Soft sweeps are the dominant mode of adaptation in the human genome

Daniel R. Schrider[*, †, 1] and Andrew D. Kern[*, †]

[*]Department of Genetics, Rutgers University, Piscataway, NJ, 08854, USA

[†]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, NJ, 08554, USA

[1]Corresponding author: Department of Genetics, Rutgers University, 604 Allison Rd., Piscataway, NJ 08854. E-mail: dan.schrider@rutgers.edu

Keywords: adaptation, selective sweeps, soft sweeps, machine learning, population genomics

Running title: Adaptation via soft sweeps is prevalent in humans

ABSTRACT

The degree to which adaptation in recent human evolution shapes genetic variation remains controversial. This is in part due to the limited evidence in humans for classic "hard selective sweeps," wherein a novel beneficial mutation rapidly sweeps through a population to fixation. However, positive selection may often proceed via "soft sweeps" acting on mutations already present within a population. Here we examine recent positive selection across six human populations using a powerful machine learning approach that is sensitive to both hard and soft sweeps. We found evidence that soft sweeps are widespread and account for the vast majority of recent human adaptation. Surprisingly, our results also suggest that linked positive selection affects patterns of variation across much of the genome, and may increase the frequencies of deleterious mutations. Our results also reveal insights into the role of sexual selection, cancer risk, and central nervous system development in recent human evolution.

INTRODUCTION

Spurred by the ongoing revolution in DNA sequencing capacity, human population genetic datasets have grown exponentially in size over the past five years (Auton et al. 2015; Consortium 2015). Such growth enables insight into the evolutionary histories of human populations with hitherto unrivaled precision. A central question in the study of human evolution is the extent to which adaptation has driven recent evolution and affected patterns of genetic diversity (Akey 2009). This can be addressed by scanning genomic data for evidence of selective sweeps, wherein a beneficial mutation is favored by natural selection and therefore rapidly increases in frequency within a population. Such selective sweeps leave a characteristic footprint in variation; they create a valley of diversity around the selected site (Maynard Smith and Haigh 1974;

36 Kaplan et al. 1989; Stephan et al. 1992), a deficit of both low- and high-frequency derived alleles
37 at linked sites (Fay and Wu 2000), and an increase in linkage disequilibrium in flanking regions
38 (Kim and Nielsen 2004). Thus there are multiple population genetic signals to exploit.
39 Accordingly numerous theoretical and methodological advances (Kaplan et al. 1989; Stephan et
40 al. 1992; Fu 1997; Kim and Stephan 2002; Nielsen et al. 2005b; Voight et al. 2006) in the study
41 of selective sweeps have given researchers the ability to uncover the genetic basis of adaptation
42 on a genome-wide scale.

43 There are two complimentary approaches to studying the impact of adaptive evolution on
44 genetic variation. The first approach aims to infer genome-wide rates of adaptive evolution by
45 estimating the mean effects of selective sweeps across the genome (Wiehe and Stephan 1993;
46 Kern et al. 2002; Andolfatto 2007; Jensen et al. 2008; Hernandez et al. 2011; Sattath et al. 2011).
47 Such approaches may estimate the rates of sweeps or their effects with respect to the genomic
48 background, but do not focus on the targets of sweeps themselves. An alternative approach is to
49 focus on finding individual selective sweeps throughout the genome, and in so doing characterize
50 specific cases of the adaptation with hopes of gaining general insight into the adaptive process
51 (Sabeti et al. 2002; Voight et al. 2006; Williamson et al. 2007). Such approaches have yielded
52 valuable insights into the recent evolutionary histories of natural populations in various species,
53 such as revealing the pervasive impact of adaptive evolution on polymorphism in *Drosophila*
54 *melanogaster* (Begun et al. 2007; Macpherson et al. 2007; Langley et al. 2012; Lee et al. 2013;
55 Garud et al. 2015). In humans, the picture remains less clear: while scans for selective sweeps
56 have discovered numerous compelling candidates for strong positive selection (e.g. Ruwende et
57 al. 1995; Stephens et al. 1998; Tishkoff et al. 2007; Bryk et al. 2008; Huerta-Sánchez et al.
58 2014), some recent studies have suggested that the impact of adaptation on patterns of variation

59 genome-wide is quite limited (Hernandez et al. 2011; Lohmueller et al. 2011). Conversely, Enard
60 et al. (2014) argue that the genome-wide reduction in diversity around substitutions is driven in
61 part by positive selection.

62 One possible explanation for the difficulty in characterizing the contributions of adaptive
63 and non-adaptive forces in human populations is that genetic hitchhiking effects may be muted
64 by human demographic history. Many human populations appear to have experienced
65 bottlenecks and/or recent growth (Marth et al. 2004; Fagundes et al. 2007; Gravel et al. 2011;
66 Auton et al. 2015), which cause much of the genome to resemble selective sweeps (Nielsen et al.
67 2005b). Moreover, positive selection has historically been modeled as the process of a *de novo*
68 beneficial mutation rapidly sweeping to fixation, a process now referred to as a hard sweep.
69 However selection may act on previously segregating neutral or weakly deleterious variants (Orr
70 and Betancourt 2001; Innan and Kim 2004). Selection on standing variation will produce
71 qualitatively different skews in linkage disequilibrium and allele frequencies, along with a
72 shallower valley in diversity (Hermisson and Pennings 2005; Przeworski et al. 2005; Berg and
73 Coop 2015; Schrider et al. 2015)—such an event is thus referred to as a soft sweep. If selection
74 typically proceeds through soft sweeps, as may be the case in Drosophila (Garud et al. 2015),
75 then many sweeps may have been missed by previous scans that were designed to detect
76 signatures produced under a hard sweep model.

77 We sought to address the controversy over the impact of adaptation on human genomic
78 variation by conducting a genome-wide scan for both hard and soft selective sweeps across
79 human populations. We previously developed S/HIC (Soft/Hard Inference through
80 Classification), a machine learning method capable of detecting completed sweeps and inferring
81 their mode of selection with unparalleled accuracy and robustness to non-equilibrium

82 demography (Schrider and Kern 2016). Here we apply S/HIC to uncover hard and soft sweeps in
83 six population samples from the 1000 Genomes Project (Auton et al. 2015), thereby performing
84 the most comprehensive investigation of completed selective sweeps in humans to date.
85 Surprisingly, our results suggest that patterns of polymorphism across much of the human
86 genome may be affected by linked positive selection—primarily soft sweeps. Moreover, we find
87 evidence that the mode of selection differs substantially across populations, with non-African
88 populations adapting via hard sweeps to a much greater extent than African populations. Finally,
89 we investigate the biological targets of selection in recent human evolution, with particular
90 processes such as immunity, cancer, and sexual reproduction playing outsized roles.
91

92 **RESULTS**

93 We set out to detect completed hard and soft selective sweeps in six populations from Phase 3 of
94 the 1000 Genomes Project: two West-African populations (YRI and GWD from Yoruba and The
95 Gambia, respectively), one East-African population (LWK from Kenya), one European
96 population (CEU, from Utah, USA), one East Asian population (JPT from Japan), and one from
97 the Americas (PEL from Peru). For each population we trained and applied a S/HIC classifier to
98 identify hard and soft selective sweeps across the genome (Methods), distinguishing them from
99 neutrally evolving regions as well as those linked to sweeps (Schrider and Kern 2016). Briefly,
100 S/HIC is a machine learning method that leverages spatial patterns of a variety of statistics across
101 a large genomic window in order to infer the mode of evolution at the center of the window. We
102 previously showed that S/HIC is exceptionally robust to the confounding effect of linked
103 selection (i.e. the "soft shoulder" effect; Schrider et al. 2015), as well as non-equilibrium
104 demographic histories, making it well suited for a survey of positive selection in humans. We

105 also assessed the accuracy of our classifiers on simulated test data with the same demographic
106 history used to generate training data, finding that S/HIC achieved good power for each
107 demographic history, with somewhat higher accuracy for histories inferred from the African than
108 non-African populations (Figure S1). We also examined values of Garud et al.'s (2015) H_{12} and
109 H_2/H_1 within windows classified by S/HIC as hard, soft, or neutral, noting that as expected, H_{12}
110 is higher in sweeps than neutral regions, while H_2/H_1 is higher for soft sweeps than hard sweeps
111 (Figure S2). Below, we begin with a brief overview of the broad patterns of adaptation we
112 observe across populations, before discussing genomic features and biological pathways with a
113 strong enrichment of selective sweeps, as well as compelling novel candidates for recently
114 completed selective sweeps.
115

116 **The majority of sweeps in humans resemble selection on standing variation**

117 We found a total of 1,927 distinct selective sweeps merged across all six populations (Methods).
118 190 (9.9%) of these are present in all populations, 59 (3.1%) are shared among the African
119 populations, 71 (3.7%) are shared among the non-African populations, and 701 (36.4%) are
120 population-specific. The remaining 906 (47.0%) sweeps were present in more than one
121 population but do not fit into any of the categories above. We observe that across populations,
122 the vast majority (1,776, or 92.2%) of sweeps were classified as soft. While hard sweeps appear
123 to be quite rare globally, the fraction of hard sweeps is significantly higher in non-African than
124 African populations (Table 1). For example, when comparing PEL to GWD, we observe a
125 significantly higher fraction of hard sweeps in PEL (4.7% versus 1.6%; p=0.05). For each other
126 African vs. non-African comparison we see an even greater (and more significant) disparity.
127 Further, we observe a suggestive correlation between the fraction sweeps in a population that

128 were classified as soft and the harmonic mean of its population size within the last 4*N*
129 generations (Pearson's ρ=-0.96; Methods). Though taken at face value this correlation appears to
130 be highly significant, we note that due to the six populations' shared evolutionary history a
131 statistical test of this correlation would be invalid.

132 Comparing our results to those of previous scans we find that 519 of S/HIC's sweep calls
133 (26.9%) have previously been identified according to dbPSHP, a database of candidate regions
134 for recent positive selection across human populations (Li et al. 2013). This accounts for 10.9%
135 of the loci in the dbPSHP set (ignoring regions not classified by S/HIC). The remaining 1,408
136 sweeps called by S/HIC (73.1% of calls) represent potentially novel selective sweeps. There are
137 several possible explanations for the modest overlap between our set of sweep candidates and
138 those in dbPSHP. First, the sweep candidates in dbPSHP have been identified by a variety of
139 methods, some of which are designed to detect selective scenarios other than completed sweeps
140 (e.g. partial sweeps, spatially varying selection). Second, when comparing results from methods
141 designed to detect the same type of sweeps, the intersection between studies is often fairly small
142 (Akey 2009). The presence of false positives may in part be responsible for the imperfect
143 concordance between scans.

144

145 **Selective sweeps preferentially target genes involved in cancer and viral infection**

146 Examining the locations of selective sweeps across the genome, we find that regions classified as
147 selective sweeps are significantly overrepresented for both coding sequence and untranslated
148 regions (q<0.05 in several populations for hard sweeps, and each population for soft sweeps;
149 Figure 1A, B; Table S1), relative to data sets with permuted classifications (see Methods).
150 Enrichment for transcription factor binding sites was less pronounced, and only significant in

151 soft sweeps for the three African populations along with PEL. The most striking result we
152 observed was a dramatic enrichment of sweep windows for mutations in the COSMIC data set of
153 somatic mutations that have been observed in cancer cells (Forbes et al. 2015) and may therefore
154 play a role in tumor suppression/progression. Averaged across populations, the number of
155 COSMIC mutations found in soft sweeps represents a 3.7-fold increase relative to that observed
156 in permuted data sets; this enrichment was significant in each population, and peaked at 4.5-fold
157 in PEL. For hard sweeps, this enrichment was 12-fold on average, reaching as high as 21-fold in
158 CEU, though this was the only population for which the enrichment was statistically significant.
159 We also observed a sizeable overrepresentation of genes encoding virus-interacting proteins
160 (VIPs) curated by (Enard et al. 2016) in soft sweeps, with a 1.9-fold increase relative to
161 permuted sets (averaged across populations). VIPs show a similar magnitude of enrichment in
162 hard sweeps for some populations, but does not achieve significance at $q<0.05$.
163

164 **Selective sweeps increase linked deleterious variation**

165 Because S/HIC not only detects selective sweeps, but also attempts to identify regions of the
166 genome that appear to be linked to recent sweeps, our classifications allow us to examine the
167 effect of linked selection in a principled way. We found that while a minority of genomic
168 windows were classified as selective sweeps (7.6% on average across all populations), a large
169 fraction of windows were classified as linked to a completed selected sweep, either hard or soft
170 (56.4% on average). These estimates range from 41.5% in JPT to 74.0% in GWD (Figure 2).

171 We also asked whether selective sweeps have a detectable impact on linked deleterious
172 variation. As beneficial alleles increase in frequency in a population, they may carry along with
173 them linked deleterious polymorphisms as hitchhikers, potentially increasing the frequency of

174 deleterious variants over what would be expected given mutation-selection-drift equilibrium
175 (Birky and Walsh 1988; Hartfield and Otto 2011). To this end we asked whether relatively
176 common candidate deleterious mutations were enriched in regions classified as either hard-
177 linked or soft-linked. Indeed, we observed a fairly subtle but significant overrepresentation of
178 SNPs with derived allele frequencies of at least 0.01 but predicted to by damaging by SIFT
179 (Kumar et al. 2009) in both the hard-linked (mean enrichment across populations: 1.3-fold) and
180 soft-linked (mean enrichment: 1.1-fold) classes for most populations (Fig 1C, D; Table S1). We
181 find a similar enrichment in these sweep-linked classes of common SNPs in regions inferred to
182 be conserved across primates according to phastCons (Siepel et al. 2005). Phenotype-associated
183 variants from the GWAS catalogue (Welter et al. 2014) were also significantly overrepresented
184 sweep-linked regions in several populations (Fig 1C, D).
185

186 **Sexual reproduction, the central nervous system, and immunity are targets of recent**
187 **sweeps**

188 In order to determine if positive selection preferentially acts on particular organismal functions,
189 we asked which Gene Ontology (GO) terms were enriched in our sweep calls relative to the
190 permuted data (Methods). In soft sweeps, we found a sizeable and significant enrichment
191 (q<0.05) of terms related to sperm development, structure, and function. Terms enriched in at
192 least one population include spermatogenesis (4.4-fold enrichment averaged across populations),
193 "sperm flagellum" (3.9-fold enrichment), and "sperm-egg recognition" (3.3-fold enrichment).
194 These terms were not significantly enriched for hard sweeps in any population, though this may
195 be due in part to reduced power because of the smaller number of sweeps.

196 We also observed an overrepresentation of genes involved in the "glutamate receptor
197 signaling pathway" in our soft sweep sets (q<0.05 for each population; 4-fold mean enrichment).
198 Glutamate receptors are the primary excitatory neurotransmitter in the central nervous system,
199 and important for both proper brain development and function (Luján et al. 2005). Indeed, soft
200 sweeps are enriched for "central nervous system development" in multiple populations (1.6-fold
201 mean enrichment). Numerous terms related to immune response, especially adaptive immunity,
202 were also significantly enriched among soft sweeps. These, and the many other GO terms
203 overrepresented in our sweep sets are listed in Table S2. We also found numerous KEGG
204 pathways enriched for sweeps, including many pathways related to immunity, cancer
205 progression/tumor suppression (see Table S2 for full list).
206
207 **Positive selection on interacting gene pairs**
208 We examined three types of gene interaction networks: protein-protein interactions (PPIs),
209 transcription factor-target gene interactions, and genetic interactions where one gene modifies
210 the effect of another (Methods). Interestingly, we observed a dramatic enrichment of sweeps in
211 genes that encode proteins that physically interact with one another (Figure 3A–B): if a gene
212 experienced a soft sweep, genes that interact with this gene were on average 3.3 times more
213 likely to have experienced a soft sweep than expected by chance (p<0.0001 for each population;
214 Figure 3B). Despite the smaller number of candidate regions, we found a significant enrichment
215 for PPIs in hard sweeps, though this was only significant in for non-African populations (4.0-fold
216 enrichment averaged across populations; p<0.05 in CEU, JPT, YRI; Figure 3A). For
217 transcription factor-target interactions, we observe no overrepresentation of soft sweeps, but a
218 significant enrichment of hard sweeps in non-African populations (p<0.05 for each; 8.5-fold

219 enrichment on average; Figure 3C–D). There were no populations exhibiting an
220 overrepresentation of pairs of genes with genetic interactions and experiencing sweeps of either
221 type (Figure 3E–F).
222
223 **Examples of novel selective sweep candidates**
224 In this section we describe several sweep candidates that exemplify the set of sweeps, and
225 functions of putative targets of selection, that we were able to detect. As discussed above, our
226 sets of sweeps were highly enriched for glutamate receptor-encoding genes. In Figure S3, we
227 show a sweep candidate region on chromosome 4 that encompasses the glutamate receptor gene
228 *GRIA2*. This sweep was previously detected in non-African populations by Pickrell et al. (2009),
229 who did not find any evidence of selection in Africa. However, S/HIC infers that this region has
230 experienced a soft sweep that is found in GWD and YRI, as well as the non-African populations.
231 Consistent with this, Europeans, Asians, and African populations show a reduction in π, a trough
232 in Tajima's *D* (Tajima 1989), and a peak in Nielsen et al.'s SweepFinder composite likelihood
233 ratio (CLR) test statistic, which captures regions that appear to be at the epicenter of the spatial
234 skew in the SFS expected around sweeps (Nielsen et al. 2005b). Intriguingly, *GRIA2* interacts
235 with the *GRID2* glutamate receptor gene (Kohda et al. 2003), which itself is classified as a soft
236 sweep in CEU, LWK, PEL, and GWD. The remaining glutamate receptors overlapping identified
237 sweeps are *GRIA4*, *GRID1*, *GRIK1*, *GRIK3*, *GRM2*, and *GRM7*. Of these genes, *GRIA4* and
238 *GRID2* were shown by Liu et al. (2012) to have evolved a human-specific developmental
239 expression profile.
240 Figure 4 shows a region on chromosome 9 that exhibits strong evidence of a previously
241 undetected hard sweep in each of our six populations. This region contains several members of

242 the spermatogenesis associated 31 gene family: *SPATA31B1*, *SPATA31D1*, *SPATA31D3*, and
243 *SPATA31D4*. Across populations this region shows dramatic valleys in π and Tajima's *D*, as well
244 as an elevated CLR near the center of the sweep window. These genes are highly testis-specific
245 according to data from the GTEx project (Lonsdale et al. 2013), and male mice are infertile when
246 lacking Spata31, another member of these gene family (Wu et al. 2015). Figure 4 also shows that
247 each of these genes overlaps a cluster of non-repetitive piRNAs (data from piRBase; Zhang et al.
248 2014). Also near this region is *DDX10P2*, which GENCODE annotates as a processed
249 pseudogene (Pei et al. 2012). *DDX10P2*, which is located at the center of the CLR peak for CEU,
250 is expressed with a high degree of testis-specificity according to GTEx data, similar to the
251 neighboring *SPATA31* genes. A BLAT search (Kent 2002) revealed that this putative
252 pseudogene exhibits 99.5% sequence identity to the orthologous sequence in chimpanzees. The
253 parent gene of *DDX10P2*, *DDX10*, is expressed in many tissues, but shows highest expression in
254 the testis.

255 On chromosome 11 we detected what appear to be several novel soft sweeps present in
256 and upstream of *CADM1* (cell adhesion molecule 1; Figure 5), one of which is present in each
257 population. This gene is essential for spermatogenesis in mice (Van Der Weyden et al. 2006),
258 and is also a tumor suppressor that is hypermethelated in various cancers (Kuramochi et al. 2001;
259 Allinen et al. 2002; Fukuhara et al. 2002), as it works with the adaptive immune system to
260 suppress metastasis (Faraji et al. 2012). *CADM1* is also active in the brain where it is involved in
261 synaptic adhesion and has been linked to autism (Zhiling et al. 2008; Fujita et al. 2010). *CADM1*
262 forms a complex with those other genes: the GABA receptor *GABBR2*, which has a soft sweep in
263 YRI, and *MUPP1*, which has a soft sweep found in each population; this complex appears to
264 localize to Purkinje cell dendrites (Fujita et al. 2012). Thus, this example encompasses many of

265 the functions that we find are highly enriched across our sweep sets: adaptation in multiple
266 interacting genes (one of which is a neurotransmitter), spermatogenesis, and tumor suppression
267 (via adaptive immunity).
268

269 **DISCUSSION**

270 Understanding the history of human adaptation at the genetic level is a central goal of population
271 genomics and human evolutionary biology. Accordingly, since the completion of the human
272 genome assembly (Lander et al. 2001) and subsequent proliferation of population genomic data,
273 numerous genome-wide scans for selection have been conducted using differing methodologies
274 (Sabeti et al. 2002; Voight et al. 2006; Sabeti et al. 2007; Pickrell et al. 2009; Field et al. 2016).
275 The majority of these studies searched primarily for partial selective sweeps—the signature of a
276 beneficial mutation currently sweeping through a population (see Williamson et al. 2007 for a
277 notable exception)—and rightly so, as these sweeps can reveal the targets of ongoing adaptation
278 in human populations. However, because the sojourn of an adaptive mutation to fixation should
279 be rapid, the success of efforts to detect ongoing selection implies the presence of a larger
280 number of recently completed sweeps. We have therefore focused on completed sweeps in order
281 to complement previous studies and to construct a more comprehensive catalogue of the loci
282 underpinning recent human adaptation. Using a powerful and robust machine learning method
283 that we have recently introduced (S/HIC; Schrider and Kern 2016) for finding completed
284 selective sweeps, we performed a genome-wide search for the targets of recent positive selection
285 in six human populations. Furthermore, we sought to determine the mode of positive selection,
286 distinguishing between selection on *de novo* mutations and on previously standing variation.
287

288 **Soft sweeps dominate human adaptation**

289 Perhaps our most consequential result is the finding that the majority of our candidate sweeps
290 resemble soft sweeps on standing variation. This result implies that adaptation in humans may
291 not be mutation-limited (Gillespie 1991; Karasov et al. 2010): rather than waiting for a novel
292 mutation to arise, human populations may often be able to respond via selection on previously
293 segregating polymorphisms, thereby more rapidly responding to novel environmental challenges.
294 This may be surprising given the apparently small effective population size and low nucleotide
295 diversity levels in humans. However, if the mutational target for the trait to be selected on is
296 fairly large, then the probability of a population harboring a mutation affecting that trait may be
297 appreciable.

298 While soft sweeps appear to be the dominant mode of selection globally, there is a
299 significant increase in the proportion of putative hard sweeps in non-African populations relative
300 to African populations. This is consistent with theoretical expectations, as larger populations
301 have more standing variation for selection to act on (Hermisson and Pennings 2005). Moreover,
302 the human migration out of Africa was associated with a severe population bottleneck (Marth et
303 al. 2004; Fagundes et al. 2007). Soft selective sweeps may be "hardened" by a reduction in
304 population size, which can result in the stochastic loss of some genetic backgrounds harboring
305 the adaptive allele so that only a single haplotype reaches fixation (Wilson et al. 2014). Thus,
306 though one might expect selection on segregating neutral or nearly neutral variation when a
307 population enters a new environment with novel selective pressures, if the migration event is
308 accompanied by a bottleneck then the population may experience a somewhat counterintuitive
309 increase in the proportion of hard sweeps. Moreover, the causal relationship between population
310 size and mode of adaptation may not be unidirectional. As Orr and Unckless (2014) have shown

311 in the context of evolutionary rescue, when faced with a changing environment, a population
312 which does not harbor standing variation that is beneficial may experience a more protracted
313 decline in size while it waits for an adaptive *de novo* mutation.

314 Our genome-wide results amplify results of earlier studies that by design have tried to
315 infer the mode of adaptation in a smaller number of targeted loci. For instance Peter *et al.* (2012)
316 attempted to infer the mode of adaptation among 7 loci previously identified to be under
317 selection in human populations. They report that half of the loci that they could confidently
318 classify supported selection on standing variation. In *Drosophila melanogaster,* when looking
319 among strong outliers of haplotype homozgosity, Garud *et al.* (2015) found that patterns of
320 variation in those regions were consistent with recent soft selective sweeps. Our finding, that the
321 vast majority of sweeps in human populations are soft sweeps, thus underscores the ubiquity of
322 selection from standing variation in natural populations. Indeed it seems plausible that adaptation
323 from standing variation might be the rule, rather than the exception.

324 There are two caveats affecting our ability to discriminate between selection on standing
325 variation and on *de novo* mutations. First, while we have trained our classifier to detect soft
326 sweeps on previously segregating mutations, soft sweeps may also occur via recurrent mutation
327 to the adaptive allele (Pennings and Hermisson 2006b; Pennings and Hermisson 2006a). Though
328 there are some qualitative differences between these two models of soft sweeps (Berg and Coop
329 2015; Schrider et al. 2015), these are fairly subtle in comparison to the differences between the
330 other models we consider. Thus, our classifiers may have sensitivity to both types of sweeps. If
331 this is so, then some of the soft sweeps that we detect may result from recurrent mutation.
332 Additionally, gene conversion during a sweep can transfer the adaptive mutation on to new
333 genetic backgrounds (Jones and Wakeley 2008), thereby "softening" the sweep (Schrider et al.

334 2015). This implies that selection on a single de novo mutation could sometimes appear to be a
335 soft sweep in our classification. In any case, our finding that most sweeps in humans do not
336 appear to be hard sweeps underscores the importance of using methods that are sensitive to soft
337 sweeps.
338
339 **Extensive impact of linked positive selection**
340 Our analysis demonstrates that the impact of linked positive selection on genetic variation is
341 considerable, with roughly half of the genome classified by S/HIC as being influenced by a
342 nearby sweep. This result has important implications for efforts to infer demographic histories
343 from patterns of genetic polymorphism, as most inference methods hinge on the assumption of
344 neutrality. Indeed, we have recently shown that linked positive selection has the potential to
345 severely confound demographic inferences (Schrider et al. 2016). Similarly, Ewing and Jensen
346 (2016) have found that background selection, the effect of purifying selection on polymorphism
347 at linked unselected sites (Charlesworth et al. 1993), can also bias demographic estimates. One
348 strategy is to use only those polymorphisms that are distant from genes and conserved noncoding
349 elements to mitigate these effects (Gazave et al. 2014). One could further supplement such an
350 approach by using S/HIC to directly ask which intergenic regions are unaffected by hitchhiking
351 in order to further diminish the bias introduced by linked selection.
352 If linked positive selection affects much of the genome, then that implies that the
353 frequencies of many neutral or weakly deleterious mutations may be altered by genetic draft
354 (Gillespie 2000). That is to say, deleterious mutations that happen to reside on chromosomes that
355 begin to sweep may be able to reach higher frequencies than expected from mutation-selection-
356 drift equilibrium. Consistent with this, we observe a slight but significant excess of potentially

357 deleterious polymorphisms in windows classified as linked to selective sweeps. Previously, Chun
358 and Fay (2011) found evidence that the ratio of deleterious to neutral polymorphisms is elevated
359 in sweep regions, concluding that hitchhiking carries linked deleterious variants to higher
360 frequencies. Our finding that SNPs from the GWAS catalogue are also enriched regions linked to
361 selective sweeps lends further support to this hypothesis. Indeed, several compelling examples of
362 hitchhiking mutations known or suspected of causing disease have been described in the
363 literature (Helgason et al. 2007; Chun and Fay 2011; Huff et al. 2012). It seems that the
364 phenomenon of deleterious alleles hitchhiking along with strongly beneficial alleles is not
365 restricted to humans: a recent study also uncovered evidence that selection during domestication
366 increased the frequency of deleterious polymorphisms in dogs (Marsden et al. 2016).
367

368 **Targets of recent human selective sweeps**

369 Our catalogue of sweep candidates allowed us to characterize the biological functions that are
370 overrepresented in sweeps. Notably, we found a strong excess of spermatogenesis genes within
371 sweep regions, a phenomenon previously observed by Voight et al. (2006). This signature may
372 be a result of sexual selection, sexual conflict, and/or sperm competition (Swanson and Vacquier
373 2002). We also observed a significant enrichment of cancer-related genes among our sweep
374 candidates. Nielsen et al. (2005a) found a similar enrichment of candidate genes under selection
375 related to cancer when examining protein divergence between humans and chimpanzees. These
376 authors found that some of these genes are also involved in spermatogenesis (much like our
377 *CADM1* example), and concluded that genomic conflict between tumor suppression and the
378 advantage of avoiding apoptosis during spermatogenesis may explain the selection on cancer
379 genes. An alternative (and non-mutually exclusive) explanation is that the increase in longevity

380 along the human lineage has created an immense selective pressure to reduce the rate of cancer
381 progression by orders of magnitude (Nunney and Muir 2015).

382 We also observed a significant excess of glutamate receptor genes targeted by sweeps,
383 suggesting that these loci may underlie some of the dramatic neurological changes that have
384 occurred along the human lineage. Consistent with this, we previously found evidence
385 suggesting some of these glutamate receptor genes may have recently gained novel regulatory
386 elements in humans (Schrider and Kern 2015). The most striking examples of glutamate
387 receptors experiencing sweeps are *GRIA2* and *GRID2*, which show strong signatures of selection
388 in multiple populations and physically interact with one another. The action of positive selection
389 on multiple members of the protein complex appears to be a general phenomenon (Figure 3). For
390 a more in-depth examination of positive selection in the PPI network, see Qian et al. (2015), who
391 found that genes in candidate regions for positive selection were more likely to lie close together
392 in the PPI network.

393

394 **Conclusions**

395 Our investigation has revealed several valuable insights into the adaptive process in human
396 populations. The success of our approach exemplifies the potential of machine learning methods
397 to elucidate the adaptive process in humans and other species (Fan et al. 2016). To date several
398 machine learning methods have been devised to detect selective sweeps (Pavlidis et al. 2010; Lin
399 et al. 2011; Ronen et al. 2013; Pybus et al. 2015; Sheehan and Song 2016), and they tend to
400 substantially outperform more traditional approaches (see Schrider and Kern 2016). We suspect
401 that machine learning could be used to make important inroads in answering a variety of
402 evolutionary questions.

403 Finally, Hernandez et al. (2011) argued that hard selective sweeps might be rare in human
404 populations, and instead suggested that the majority of adaptation might be a consequence of
405 selection on standing variation or selection on polygenic traits. We here find direct evidence that
406 indeed this is the case—the vast bulk of human adaptation is occurring as a consequence of soft
407 sweeps. Our observation thus reconciles Hernandez et al.'s findings with those of Enard et al.,
408 who conclude that the reduction in diversity around amino acid substitutions is caused by
409 widespread selective sweeps (Enard et al. 2014). Moreover, while our scan leveraged a method
410 that performs very well in detecting both hard and soft sweeps, it was not trained to detect cases
411 of polygenic selection (e.g. Berg and Coop 2014). It is fair to assume that a large majority of
412 phenotypes are determined by multiple loci, thus polygenic selection should be expected to be
413 common. If that were the case, then it could very well be that an even larger portion of genetic
414 variation is influenced by natural selection and its linked effects throughout the genome.
415

416 **METHODS**

417 **Sequence and annotation data**

418 We downloaded phased genotype data from Phase 3 of the 1000 Genomes Project (Auton et al.
419 2015). This data set consists of 26 population samples from Africa, East Asia, South Asia,
420 Europe, and the Americas. We wished to include only populations where the influence of
421 admixture/migration on genetic variation appeared to be minimal, while still allowing us to
422 characterize selection across multiple continents. We therefore chose to scan the following
423 populations for selective sweeps: the GWD (Gambians in Western Divisions in The Gambia) and
424 YRI (Yoruba in Ibadan, Nigeria) populations from West Africa, LWK (Luhya in Webuye,
425 Kenya) from East Africa, JPT (Japanese in Tokyo, Japan) from Asia, CEU (Utah residents with

426 Northern and Western European Ancestry) from Europe, and PEL (Peruvians from Lima, Peru)
427 from the Americas. Examining Auton et al.'s results from running ADMIXTURE (Alexander et
428 al. 2009), we see that for most values of *K*, each of these populations appears to correspond
429 primarily to a single ancestral population rather than displaying multiple clusters of ancestry (see
430 Extended Data Figure 5 from Auton et al. 2015). One exception may be the PEL population, but
431 among the highly admixed American samples it appears to exhibit the smallest amount of
432 possible mixed ancestry (for most values of *K*), so we retained this population in order to have
433 some representation from the Americas. We opted not to examine any South Asian population,
434 as for each of these samples ADMIXTURE inferred evidence of ancestry from three or more
435 ancestral populations.

436 We downloaded numerous annotation data sets containing genomic features to test for
437 enrichment/depletion of selective sweeps and perform other downstream analyses. These
438 included GENCODE gene model release 19 (Harrow et al. 2012) including pseudogenes (Pei et
439 al. 2012), virus-interacting proteins from Enard et al. (2016), enhancers gained or along the
440 human lineage since diverging from Old World monkeys (Cotney et al. 2013), and SIFT's
441 (Kumar et al. 2009) predictions of damaging amino acid polymorphisms from dbNSFP version
442 3.2a (Liu et al. 2016). We obtained Gene Ontology (GO) annotations from ENSEMBL release
443 75 (Yates et al. 2016). We also downloaded coordinates of previously identified selective sweeps
444 from dbPSHP (Li et al. 2013).

445 We used the UCSC Table Browser (Karolchik et al. 2004) to obtain the following data
446 sets: phenotype-associated SNPs from the GWAS Catalog (accessed Apr 12, 2016; Welter et al.
447 2014), ClinVar pathogenic SNPs and indels ≤ 20 bp in length (Apr 26, 2016; Landrum et al.
448 2016), COSMIC somatic mutations in cancer (accessed Feb 25, 2014; Forbes et al. 2015),

449 phastCons elements conserved across primates (accessed Jun 2, 2013; Siepel et al. 2005),
450 ENCODE transcription factor binding sites version 3 (accessed Aug 25, 2013; Dunham et al.
451 2012), tables of genes and SNPs implicated in Mendelian phenotypes from OMIM (accessed
452 May 2, 2016; Amberger et al. 2015), and KEGG pathway annotations (accessed Apr 27, 2016;
453 Kanehisa et al. 2015). For each of these data sets we used GRCh37/hg19 coordinates.

454 In order to examine the prevalence of selective sweeps within interacting gene networks,
455 we downloaded physical and genetic interactions from BioGRID version 3.4.136 (Chatr-
456 Aryamontri et al. 2015). Our set of genetic interactions consisted of those annotated as "synthetic
457 genetic interaction defined by inequality," "suppressive genetic interaction defined by
458 inequality," or "additive genetic interaction defined by inequality." Physical interactions
459 included those annotated as "direct interaction," "association," or "physical association." We
460 extracted transcription factor-target interactions from ORegAnno (accessed Dec 22, 2015;
461 Griffith et al. 2008), retaining only interacting pairs where the ENSEMBL gene identifier were
462 provided for both genes in order to avoid ambiguity.

463

464 **Building classifiers to detect selective sweeps**

465 To detect sweeps we used S/HIC (https://github.com/kern-lab/shIC), a machine learning
466 approach we previously described and showed to be remarkably powerful and robust to non-
467 equilibrium demography (Schrider and Kern 2016). Briefly, the S/HIC machine learning
468 approach leverages spatial patterns (along a genome) of a variety of population genetic summary
469 statistics to classify genomic windows as being the target of a completed hard sweep (hard),
470 being closely linked to a hard sweep (hard-linked), a completed soft sweep (soft), linked to a soft
471 sweep (soft-linked), or evolving neutrally (neutral). While this classification approach allows

472 inference when considering a large number of features jointly, it necessitates training from a
473 large number of data instances known to belong to each class. Because the number of genomic
474 windows known to belong to each our five classes is limited, we must rely on simulation to
475 generate our training data. To this end we used the program discoal (Kern and Schrider 2016) to
476 simulate large chromosomal regions, subdivided into 11 sub-windows. Training examples for the
477 hard class experienced a hard sweep in the center of the central sub-window (i.e. the 6th
478 window), while examples for the hard-linked class experienced a hard sweep in the center of one
479 of the remaining sub-windows (selected randomly). Analogous simulations with soft sweeps
480 were generated for the soft and soft-linked classes, respectively. Finally, neutrally evolving
481 examples did not experience any selective sweep.

482 We sought to train a classifier for each population under a demographic model that offers
483 a better approximation to the population size history than the standard neutral model. For this we
484 used Auton et al.'s (2015) population histories inferred by PSMC (Li and Durbin 2011). The
485 1000 Genomes Project's PSMC output did not contain estimates of θ, the population mutation
486 rate parameter. Thus for each population we conducted a grid search by simulating genomic
487 windows with the appropriate sample size under each demographic model with varying values of
488 θ=4NuL (where L is the length of the locus, which we set to 100 kb); the grid of θ values raged
489 from 10 to 250, examining multiples of 10. For each value of θ, we compared the values of π
490 (Nei and Li 1979), $\hat{\theta}_w$ (Watterson 1975), $\hat{\theta}_H$ (Fay and Wu 2000), H_2/H_1 (Garud et al. 2015), and
491 Z_{nS} (Kelly 1997) from 1000 simulations to those from 1000 randomly selected genomic loci
492 (calculated as described below), calculating the mean of each statistic in the real and simulated
493 datasets. We chose as the final values of θ that for which the sum of the percent deviations of the
494 simulated from the observed means of each statistic was minimized. This estimate of θ allowed

us to calculate estimated population sizes and times scaled by the number generations for each time point in the history inferred by PSMC. The harmonic mean of each population's size was calculated by taking the estimated population size for each of the last 4*N* generations. We note that these models may not accurately capture the demographic histories of the populations we examined due to the confounding effects of positive (Schrider et al. 2016) and negative (Ewing and Jensen 2016) selection. However, because of S/HIC's robustness to demographic misspecification, we do not expect this to severely impact our analysis (Schrider and Kern 2016).

For each population we simulated a total of 2000 regions for each of our five classes. For simulations involving sweeps, we drew the selection coefficient from $U(0.005, 0.1)$, the sweep completion time from $U(0, 2000)$, the initial selected frequency for soft sweeps from $U(1/N, 0.2)$. We drew values of θ uniformly from a range spanning exactly one order of magnitude, specified so that the mean value of θ was equal to that estimated for the population as described above. We drew recombination rates from an exponential distribution with mean 1×10^{-8}, truncated at triple the mean due to memory constraints. The simulation program discoal requires some of these parameters to be scaled by the present-day effective population size; we did this by taking the mean value of θ and dividing by $4uL$, where u was set to 1.2×10^{-8} (Kong et al. 2012). The full command lines we used to generate 1.1 Mb regions (to be subdivided into 11 windows each 100 kb in length) for each population are shown in Table S3. We also simulated 1000 test examples for each population in the same manner as for the training data.

Our feature vector for each simulated region examined the spatial patters (following Schrider and Kern 2016) of each of the following statistics: π (Nei and Li 1979), $\hat{\theta}_w$ (Watterson 1975), $\hat{\theta}_H$ (Fay and Wu 2000), the number of distinct haplotypes, average haplotype homozygosity, Garud et al.'s (2015) H_{12} and H_2/H_1 statistics, Z_{nS}, ω (Kim and Nielsen 2004), and

518 the maximum frequency of derived mutations (Li 2011). Before calculating these summary
519 statistics we masked a number of sites within each simulation by randomly selecting a 1.1 Mb
520 region from our empirical windows sampled throughout the genome and masking the same
521 regions in the simulated window as were masked in the genomic window (see below). Thus our
522 simulated windows exhibit the same distribution of regions of missing data as the windows to
523 which we applied our classifiers. We then used S/HIC to train extra-trees classifiers (Geurts et al.
524 2006), one for each population.
525
526 **Classifying genomic windows in each population**
527 Having trained our classifiers, we then applied them to genomic data from the corresponding
528 population. We inferred ancestral states of polymorphisms and masked inaccessible sites
529 (whether polymorphic or not) in the same manner as described previously (Schrider and Kern
530 2016). We then used S/HIC to classify the central 100 kb sub-window of 1.1 Mb windows across
531 the autosomes, while taking the stringent approach of omitting those for which any sub-window
532 was less than 25% accessible, before sliding 100 kb downstream to examine the next window.
533 We also removed windows where any of the three central sub-windows had a mean
534 recombination rate of zero (using data from Kong et al. 2010). Importantly, for each retained 1.1
535 Mb window, we recorded the locations of all sites deemed inaccessible for use in masking our
536 training data (see above). In total we classified 13,968 windows, accounting for 48.5% of the
537 assembled autosomes.
538 In order to count the number of distinct sweep candidates found within our set of
539 populations , we simply merged all 100 kb windows classified as a sweep of either type that were
540 located either at the exact same coordinates or adjacent to one another, repeating this until no

541 more sweep regions could be merged. If all constituent windows were classified as soft, we
542 counted the sweep as soft; otherwise we counted it as a hard sweep. We used a similar approach
543 but examining classifications from only one population at a time in order to count the number of
544 sweeps of each type in that population. If a gene found within a sweep window identified by
545 S/HIC was not found in an entry of dbPSHP (Li et al. 2013), we referred to it as a novel sweep.
546 Visualization of sweep candidates was performed using the UCSC Genome Browser (Kent et al.
547 2002) , along with custom tracks showing values of various population genetic summary
548 statistics and selection scan scores for the CEU, YRI, and JPT populations from the Human
549 Positive Selection Browser (Pybus et al. 2013). Our classification results are available at
550 http://kerndev.rutgers.edu/~dan/shicHumanScan/.
551

552 **Permutation tests for enrichment of annotation features in sweeps**

553 To determine whether certain annotation features were enriched within any of our five classes,
554 we carefully designed a permutation test to account for the subset of the genome that we
555 examined with S/HIC, as well as the spatial correlation of S/HIC's classifications (i.e. adjacent
556 windows are especially likely to receive the same classification). Briefly, the permutation
557 algorithm begins by examining our classification results for a given population and keeping track
558 of the length of runs of consecutive windows assigned to each class. The permutation algorithm
559 then selects a chromosome, and begins at its first classified window (i.e. not removed by data
560 filtering). A run length and associated class assignment is then randomly drawn without
561 replacement. This process continues until the end of the chromosome, and then another
562 chromosome is selected until the end of the final chromosome is reached, at which point the
563 permutation has been completed. We then repeated this permutation procedure 10,000 times for

564 each population. Note that this process preserves the run length distribution of our classifications
565 while permuting them across the set of genomic windows that had enough unmasked data to be
566 included in our scan.

567 After constructing our permuted data sets, we conducted one-sided enrichment tests by
568 counting the number of base pairs in the intersection between the S/HIC class of interest and the
569 annotation feature of interest, and comparing this number to its distribution among the permuted
570 data sets. The fraction of permuted data sets where this intersect was greater than or equal to that
571 observed for the real data is the *p*-value. Because we tested each of S/HIC's five classes for
572 enrichment of a fairly large number of genomic features (Table S1), we corrected for multiple
573 testing using false discovery rate *q*-values following Storey (2002). When testing GO terms and
574 KEGG pathways for enrichment, we considered only the hard and soft sweep classes, corrected
575 for calculating *q*-values separately for each class.

576 We also asked whether the number of pairs of interacting genes both overlapping
577 windows classified as sweeps was greater than in our permuted data sets. To ensure that our
578 results were not inflated by the spatial clustering of interacting genes, we only counted
579 interacting pairs overlapping sweep windows if they were separated by at least 10 Mb or on
580 separate chromosomes. In addition, if we observed an interaction between two genes, *A* and *B*,
581 that each overlapped sweeps, and a third sweep candidate gene, *C*, was found, to avoid
582 redundancy we counted at most one interaction between *A* and *C* and *B* and *C*, even if *C* was
583 found interact with both other genes. As with GO and KEGG terms, we only searched the hard
584 and soft classes for enrichments before calculating one-sided q-values as described above.

585

586 **ACKNOWLEDGMENTS**

587 We thank Matthew Hahn for feedback on the manuscript. A.D.K. was supported in part by NIH
588 award no. R01GM078204.

589

590 **DISCLOSURE DECLARATION**

591 The authors have no conflicts of interest to declare.

592

593 **REFERENCES**

594

595 Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go
596 from here? *Genome Res* **19**: 711-722.
597 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
598 unrelated individuals. *Genome Res* **19**: 1655-1664.
599 Allinen M, Peri L, Kujala S, Lahti‐Domenici J, Outila K, Karppinen SM, Launonen V, Winqvist
600 R. 2002. Analysis of 11q21–24 loss of heterozygosity candidate target genes in breast
601 cancer: indications of TSLC1 promoter hypermethylation. *Genes Chromosomes Cancer*
602 **34**: 384-389.
603 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM. org: Online
604 Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic
605 disorders. *Nucleic Acids Res* **43**: D789-D798.
606 Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the
607 Drosophila melanogaster genome. *Genome Res* **17**: 1755-1762.
608 Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy
609 S, McVean GA, Abecasis GR et al. 2015. A global reference for human genetic variation.
610 *Nature* **526**: 68-74.
611 Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD,
612 Kern AD, Dewey CN. 2007. Population genomics: whole-genome analysis of
613 polymorphism and divergence in Drosophila simulans. *PLoS Biol* **5**: e310.
614 Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet* **10**:
615 e1004412.
616 Berg JJ, Coop G. 2015. A Coalescent Model for a Sweep of a Unique Standing Variant. *Genetics*
617 **201**: 707-725.
618 Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proceedings of*
619 *the National Academy of Sciences* **85**: 6414-6418.
620 Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S. 2008. Positive
621 selection in East Asians for an EDAR allele that enhances NF-κB activation. *PLoS ONE*
622 **3**: e2209.
623 Charlesworth B, Morgan M, Charlesworth D. 1993. The effect of deleterious mutations on
624 neutral molecular variation. *Genetics* **134**: 1289-1303.

625 Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C,
626 Breitkreutz A, Kolas N, O'Donnell L. 2015. The BioGRID interaction database: 2015
627 update. *Nucleic Acids Res* **43**: D470-D478.
628 Chun S, Fay JC. 2011. Evidence for hitchhiking of deleterious mutations within the human
629 genome. *PLoS Genet* **7**: e1002240.
630 Consortium UK. 2015. The UK10K project identifies rare variants in health and disease. *Nature*
631 **526**: 82-90.
632 Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP.
633 2013. The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic
634 Limb. *Cell* **154**: 185-196.
635 Dunham I Kundaje A Aldred SF Collins PJ Davis C Doyle F Epstein CB Frietze S Harrow J
636 Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome.
637 *Nature* **489**: 57-74.
638 Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein
639 adaptation in mammals. *eLife* **5**: e12469.
640 Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human
641 evolution. *Genome Res*.
642 Ewing GB, Jensen JD. 2016. The consequences of not accounting for background selection in
643 demographic inference. *Mol Ecol* **25**: 135-141.
644 Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L.
645 2007. Statistical evaluation of alternative models of human evolution. *Proceedings of the*
646 *National Academy of Sciences* **104**: 17614-17619.
647 Fan S, Hansen ME, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of recent
648 human adaptation. *Science* **354**: 54-59.
649 Faraji F, Pang Y, Walker RC, Borges RN, Yang L, Hunter KW. 2012. Cadm1 is a metastasis
650 susceptibility gene that suppresses metastasis by modifying tumor interaction with the
651 cell-mediated immunity. *PLoS Genet* **8**: e1002926.
652 Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405-
653 1413.
654 Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P,
655 McCarthy MI et al. 2016. Detection of human adaptation during the past 2000 years.
656 *Science* doi:10.1126/science.aag0776.
657 Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S,
658 Cole C, Ward S. 2015. COSMIC: exploring the world's knowledge of somatic mutations
659 in human cancer. *Nucleic Acids Res* **43**: D805-D811.
660 Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking
661 and background selection. *Genetics* **147**: 915-925.
662 Fujita E, Dai H, Tanabe Y, Zhiling Y, Yamagata T, Miyakawa T, Tanokura M, Momoi M,
663 Momoi T. 2010. Autism spectrum disorder is related to endoplasmic reticulum stress
664 induced by mutations in the synaptic cell adhesion molecule, CADM1. *Cell death &*
665 *disease* **1**: e47.
666 Fujita E, Tanabe Y, Imhof BA, Momoi MY, Momoi T. 2012. A complex of synaptic adhesion
667 molecule CADM1, a molecule related to autism spectrum disorder, with MUPP1 in the
668 cerebellum. *J Neurochem* **123**: 886-894.

669 Fukuhara H, Kuramochi M, Fukami T, Kasahara K, Furuhata M, Nobukuni T, Maruyama T,
670 Isogai K, Sekiya T, Shuin T. 2002. Promoter methylation of TSLC1 and tumor
671 suppression by its gene product in human prostate cancer. *Jpn J Cancer Res* **93**: 605-609.
672 Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North
673 American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet* **11**:
674 e1005004.
675 Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, Boerwinkle E, Gibbs RA, Sing CF,
676 Clark AG et al. 2014. Neutral genomic regions refine models of recent rapid human
677 population growth. *Proceedings of the National Academy of Sciences* **111**: 757-762.
678 Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. *Machine Learning* **63**: 3-42.
679 Gillespie JH. 1991. *The causes of molecular evolution*. Oxford University Press, Oxford.
680 Gillespie JH. 2000. Genetic drift in an infinite population: the pseudohitchhiking model.
681 *Genetics* **155**: 909-919.
682 Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA,
683 Bustamante CD, Altshuler DL. 2011. Demographic history and rare allele sharing among
684 human populations. *Proceedings of the National Academy of Sciences* **108**: 11983-11988.
685 Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC,
686 Bilenky M, Haeussler M. 2008. ORegAnno: an open-access community-driven resource
687 for regulatory annotation. *Nucleic Acids Res* **36**: D107-D113.
688 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell
689 D, Zadissa A, Searle S. 2012. GENCODE: The reference human genome annotation for
690 The ENCODE Project. *Genome Res* **22**: 1760-1774.
691 Hartfield M, Otto SP. 2011. Recombination and hitchhiking of deleterious alleles. *Evolution* **65**:
692 2421-2434.
693 Helgason A, Pálsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, Adeyemo A,
694 Chen Y, Chen G, Reynisdottir I. 2007. Refining the impact of TCF7L2 gene variants on
695 type 2 diabetes and adaptive evolution. *Nat Genet* **39**: 218-225.
696 Hermisson J, Pennings PS. 2005. Soft sweeps molecular population genetics of adaptation from
697 standing genetic variation. *Genetics* **169**: 2335-2352.
698 Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M.
699 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920-
700 924.
701 Huerta-Sánchez E, Jin X, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel
702 M, Ni P. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like
703 DNA. *Nature* **512**: 194-197.
704 Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, Hakonarson H,
705 Whiting A, Davis CT, Wu W. 2012. Crohn's disease and genetic hitchhiking at IBD5.
706 *Mol Biol Evol* **29**: 101-111.
707 Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a
708 domestication event. *Proc Natl Acad Sci U S A* **101**: 10667-10672.
709 Jensen JD, Thornton KR, Andolfatto P. 2008. An approximate Bayesian estimator suggests
710 strong, recurrent selective sweeps in Drosophila. *PLoS Genet* **4**: e1000198.
711 Jones DA, Wakeley J. 2008. The influence of gene conversion on linkage disequilibrium around
712 a selective sweep. *Genetics* **180**: 1251-1259.
713 Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2015. KEGG as a reference
714 resource for gene and protein annotation. *Nucleic Acids Res*: gkv1070.

715 Kaplan NL, Hudson R, Langley C. 1989. The" hitchhiking effect" revisited. *Genetics* **123**: 887-
716 899.
717 Karasov T, Messer PW, Petrov DA. 2010. Evidence that adaptation in Drosophila is not limited
718 by mutation at single sites. *PLoS Genet* **6**: e1000924.
719 Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The
720 UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-D496.
721 Kelly JK. 1997. A test of neutrality based on interlocus associations. *Genetics* **146**: 1197-1206.
722 Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12**: 656-664.
723 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The
724 human genome browser at UCSC. *Genome Res* **12**: 996-1006.
725 Kern AD, Jones CD, Begun DJ. 2002. Genomic effects of nucleotide substitutions in Drosophila
726 simulans. *Genetics* **162**: 1753-1761.
727 Kern AD, Schrider DR. 2016. discoal: flexible coalescent simulations with selection.
728 *Bioinformatics*: btw556.
729 Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics*
730 **167**: 1513-1524.
731 Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining
732 chromosome. *Genetics* **160**: 765-777.
733 Kohda K, Kamiya Y, Matsuda S, Kato K, Umemori H, Yuzaki M. 2003. Heteromer formation of
734 δ2 glutamate receptors with AMPA or kainate receptors. *Molecular Brain Research* **110**:
735 27-37.
736 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA,
737 Sigurdsson A, Jonasdottir A, Jonasdottir A. 2012. Rate of *de novo* mutations and the
738 importance of father's age to disease risk. *Nature* **488**: 471-475.
739 Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB,
740 Jonasdottir A, Gylfason A, Kristinsson KT et al. 2010. Fine-scale recombination rate
741 differences between sexes, populations and individuals. *Nature* **467**: 1099-1103.
742 Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants
743 on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.
744 Kuramochi M, Fukuhara H, Nobukuni T, Kanbe T, Maruyama T, Ghosh HP, Pletcher M,
745 Isomura M, Onizuka M, Kitamura T. 2001. TSLC1 is a tumor-suppressor gene in human
746 non-small-cell lung cancer. *Nat Genet* **27**: 427-430.
747 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle
748 M, FitzHugh W. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:
749 860-921.
750 Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D,
751 Hoover J. 2016. ClinVar: public archive of interpretations of clinically relevant variants.
752 *Nucleic Acids Res* **44**: D862-D868.
753 Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C,
754 Corbett-Detig RB, Kolaczkowski B. 2012. Genomic variation in natural populations of
755 *Drosophila melanogaster*. *Genetics* **192**: 533-598.
756 Lee YCG, Langley CH, Begun DJ. 2013. Differential strengths of positive selection revealed by
757 hitchhiking effects at small physical scales in Drosophila melanogaster. *Mol Biol Evol*:
758 mst270.
759 Li H. 2011. A new test for detecting recent positive selection that is free from the confounding
760 impacts of demography. *Mol Biol Evol* **28**: 365-375.

761 Li H, Durbin R. 2011. Inference of human population history from individual whole-genome
762 sequences. *Nature* **475**: 493-496.
763 Li MJ, Wang LY, Xia Z, Wong MP, Sham PC, Wang J. 2013. dbPSHP: a database of recent
764 positive selection across human populations. *Nucleic Acids Res*: gkt1052.
765 Lin K, Li H, Schlötterer C, Futschik A. 2011. Distinguishing positive selection from neutral
766 evolution: boosting the performance of summary statistics. *Genetics* **187**: 229-244.
767 Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y. 2012.
768 Extension of cortical synaptic development distinguishes humans from chimpanzees and
769 macaques. *Genome Res* **22**: 611-622.
770 Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3. 0: A One‐Stop Database of Functional
771 Predictions and Annotations for Human Nonsynonymous and Splice‐Site SNVs. *Hum*
772 *Mutat* **37**: 235-241.
773 Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G,
774 Huerta-Sanchez E, Feder AF, Grarup N. 2011. Natural selection affects multiple aspects
775 of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**:
776 e1002326.
777 Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F,
778 Young N. 2013. The genotype-tissue expression (GTEx) project. *Nat Genet* **45**: 580-585.
779 Luján R, Shigemoto R, Lopez-Bendito G. 2005. Glutamate and GABA receptor signalling in the
780 developing brain. *Neuroscience* **130**: 567-580.
781 Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence
782 between nonsynonymous divergence and neutral polymorphism reveals extensive
783 adaptation in Drosophila. *Genetics* **177**: 2083-2099.
784 Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, Marques-
785 Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. 2016. Bottlenecks and selective
786 sweeps during domestication have increased deleterious genetic variation in dogs.
787 *Proceedings of the National Academy of Sciences* **113**: 152-157.
788 Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-
789 wide human variation data reveals signals of differential demographic history in three
790 large world populations. *Genetics* **166**: 351-372.
791 Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23-
792 35.
793 Nei M, Li W-H. 1979. Mathematical model for studying genetic variation in terms of restriction
794 endonucleases. *Proceedings of the National Academy of Sciences* **76**: 5269-5273.
795 Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A,
796 Tanenbaum DM, Civello D, White TJ. 2005a. A scan for positively selected genes in the
797 genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.
798 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005b. Genomic scans
799 for selective sweeps using SNP data. *Genome Res* **15**: 1566-1575.
800 Nunney L, Muir B. 2015. Peto's paradox and the hallmarks of cancer: constructing an
801 evolutionary framework for understanding the incidence of cancer. *Phil Trans R Soc B*
802 **370**: 20150161.
803 Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic
804 variation. *Genetics* **157**: 875-884.
805 Orr HA, Unckless RL. 2014. The population genetics of evolutionary rescue. *PLoS Genet* **10**:
806 e1004551.

807 Pavlidis P, Jensen JD, Stephan W. 2010. Searching for footprints of positive selection in whole-
808 genome SNP data from nonequilibrium populations. *Genetics* **185**: 907-922.
809 Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer
810 A, Diekhans M. 2012. The GENCODE pseudogene resource. *Genome Biol* **13**: 1.
811 Pennings PS, Hermisson J. 2006a. Soft sweeps II—molecular population genetics of adaptation
812 from recurrent mutation or migration. *Mol Biol Evol* **23**: 1076-1084.
813 Pennings PS, Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from
814 recurrent mutation. *PLoS Genet* **2**: e186.
815 Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS,
816 Myers RM, Feldman MW. 2009. Signals of recent positive selection in a worldwide
817 sample of human populations. *Genome Res* **19**: 826-837.
818 Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic
819 variation. *Evolution* **59**: 2312-2323.
820 Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Laayouni H,
821 Bertranpetit J, Engelken J. 2013. 1000 Genomes Selection Browser 1.0: a genome
822 browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids*
823 *Res*: gkt1188.
824 Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J. 2015.
825 Hierarchical boosting: a machine-learning framework to detect and classify hard selective
826 sweeps in human populations. *Bioinformatics*: btv493.
827 Qian W, Zhou H, Tang K. 2015. Recent Coselection in Human Populations Revealed by
828 Protein–Protein Interaction Network. *Genome Biol Evol* **7**: 136-153.
829 Ronen R, Udpa N, Halperin E, Bafna V. 2013. Learning natural selection from the site frequency
830 spectrum. *Genetics* **195**: 181-193.
831 Ruwende C, Khoo S, Snow R, Yates S, Kwiatkowski D, Gupta S, Warn P, Allsopp C, Gilbert S,
832 Peschu N. 1995. Natural selection of hemi-and heterozygotes for G6PD deficiency in
833 Africa by resistance to severe malaria. *Nature* **376**: 246-249.
834 Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV,
835 Patterson NJ, McDonald GJ. 2002. Detecting recent positive selection in the human
836 genome from haplotype structure. *Nature* **419**: 832-837.
837 Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll
838 SA, Gaudet R. 2007. Genome-wide detection and characterization of positive selection in
839 human populations. *Nature* **449**: 913-918.
840 Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. 2011. Pervasive adaptive protein evolution
841 apparent in diversity patterns around amino acid substitutions in Drosophila simulans.
842 *PLoS Genet* **7**: e1001302.
843 Schrider D, Shanku A, Kern A. 2016. Effects of Linked Selective Sweeps on Demographic
844 Inference and Model Selection. *Genetics* doi:10.1534/genetics.116.190223.
845 Schrider DR, Kern AD. 2015. Inferring selective constraint from population genomic data
846 suggests recent regulatory turnover in the human brain. *Genome Biol Evol* **7**: 3511-3528.
847 Schrider DR, Kern AD. 2016. S/HIC: Robust Identification of Soft and Hard Sweeps Using
848 Machine Learning. *PLoS Genet* **12**: e1005928.
849 Schrider DR, Mendes FK, Hahn MW, Kern AD. 2015. Soft shoulders ahead: spurious signatures
850 of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**: 267-
851 284.

852 Sheehan S, Song YS. 2016. Deep learning for population genetic inference. *PLoS Comput Biol*
853 **12**: e1004845.
854 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J,
855 Hillier LW, Richards S. 2005. Evolutionarily conserved elements in vertebrate, insect,
856 worm, and yeast genomes. *Genome Res* **15**: 1034-1050.
857 Stephan W, Wiehe TH, Lenz MW. 1992. The effect of strongly selected substitutions on neutral
858 polymorphism: analytical results based on diffusion theory. *Theor Popul Biol* **41**: 237-
859 254.
860 Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley
861 GA, Allikmets R, Schriml L. 1998. Dating the origin of the CCR5-Δ32 AIDS-resistance
862 allele by the coalescence of haplotypes. *The American Journal of Human Genetics* **62**:
863 1507-1515.
864 Storey JD. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical*
865 *Society: Series B (Statistical Methodology)* **64**: 479-498.
866 Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nature Reviews*
867 *Genetics* **3**: 137-144.
868 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
869 polymorphism. *Genetics* **123**: 585-595.
870 Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen
871 HM, Hirbo JB, Osman M. 2007. Convergent adaptation of human lactase persistence in
872 Africa and Europe. *Nat Genet* **39**: 31-40.
873 Van Der Weyden L, Arends MJ, Chausiaux OE, Ellis PJ, Lange UC, Surani MA, Affara N,
874 Murakami Y, Adams DJ, Bradley A. 2006. Loss of TSLC1 causes male infertility due to
875 a defect at the spermatid stage of spermatogenesis. *Mol Cell Biol* **26**: 3595-3609.
876 Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the
877 human genome. *PLoS Biol* **4**: e72.
878 Watterson G. 1975. On the number of segregating sites in genetical models without
879 recombination. *Theor Popul Biol* **7**: 256-276.
880 Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T,
881 Hindorff L. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait
882 associations. *Nucleic Acids Res* **42**: D1001-D1006.
883 Wiehe T, Stephan W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA
884 polymorphism data from Drosophila melanogaster. *Mol Biol Evol* **10**: 842-854.
885 Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007.
886 Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3**: e90.
887 Wilson BA, Petrov DA, Messer PW. 2014. Soft selective sweeps in complex demographic
888 scenarios. *Genetics* **198**: 669-684.
889 Wu YY, Yang Y, Xu YD, Yu HL. 2015. Targeted disruption of the spermatid‑specific gene
890 Spata31 causes male infertility. *Mol Reprod Dev* **82**: 432-440.
891 Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham
892 P, Fitzgerald S, Gil L. 2016. Ensembl 2016. *Nucleic Acids Res* **44**: D710-D716.
893 Zhang P, Si X, Skogerbø G, Wang J, Cui D, Li Y, Sun X, Liu L, Sun B, Chen R. 2014. piRBase:
894 a web resource assisting piRNA functional study. *Database* **2014**: bau110.
895 Zhiling Y, Fujita E, Tanabe Y, Yamagata T, Momoi T, Momoi MY. 2008. Mutations in the gene
896 encoding CADM1 are associated with autism spectrum disorder. *Biochem Biophys Res*
897 *Commun* **377**: 926-929.

900 **Table 1: Number of sweeps of each type detected in each population sample.**

901

Population	# of Hard Sweeps	# of Soft Sweeps
JPT	61 (5.8%)	998 (94.2%)
CEU	66 (6.5%)	947 (93.5%)
PEL	32 (4.7%)	655 (95.3%)
GWD	5 (0.6%)	795 (99.4%)
YRI	13 (1.6%)	797 (98.4%)
LWK	3 (0.4%)	805 (99.6%)

902

903

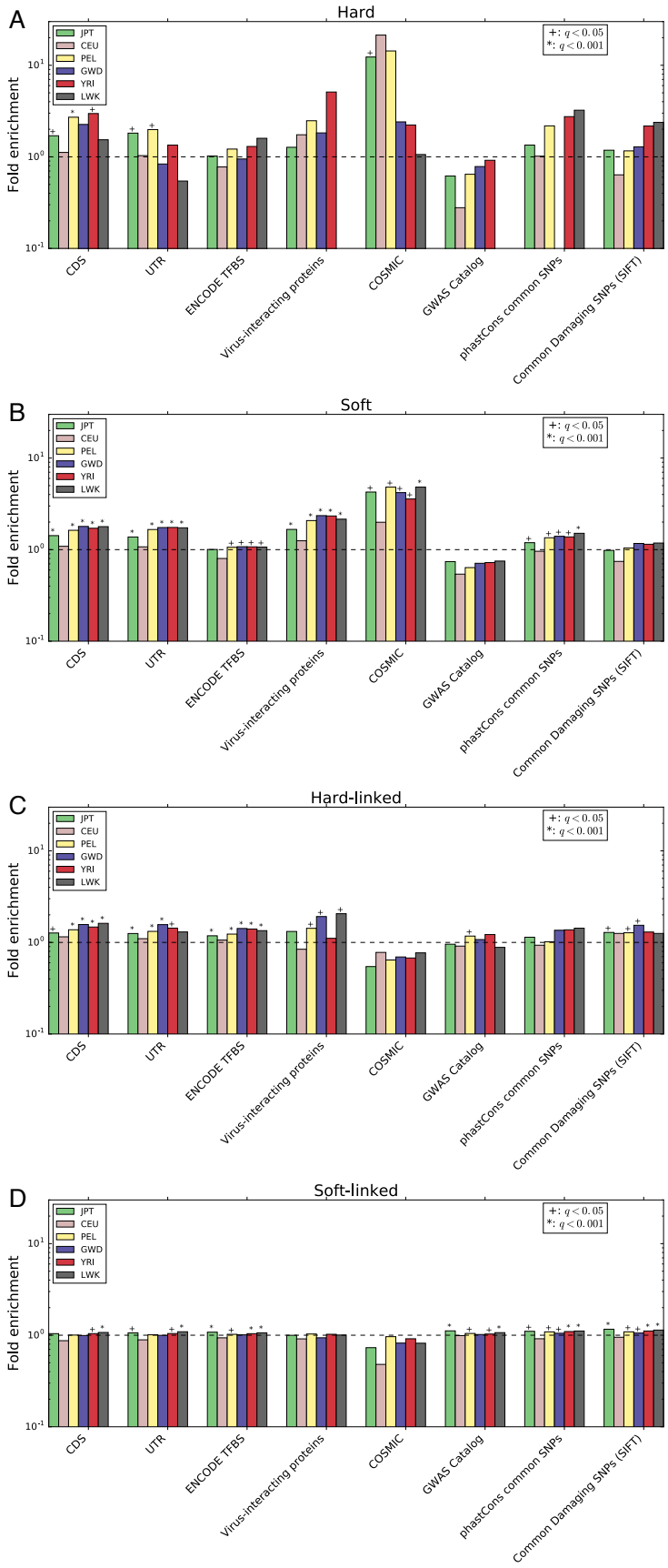

Figure 1: Enrichment of various annotation features in regions classified as sweeps or linked to sweeps relative. The fold enrichment is the ratio of the number of base pairs in the intersection between windows assigned to a given class and an annotation feature divided by the mean of this intersection across the permuted data sets (Methods). This was calculated separately for each population. (A) Enrichment of elements in windows classified as hard sweeps. (B) Same as A, but for soft sweeps. (C) Enrichment of elements in windows classified as affected by linked hard sweeps. (D) Linked soft sweeps.

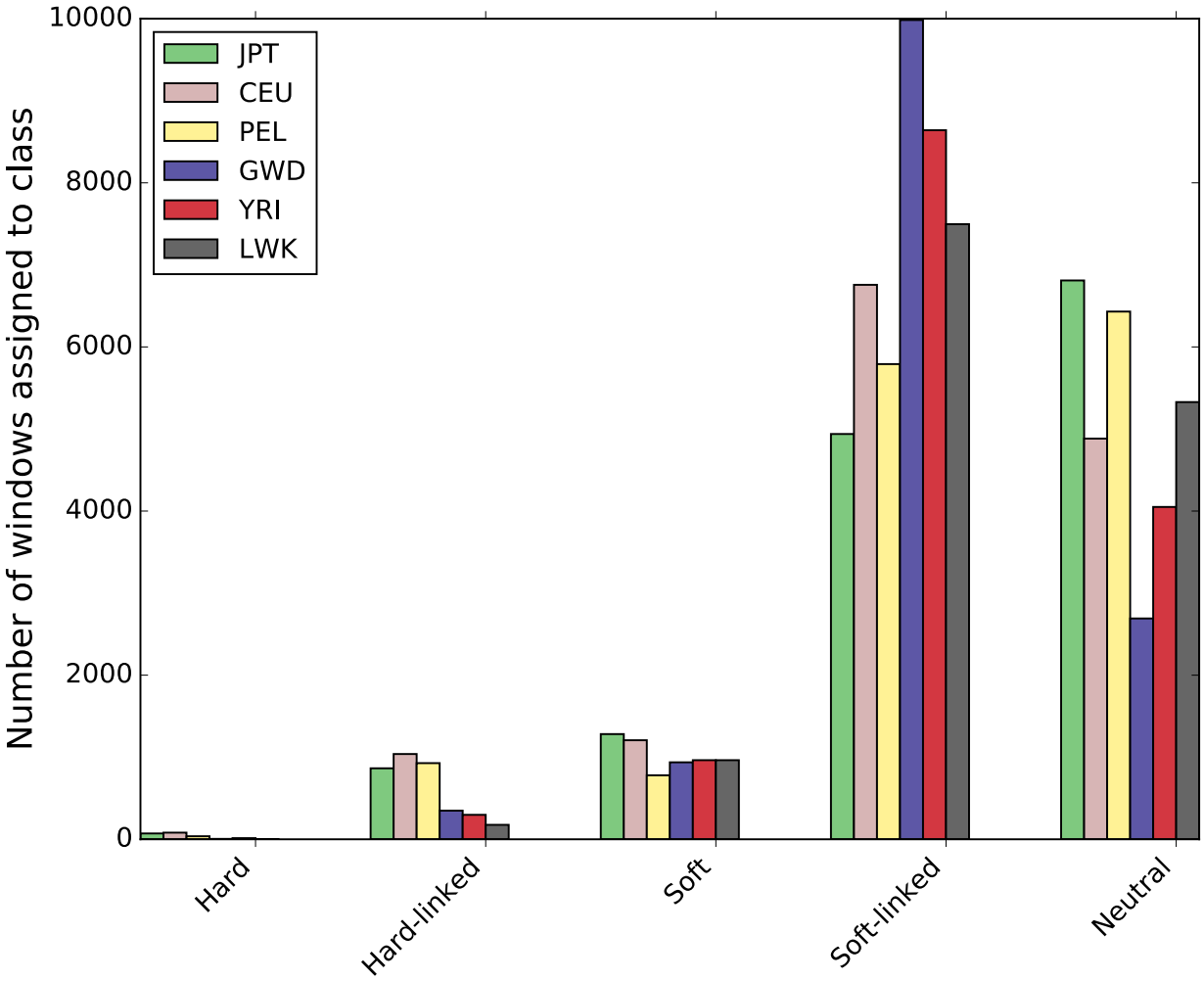

Figure 2: The number of windows assigned to each class by S/HIC in each population.

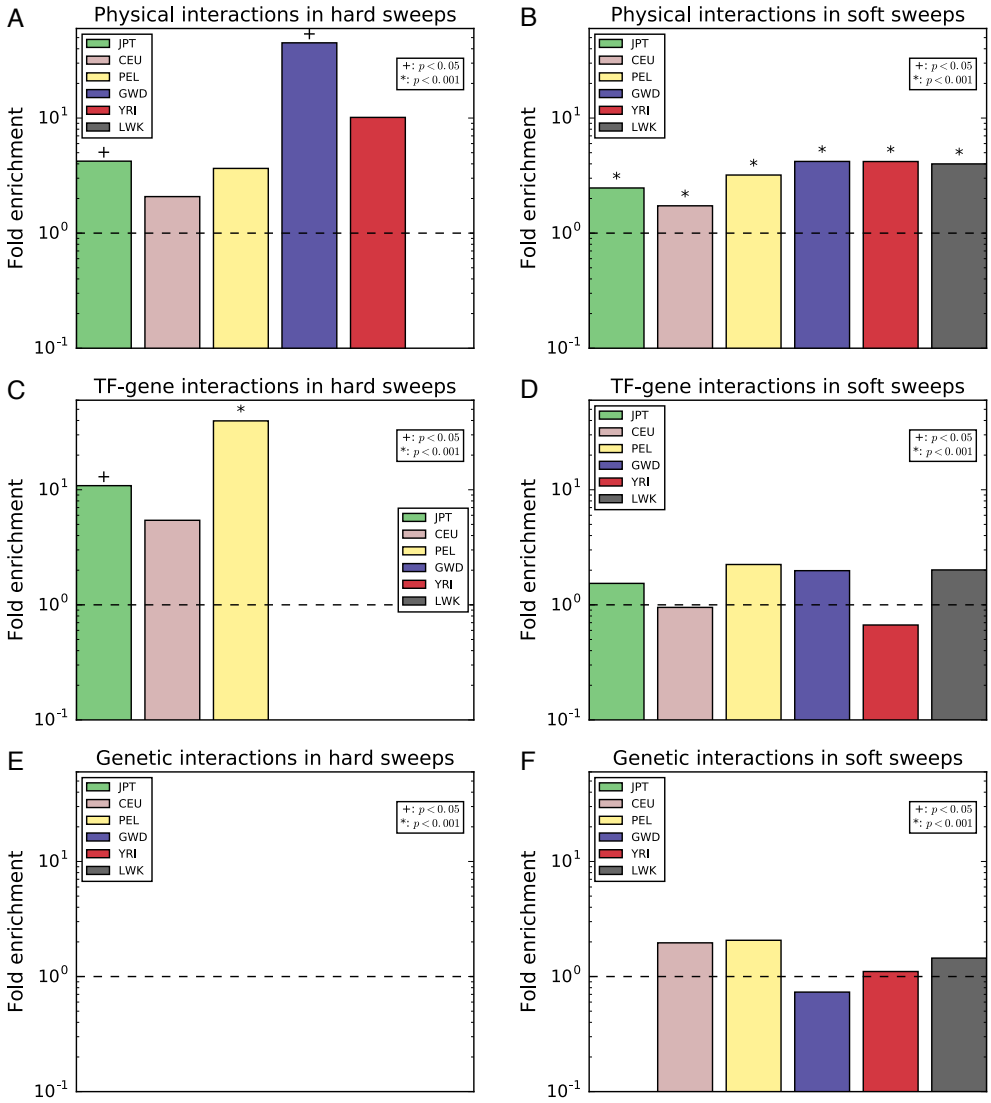

Figure 3: Enrichment of pairs of interacting genes each falling within a window classified as a sweep. The fold enrichment is the ratio of the number of pairs of interacting genes overlapping a window classified as a sweep of a given type divided by the mean of this number across the permuted data sets (Methods). This was calculated separately for each population. When no pairs of interacting sweep genes were observed in our true data set or a population, no bar was drawn. (A) Enrichment of pairs of genes encoding protein products that physically interact with each other (data from BioGRID) and both overlap hard sweep windows. (B) Same as A, but for soft sweeps. (C) Enrichment of pairs of genes, one of which is encodes a transcription factor that affects expression of the other (data from ORegAnno), where both overlap hard sweep windows. (D) Same as D, but for soft sweeps. (E) Enrichment of pairs of genes for which a genetic interaction has been observed (data from BioGRID) and both overlap hard sweep windows. (F) Same as E, but for soft sweeps.

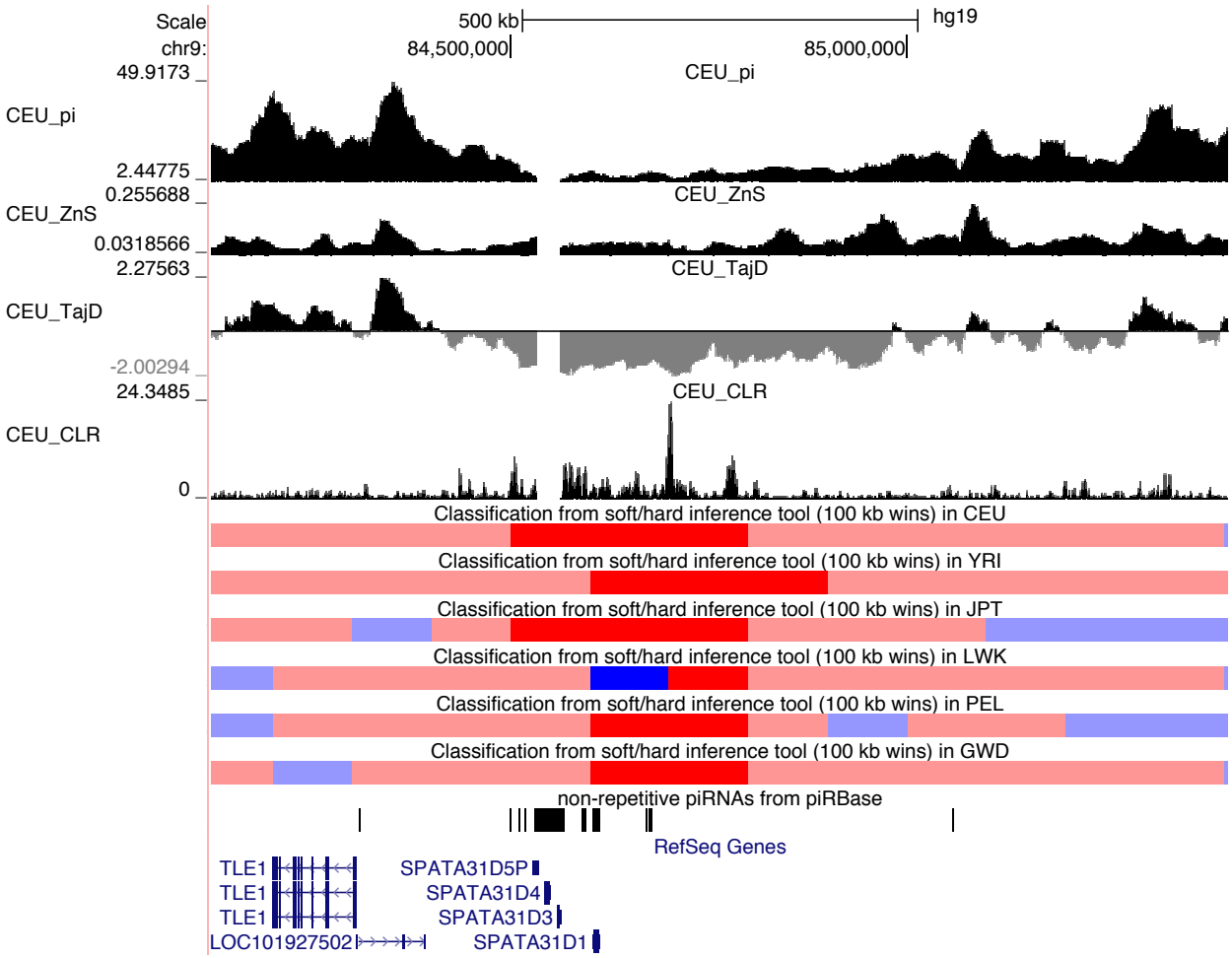

Figure 4: Hard selective sweep near several *SPATA31* spermatogenesis-associated genes. The S/HIC classification tracks show the raw classifier output for each population (red=hard sweep, blue=soft sweep, light red=hard-linked, light blue=soft-linked, black=neutral). We also show the values of various population genetic summary and test statistics (π, Tajima's D, Kelly's Z_{nS}, and the SweepFinder composite likelihood ratio, or CLR). To avoid clutter, we only show statistics from CEU.

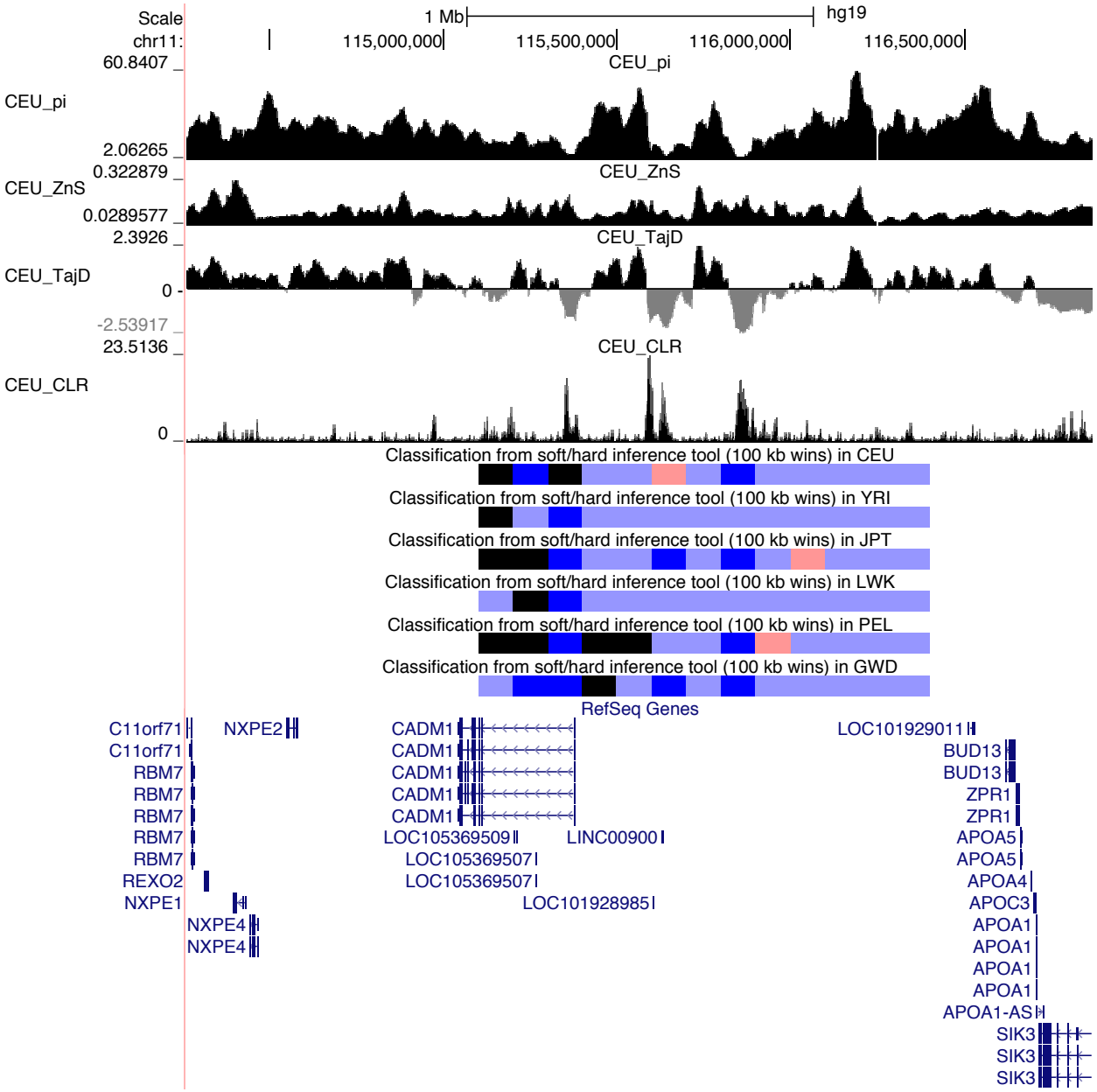

Figure 5: Soft selective sweeps near *CADM1*. The same tracks are shown as in Figure 4.

953 **SUPPLEMENTARY FIGURE AND TABLE LEGENDS**
954
955 **Figure S1: Heatmaps showing the accuracy of our six classifiers on test data, one for each**
956 **population.** On the *y*-axis, we show the location of the sweep relative to the classified window
957 (i.e. the central sub-window), with the exception of the "Neutral" case where there is no sweep.
958 The test data were simulated under the same demographic models used for training. On the *x*-
959 axis we show the class inferred by S/HIC. A perfect classifier would infer "Hard" for 100% test
960 instance where a hard sweep is in the focal sub-window (and analogously for soft sweeps),
961 "Hard-linked" for 100% of cases where a hard sweep occurs elsewhere (and analogously for soft
962 sweeps not located in the central sub-window), and "Neutral" for 100% of cases with no sweep.
963 (A) Test results for CEU. (B) GWD. (C) JPT. (D) LWK. (E) PEL. (F) YRI.
964
965 **Figure S2: Histograms of H_{12} and H_2/H_1 within windows classified has hard sweeps, soft**
966 **sweeps, or neutral for each population.**
967
968 **Figure S3: Soft selective sweep in *GRIA2*.** The same tracks are shown as in Figures 4 and 5.
969
970 **Table S1: Enrichment of various sequence annotations in each S/HIC class.**
971
972 **Table S2: Enrichment of annotation terms in hard and soft sweeps (only terms with q<0.05**
973 **for at least one sweep type in at least one population are shown).**
974
975 **Table S3: Example command lines used to generate training data for each population, with**
976 **a soft sweep occurring in the central sub-window.**