  
664 participation in the consortium.

665 This material is based upon work supported by the National  
666 Science Foundation Graduate Research Fellowship Program under  
667 Grant No. DGE1144152 (D.W.R.). Any opinions, findings, and  
668 conclusions or recommendations expressed in this material are those  
669 of the author(s) and do not necessarily reflect the views of the National  
670 Science Foundation. D.W.R. was also supported by the National  
671 Institutes of Health (NIH) under Ruth L. Kirschstein National  
672 Research Service Awards 4T32HG002295-14 and 5T15LM007092-27  
673 and is currently supported by NIH Grant 5R01HG010372. S.R.S. is  
674 currently supported by NIH Grants R35GM127131, R01MH101244  
675 and U01HG006500. R.C.G. is currently supported by NIH Grants  
676 R01-HG009922 and R01-HL143295 and by the Franca Sozzani  
677 Fund.

678 Data collection and sharing for this project was funded by the  
679 Alzheimer's Disease Neuroimaging Initiative (ADNI) (National  
680 Institutes of Health Grant U01 AG024904) and DOD ADNI (Department  
681 of Defense award number W81XWH-12-2-0012). ADNI is  
682 funded by the National Institute on Aging, the National Institute  
683 of Biomedical Imaging and Bioengineering, and through generous  
684 contributions from the following: AbbVie, Alzheimer's Association;  
685 Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica,  
686 Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.;  
687 Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company;  
688 EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company  
689 Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen  
690 Alzheimer Immunotherapy Research Development, LLC.; Johnson  
691 Johnson Pharmaceutical Research Development LLC.; Lumosity; Lundbeck;  
692 Merck Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research;  
693 Neurotrack Technologies; Novartis Pharmaceuticals Corporation;  
694 Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical  
695 Company; and Transition Therapeutics.

The Canadian Institutes of Health Research is providing funds to  
support ADNI clinical sites in Canada. Private sector contributions  
are facilitated by the Foundation for the National Institutes of  
Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern  
California Institute for Research and Education, and the study is  
coordinated by the Alzheimer's Therapeutic Research Institute at  
the University of Southern California. ADNI data are disseminated  
by the Laboratory for Neuro Imaging at the University of Southern  
California.

1. The ENCODE Project Consortium, et al., An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Roadmap Epigenomics Consortium, et al., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
3. MT Maurano, et al., Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. M Kircher, et al., A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
5. GR Ritchie, I Dunham, E Zeggini, P Flicek, Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
6. D Quang, Y Chen, X Xie, J Hancock, DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
7. D Lee, et al., A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
8. J Zhou, OG Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
9. I Ionita-Laza, K Mccallum, B Xu, JD Buxbaum, A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
10. YF Huang, B Gulko, A Siepel, Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
11. E Rojano, P Seoane, JAG Ranea, JR Perkins, Regulatory variants: from detection to predicting impact. *Briefings Bioinforma.* **20**, 1639–1654 (2019).
12. S Zhu, et al., Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.* **34**, 1279–1286 (2016).
13. SJ Liu, et al., CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, eaah7111 (2017).
14. PH Sudmant, et al., Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
15. PH Sudmant, et al., An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
16. HJ Abel, et al., Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv:508515* (31 December 2018).
17. D Xu, O Gokcumen, E Khurana, Loss-of-function tolerance of enhancers in the human genome. *PLoS Genet.* **16**, e1008663 (2020).
18. J Huddleston, EE Eichler, An incomplete understanding of human genetic variation. *Genetics* **202**, 1251–1254 (2016).
19. TJ Treangen, SL Salzberg, Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
20. RE Mills, et al., Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
21. J Ibn-Salem, et al., Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.* **15**, 423 (2014).
22. RC Petersen, et al., Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology* **74**, 201–209 (2010).
23. RE Handsaker, JM Korn, J Nemes, SA McCarroll, Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
24. SS Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
25. WP Kloosterman, et al., Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).
26. DG Lupiáñez, et al., Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
27. S Schoenfelder, P Fraser, Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
28. Y Guo, et al., CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).
29. EP Nora, et al., Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930–944 (2017).
30. J Zeng, et al., Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
31. S Gazal, et al., Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
32. AP Schoech, et al., Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 1–10 (2019).
33. HK Finucane, et al., Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
34. S Girirajan, et al., Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes. *PLoS Genet.* **7**, e1002334 (2011).
35. DF Conrad, et al., Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
36. EA Khramtsova, LK Davis, BE Stranger, The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).