

1 Probing the aggregated effects of purifying selection per individual on 1,380 medical  
2 phenotypes in the UK biobank

3

4 Ha My T. Vy,<sup>1¶</sup> Daniel M. Jordan,<sup>1,2¶</sup> Daniel J. Balick<sup>3,4¶</sup>, Ron Do<sup>1,2\*</sup>

5

6 <sup>1</sup> The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine  
7 at Mount Sinai, New York, NY, USA

8 <sup>2</sup> Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount  
9 Sinai, New York, NY, USA

10 <sup>3</sup> Division of Genetics, Brigham and Women's Hospital, Harvard Medical School,  
11 Boston, MA, USA

12 <sup>4</sup> Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

13

14 \* Corresponding author

15 Ron Do, PhD

16 Annenberg Building, Floor 18 Room 80A

17 1468 Madison Ave

18 New York, NY-10029

19 Phone Number: 212-241-6206

20 Fax Number: 212-849-2643

21 Email: [ron.do@mssm.edu](mailto:ron.do@mssm.edu)

22 ¶ These authors contributed equally to this work

23

## 24 **Abstract**

25           Understanding the relationship between natural selection and phenotypic  
26 variation has been a long-standing challenge in human population genetics. With the  
27 emergence of biobank-scale datasets, along with new statistical metrics to approximate  
28 strength of purifying selection at the variant level, it is now possible to correlate a proxy  
29 of individual relative fitness with a range of medical phenotypes. We calculated a per-  
30 individual deleterious load score by summing the total number of derived alleles per  
31 individual after incorporating a weight that approximates strength of purifying selection.  
32 We assessed four methods for the weight, including GERP, phyloP, CADD, and fitcons.  
33 By quantitatively tracking each of these scores with the site frequency spectrum, we  
34 identified phyloP as the most appropriate weight. The phyloP-weighted load score was  
35 then calculated across 15,129,142 variants in 335,161 individuals from the UK Biobank  
36 and tested for association on 1,380 medical phenotypes. After accounting for multiple  
37 test correction, we observed a strong association of the load score amongst coding  
38 sites only on 27 traits including body mass, adiposity and metabolic rate. We further  
39 observed that the association signals were driven by common variants (derived allele  
40 frequency > 5%) with high phyloP score (phyloP > 2). Finally, through permutation  
41 analyses, we showed that the load score amongst coding sites had an excess of  
42 nominally significant associations on many medical phenotypes. These results suggest  
43 a broad impact of deleterious load on medical phenotypes and highlight the deleterious  
44 load score as a tool to disentangle the complex relationship between natural selection  
45 and medical phenotypes.

46

## 47 **Author summary**

48           This study aims to augment our understanding between the complex relation  
49 between natural selection and human phenotypic variation. We developed a load score  
50 to approximate the relative fitness of an individual and correlate it with a set of medical  
51 phenotypes. Association tests between the load score amongst coding sites and 1,380  
52 phenotypes in a sample of 335,161 individuals from the UK Biobank showed a strong  
53 association with 27 traits including body mass, adiposity and metabolic rate.  
54 Furthermore, an excess of nominal associations at suggestive levels was observed  
55 between the load score amongst coding sites and medical phenotypes than would be  
56 expected under a null model. These results suggest that the aggregate effect of  
57 deleterious mutations as measured by the load score has a broad effect on human  
58 phenotypes.

59

## 60 **Introduction**

61           One of the primary questions of interest in the study of human population  
62 genetics is the relation between natural selection and the evolution of human  
63 phenotypes, from quantitative traits to complex disease. With the emergence of  
64 biobank-scale datasets, along with new statistical metrics to approximate strength of  
65 purifying selection at the variant level, it is now possible to both estimate the net impact  
66 of deleterious mutations for each individual in a large population sample and correlate it  
67 to a range of medical phenotypes exhibited by that individual. This provides an  
68 opportunity to simultaneously study the genetics of individuals within a relatively

69 homogenous population and the potential impact of natural selection on annotated  
70 phenotypes.

71         A large body of literature exists on evolution and estimation of the deleterious  
72 mutation load from human population samples, with particular emphasis on cross-  
73 ancestry comparisons [1-6]. Rather than a comparison between human populations, we  
74 aimed to assess the distribution of deleterious loads—the sum of all purifying selective  
75 effects in each individual’s genome—within a single human population. While the  
76 mutation load generally represents a population-wide average of this quantity, we  
77 estimated the same object for each individual in the population to produce a “load  
78 score” that counts the net effect of deleterious variation in each individual’s genome, a  
79 count of derived alleles weighted by an estimate of the selective disadvantage for each  
80 variant. When compared to the mean of the population, this per-individual load score  
81 can be interpreted as a component of the relative fitness of each individual.

82         In this study, we aim to augment our understanding of the relation between  
83 natural selection and human phenotypes by focusing on the net impact of purifying  
84 selection on the fitness of each individual, and correlating this quantity to the set of  
85 phenotypes acting on that individual. Previously this has been difficult for two reasons:  
86 first, we do not have a direct measure of the fitness of individual humans that can be  
87 estimated from genetic information, and second, there were no large databases  
88 available to quantify the wide range of phenotypes possessed by each individual.  
89 Biobank-scale datasets that contain both individual genotypes and phenotypes, such as  
90 the UK Biobank [7, 8], finally provides access to both large-scale phenotypic  
91 descriptions of each individual and some part of their genetic sequence.

92           We ventured to apply computational tools that predict aspects of purifying  
93 selection for individual alleles to published genotypes of 335,161 white British  
94 individuals from the UK Biobank to estimate the fitness impact of derived variation  
95 present in each imputed genome in this sample. Most of the variation in the sample  
96 exists at appreciable frequencies, and is likely under relatively small selective  
97 disadvantage, but in aggregate the fitness impact can be substantial. Using this  
98 representation of each individual's relative fitness, we probed correlations between the  
99 impact of common deleterious variation in an individual's genome and their personal  
100 phenotypic makeup. This provides a different lens into questions about the relation  
101 between fitness vis-à-vis mutation load and human traits by looking at a per-individual  
102 measure correlated to fitness, rather than focusing on the distribution of selective effects  
103 in the population as a whole. This allows us to ask which phenotypes, if any, are highly  
104 correlated to the aggregation of deleterious variation, and probe the relation between  
105 the ensemble of phenotypes and fitness loss due to common variation in individuals.

106

## 107 **Results**

### 108 **Comparison of four deleteriousness prediction scoring methods**

109           The additive effects of deleterious variation can be quantified in aggregate by a  
110 genome-wide score representing the net action of purifying selection on an individual  
111 under the assumption that effects of individual variants can be summed additively.

112 Multiple methods have been developed to characterize purifying selection, including  
113 methods that predict deleterious selection acting on the level of a single allele (fitCons  
114 [9], FATHMM-MKL [10], deltaSVM [11], Funseq2 [12]), methods that measure

115 evolutionary conservation (phyloP [13], phastCons [14], GERP++ [15], SiPhy [16]) and  
116 methods that predict the effect of an allele on molecular function (CADD [17], DANN  
117 [18], GenoCanyon [19], Eigen and EigenPC [20]). Although these scores are formulated  
118 as tests for strong selection or for molecular function, rather than as estimates of the  
119 strength of selection, they are also correlated to the strength of selection, and are often  
120 used as proxies for strength of selection [4, 21, 22]. In this study, we compared the  
121 predicted deleteriousness of alleles for four widely used scoring methods that  
122 approximate deleteriousness of a variant-- GERP++, phyloP, CADD, and fitCons-- with  
123 their effects on allelic frequency in human population genetic data to select the most  
124 appropriate measure for computation of the additive load.

125 Under negative or purifying selection, natural selection acts to reduce the  
126 population frequency of deleterious mutations. This effect is more able to overcome  
127 genetic drift as the strength of selection increases. As a result, we expect to observe a  
128 higher number of rare alleles and a lower number of common alleles in regions of the  
129 genome that are under negative selection, relative to putatively neutral regions. This  
130 can be seen as a shift of the allele frequency spectrum (AFS) towards rare alleles, with  
131 a steeper slope of the AFS indicating stronger purifying selection. We evaluated the  
132 extent to which each scoring method captures the deleteriousness of an allele by  
133 grouping alleles by the scores provided by each method and measuring the slope of the  
134 resulting AFS. The more strongly a score is related to the strength of selection, the  
135 more marked the increase in slope will be for high-scoring alleles relative to lower  
136 scoring alleles.

137

138 **Figure 1. Derived allele frequency spectra of different score categories for each deleteriousness**  
139 **prediction scoring method.**

140 For each scoring method, polymorphic sites are grouped into score intervals by the value of the score  
141 annotated at the sites. Each solid line represents derived allele frequency spectrum of polymorphic sites  
142 belonging to one score interval and three dashed lines represent derived allele frequency spectra of three  
143 control categories: synonymous (syn), missense (mis), and loss of function (LOF) variants.

144

145 We evaluated this correlation using whole genome sequencing data from a non-  
146 Finnish European population in the Genome Aggregation Database (gnomAD) [23]. For  
147 each scoring method, we grouped alleles by score and compared the non-normalized  
148 derived allele frequency spectra for each group (see Methods). Fig 1 plots the log of the  
149 derived allele frequency (DAF), and shows a consistent pattern across all scores: the  
150 higher the score, the steeper the slope of the log DAF (S1 Table). This indicates that,  
151 for all scores shown, higher scores are associated with sites under stronger negative  
152 selection, as expected. While all four scores show this pattern, CADD and phyloP show  
153 clearer separations between DAF spectra than fitCons and GERP++. In the case of  
154 fitCons, this underperformance is likely due to its incorporation of functional genomic  
155 signatures that may increase its performance at identifying functional regions, but  
156 detract from its performance at identifying sites under purifying selection. In the case of  
157 GERP++, the underperformance is more surprising, since GERP++ and phyloP are very  
158 similar methods. The difference in performance may be explained by differences in how  
159 the final scores computed by the two methods are defined, or by the fact that the scores  
160 were calculated using two different multiple sequence alignments: the phyloP scores  
161 were calculated from UCSC's alignment of 100 vertebrate sequences generated using

162 the MultiZ method [24], while the GERP++ scores were calculated from the Ensembl  
163 alignment of 111 mammalian sequences generated using the EPO-Extended method  
164 [25]. For comparison, we also calculated DAF spectra for synonymous sites, missense  
165 sites, and loss of function sites (LOF). Based on this comparison, phyloP scores  
166 between 5 and 7.5 appear to be similarly deleterious to missense sites (nearly the same  
167 DAF slope) and phyloP scores greater than 7.5 appear to be under similar purifying  
168 selection to LOF sites. The equivalent numbers for CADD are 20 to 25 for missense and  
169 greater than 30 for LOF.

170 To further compare between CADD and phyloP, we examined the DAF  
171 distribution for protein coding variants and noncoding variants separately. Both scores  
172 performed similarly for coding variants, but phyloP showed better separation in  
173 noncoding variants (S1 Fig). This is as expected, since CADD uses more features when  
174 scoring coding variants than noncoding variants, while the phyloP method is identical for  
175 coding and noncoding sites. For this reason, we concluded that phyloP has the most  
176 consistent relationship between score and strength of negative selection, and selected  
177 phyloP as our weight for our load score computation.

178

### 179 **Per-individual load scores in UK Biobank**

180 We calculated a per-individual deleterious load score by summing the total  
181 number of derived alleles per individual, weighting each derived allele by its phyloP  
182 score to account for the strength of purifying selection. We considered three load  
183 scores: a genome-wide load score, a coding-specific load score, and a non-coding-  
184 specific load score. Each score was computed for 335,161 unrelated, white-British



185 ancestry individuals in the UK Biobank using 6,774,062 variants from imputed  
186 genotypes (95,850 coding and 6,678,212 non-coding) with positive phyloP scores  
187 (positive scores denote uniform purifying selection, while negative scores denote clade-  
188 specific selection). The observed population distribution across all sampled individuals  
189 appear very close to normal for each of our three scores (Kolmogorov-Smirnov Tests P-  
190 values = 0.32; 0.55; 0.20 for all variants, non-coding, and coding, respectively, Fig 2).  
191 This is the expected result if the phyloP scores of derived alleles are identically  
192 distributed across the entire population, due to the Central Limit Theorem. By contrast, if  
193 the white-British population contained distinct subpopulations with dramatically different  
194 distributions of phyloP scores among derived alleles, we would expect to see a sum of  
195 multiple normal distributions with different means, resulting in a skewed or multi-modal  
196 distribution.

197

## 198 **Figure 2. Distribution of load score.**

199 Histogram of three load scores computed from three sets of variants: coding variants (coding load score),  
200 non-coding variants (non-coding load score), and both coding and non-coding variants (genome-wide  
201 load score). Each load score was computed for 335,161 unrelated, white-British ancestry individuals.

202

## 203 **Significant association between load score of coding variants and** 204 **anthropometric and metabolic traits**

205 To explore the overall effect of deleterious mutations on specific clinically  
206 measured phenotypes, we tested the association of each of the three load scores  
207 (genome-wide, coding and non-coding) with 1,380 traits, after adjusting for age, sex,  
208 genotyping chip, and assessment center. To account for potential confounders, we

209 further included a set of geographical and socioeconomic variables available in the UK  
210 Biobank data as additional covariates (S2 Table). We note that many of these variables  
211 are significantly associated with the load score but the effects are small. Nonetheless,  
212 careful consideration was taken to add these as covariates in our association tests (S2  
213 Table).

214 We discovered no phenotype significantly associated with either the genome-  
215 wide load score or non-coding load score (Bonferroni P value threshold =  $1.2 \times 10^{-5}$ ).  
216 However, 27 traits were significantly associated with the load score calculated from  
217 coding SNPs; these included body mass, metabolic rate, and several adiposity traits  
218 such as body mass index and waist circumference (Table 1). Some of these traits have  
219 been found under directional selection in contemporary populations [26-28].

220

221

222

223

224

**Table 1. Association between coding load score and 27 traits.**

Trait	Sample size	beta	SE	P-value
Arm fat-free mass (right)	329238	-0.0073	0.0012	$6.61 \times 10^{-10}$
Arm fat-free mass (left)	329182	-0.0074	0.0012	$7.93 \times 10^{-10}$
Arm predicted mass (left)	329170	-0.0073	0.0012	$1.12 \times 10^{-9}$
Leg fat mass (left)	329295	-0.0090	0.0015	$1.39 \times 10^{-9}$
Basal metabolic rate	329326	-0.0074	0.0012	$2.83 \times 10^{-9}$
Arm predicted mass (right)	329235	-0.0069	0.0012	$3.76 \times 10^{-9}$
Weight	334221	-0.0098	0.0017	$3.98 \times 10^{-9}$
Whole body water mass	329333	-0.0068	0.0012	$8.68 \times 10^{-9}$
Leg fat mass (right)	329311	-0.0086	0.0015	$1.02 \times 10^{-8}$
Whole body fat-free mass	329306	-0.0067	0.0012	$1.24 \times 10^{-8}$
Whole body fat mass	328780	-0.0103	0.0018	$1.80 \times 10^{-8}$
Leg fat percentage (left)	329297	-0.0064	0.0011	$2.66 \times 10^{-8}$
Trunk fat-free mass	329057	-0.0065	0.0012	$3.55 \times 10^{-8}$
Trunk predicted mass	329019	-0.0064	0.0012	$5.07 \times 10^{-8}$
Trunk fat mass	329118	-0.0100	0.0019	$1.21 \times 10^{-7}$

Leg predicted mass (right)	329303	-0.0063	0.0012	$1.54 \times 10^{-7}$
Leg fat-free mass (right)	329303	-0.0063	0.0012	$1.96 \times 10^{-7}$
Leg fat-free mass (left)	329280	-0.0062	0.0012	$3.18 \times 10^{-7}$
Leg predicted mass (left)	329275	-0.0062	0.0012	$3.77 \times 10^{-7}$
Leg fat percentage (right)	329316	-0.0058	0.0012	$5.74 \times 10^{-7}$
Arm fat mass (right)	329242	-0.0092	0.0018	$6.38 \times 10^{-7}$
Body mass index (BMI)	334097	-0.0092	0.0019	$7.25 \times 10^{-7}$
Arm fat mass (left)	329188	-0.0090	0.0018	$1.15 \times 10^{-6}$
Waist circumference	334612	-0.0080	0.0016	$1.35 \times 10^{-6}$
Body fat percentage	329134	-0.0066	0.0014	$2.98 \times 10^{-6}$
Hip circumference	334579	-0.0088	0.0019	$3.29 \times 10^{-6}$
Impedance of arm (left)	329313	0.0061	0.0013	$5.25 \times 10^{-6}$

225 SE: standard error

226 Stratification by derived allele frequency showed that these association signals  
227 are more pronounced when limiting to variants that are common (DAF > 5%) but not  
228 close to fixation (DAF < 70%), while stratification by phyloP score shows that they are  
229 more pronounced when limiting to variants with higher phyloP scores (phyloP>2, S3 and  
230 S4 Tables). We therefore performed an additional stratification analysis by both DAF  
231 and phyloP score (S5 Table). We observed that the signals are mostly driven by  
232 common variants ( $5 \leq \text{DAF} < 70\%$ ) with higher phyloP score (phyloP>2). This class of  
233 variants notably contributes a large fraction (mean: 0.38 and sd: 0.005 per individual)  
234 towards the per individual coding load score. This analysis necessarily excludes  
235 extremely rare alleles, which are not well captured by the process of genotyping and  
236 imputation. It is not clear how significant the aggregate contribution of these alleles to  
237 the per individual load score would be.

238 To assess the effect of our weighting procedure, we calculated an unweighted  
239 load score, the per-individual mutation burden, that simply counts derived alleles with no  
240 reference to phyloP or other measures of selection. When using this score, all  
241 significant association signals observed for the coding load score disappeared and no

242 significant association for genome-wide and non-coding unweighted score was detected  
243 (S2 Fig). We further tested the associations with burden scores while restricting to only  
244 rare variants (DAF < 5%) or only common variants (5% < DAF < 70%, S2 Fig), however  
245 no significant association was observed. This is likely due to the domination of the  
246 mutation burden by alleles under effectively no purifying selection, highlighting the need  
247 for a weighting scheme to identify correlations to the relative per-individual fitness.

248 To assess whether the observed significant associations are sensitive to  
249 reference bias, we included as a covariate the number of non-reference sites per  
250 individual in our association testing for the top results in Table 2 (S6 Table). Association  
251 results were very consistent, suggesting that reference bias is not likely a confounder.  
252 Similarly, associations between the phenotypes and load score remain significant when  
253 restricted to variants at which reference alleles are the same as predicted ancestral  
254 alleles (S7 Table). We also re-computed load scores using phyloPNH scores, which are  
255 phyloP scores calculated without human reference genome [4], and obtained similar but  
256 slightly less significant results, with all the 27 phenotypes yielded p-value <  $6.13 \times 10^{-4}$   
257 (S8 Table).

258

## 259 **Associations with coding load score are enriched for nominal associations with** 260 **disease**

261 Phenome wide association test results showed that no single disease is significantly  
262 associated with the load score (all  $P > 0.05$  after accounting for multiple tests using  
263 Bonferroni correction). However, rather than the load score having a strong effect on a  
264 single disease, we hypothesized that the load score may have subtle effects on many

265 diseases, leading to an excess of weak associations that do not individually reach  
266 statistical significance. To test this hypothesis, we compared the number of phenotypes  
267 nominally associated with the load score (p-value < 0.05 without multiple test correction)  
268 to a null distribution generated by random permutation of individual load score values  
269 (Methods). For this analysis, we restricted to associations with clinical phenotypes  
270 defined by phecodes. Out of 539 phecodes, 46, 24, and 27 phecodes (S9 Table) were  
271 found to be nominally associated with coding load score, non-coding load score, and  
272 genome-wide load score respectively. The number of nominally significant associations  
273 for coding load score is significantly larger than the expected number under the null  
274 model (P=0.005), supporting this hypothesis (Fig 3). However, this analysis was not  
275 statistically significant for the genome-wide load score and the non-coding load score  
276 (P>0.05) (S3 Fig), suggesting that diseases are largely correlated to the effect of  
277 variants in coding regions. We repeated the permutation analysis for the unweighted  
278 burden score as negative controls. As expected, enrichment of weak association  
279 between burden scores and diseases are not statistically significant (S4 Fig).

280

281 **Figure 3. Enrichment of clinical phenotypes nominally associated with coding load score.** Null  
282 distribution of the number of clinical phenotypes weakly associated with coding load score was obtained  
283 from 2,000 permutations in total. For each permutation, coding load score was shuffled randomly among  
284 335,161 samples and the number of association was the count of phenotypes which yielded a p-value <  
285 0.05 in the association tests between permuted load score and 539 phecodes. Red dashed line indicates  
286 the observed number of clinical phenotypes nominally associated with coding load score (n = 46).

287

288 **Discussion**

289 In this study, we have described a polygenic load score that estimates the deleterious  
290 load carried by an individual, and applied this score to 335,161 white British individuals  
291 from the UK Biobank. Our analysis produced two major results: First, while we found no  
292 significant associations between individual medical phenotypes and the genome-wide  
293 load score, we found that more phenotypes are nominally associated with the coding  
294 load score than would be expected under a null model (Figure 3). This suggests that the  
295 deleterious load has a broad effect on the human phenome, rather than being  
296 specifically associated with a small number of phenotypes. This is consistent with  
297 Fisher's Geometric Model of fitness, which proposes that the fitness of a population is  
298 determined by overall phenotypic distance from a theoretical optimal point in a  
299 phenotype space that potentially encompasses the organism's entire phenome [29, 30].  
300 Second, by restricting to protein coding variation, we found significant associations  
301 between the coding-only deleterious load score and a variety of adiposity phenotypes,  
302 along with other anthropometric phenotypes and phenotypes related to metabolic rate.  
303 This suggests that adiposity may be under polygenic selection driven by a large number  
304 of coding variants in humans. This is consistent with previous results obtained from the  
305 UK Biobank using an unrelated methodology [26]. We found no similar associations with  
306 the noncoding deleterious load score, which is in contrast to numerous studies finding  
307 significant genetic associations in noncoding regions, including associations with the  
308 same adiposity traits we found associated with our coding load score. Since our derived  
309 allele frequency spectrum analysis (Fig. 1) suggests that sites with higher phyloP scores  
310 are under purifying selection in noncoding regions as well as coding, the lack of  
311 significance in non-coding regions cannot be interpreted as a lack of purifying selection

312 in these regions or poor sensitivity to selection in these regions. It may instead indicate  
313 that selection acts on phenotype associations in noncoding regions in a different way  
314 from how it acts in coding regions, possibly due to the small effect size of individual  
315 noncoding variants.

316

317       There are several limitations to our method. We computed the load score from  
318 imputed genotypes rather than sequenced whole genomes, which gives us little  
319 information about extremely rare variants in the population, masking potentially large  
320 contributions to the load from variants under the strongest selection. As a future topic of  
321 research, the same methodology can be applied to include rare variants, which would  
322 shed light on the relative contribution of common and rare variation to the phenotypic  
323 associations of load. Previous studies have shown that rare variation contributes  
324 substantially to differences in deleterious load between human populations, so we may  
325 expect it to have a significant impact on individual load in this context as well [1, 2].  
326 Furthermore, the phyloP score used to estimate the deleteriousness of alleles measures  
327 only the likelihood that a site is evolving under constraint in vertebrates, and is not a  
328 direct estimate of the selective effect of a variant in humans. It is possible that the use of  
329 vertebrate-level conservation has reduced our ability to identify recent selection on  
330 human phenotypes, particularly those that are human specific. However, the fact that  
331 selection on adiposity traits was also detected by a method [26, 27] that does not rely  
332 on phyloP suggests that this result is not spurious. This feature of the phyloP score also  
333 makes it difficult to measure the effect of dominant or recessive selection, which may  
334 contain additional important insights. Finally, we did not incorporate any measure of

335 positive selection in the computation of the load score. Scores similar to phyloP that  
336 could be used to detect positive selection do exist, but they rely on measures of  
337 nucleotide diversity and haplotype structure within larger regions of the genome, and  
338 are difficult to apply to single nucleotides as would be required to incorporate them into  
339 this analysis [31]. Methods to detect positive selection in the human lineage on finer  
340 scales are an area of ongoing research, and such methods could be incorporated into  
341 this approach in the future. All these methodological constraints limit the range of  
342 variation we identify as contributing to an individual's burden score, and this limitation  
343 may be biased with respect to trait associations. In particular, we might expect rare  
344 variation or positive selection to reveal a different set of trait associations than the ones  
345 we find here by investigating common variation under purifying selection. This could  
346 potentially expand the scope of associated phenotypes beyond adiposity traits, a  
347 possibility that is also supported by the presence of nominal associations with many  
348 phenotypes unrelated to adiposity (S9 Table). Nevertheless, we do expect common  
349 variation under purifying selection to underlie a large fraction of common disease  
350 phenotypes, and therefore to provide valuable insights about the action of natural  
351 selection in humans.

352         One potentially exciting application for this approach is applying it to different  
353 populations to discover of population-specific insights into phenotypic associations with  
354 deleterious load. Since PhyloP scores can be calculated without any reference to  
355 specific human populations [4], there is no reason in principle that this method could not  
356 be applied to biobank data from other populations, given a sufficient number of  
357 samples. However, a few cautions are necessary. First, it is well known that



358 comparisons of genetic associations and polygenic risk are unreliable across different  
359 ancestries [32], that signals of polygenic selection can easily be confounded by  
360 population structure or admixture [33, 34], and that mutation load specifically differs  
361 substantially between populations based on their demographic history [1, 2]. This  
362 makes it difficult to compare load scores directly between individuals of different  
363 ancestry, and also would likely make it difficult to apply this approach to admixed  
364 populations or populations with heterogeneous ancestry. Second, the approach of  
365 genotyping and imputation is entirely dependent on the availability of appropriate  
366 genotyping arrays and imputation panels, neither of which is necessarily available for all  
367 populations. It will be essential to use sequencing data for any population that is not well  
368 represented in these resources. Finally, many traits are strongly influenced by social,  
369 cultural, and environmental factors which may differ dramatically across populations,  
370 resulting in differences between populations that are not necessarily related to natural  
371 selection in a straightforward way. This is certainly true of the adiposity traits we identify  
372 in this study. Results of such studies should therefore be interpreted with caution.

373         The deleterious load score presented here provides a new approach to  
374 investigate the complex relationship between natural selection acting on individuals,  
375 individual medical phenotypes, and the human phenome at large. We expect that as the  
376 available biobank data continues to grow in size and scope, this method can be applied  
377 to larger and more diverse populations to gain additional insights into how load varies  
378 between different populations, possibly empowering population-specific medical  
379 discoveries with deleterious load.

380

## 381 **Material and Methods**

### 382 **The dependence of derived allele frequency on deleteriousness score**

383 We evaluated the dependence of derived allele frequency of single nucleotide  
384 polymorphisms (SNPs) discovered in the whole genome sequences of 7,509 non-  
385 Finnish European individuals in the GnomAD data set [23] on each of the four candidate  
386 annotations for the presence of purifying selection: GERP++ [15], phyloP [13], CADD  
387 [17], and fitcons [9]. 88,060,485 SNPs with less than one percent of missing data were  
388 considered. Functional effects and deleterious scores at each SNP were annotated  
389 using Whole Genome Sequence Annotator (WGSA) v0.7 [35]. We used functional  
390 effects annotated by Variant Effect Predictor (VEP) and determined derived and  
391 ancestral allele status based on the six-way EPO (Enredo, Pecan, Ortheus) multiple  
392 alignments of primate species.

393 For each deleteriousness score, we divided the SNPs into multiple groups with  
394 arbitrarily defined intervals based on the range of each score. The intervals used were:  
395 (0 ~ 0.2, 0.2 ~ 0.4, 0.4 ~ 0.6, 0.6 ~ 0.8) for fitcons, (-10 ~ -7.5, -7.5 ~ -5, -5 ~ -2.5, -2.5 ~  
396 0, 0 ~ 2.5, 2.5 ~ 5, 5 ~ 7.5, 7.5 ~ 10) for GERP, (0 ~ 5, 5 ~ 10, 10 ~ 15, 15 ~ 20, 20 ~  
397 25, 25 ~ 30, 30 ~ 35) for CADD, and (-5 ~ -2.5, -2.5 ~ 0, 0 ~ 2.5, 2.5 ~ 5.0, 5.0 ~ 7.5,  
398 7.5~ 10) for phyloP.

399

### 400 **Load score calculation**

401 The load score of each individual was calculated by adding up the number  
402 (dosage in case of imputed SNPs) of derived alleles at each SNP, weighted by the  
403 phyloP score at that site, across the entire genome. Derived alleles were determined

404 based on the six-way EPO alignment, as described above. Since we are focusing on  
405 the effect of purifying selection, only SNPs with positive phyloP score (positive scores  
406 denote uniform purifying selection, while negative scores denote clade-specific  
407 selection) were included. In this paper, we computed three load scores using three  
408 different SNP sets: the coding load score summed only over coding variants, the non-  
409 coding load score summed only over non-coding variants, and the genome-wide load  
410 score computed from both coding and non-coding variants. All load scores were  
411 computed using PRSice-2 software [36] under an additive model. Coding and  
412 noncoding variants were defined based on VEP annotation.

413

#### 414 **Genotypic and phenotypic data**

415 The UK Biobank consists of genotype, phenotype, and demographic data of  
416 more than 500,000 individuals recruited across the United Kingdom. Individual  
417 genotypes were generated from either the Affymetrix Axiom UK Biobank array  
418 (~450,000 individuals) or the UK BiLEVE array (~50,000 individuals), each contains  
419 ~0.9 million markers. Additional variants were then imputed using the Haplotype  
420 Reference Consortium (HRC) combined with the UK10K haplotype resource, with a  
421 total of ~96 million variants available in the latest released imputed data (version 3). To  
422 compute per-individual load scores, we restricted to variants with imputation quality  
423 INFO score  $\geq 0.9$ . We excluded samples that were outliers in heterozygosity or  
424 missing rates, samples with putative sex chromosome aneuploidy, and samples with  
425 self-reported non-white British ancestry. We also excluded one individual from each pair  
426 of samples with relatedness up to the third degree. This produced a subsample of

427 335,161 individuals. All information used to exclude samples is included in the UK  
428 Biobank resource page.

429 UK Biobank provides a wide range of medical phenotypes from base line  
430 assessment, biochemical assays, dietary questionnaire, and health records. In the  
431 present study, we focused on 2,419 phenotypes which had been selected for heritability  
432 estimation by the Neale group ([http://www.nealelab.is/blog/2017/9/15/heritability-of-](http://www.nealelab.is/blog/2017/9/15/heritability-of-2000-traits-and-disorders-in-the-uk-biobank)  
433 [2000-traits-and-disorders-in-the-uk-biobank](http://www.nealelab.is/blog/2017/9/15/heritability-of-2000-traits-and-disorders-in-the-uk-biobank)). This subgroup covers phenotypes in most  
434 of the core categories, including early life and reproductive factors, family history,  
435 cognitive function, physical measures, lifestyle and health outcomes.

436

#### 437 **Phenotype processing and association tests**

438 Among the 2,419 phenotypes considered in our analysis, 619 phenotypes are  
439 international classification of disease (ICD-10) codes from electric health records. We  
440 converted ICD codes (including ICD-9 and ICD-10 codes) into phecodes using Phecode  
441 Maps 1.2 [37, 38]. This resulted in 1,677 unique phecodes in total. Of these, 539  
442 phecodes with the number of cases greater than 500 were selected for phenome-wide  
443 association testing.

444 The remaining 1,800 phenotypes (2,419 – 619 ICD codes) were pre-processed  
445 using PHESANT [39], a package designed to process phenotypes and run phenome  
446 scans in UK Biobank. The PHESANT pipeline loads each input phenotype as  
447 continuous, integer, or categorical based on the information in the UK Biobank data  
448 dictionary; preprocesses and re-categorizes the phenotype data based on predefined  
449 rules; and assigns them into one of the four data types: continuous, ordered categorical,

450 unordered categorical and binary. Of these 1,800 phenotypes, we only considered  
451 those with a minimum number of cases or controls equal to 500 and a minimum number  
452 of individuals equal to 5,000. This resulted in 841 phenotypes: 75 continuous, 104  
453 ordered categorical, 36 unordered categorical, and 626 binary. In total, 1,380  
454 phenotypes was included in our association analysis.

455         The association between load score and each phenotype was tested using a  
456 regression test in PHESANT: linear regression / lm R function for continuous, ordered  
457 logistic regression / polr R function for ordered categorical, multinomial logistic  
458 regression / multinom R function for unordered categorical, and binomial regression /  
459 glm R function with family = binomial for binary. Besides the commonly used covariates  
460 of age, sex, genotype chip, assessment center and 40 principal components, we added  
461 five variables as covariates in all association tests that might denote population  
462 structure: birth location, home area population density, Townsend deprivation index,  
463 and UK deprivation index.

464

#### 465 **Association of 27 adiposity traits with coding load scores stratified by phyloP** 466 **score and derived allele frequency**

467         To explore which variants drive the association between the 27 adiposity traits  
468 and the coding load score, we stratified variants by derived allele frequency (rare  
469 variants, 0 to 0.05; intermediate frequency variants, 0.05 to 0.3; common variants, 0.3 to  
470 0.7; and variants near fixation, 0.7 to 1) and phyloP score (0 to 2, 2 to 4, 4 to 6, 6 to 8,  
471 and 8 to 10). Simultaneous stratification was performed with four groups of SNPs:

472 DAF<0.05 and phyloP $\leq$ 2, DAF<0.05 and phyloP>2, DAF $\geq$ 0.05 and phyloP $\leq$ 2, DAF $\geq$ 0.05  
473 and phyloP>2.

474

## 475 **Permutation of phenome-wide association analysis and creation of null** 476 **distribution**

477 A null distribution of the number of clinical phenotypes weakly associated with  
478 load score was created by repeatedly running the association test between load scores  
479 and phenotypes after randomly shuffling the load scores of individuals within the tested  
480 sample. The phenotypes included in this permutation analysis were all 539 phecodes.  
481 The same set of covariates used in phenome-wide association study (PHEWAS) tests  
482 above was applied. For each permutation, the number of phenotypes nominally  
483 associated with the load score (p-value<0.05) was then computed. The permutation p-  
484 value was calculated as the fraction of permutations for which the number of nominally  
485 associated traits was at least as large as the observed number of nominally associated  
486 traits.

487

## 488 **Acknowledgments**

489 This research has been conducted using the UK Biobank Resource under  
490 Application Number '16218'.

## 491 **Funding**

492 RD is supported by R35GM124836 from the National Institute of General Medical  
493 Sciences of the National Institutes of Health, and R01HL139865 from the National  
494 Heart, Lung, and Blood Institute of the National Institutes of Health.

495 **Conflict of interest**

496 RD has received research support from AstraZeneca and Goldfinch Bio, being a  
497 scientific co-founder and equity holder for Pensieve Health and being a consultant for  
498 Variant Bio, all not related to this work.

499

500

## 501 References

502

503 1. Henn BM, Botigue LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load  
504 in human genomes. *Nat Rev Genet.* 2015;16(6):333-43. Epub 2015/05/13. doi:

505 10.1038/nrg3931. PubMed PMID: 25963372; PubMed Central PMCID: PMC4959039.

506 2. Henn BM, Botigue LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from  
507 sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U S*  
508 *A.* 2016;113(4):E440-9. Epub 2015/12/30. doi: 10.1073/pnas.1510805112. PubMed PMID:  
509 26712023; PubMed Central PMCID: PMC4743782.

510 3. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, et al.  
511 Proportionally more deleterious genetic variation in European than in African populations.  
512 *Nature.* 2008;451(7181):994-7. Epub 2008/02/22. doi: 10.1038/nature06611. PubMed PMID:  
513 18288194; PubMed Central PMCID: PMC2923434.

514 4. Fu W, Gittelman RM, Bamshad MJ, Akey JM. Characteristics of neutral and deleterious  
515 protein-coding variation among individuals and populations. *The American Journal of Human*  
516 *Genetics.* 2014;95(4):421-36.

517 5. Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been  
518 less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet.*  
519 2015;47(2):126-31. Epub 2015/01/13. doi: 10.1038/ng.3186. PubMed PMID: 25581429;  
520 PubMed Central PMCID: PMC4310772.

521 6. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is  
522 insensitive to recent population history. *Nat Genet.* 2014;46(3):220-4. Epub 2014/02/11. doi:  
523 10.1038/ng.2896. PubMed PMID: 24509481; PubMed Central PMCID: PMC3953611.

524 7. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open  
525 access resource for identifying the causes of a wide range of complex diseases of middle and  
526 old age. *PLoS Med.* 2015;12(3):e1001779. Epub 2015/04/01. doi:

527 10.1371/journal.pmed.1001779. PubMed PMID: 25826379; PubMed Central PMCID:  
528 PMC4380465.

529 8. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank  
530 resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203-9. Epub  
531 2018/10/12. doi: 10.1038/s41586-018-0579-z. PubMed PMID: 30305743; PubMed Central  
532 PMCID: PMC6786975.

533 9. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness  
534 consequences for point mutations across the human genome. *Nat Genet.* 2015;47(3):276-83.  
535 Epub 2015/01/20. doi: 10.1038/ng.3196. PubMed PMID: 25599402; PubMed Central PMCID:  
536 PMC4342276.

537 10. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative  
538 approach to predicting the functional effects of non-coding and coding sequence variation.  
539 *Bioinformatics.* 2015;31(10):1536-43. Epub 2015/01/15. doi: 10.1093/bioinformatics/btv009.  
540 PubMed PMID: 25583119; PubMed Central PMCID: PMC4426838.

541 11. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to  
542 predict the impact of regulatory variants from DNA sequence. *Nat Genet.* 2015;47(8):955-61.  
543 Epub 2015/06/16. doi: 10.1038/ng.3331. PubMed PMID: 26075791; PubMed Central PMCID:  
544 PMC4520745.



- 545 12. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing  
546 noncoding regulatory variants in cancer. *Genome Biol.* 2014;15(10):480. Epub 2014/10/03. doi:  
547 10.1186/s13059-014-0480-5. PubMed PMID: 25273974; PubMed Central PMCID:  
548 PMCPMC4203974.
- 549 13. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution  
550 rates on mammalian phylogenies. *Genome Res.* 2010;20(1):110-21. Epub 2009/10/28. doi:  
551 10.1101/gr.097857.109. PubMed PMID: 19858363; PubMed Central PMCID: PMCPMC2798823.
- 552 14. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al.  
553 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome*  
554 *Res.* 2005;15(8):1034-50. Epub 2005/07/19. doi: 10.1101/gr.3715005. PubMed PMID:  
555 16024819; PubMed Central PMCID: PMCPMC1182216.
- 556 15. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high  
557 fraction of the human genome to be under selective constraint using GERP++. *PLOS*  
558 *Computational Biology.* 2010;6(12).
- 559 16. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel  
560 constrained elements by exploiting biased substitution patterns. *Bioinformatics.*  
561 2009;25(12):i54-62. Epub 2009/05/30. doi: 10.1093/bioinformatics/btp190. PubMed PMID:  
562 19478016; PubMed Central PMCID: PMCPMC2687944.
- 563 17. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework  
564 for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-  
565 5. Epub 2014/02/04. doi: 10.1038/ng.2892. PubMed PMID: 24487276; PubMed Central PMCID:  
566 PMCPMC3992975.
- 567 18. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the  
568 pathogenicity of genetic variants. *Bioinformatics.* 2015;31(5):761-3. Epub 2014/10/24. doi:  
569 10.1093/bioinformatics/btu703. PubMed PMID: 25338716; PubMed Central PMCID:  
570 PMCPMC4341060.
- 571 19. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. A statistical framework to predict  
572 functional non-coding regions in the human genome through integrated analysis of annotation  
573 data. *Sci Rep.* 2015;5:10576.
- 574 20. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional  
575 genomic annotations for coding and noncoding variants. *Nat Genet.* 2016;48(2):214-20. Epub  
576 2016/01/05. doi: 10.1038/ng.3477. PubMed PMID: 26727659; PubMed Central PMCID:  
577 PMCPMC4731313.
- 578 21. Racimo F, Schraiber JG. Approximation to the distribution of fitness effects across  
579 functional categories in human segregating polymorphisms. *PLoS Genet.*  
580 2014;10(11):e1004697. Epub 2014/11/07. doi: 10.1371/journal.pgen.1004697. PubMed PMID:  
581 25375159; PubMed Central PMCID: PMCPMC4222666.
- 582 22. Huang YF, Siepel A. Estimation of allele-specific fitness effects across human protein-  
583 coding sequences and implications for disease. *Genome Res.* 2019;29(8):1310-21. Epub  
584 2019/06/30. doi: 10.1101/gr.245522.118. PubMed PMID: 31249063; PubMed Central PMCID:  
585 PMCPMC6673719.
- 586 23. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The  
587 mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.*  
588 2020;581(7809):434-43.

- 589 24. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC  
590 Genome Browser database: 2016 update. *Nucleic Acids Res.* 2016;44(D1):D717-25. Epub  
591 2015/11/22. doi: 10.1093/nar/gkv1275. PubMed PMID: 26590259; PubMed Central PMCID:  
592 PMCPMC4702902.
- 593 25. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl  
594 comparative genomics resources. *Database (Oxford)*. 2016;2016. Epub 2016/05/04. doi:  
595 10.1093/database/baw053. PubMed PMID: 27141089; PubMed Central PMCID:  
596 PMCPMC4852398.
- 597 26. Sanjak JS, Sidorenko J, Robinson MR, Thornton KR, Visscher PM. Evidence of directional  
598 and stabilizing selection in contemporary humans. *Proc Natl Acad Sci U S A.* 2018;115(1):151-6.  
599 Epub 2017/12/20. doi: 10.1073/pnas.1707227114. PubMed PMID: 29255044; PubMed Central  
600 PMCID: PMCPMC5776788.
- 601 27. Byars SG, Ewbank D, Govindaraju DR, Stearns SC. Colloquium papers: Natural selection  
602 in a contemporary human population. *Proc Natl Acad Sci U S A.* 2010;107 Suppl 1(suppl  
603 1):1787-92. Epub 2009/10/28. doi: 10.1073/pnas.0906199106. PubMed PMID: 19858476;  
604 PubMed Central PMCID: PMCPMC2868295.
- 605 28. Beauchamp JP. Genetic evidence for natural selection in humans in the contemporary  
606 United States. *Proc Natl Acad Sci U S A.* 2016;113(28):7774-9. Epub 2016/07/13. doi:  
607 10.1073/pnas.1600398113. PubMed PMID: 27402742; PubMed Central PMCID:  
608 PMCPMC4948342.
- 609 29. Sella G, Barton NH. Thinking about the evolution of complex traits in the era of genome-  
610 wide association studies. *Annu Rev Genom Hum Genet.* 2019;20:461-93.
- 611 30. Fisher RA. *The genetical theory of natural selection.* Oxford, UK: Clarendon Press; 1930.
- 612 31. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. *BMC*  
613 *Biol.* 2017;15(1):98. Epub 2017/11/01. doi: 10.1186/s12915-017-0434-y. PubMed PMID:  
614 29084517; PubMed Central PMCID: PMCPMC5662103.
- 615 32. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human  
616 Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum*  
617 *Genet.* 2017;100(4):635-49. Epub 2017/04/04. doi: 10.1016/j.ajhg.2017.03.004. PubMed PMID:  
618 28366442; PubMed Central PMCID: PMCPMC5384097.
- 619 33. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al.  
620 Reduced signal for polygenic adaptation of height in UK Biobank. *Elife.* 2019;8. Epub  
621 2019/03/22. doi: 10.7554/eLife.39725. PubMed PMID: 30895923; PubMed Central PMCID:  
622 PMCPMC6428572.
- 623 34. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic  
624 adaptation on height is overestimated due to uncorrected stratification in genome-wide  
625 association studies. *Elife.* 2019;8. Epub 2019/03/22. doi: 10.7554/eLife.39702. PubMed PMID:  
626 30895926; PubMed Central PMCID: PMCPMC6428571.
- 627 35. Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, et al. WGS: an annotation pipeline  
628 for human genome sequencing studies. *J Med Genet.* 2016;53(2):111-2. Epub 2015/09/24. doi:  
629 10.1136/jmedgenet-2015-103423. PubMed PMID: 26395054; PubMed Central PMCID:  
630 PMCPMC5124490.

- 631 36. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data.  
632 Gigascience. 2019;8(7). Epub 2019/07/16. doi: 10.1093/gigascience/giz082. PubMed PMID:  
633 31307061; PubMed Central PMCID: PMC6629542.
- 634 37. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-  
635 CM codes to phecodes: workflow development and initial evaluation. JMIR Medical Informatics.  
636 2019;7(4):e14325.
- 637 38. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al.  
638 Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide  
639 association studies in the electronic health record. PloS one. 2017;12(7):e0175508.
- 640 39. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile:  
641 PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol.  
642 2018;47(1):29-35. Epub 2017/10/19. doi: 10.1093/ije/dyx204. PubMed PMID: 29040602;  
643 PubMed Central PMCID: PMC6629542.
- 644
- 645

646

647 **Supporting information**

648

649 **S1 Fig. Derived allele frequency spectra of coding and non-coding variants for**

650 **different CADD and phyloP score categories.** Top row: Derived allele frequency  
651 spectrum of coding variants. Bottom row: Derived allele frequency spectrum of non-  
652 coding variants. Each solid line represents derived allele frequency spectrum of  
653 polymorphic sites belonging to one score category and three dashed lines represent  
654 derived allele frequency spectra of three control categories: synonymous (syn),  
655 missense (mis), and loss of function (LOF) variants.

656

657 **S2 Fig. Phenotypic association of load (phyloP-weighted) and burden**

658 **(unweighted) scores.** Quantile-quantile plot of  $-\log_{10}$  p-values for the phenotypic  
659 association of A) load scores (weighted by phyloP); B) burden scores (unweighted); C)  
660 burden scores restricted to rare variants ( $DAF < 5\%$ ); and D) burden scores restricted to  
661 common variants ( $5\% \leq DAF < 70\%$ ).

662

663 **S3 Fig: Enrichment of clinical phenotypes nominally associated with genome-**

664 **wide load score and non-coding load score.** Null distribution of the number of  
665 clinical phenotypes weakly associated with genome-wide load score (left) and non-  
666 coding load score (right) was obtained from 2,000 permutations each. For each  
667 permutation, the load score was shuffled randomly among 335,161 samples and the  
668 number of associations on the x-axis was the count of phenotypes which yielded p-  
669 value  $< 0.05$  in the association tests between the permuted load score and 539  
670 phecodes. The red dashed lines indicates the observed number of clinical phenotypes

671 nominally associated with genome-wide load score (n = 27, left) and non-coding load  
672 score (n = 24, right).

673

674 **S4 Fig: Enrichment of clinical phenotypes nominally associated with burden**

675 **scores.** Null distributions of the number of clinical phenotypes weakly associated with  
676 burden scores were obtained using the same procedure to obtain the null distributions  
677 for load scores (Figure 3 and Figure S3). The red dashed lines indicates the observed  
678 number of clinical phenotypes nominally associated with genome-wide burden score (n  
679 = 22, left), coding burden score (n=20, middle), and non-coding load score (n = 20,  
680 right).

681

682

683 **S1 Table.** Linear regression between slopes and score categories.

684 **S2 Table.** Association between load score and the first 10 principal components.

685 **S3 Table.** Derived allele frequency stratification analysis.

686 **S4 Table.** phyloP score stratification analysis.

687 **S5 Table.** phyloP score and DAF stratification analysis.

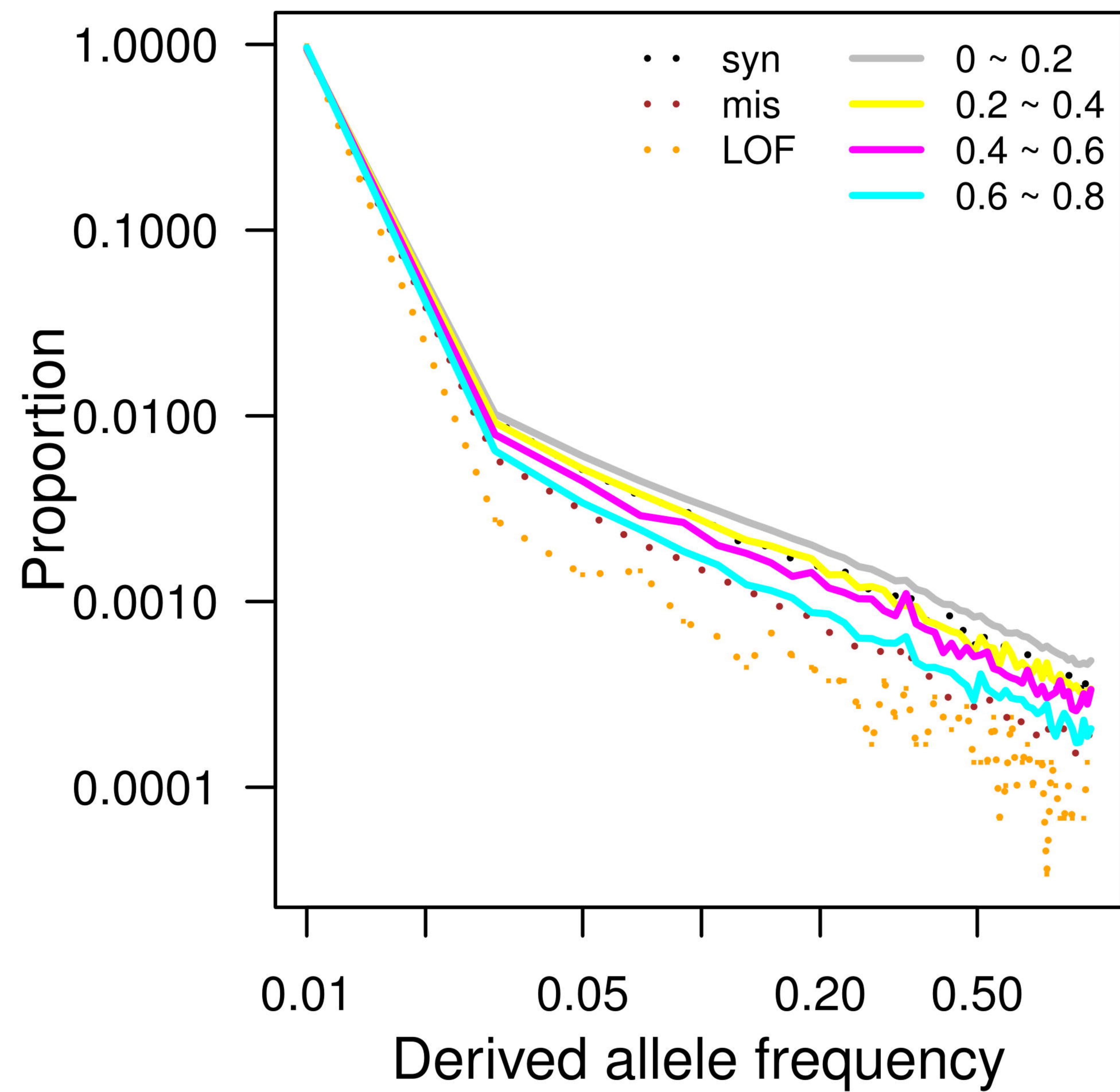
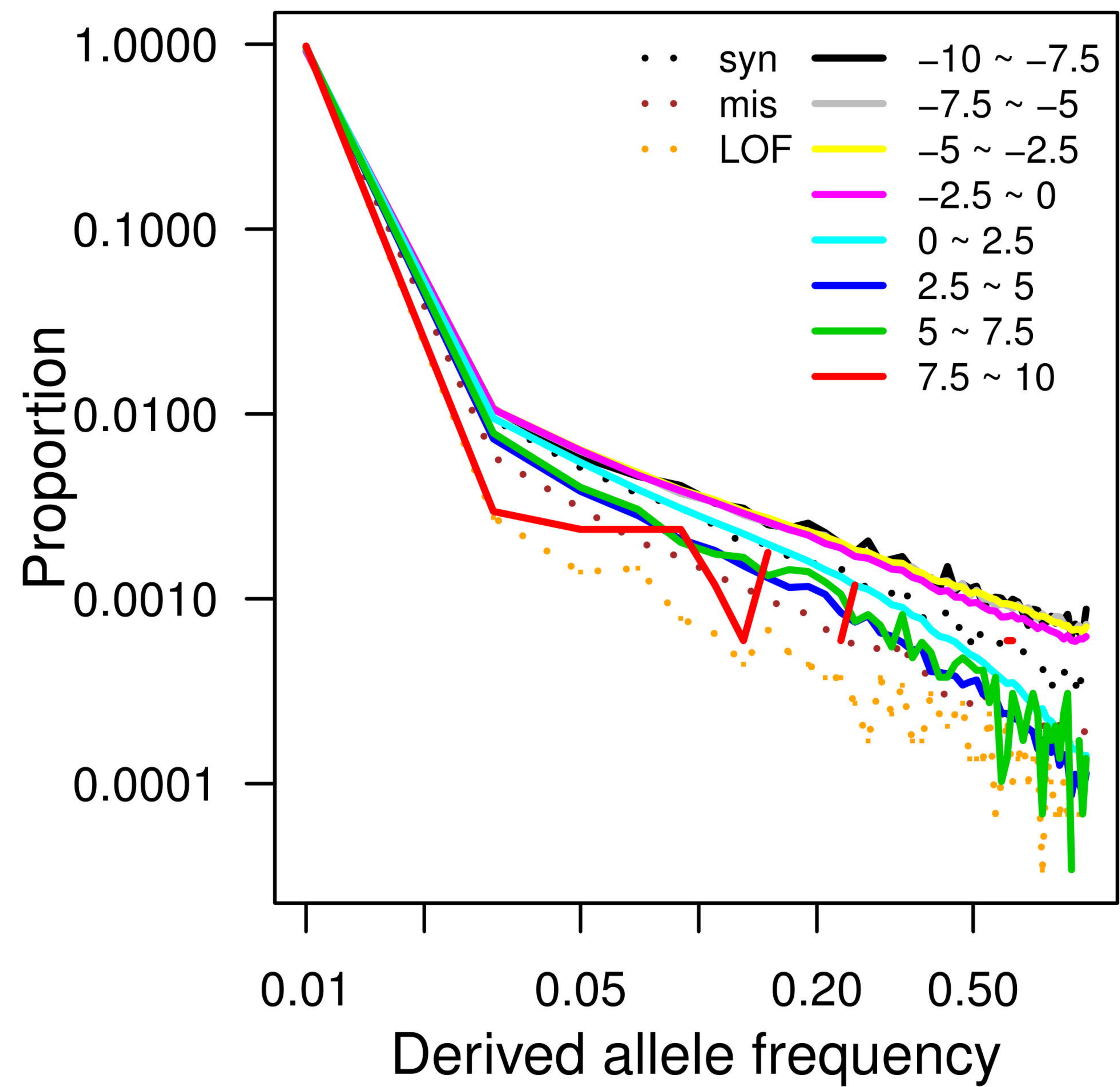
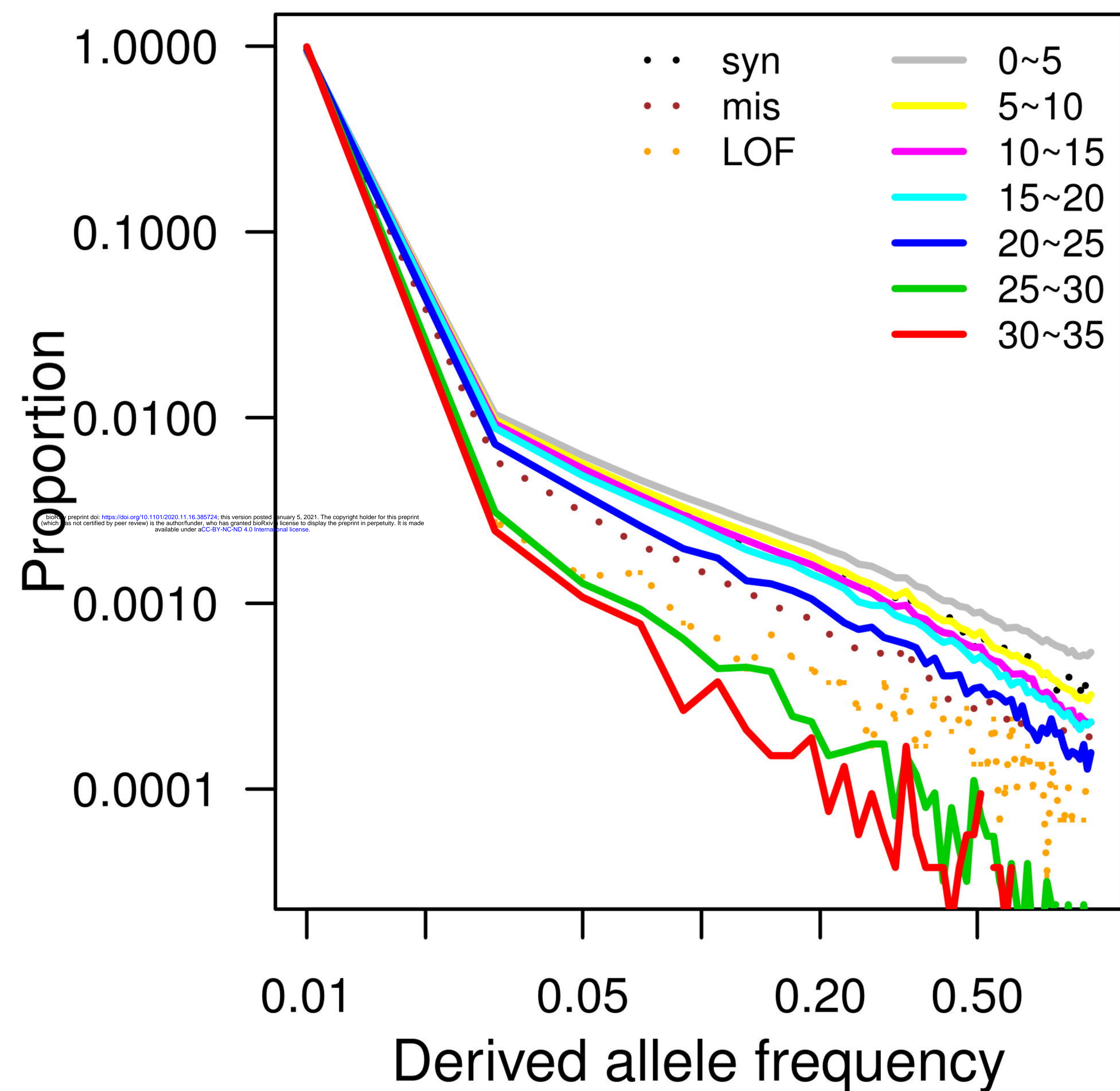
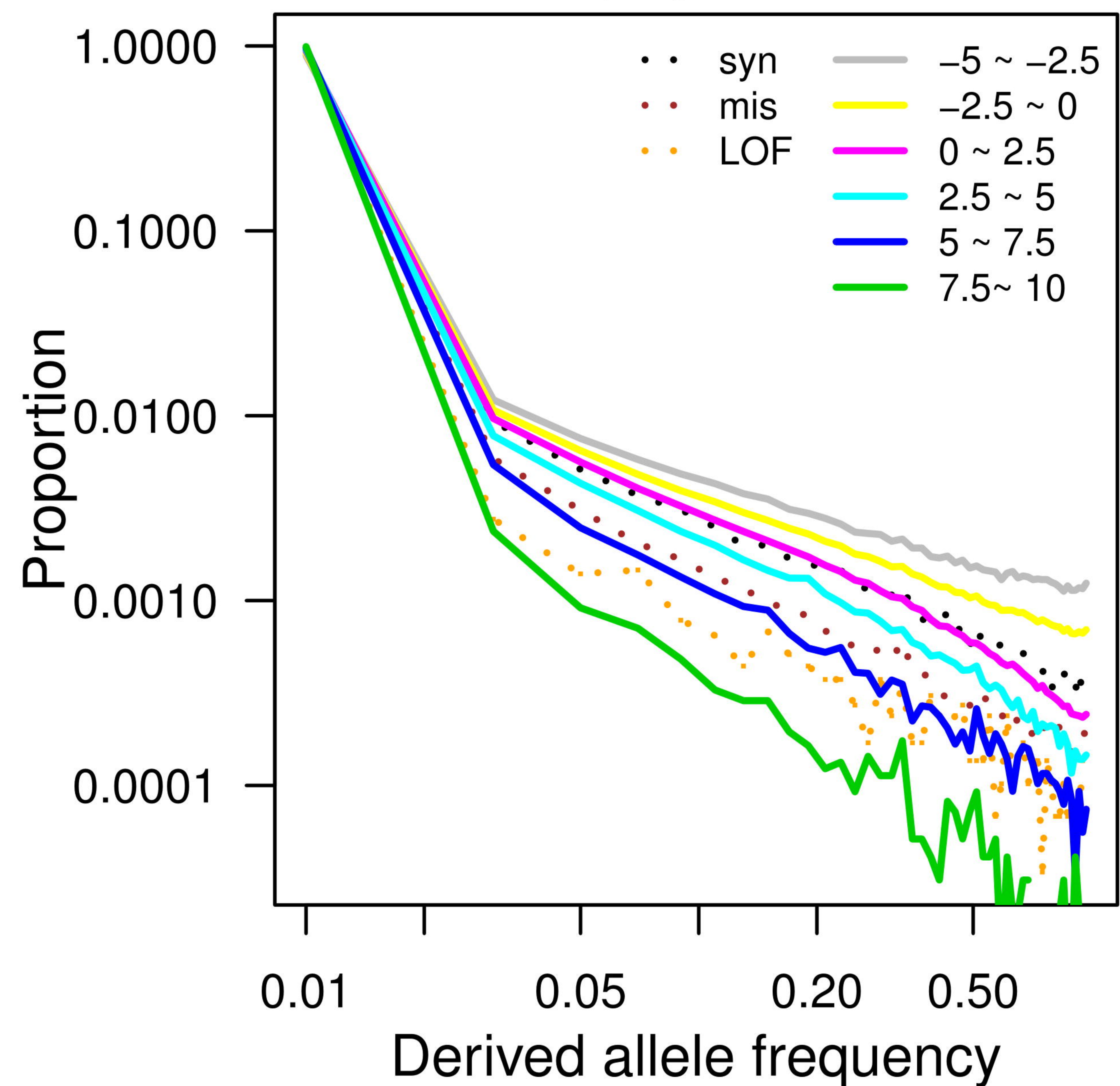
688 **S6 Table.** Association between coding load score and the 27 phenotypes with number  
689 of non-reference variants included as a covariate.

690 **S7 Table.** Association between load scores restricted to sites where the human genome  
691 reference allele is the ancestral allele and 27 phenotypes.

692 **S8 Table.** Association between coding load scores computed from phyloPNH and 27  
693 phenotypes.

694 **S9 Table.** Clinical phenotypes weekly associated with load scores.

695

**fitCons****GERP****CADD****phyloP**

Count

100

50

0

10

20

30

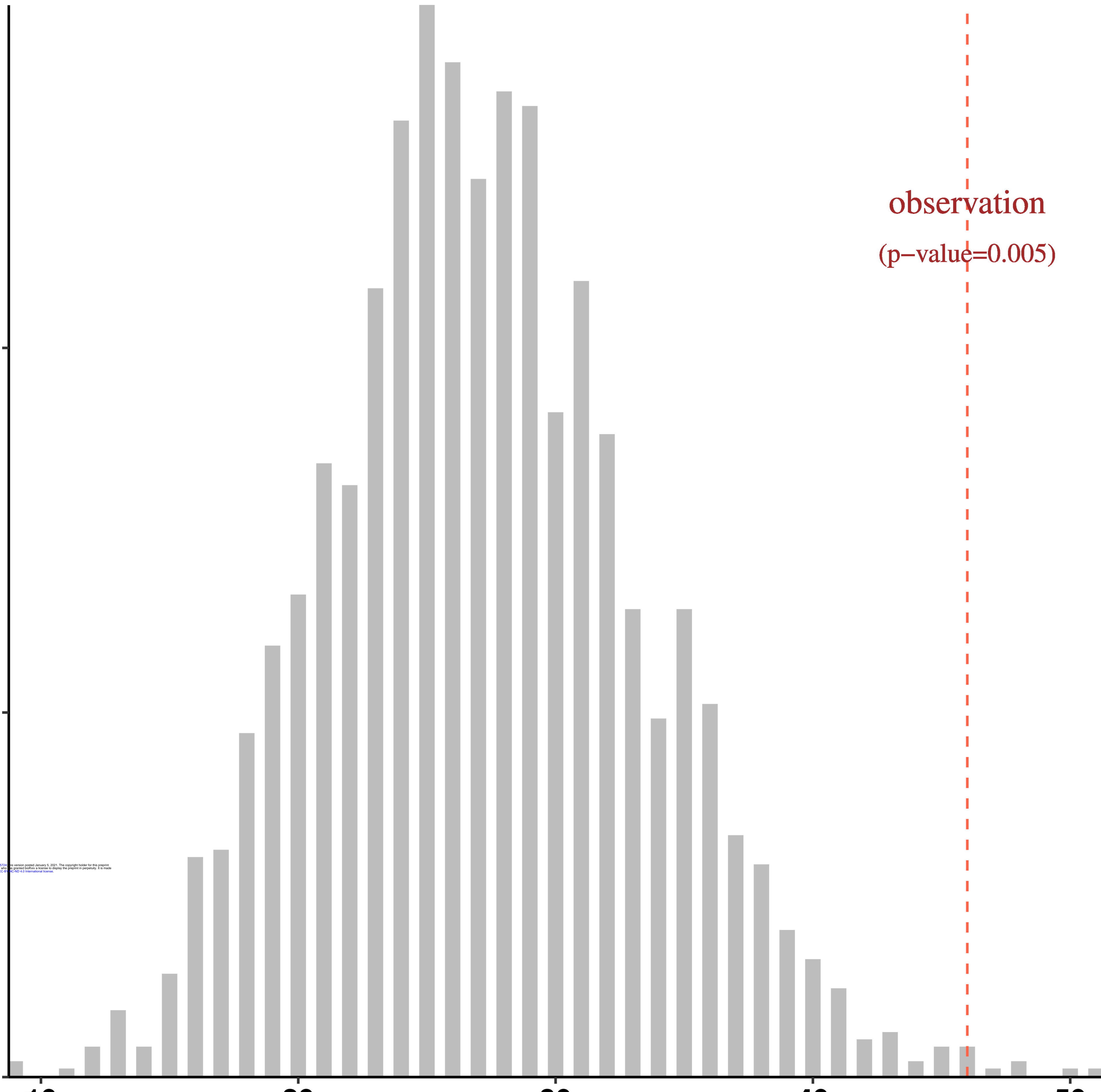
40

50

Number of association

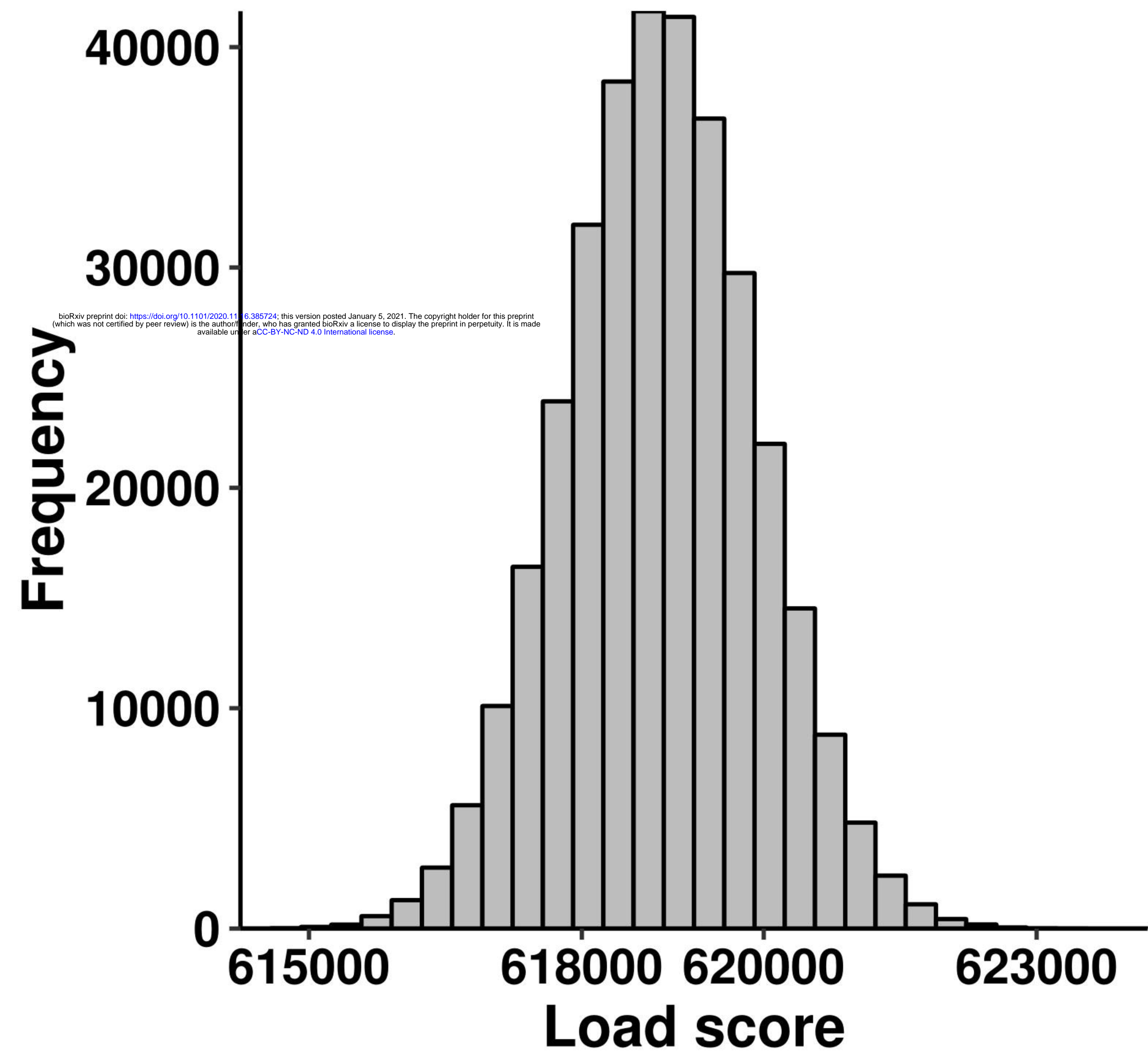
observation  
(p-value=0.005)

bioRxiv preprint doi: <https://doi.org/10.1101/2020.11.16.367245>; this version posted January 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

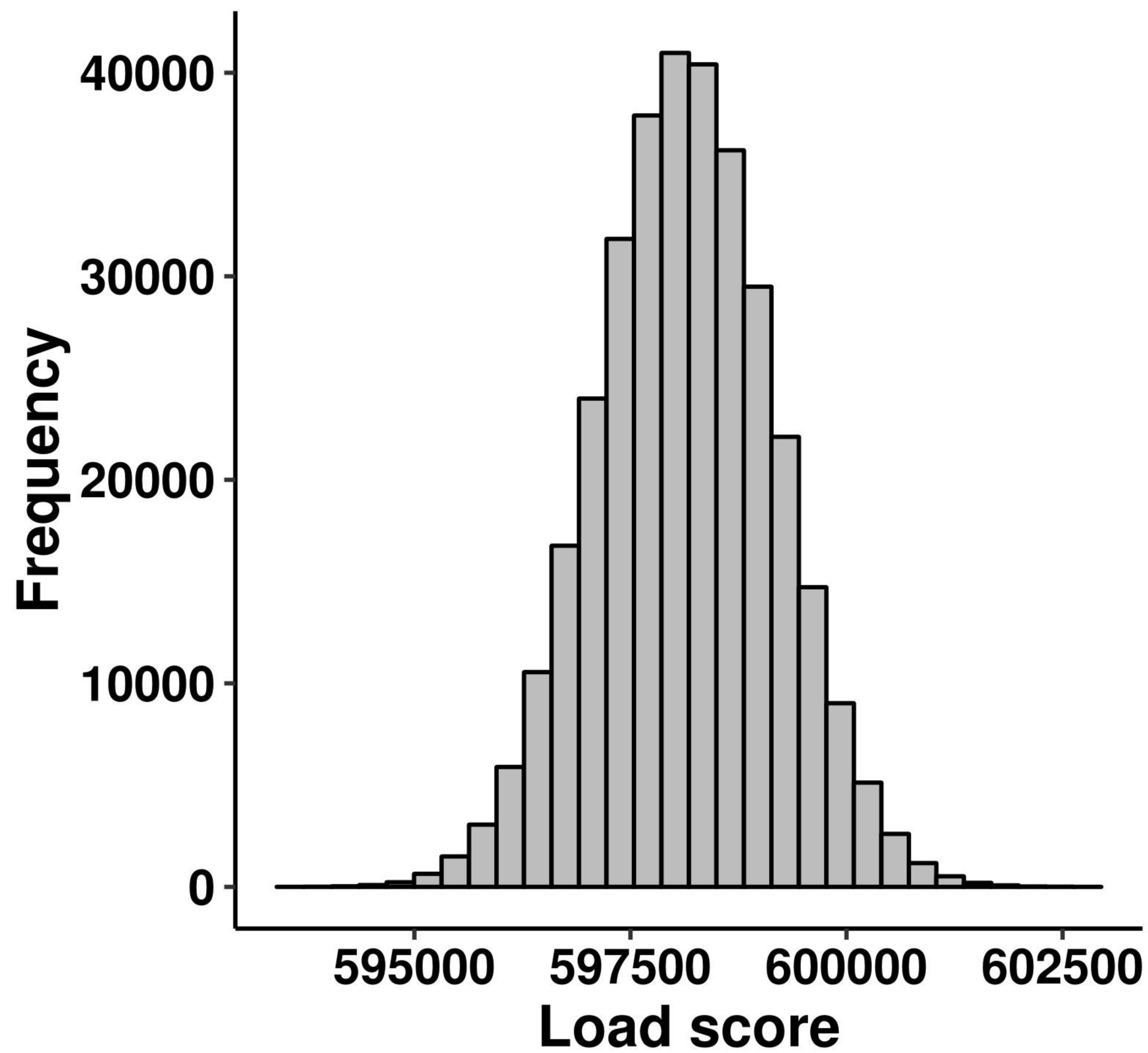




# Genome-wide load score



# Non-coding load score



# Coding load score

