Using singleton densities to detect recent selection in *Bos taurus*

Matthew Hartfield[1,2], Nina Aagaard Poulsen[3], Bernt Guldbrandtsen[4,5,6], Thomas Bataillon[1]

1) Bioinformatics Research Centre, Aarhus University, DK–8000 Aarhus, Denmark.

2) Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK.

3) Department of Food Science, Aarhus University, Agro Food Park 48, 8200 Aarhus N, Denmark.

4) Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK–8830 Tjele, Denmark.

5) Rheinische Friedrich-Wilhelms-Universität Bonn, Institut für Tierwissenschaften, Tierzucht, Endenicher Allee 15, D-53115 Bonn, Germany.

6) Department of Veterinary Sciences, Copenhagen University, DK–1870 Frederiksberg C, Denmark.

Corresponding author: Matthew Hartfield (m.hartfield@ed.ac.uk).

Key words: selection, genomics, *Bos taurus*, milk protein, milk fat, stature.

Abstract: Many quantitative traits are subject to polygenic selection, where several genomic regions undergo small, simultaneous changes in allele frequency that collectively alter a phenotype. The widespread availability of genome data, along with novel statistical techniques, has made it easier to detect these changes. We apply one such method, the 'Singleton Density Score', to the Holstein breed of *Bos taurus* to detect recent selection (arising up to around 740 years ago). We identify several genes as candidates for targets of recent selection, including some relating to cell regulation, catabolic processes, neural-cell adhesion and immunity. We do not find strong evidence that three traits that are important to humans – milk protein content, milk fat content, and stature – have been subject to directional selection. Simulations demonstrate that since *B. taurus* recently experienced a population bottleneck, singletons are depleted so the power of SDS methods are reduced. These results inform on which genes underlie recent genetic change in *B. taurus*, while providing information on how polygenic selection can be best investigated in future studies.

Impact statement: Many traits of ecological or economic importance (including height, disease propensity, climatic adaptation) are ‘polygenic’. That is, they are affected by a large number of genetic variants, with each one only making a small contribution to a trait, but collectively influence variation. As selection acts on all of these variants simultaneously, it only changes the frequency of each one by a small amount, making it hard to detect such selection from genome data. This situation has changed in recent years ,with the proliferation of whole–genome data from many individuals, along with the development of methods to detect the subtle effects of polygenic selection. Here, we use data from 102 genomes from domesticated cattle (*Bos taurus*) that has experienced intense artificial selection since domestication, and test if we can detect signatures of recent selection (arising up to 740 years ago). Domesticated species are appealing for this kind of study, as they are subject to extensive genome sequencing studies, and genetic variants can be related to traits under selection. We carried out our analysis in two parts. We first performed a genome–wide scan to find individual genetic regions that show signatures of recent selection. We identify some relating to cell regulation, catabolic processes, neural-cell adhesion and immunity. In the second part, we then analysed genetic regions associated with three key traits: milk protein content, milk fat content, and stature. We tested if these regions collectively showed a signature of selection, but did not find a significant result in either case. Simulations suggest that the domestication history of cattle affected the power of these methods. We end with a discussion on how to best detect polygenic selection in future studies.

Introduction

Determining which genomic regions have been subject to selection is a major research goal in evolutionary genetics. Traditional methods have focused on detecting strong selection affecting individual genes (Nielsen, 2005; Vitti *et al.*, 2013; Stephan, 2019). An alternative process is 'polygenic selection', where many loci contribute to genetic variation in a trait, so selection acting on it is expected to generate small and simultaneous allele frequency changes at multiple loci (Pritchard & Di Rienzo, 2010; Pritchard *et al.*, 2010). Many polygenic models have been formulated to account for both the response to phenotypic selection, and the maintenance of genetic variance in quantitative traits [reviewed by (Sella & Barton, 2019; Barghi *et al.*, 2020)]. Among them is Fisher's infinitesimal model, which is important for its historical role in uniting population and quantitative genetics, and its recent renaissance in the context of genome–wide association studies (Fisher, 1918; Barton & Keightley, 2002; Barton *et al.*, 2017; Charlesworth & Edwards, 2018; Visscher & Goddard, 2019). However, whereas it has been possible to identify which genetic regions contribute to trait variation, it has historically been hard to infer which alleles have been involved in the polygenic selection response. Extensive theoretical studies of how alleles at multiple loci act when a population adapts to a new optimum generally find that 'large–effect' alleles, which strongly affect a trait, are the first to spread and fix while 'small–effect' alleles take much longer to reach high frequencies (de Vladar & Barton, 2014; Wollstein & Stephan, 2014; Jain & Stephan, 2015, 2017a, 2017b; Stetter *et al.*, 2018; Thornton, 2019; Hayward & Sella, 2019). Furthermore, if epistasis exists between variants, many selected alleles do not reach fixation as they eventually become deleterious (de Vladar & Barton, 2014; Jain & Stephan, 2017b).

81 The spread of large–effect alleles may also be impeded if a faster adaptive response
82 can be otherwise realised through changes at many small–effect alleles (Lande,
83 1983; Chevin & Hospital, 2008; Pavlidis *et al.*, 2012; Chevin, 2019). Alternatively, if
84 the optimum shift is sufficiently big, then large-effect mutations that first go to fixation
85 can subsequently be replaced by small–effect variants over longer timescales (on the
86 order of the population size; Hayward and Sella (2019)). Overall, only a small
87 proportion of loci affected by polygenic selection are expected to fix sufficiently
88 quickly to leave selection signatures in genomic data (Pavlidis *et al.*, 2012; Thornton,
89 2019).

90 Due to this difficulty, earlier methods for detecting polygenic selection focused
91 on cases where selection favours distinct phenotypes in different populations, so trait
92 differentiation amongst populations will be greater than expected under neutral drift.
93 Tests for this form of selection relied on comparing Q_{st} and F_{st} statistics, which
94 respectively measured mean genetic differentiation at the trait itself and a set of
95 neutral loci (Whitlock, 2008; Le Corre & Kremer, 2012; Savolainen *et al.*, 2013). Yet
96 these methods do not determine which genomic regions are subject to selection. This
97 situation has now changed with the increased number of genome–wide association
98 study (GWAS) data that link genotypes and phenotypes, as exemplified by the
99 development of large cohort studies [e.g., the UK Biobank; Bycroft et al. (2018)]. The
100 release of these data spurred a series of studies and new methods designed
101 specifically to detect polygenic selection. These methods usually involve determining
102 which SNPs affecting a phenotype show correlated changes in frequency (Berg &
103 Coop, 2014; Racimo *et al.*, 2018; Sanjak *et al.*, 2018; Josephs *et al.*, 2019; Berg *et*
104 *al.*, 2019a, 2019b; Uricchio *et al.*, 2019; Edge & Coop, 2019; Kreiner *et al.*, 2020;

105 Wieters *et al.*, 2021; Gramlich *et al.*, 2021); which sets of alleles are associated with
106 certain environmental or climatic variations (Coop *et al.*, 2010; Turchin *et al.*, 2012;
107 Robinson *et al.*, 2015; Yeaman *et al.*, 2016; Exposito-Alonso *et al.*, 2018; Zan &
108 Carlborg, 2018; Exposito-Alonso *et al.*, 2019; MacLachlan *et al.*, 2021; Ehrlich *et al.*,
109 2021; Fuhrmann *et al.*, 2021; Rowan *et al.*, 2021); or determining which SNPs or
110 genetic regions explain a large fraction of phenotypic variance and trait heritability
111 (Zhou *et al.*, 2013; Yang *et al.*, 2015; Gazal *et al.*, 2017; Zeng *et al.*, 2018; Schoech
112 *et al.*, 2019; Exposito-Alonso *et al.*, 2020; Duntsch *et al.*, 2020; Zeng *et al.*, 2021).
113 Some of these approaches use overlapping methods.

114 Detecting recent polygenic selection is much harder, as long periods of time
115 (number of generations on the order of the population size; Hayward and Sella, 2019;
116 Thornton, 2019) may be needed to cause detectable frequency changes in alleles
117 with small effect sizes. Over shorter timescales, these frequency changes are
118 expected to be more modest and harder to detect (Stephan, 2016; Jain & Stephan,
119 2017a). A recent breakthrough in detecting these subtle changes was the
120 development of the 'Singleton Density Score' (SDS), a statistic tailored to detect
121 recent and coordinated allele frequency changes over many SNPs (Field *et al.*,
122 2016). Recent selection at a locus favouring one variant will lead to a reduction in the
123 number of singletons (i.e., variants that are only observed once) around it. The SDS
124 detects regions that exhibit a reduction in the density of singletons, to determine
125 candidate regions that have been subject to recent selection. Using this approach,
126 Field et al. (2016) found correlations between SDS scores at SNPs and their
127 associated GWAS effect sizes for several polygenic traits in the modern UK human
128 population, including increased height, infant head circumference and fasting insulin.

129 Their findings suggested that these traits have been subject to recent selection
130 during the most recent 75 or so generations (about 2,000 years). However, these
131 (and other) results that detect selection for increased height may instead reflect
132 previously unaccounted–for population structure (Novembre and Barton, 2018;
133 Barton et al., 2019; Sohail et al., 2019; Berg et al., 2019; Uricchio et al., 2019; Edge
134 and Coop, 2019).

135 The SDS method is ideally suited to organisms where large amount of whole-
136 genome data are available, along with QTL or GWAS information that link genotypes
137 to phenotypes, Domesticated species are attractive systems for studying recent
138 selection, as selected phenotypes are often already known and these species are
139 subject to large–scale sequencing studies. Investigating the genetic architecture
140 underlying rapid selection in these species is also important to determine how they
141 respond to agricultural practices, and uncover selection targets that can be used to
142 improve breeding programs (Georges *et al.*, 2018). Domestic cattle *Bos taurus* has
143 been subject to intensive genomics analyses to improve artificial selection for traits
144 that are important for human use, including milk protein yield, milk fat content, and
145 stature (Hayes *et al.*, 2009; Meuwissen *et al.*, 2013; Wray *et al.*, 2019). These traits
146 are influenced in part by an individual's genome, with significant heritability estimates
147 being recorded, some as high as 80% (Soyeurt *et al.*, 2007; Haile-Mariam *et al.*,
148 2013; Buitenhuis *et al.*, 2016). Previous selection scans on *B. taurus* reported
149 individual regions that were likely to be subject to recent selection, some of which
150 were close to genetic regions for stature and milk protein content (Lemay *et al.*, 2009;
151 MacEachern *et al.*, 2009; Qanbari *et al.*, 2010; Boitard & Rocha, 2013; Qanbari *et al.*,
152 2014; Zhao *et al.*, 2015; Boitard *et al.*, 2016a; Bouwman *et al.*, 2018). Recently,

153 Rowan *et al.* (2021) introduced a method to detect polygenic environmental
154 adaptation. However, stature and milk protein content are polygenic traits, with
155 several genetic regions and QTLs associated with each (Lemay *et al.*, 2009; Boitard
156 *et al.*, 2016a; Bouwman *et al.*, 2018; van den Berg *et al.*, 2020). There has also yet to
157 be a formal test of whether these traits show evidence of polygenic selection.
158 Here, we applied the SDS method to whole–autosome sequencing data from
159 102 *B. taurus* Holstein individuals. We first determined genetic regions that have
160 been subject to recent directional selection, and subsequently tested whether
161 evidence exists for recent selection acting on a set of QTLs underlying either milk
162 protein content, milk fat content, or stature in this breed.
163

164 ## ***Results***

165 ### *Methods outline*

166 We filtered the data to retain only bi–allelic SNPs that had a sensible level of
167 coverage and did not lie in putatively over–assembled regions (i.e., duplicated
168 sections that caused many reads to assemble at a specific genetic location). Over–
169 assembled regions appear as highly heterozygous with elevated coverage, and can
170 exhibit false signatures of recent selection. We also obtained a set of singletons and
171 filtered them to retain high-quality variants where both alleles were equally well
172 covered to remove potentially erroneous calls. We polarised test SNPs using
173 outgroup sequences and applied the SDS test of Field et al. (2016) to detect recent
174 selection, with increased SDS values reflecting selection favouring derived SNPs
175 over ancestral variants. We standardised SDS scores with those of a similar
176 frequency, so they are normally distributed [similar normalisation was also carried out

177 by Field et al. (2016)]. These values are denoted sSDS for 'standardised SDS'.
178 Further details are available in the *Methods* section.
179

180 ## *Estimating timescale of selection*

181 We first determined the timescale over which we expect to detect selection in
182 *B. taurus* using the SDS method. SDS measures the changes in singleton numbers
183 around putatively selected SNPs, relative to background numbers in the absence of
184 selection. As singletons arise on the tips of the underlying gene trees, the average tip
185 length in the genealogy of sequenced samples determines the timescale over which
186 the SDS detects a signal (Field et al., 2016). As more haploid genomes are included
187 in the study, the time to first coalescence between two samples decreases, reducing
188 the tip lengths and therefore shortening the timescale over which SDS detects
189 selection (Field *et al.*, 2016). We hence simulate tip-ages over a range of sample
190 sizes to investigate how this timescale changes accordingly.
191 To calculate the mean tip age, we simulated gene genealogies under two
192 scenarios. We first simulated the Holstein population demography inferred by Boitard
193 et al. (2016b), which suggested that this population experienced a sudden decline in
194 effective population size (N_e) since domestication, but with a present–day N_e (~793)
195 that is much larger than that inferred from pedigree data [~49; Sørensen et al.
196 (2005)] or from temporal variation in SNP frequencies (~48; Jiménez–Mena et al.
197 2016). Hence, we also simulated genealogies under a second model that used the
198 Boitard et al. (2016b) demographic model, but with the present–day N_e set to 49.
199 These scenarios will be referred to as the 'High N_0' and 'Low N_0' models,
200 respectively.

201 Figure 1 shows simulation results. Depending on the assumed present–day
202 N_e, the tip length in our sample of 204 alleles (i.e., assuming two per diploid
203 individual) goes back either 65 or 148 generations. Assuming 5 years per generation
204 (Boitard *et al.*, 2016b), this timescale corresponds to between 325 and 740 years
205 ago. Since *B. taurus* domestication started around 10,000 years ago (Zeder, 2008)
206 the sample size used in this study will only capture selection acting in the very recent
207 past that is more relevant for breed formation, rather than selection during *B. taurus*
208 domestication. Sample sizes and tip-ages are linearly related on a log-log scale,
209 meaning that an increase in sample size will greatly decrease the timescale over
210 which SDS detects selection. For example, with 500 haplotypes then SDS will detect
211 selection acting no more than 50 generations ago, depending on the underlying
212 demographic model.
213 We will focus on detecting selection signatures assuming the high N_0 model.
214 Results using the low N_0 model to calibrate scores were broadly similar. They are
215 outlined in the Supplementary Text; we will highlight when differences arise.
216

217 *Genome–wide sSDS*

218 Figure 2 plots sSDS values (at SNPs with minor allele frequency greater than
219 5%) across all autosomes, excluding chromosome 25 (due to an insufficient number
220 of singletons needed to obtain SDS scores after filtering). Many SNPs have elevated
221 sSDS scores (158 SNPs at $FDR < 0.05$; 306 for the low N_0 model). Several regions
222 contain SNPs with significantly high sSDS values (Bonferroni–corrected nominal $P <$
223 0.05; actual $P < \sim 2.7 \times 10^{-8}$). To further investigate potential selection targets, we
224 looked for genes that either overlapped significant SNPs or lay 10kb up– or

225 downstream of them. Linkage disequilibrium (LD), as measured by r^2, decays to
226 around 0.2 over 50kb in Danish Holstein breeds (Buitenhuis *et al.*, 2016), so genes
227 within 10kb should be in LD with regions harbouring high sSDS scores. Table 1 lists
228 these genes, with more targets present under the low N_0 model. Most of these genes
229 are of unknown function (as listed on UniProt); the list also includes an snRNA.
230 *PPM1L* is involved with cellular regulation and the activation of stress–activated
231 protein kinases. *TDO2* is involved in tryptophan–related catabolic processes, while
232 *NTM* is implicated in neural cell adhesion. SNPs with significantly elevated scores
233 are also found on chromosome 23 near the MHC region, which may reflect over–
234 dominant selection. All Bonferroni–significant SNPs were removed from subsequent
235 tests of recent polygenic selection to prevent directional selection from skewing the
236 underlying sSDS distributions. Figure S1 shows results for the low N_0 model.
237

238 *Testing for polygenic selection acting on milk protein and stature*

239 If polygenic selection were acting on specific traits, we expect a positive
240 correlation between the effect size of variant underpinning it, and selection acting on
241 it as measured by sSDS. We collated sSDS scores of SNPs that lie close to QTLs
242 reported for either milk fat percentage, milk protein percentage (van den Berg *et al.*,
243 2020), or those that lie close to stature QTLs (Bouwman *et al.*, 2018). The latter were
244 inferred from a meta–analysis of GWAS studies conducted in seven Holstein
245 populations, but not every QTL had an effect size reported in each population. We
246 hence investigated two overlapping consensus QTL sets, where an effect size was
247 either reported in at least 6 of 7 populations (yielding 42 QTLs with sSDS scores
248 associated with them), or where effect sizes were reported in at least 5 of 7

249 populations (58 QTLs had sSDS scores). We re-polarised SDS scores so that a
250 positive score reflected a trait-increasing effect; we denote these values 'tSDS'
251 following Field *et al.* (2016). We then determine if there was a positive correlation
252 between the absolute $\log_{10}$-value of the QTL *P*-value (a proxy for the effect size) and
253 tSDS.

254 Figure 3 shows the relationship between QTL *P*-values and tSDS for SNPs
255 that lie close to QTLs. Although positive trends are observed, they all exhibit non-
256 significant regression slopes (milk fat percentage $P = 0.596$; milk protein percentage
257 $P = 0.873$; stature from 6 breeds $P = 0.558$; stature from 5 breeds $P = 0.344$).
258 Relationships remain non-significant after removing an outlier point for the milk traits
259 whose QTL has an extremely low *P*-value (Figure S2), and also under the low N_0
260 model (Figure S3; see figure legends for regression *P*-values).

261 sSDS (and tSDS) can become correlated along the genome if focal SNPs are
262 in LD with one another, which was not accounted for in the preceding analyses. To
263 determine whether LD could have impacted upon these correlations, we randomly
264 subsampled sSDS scores from SNPs that shared the same chromosome and bin of
265 derived-allele frequency as the SNPs used in the above analyses, and re-polarised
266 them to transform them into tSDS values. We then determined the *F* statistic
267 associated with the linear model fit for these permuted values to determine whether
268 that for the true data was significantly elevated (see Methods for details). In all cases,
269 the observed value was not significantly higher than for permuted values (see Figure
270 S4 for histograms and exact *P*–values, which all exceed 0.05). We therefore
271 conclude that these QTL datasets do not harbour SNPs with significantly different
272 tSDS scores compared to the rest of the genome.

Discussion

Summary of results

We analysed an extensive *B. taurus* genomic dataset to identify signatures of recent selection in the Holstein breed, and to determine whether the data contained a signal of polygenic selection acting on milk proteins and QTLs underlying phenotypic variation in stature. Given the sample size and the demographic history of Holsteins, the SDS method can detect very recent selection events arising no more than approximately 740 years ago (Figure 1). A whole–genome scan for sSDS scores identified several targets of recent directional selection that overlap or lie close to protein–coding genes (Figure 2; Table 1). The genes whose functions are known are involved in protein regulation, catabolic processes, and neural-cell adhesion. Significant values were also observed near the MHC region. We subsequently investigated whether either milk protein genes or SNPs near stature QTLs collectively showed evidence of polygenic selection. We did so by testing whether there is a relationship between the QTL effect size, as measured by its *P*-value, and tSDS values to SNPs near them. However, no relationship was observed, even after performing a permutation test (Figures 3, S2-S4). Hence, while sSDS could reveal specific instances of recent selection, tests based on collective scores of variants associated with known selected traits yielded no signal of polygenic selection.

Potential reasons for a lack of polygenic selection signal

Impact of Holstein demographic history

While the SDS method detected individual candidate genes for very recent selection, we were unable to find strong evidence for polygenic selection acting on

297 three traits that were subject to artificial selection since domestication. This result is *a*
298 *priori* surprising, given that these traits have been subject to recent intense artificial
299 selection. Recent studies generally find non-zero heritability estimates for them,
300 indicating that there should be the potential for genetic variants underpinning them to
301 change in response to artificial selection (Soyeurt *et al.*, 2007; Haile-Mariam *et al.*,
302 2013; Buitenhuis *et al.*, 2016). In addition, the ratio of the mutation and recombination
303 rates in cattle is just over three (Boitard *et al.*, 2016b; Harland *et al.*, 2018), indicating
304 that several informative SNPs exist per haplotypes that should improve the power of
305 the SDS method [in contrast, this ratio is approximately equal to one in humans
306 (Field *et al.*, 2016)].

307 One potential reason for this lack of signal is due to the population history of
308 *Bos taurus*. The effective population size of many *B. taurus* breeds appears to have
309 undergone a decline since domestication (Sørensen et al., 2005; Boitard et al.,
310 2016b), which likely reflects successive bottlenecks due to domestication, breed
311 formation and intense recent selection. Population size reductions are known to
312 reduce the number of low-frequency variants and increases the prevalence of
313 intermediate-frequency variants (Harpending *et al.*, 1998), which can affect the power
314 of the SDS method. To understand if the history of *B. taurus* affects the detection of
315 recent selection in Holstein cattle using SDS, we ran coalescent simulations to
316 determine its ability to detect ongoing selection, given realistic Holstein population
317 history and genetic parameters (see Methods for details). We simulated a partial
318 sweep occurring in the middle of a 10Mb region, either assuming a mutation rate in
319 line with what has been inferred for Holstein, or one 10-fold higher to replicate
320 diversity expected in a genetic region with an elevated mutation rate.

321 For the standard mutation rate, no SDS scores were produced for any
322 simulations. After inspecting the simulation results, we see that there is a large skew
323 in the distribution of singleton numbers per individual with a large number of
324 individuals (over 20 on average) that do not carry singletons at the end of
325 simulations, preventing the calculation of a local SDS score (Figure 4). This fraction
326 remained the same irrespective of whether the simulated region was neutral or
327 subject to selection; the main effect of a sweep was to reduce the mean number of
328 singletons per individual, which is the signal measured by SDS (Field *et al.*, 2016).
329 This reduction in overall singleton numbers is consistent with the known effects of
330 population size contraction on reducing tip lengths (Harpending *et al.*, 1998).
331 With a 10-fold higher mutation rate, there were fewer cases where no
332 individual harboured singletons (Figure 4). Accordingly, SDS scores could be
333 calculated for 65 and 66 out of 100 simulations for the neutral and selective cases
334 respectively. In these cases, sSDS values were significantly higher in the selected
335 case than for the neutral case (Figure 5; two-sided Wilcox Test $P = 1.1\text{x}10^{-5}$).
336 However, note that sSDS values is less than one for the selected case, which does
337 not exceed the FDR threshold in our study (for the high N_0 case, the smallest sSDS
338 value with FDR < 0.05 is 4.46).
339 Overall, these simulations are consistent with population size reductions in *B.*
340 *taurus* reducing the number of singletons, which limits its ability to detect partial
341 sweeps. SDS is more likely to detect signals in regions of elevated mutation rate,
342 suggesting there is likely an ascertainment bias in where signals were detected;
343 variants contributing to polygenic selection are more likely to be detected by SDS in
344 regions of elevated mutation. The reduction in singletons does not just affect the

345 ability of SDS to detect sweeps overall, it also reduces the power to investigate SDS
346 values in telomeric regions. SDS values are calculated using the distance up- and
347 downstream from a SNP to the nearest singleton, and are undefined if a certain
348 number of samples do not harbour singletons in either direction (Field *et al.*, 2016).
349 SDS values are hence less likely to be defined in telomeric regions, as it is generally
350 less feasible to observe singletons up until the end of the chromosome. This problem
351 is exacerbated if there are few singletons overall.
352
353 *Other potential reasons for a lack of signal*
354 Another potential reason for a lack of signal is that the selection response on
355 these traits may have been driven by large-effect variants that have already fixed in
356 the population, with a smaller contribution from small-effect mutations. Theoretical
357 models have shown that more major–effect QTLs are likely to fix if the population lies
358 further from a fitness optimum (Lande, 1983; Jain & Stephan, 2017b; Thornton,
359 2019). Domesticated species, which experience strong and sustained directional
360 artificial selection, especially in recent generations, could thereby fix more adaptive
361 mutation via sweep–like processes compared to populations evolving in more stable
362 environments (Lande, 1983; Jain & Stephan, 2017a). Furthermore, once a population
363 has adapted to a new environment (the domestication phenotype in this case), then
364 any remaining major–effect mutations are likely to be superseded by variants with
365 weaker effects, which are harder to detect (Hayward & Sella, 2019). The response to
366 polygenic selection will be further weakened in smaller populations (John & Stephan,
367 2020), which could be a factor given the reduced effective population sizes of *B.*
368 *taurus* (Sørensen *et al.*, 2005; Boitard *et al.*, 2016b). There is some evidence of this

369 explanation; selective sweeps signatures are associated with stature QTLs
370 (Bouwman *et al.*, 2018), and the study of van den Berg et al. (2020) was more likely
371 to identify milk QTLs that had a moderate to high minor allele frequency, suggesting
372 reduced power to detect low-frequency variants that are potential contributors to
373 polygenic selection. Conversely, the stature meta-analysis by Bouwman et al. (2018)
374 found significant SNPs that explained up to 13.8% of the variance in stature, which is
375 similar to that explained by significant SNPs for human height (16%), which is a
376 classic trait for polygenic selection studies. Hence, there may be sufficient polygenic
377 SNPs present to test for polygenic selection, but the power will still be reduced due to
378 the demographic history of Holstein cattle.

379 Potential solutions to increase power include increasing sample sizes; using
380 alternative methods; or analysing different kinds of genome data to detect polygenic
381 selection. Applying SDS to a larger sample size would increase the power to detect
382 selection acting in the recent past [Figure 1; see also Field et al. (2016)], but overall
383 power will still be limited by the tip-length of neutral genealogies. Recent
384 developments in methodology involve directly inferring trees from genome data, and
385 using these to identify subtle sweep signatures associated with trait variants (Edge &
386 Coop, 2019; Speidel *et al.*, 2019; Stern *et al.*, 2021). These methods have greater
387 power to detect weakly-selected mutations that may be segregating for longer than
388 the tip-length of the population.

389 Another approach would be to look beyond sequence data and focus on gene
390 networks [reviewed by Fagny & Austerlitz (2021)]. The recently–proposed 'omnigenic'
391 model (Boyle *et al.*, 2017; Liu *et al.*, 2019) posits that variation in quantitative traits is
392 principally affected by a plethora of 'peripheral' genes that indirectly affect them,

393 rather than a limited set of ‘core’ genes that directly modify a trait. These numerous
394 peripheral genes may exert their influence via regulatory effects (e.g., gene
395 expression changes), but are also expected to be highly pleiotropic. Fully testing the
396 omnigenic model will require larger datasets and novel experimental designs (Wray
397 *et al.*, 2018). A recent example is from an experiment with *Drosophila melanogaster*,
398 where gene knockouts that do not pass a GWAS significance threshold for pupal
399 length still significantly affect it (Zhang *et al.*, 2021). There is also nascent evidence
400 that gene regulation may underlie directional polygenic selection. Boitard et al.
401 (2016a) found that some adaptive signatures of *B. taurus* are located in intergenic
402 regions; regulatory changes were also proposed to guide polygenic selection in
403 *Arabidopsis* (He *et al.*, 2016). Analyses of gene–sets associated with infection
404 responses or immunity also found evidence for polygenic selection in humans and
405 primates (Daub *et al.*, 2013, 2017; Svardal *et al.*, 2017). Immunity gene–sets might
406 be exceptional cases, as they are more likely to contain genes subject to very strong
407 selection (Castellano *et al.*, 2019). Further investigations using regulatory information
408 and a broader range of gene–sets could be a promising approach to determine the
409 impact of polygenic selection.

410 ## *Materials and Methods*

411 ### *Simulating Holstein tip ages*

412 Neutral genealogies were simulated using *msprime* (Kelleher *et al.*, 2016) to
413 determine the mean tip length, and hence the background distribution of SDS in the
414 absence of selection. We either simulated the Holstein population demography
415 inferred by Boitard et al. (2016b), rounding estimated population sizes to the nearest
416 integer, or with the present–day N_e equal to 49 (Sørensen et al., 2005). We refer to
417 these outputs as the 'High N_0' and 'Low N_0' models. 1,000 simulations were
418 performed for each number of samples, ranging from 10 to 1,050. The mean tip
419 length was calculated over all 1,000 simulations; 95% confidence intervals were
420 calculated from 1,000 bootstraps. We fitted a linear model to the log of mean tip
421 length against the log number of samples, and used it to predict the average tip age
422 for 204 alleles, which is the number of diploid haplotypes used in the study. *B. taurus*
423 are somewhat inbred (Sørensen *et al.*, 2005), which increases within–individual
424 relatedness, and could reduce the number of unique alleles (see Nordborg and
425 Donnelly (1997) for an example with self–fertilisation). Estimates of the inbreeding
426 coefficient *F* (Wright, 1951), which measure the reduction in heterozygosity, range
427 from –0.15 to 0.35, with a mean of 0.059 (Figure S5; methods outlined below). Given
428 this low mean value, we assumed two unique alleles per individual.
429

430 ### *Genome Data Extraction*

431 Whole genome sequencing for 102 Holstein bulls and cows were done by
432 Illumina and BGI short-read paired-end sequencing in various laboratories with read
433 lengths of 100 base pairs and up. No raw reads were shorter than 90 basepairs.

434 Bulls were selected for sequencing had high genetic representation in the present–
435 day Holstein population. Sequencing of close relatives was avoided. Individuals were
436 born between approximately 1980 and 2010. DNA was extracted from tissue, blood,
437 or semen samples. Read data were processed according to the 1,000 Bull Genomes
438 Project (Daetwyler *et al.*, 2014). Briefly, data were trimmed and quality filtered using
439 Trimmomatic version 0.38 (Bolger *et al.*, 2014). Reads were aligned to the ARS-
440 UCD-1.2 bovine genome assembly (Rosen *et al.*, 2018)
441 (https://sites.ualberta.ca/~stothard/1000_bull_genomes/ARS-
442 UCD1.2_Btau5.0.1Y.fa.gz) with the *B. taurus* Y chromosome assembly from BTau-
443 5.0.1 added. Alignment was performed with bwa version 0.17 (Li & Durbin, 2009).
444 Samtools (Li *et al.*, 2009) was used for sorting and marking of PCR duplicates. Base
445 qualities were recalibrated using Genome Analysis Toolkit (GATK; McKenna et al.
446 [2010]) version 3.8 using a set of known variable sites (Schnabel and Chamberlain,
447 *unpubl*). GVCF files were formed using GATK's Haplotype Caller. Genotypes were
448 called using GATK's GenotypeGVCFs.
449

450 *Initial filtering*

451 Figure S6 outlines a schematic of the data filtering. We first used *VCFtools*
452 (Danecek *et al.*, 2011) to obtain a baseline list of biallelic SNPs. We removed indels
453 and sites where the genotype was unknown in any individual. For each autosome,
454 we obtained the mean depth for each remaining site using *VCFtools*' '--site-mean-
455 depth' option. Figure S7 shows the depth distribution for these sites after initial
456 filtering. We fitted a Poisson distribution to these data that had the same mean (9.76)
457 as observed in the dataset. We determined the expected coverage range based on

458 the 99.5% quantile range of the fitted distributions, which equalled 2 to 20. We
459 subsequently removed sites that had mean coverage outside this range. We hence
460 retained 6,873,371 of 20,010,175 initial variants (all entries in each autosome VCF
461 file, including indels); we denote this dataset as the 'filtered' dataset.
462

463 *Finding putatively over–assembled regions*

464 Scaffolds of different genetic segments that each carry highly identical
465 repeated regions can be 'over–assembled', where very similar chromosome regions
466 were anchored to a single location (Chaisson *et al.*, 2015). These over–assembled
467 regions (hereafter OARs) manifest themselves in aligned data as having high levels
468 of heterozygosity, and elevated apparent coverage. If not corrected for, they can be
469 misclassified as selected sites (e.g., subject to partial sweeps or balancing selection).
470 We used a sliding window method to identify putative OARs in the dataset. For each
471 chromosome, in each window we calculated (i) the number of sites where the
472 reference allele had a frequency between 49% – 51%, (ii) the mean heterozygosity
473 for each SNP (defined as the number of heterozygotes among the 102 individuals),
474 and (iii) the mean summed allele depth. We used overlapping windows, each of size
475 500 SNPs with a step size of 10 SNPs. We first analysed all chromosomes to
476 determine the distribution of values produced per window. We then re–ran the
477 analyses, classifying windows as OARs if values for all three statistics belonged to
478 the top 99.5% of their respective distributions, merging overlapping windows. We
479 subtracted 1 from the start position of each region so that the leftmost boundary
480 would also be excluded (if using '--bed-exclude' in *VCFtools*). Figure S8 shows an
481 example where a region at the beginning of chromosome 1 was identified as an

482 OAR. Overall, 5 OARs comprising 5,880 SNPs were identified (Table S1), which
483 were subsequently masked from the rest of the pipeline.
484

485 *Calculating inbreeding coefficients*

486 Inbreeding coefficients were estimated using the '--het' option of *VCFtools*,
487 which reports *F*-statistics for each chromosome per individual. Individual *F*-values
488 (Figure S5) were calculated by taking the mean over all chromosomes, weighted by
489 the chromosome size.
490

491 *Obtaining SDS analysis inputs*

492 ***Test SNPs:*** Focal SNPs were those with an alternate allele frequency
493 between 5% and 95%, and where each genotype was observed at least once
494 amongst all samples. We then polarised values to determine the ancestral and
495 derived states. Genome data, mostly from Wu *et al.* (2018), was downloaded from
496 the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) for Bison (*Bison bison*,
497 SRR6448737, SRR6448738, SRR6448739, SRR6448740), Wisent (*Bos bonasus*,
498 SRR6448670, SRR6448682, SRR6448683, SRR6448684), Gaur (SRR6448732,
499 SRR6448733, SRR6448734, SRR6448735), Gayal (*Bos frontalis*, SRR6448578,
500 SRR6448580, SRR6448581), Banteng (*Bos javanicus*, SRR6448720 and
501 SRR6448721), and as well as data for Yak (*Bos grunniens*, SRR5641601). Ancestral
502 SNPs were determined by counting which allele was present in a majority of these
503 related species; note that some SNPs for some species carried neither the reference
504 nor the alternate alleles. We only retained SNPs where the called ancestral allele
505 matched either the reference or alternate allele in our dataset. 3,405,023 SNPs were

506 retained for testing.

507 ***Singletons:*** Raw singleton data was extracted from the filtered Holstein
508 dataset using *VCFtools*' '--singleton' option. This option identified both true singletons
509 and private doubletons (i.e., where an allele is unique to an individual but present as
510 a homozygote). Only true singletons were retained for analyses. To test whether a
511 singleton had the same coverage as the non–singleton allele, we extracted the
512 sequence depth for both alleles and retained sites satisfying the following criteria.
513 First, the total allele depth was between 2 and 20 inclusive. Second, either (a) if the
514 summed depth over both alleles exceeded 5, then the binomial probability of the
515 observed allele depth exceeded 0.1; or (b) a stricter manual cut–off was applied if the
516 total allele depth equalled 5 or less. Table S2 outlines the cut–off values used;
517 554,402 of 765,822 singletons were subsequently retained.

518 ***Other parameters:*** The SDS method requires a 'singleton observability'
519 probability, to predict how likely it is that a singleton will be detected by genome
520 sequencing. Following Field et al. (2016) we used the mean depth per individual,
521 obtained using the '--depth' option in *VCFtools.* It was also necessary to state the
522 genetic boundaries between which analyses were carried out; we used a starting
523 point of 1 and end points equal to the reported size of each autosome in *B. taurus*, as
524 obtained from the ARS–UCD 1.2 genome assembly
525 (https://www.ncbi.nlm.nih.gov/genome/?term=txid9913[orgn]).

526 Raw SDS values were calculated by fitting a gamma distribution to observed
527 singleton distances, and comparing it to the expected distribution for the neutral
528 case. We used the scripts provided by Field et al. (2016)
529 (https://github.com/yairf/SDS) to generate the expected shape values for the gamma

530 distribution for both the high and low N_0 models. Finally, we used a value of 10^{-7} to
531 initiate the search for a maximum value in likelihood space.
532

533 *Calculating SDS scores and their significance for individual SNPs*

534 Out of 3,405,023 input SNPs, we retained and assigned scores to 1,877,770
535 of these. SDS scores were not assigned to a SNP if more than 5% of individuals did
536 not harbour any singletons upstream or downstream of the SNP. This cut–off tended
537 to exclude SNPs in telomeric regions. Furthermore, SDS scores were not calculated
538 for chromosome 25 as an insufficient number of singletons were present across all
539 individuals after data filtering. Raw SDS scores were standardized using 18 bins,
540 based on alternate allele frequencies at the scored SNP (i.e., from 5% to 10%, from
541 10% to 15%, etc.). Standardised SDS scores (denoted sSDS) were obtained by
542 subtracting the bin mean score from individual measures, and dividing by the bin
543 standard deviation.

544 Statistical analyses were carried out in R (R Core Team, 2021). *P*–values for
545 each SNP were determined by calculating the probability of the observed sSDS
546 values from a standardised normal distribution. Significance was determined using a
547 Bonferroni–corrected cut–off of $0.05/(1{,}877{,}770) \approx 2.7 \times 10^{-8}$. The false–discovery
548 rate (FDR) of each SNP was calculated using the 'qvalue' package (Storey *et al.*,
549 2021); we highlighted SNPs with an FDR of less than 0.05.
550

551 *Data sources*

552 Version 104 of the GTF gene annotation file for the ARS–UCD 1.2 assembly
553 was downloaded from Ensembl (available from

554 https://www.ensembl.org/Bos_taurus/Info/Index). Bedtools v2.29.0 (Quinlan & Hall,
555 2010) was used to obtain genetic annotations 10kb up– and downstream of each
556 Bonferroni–significant SNP; overlapping windows were merged.

557 Milk fat percentage and milk protein percentage QTLs were obtained from van
558 den Berg et al. (2020). We retained only those QTLs for which effect sizes were
559 reported in both Holstein breeds that were studied (French and German Holstein),
560 and the effect was in the same direction in each. sSDS scores were then obtained for
561 the nearest SNP to each QTL. Some QTLs lie close to the telomeres, where sSDS
562 statistics were not available. These QTLs were not considered further; for each of the
563 remaining QTLs, we identified the SNP nearest to it and assigned the sSDS value at
564 that site to the QTL. Since QTL effect directions were given for the alternate allele (I.
565 v. d. Berg, personal communication), we re-polarised sSDS scores for SNPs where
566 the ancestral allele did not match the reference allele by switching its sign. We
567 additionally re-polarised scores so that a positive value indicates selection acting on
568 the trait-increasing allele (Field *et al.*, 2016). 30 of 138 milk fat QTLs had scores
569 assigned to them, while 78 of 176 milk protein QTLs had scores assigned to them.

570 Stature QTLs were obtained from Bouwman et al. (2018), which identified 164
571 QTLs in several *B. taurus* breeds, including 7 Holstein populations from different
572 countries. We initially extracted 114 QTLs for which an effect was inferred from at
573 least 5 of 7 Holstein populations. Positions were given relative to the UMD 3.1
574 assembly; we subsequently extracted sequence 100bp up– and downstream of the
575 position and remapped to ARS–UCD 1.0.25. 106 QTLs were re–aligned without
576 gaps; of the remaining 8, 4 were located close to rearrangements and discarded,
577 while the remaining 4 were kept. After removing those on chromosome 25, 107 QTLs

578 were retained for downstream analysis. We analysed this full QTL set and a subset
579 where effect sizes were reported in 6 of 7 Holstein breeds (containing 78 QTLs). We
580 obtained the sSDS value of SNPs near QTLs and re-polarised them so that positive
581 scores indicated selection on the trait-increasing allele. Overall, sSDS values were
582 assigned to 58 QTLs with effect sizes in at least 5 of 7 Holstein populations, and 42
583 QTLs with effect sizes in at least 6 of 7 Holstein populations.
584

585 *Statistical analyses of polygenic selection*

586 To determine whether evidence exists for polygenic selection acting on QTLs
587 underpinning these traits, we applied the 'lm' function in R to fit a linear model
588 between sSDS values of the nearest SNP to the QTL, and the absolute $\log_{10}$ *p*-value
589 of it as proxy for effect size. To implement the permutation test, we subsampled
590 SNPs from the genome that were from the same chromosome and derived-allele
591 frequency bin as the SNPs used in the polygenic selection tests analyses; we also
592 changed the sign of the sSDS value of the sampled SNP if this change was also
593 made to the score associated with the corresponding true SNP. We then performed a
594 linear model fit between these permuted sSDS values and the QTL *p*-values, and
595 obtained the *F*-statistic for it. 1,000 subsampled datasets were used to generate the
596 neutral distribution.
597

598 *Power simulations given Holstein demography*

599 To estimate the power of the SDS methods when applied to the Holstein
600 breed, we implemented a simulation pipeline as used in previous studies of polygenic
601 adaptation (Field *et al.*, 2016; Speidel *et al.*, 2019). We used the *simuPOP* program

602 (Peng & Kimmel, 2005) to simulate the trajectory of either a neutral variant, or a
603 partial selective sweep with homozygote selection coefficient of 0.2 (and a
604 heterozygote value of half that value). The target mutation appeared *de novo* and we
605 sampled individuals when it reached a frequency of 70%. We assumed the Holstein
606 population demography as inferred by Boitard et al. (2016). Simulated trajectories
607 were then piped into the coalescent simulation program *mbs* (Teshima & Innan,
608 2009) to generate polymorphism data for 204 haplotypes. We assumed a per-site
609 recombination rate of 3.66×10^{-9} (Boitard *et al.*, 2016b), and a mutation rate of either
610 1.21×10^{-8} as inferred from parent-offspring trios of dairy cattle (Harland *et al.*, 2018),
611 or one ten-fold higher. A region of 10Mb long was simulated, with the target mutation
612 appearing in the centre of it. Diploid genotypes were formed by randomly pairing
613 haplotypes. SDS scores were calculated using the same pipeline and parameter
614 values as for the data analyses. We also recorded the number of singletons per
615 individual, and their distribution per simulation. 100 replicate simulations were run for
616 each case. The confidence intervals for singleton distributions were calculated using
617 1,000 bootstraps. When plotting the distribution of SDS scores for the high-mutation
618 case (Figure 5), they were first normalised by subtracting the mean of the neutral
619 case from each value, then dividing by the standard deviation of neutral scores.
620

621 ***Acknowledgements.*** We would like to thank Simon Boitard for sharing his
622 results on *B. taurus* demographic inference; Simon Boitard, Jon Slate and two
623 anonymous reviewers for providing feedback on the manuscript. MH is supported by
624 a NERC Independent Research Fellowship (NE/R015686/1). NAP, BG and TB are
625 funded by a synergistic research grant from the Faculty of Science and Technology,

Aarhus University, Denmark. MH and TB also acknowledge financial support from the European Research Council under the European Union's Seventh Framework Program (FP7/20072013, ERC Grant 311341). The authors declare no conflicts of interest.

Author contributions. All authors contributed to the study design. NAP and BG provided data. MH performed the analyses and wrote the manuscript, with feedback from NAP, BG and TB.

Data archiving. Raw SDS scores and polarisation information has been deposited on Dryad (under curation, provisional DOI https://doi.org/10.5061/dryad.547d7wm8q). Data analysis and simulation scripts are available on GitHub (https://github.com/MattHartfield/CattleSDS).

References

Barghi, N., Hermisson, J. & Schlötterer, C. (2020) Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.*, **21**, 769–781.

Barton, N.H., Etheridge, A.M. & Véber, A. (2017) The infinitesimal model: Definition, derivation, and implications. *Theor. Popul. Biol.*, **118**, 50–73.

Barton, N.H., Hermisson, J. & Nordborg, M. (2019) Why structure matters. *eLife*, **8**, e45380.

Barton, N.H. & Keightley, P.D. (2002) Understanding quantitative genetic variation. *Nat. Rev. Genet.*, **3**, 11–21.

Berg, I. van den, Xiang, R., Jenko, J., Pausch, H., Boussaha, M., Schrooten, C., *et al.* (2020) Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. *Genet. Sel. Evol.*, **52**, 37.

Berg, J.J. & Coop, G. (2014) A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet.*, **10**, e1004412.

Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., *et al.* (2019a) Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*, **8**, e39725.

Berg, J.J., Zhang, X. & Coop, G. (2019b) Polygenic Adaptation has Impacted Multiple Anthropometric Traits. *bioRxiv*, 167551.

Boitard, S., Boussaha, M., Capitan, A., Rocha, D. & Servin, B. (2016a) Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds. *Genetics*, **203**, 433–450.

Boitard, S. & Rocha, D. (2013) Detection of signatures of selective sweeps in the Blonde d'Aquitaine cattle breed. *Anim. Genet.*, **44**, 579–583.

663 Boitard, S., Rodríguez, W., Jay, F., Mona, S. & Austerlitz, F. (2016b) Inferring
664 Population Size History from Large Samples of Genome-Wide Molecular Data - An
665 Approximate Bayesian Computation Approach. *PLoS Genet.*, **12**, e1005877.
666 Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for
667 Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
668 Bouwman, A.C., Daetwyler, H.D., Chamberlain, A.J., Ponce, C.H., Sargolzaei, M.,
669 Schenkel, F.S., *et al.* (2018) Meta-analysis of genome-wide association studies for
670 cattle stature identifies common genes that regulate body size in mammals. *Nat.*
671 *Genet.*, **50**, 362–367.
672 Boyle, E.A., Li, Y.I. & Pritchard, J.K. (2017) An Expanded View of Complex Traits:
673 From Polygenic to Omnigenic. *Cell*, **169**, 1177–1186.
674 Buitenhuis, B., Poulsen, N.A., Gebreyesus, G. & Larsen, L.B. (2016) Estimation of
675 genetic parameters and detection of chromosomal regions affecting the major milk
676 proteins and their post translational modifications in Danish Holstein and Danish
677 Jersey cattle. *BMC Genet.*, **17**, 114.
678 Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., *et al.* (2018)
679 The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**,
680 203–209.
681 Castellano, D., Uricchio, L.H., Munch, K. & Enard, D. (2019) Viruses rule over
682 adaptation in conserved human proteins. *bioRxiv*, 555060.
683 Chaisson, M.J.P., Wilson, R.K. & Eichler, E.E. (2015) Genetic variation and the *de*
684 *novo* assembly of human genomes. *Nat. Rev. Genet.*, **16**, 627–640.
685 Charlesworth, B. & Edwards, A.W.F. (2018) A century of variance. *Significance*, **15**,
686 20–25.

Chevin, L.-M. (2019) Selective Sweep at a QTL in a Randomly Fluctuating Environment. *Genetics*, **213**, 987–1005.

Chevin, L.-M. & Hospital, F. (2008) Selective Sweep at a Quantitative Trait Locus in the Presence of Background Genetic Variation. *Genetics*, **180**, 1645–1660.

Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K. (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Binsbergen, R. van, Brøndum, R.F., *et al.* (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.*, **46**, 858–865.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., *et al.* (2013) Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Mol. Biol. Evol.*, **30**, 1544–1558.

Daub, J.T., Moretti, S., Davydov, I.I., Excoffier, L. & Robinson-Rechavi, M. (2017) Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. *Mol. Biol. Evol.*, **34**, 1391–1402.

Duntsch, L., Tomotani, B.M., Villemereuil, P. de, Brekke, P., Lee, K.D., Ewen, J.G., *et al.* (2020) Polygenic basis for adaptive morphological variation in a threatened Aotearoa | New Zealand bird, the hihi (*Notiomystis cincta*). *Proc. R. Soc. B.*, **287**, 20200948.

Edge, M.D. & Coop, G. (2019) Reconstructing the History of Polygenic Scores Using Coalescent Trees. *Genetics*, **211**, 235–262.

Ehrlich, M.A., Wagner, D.N., Oleksiak, M.F. & Crawford, D.L. (2021) Polygenic

Selection within a Single Generation Leads to Subtle Divergence among Ecological Niches. *Genome Biol. Evol.*, **13**, evaa257.

Exposito-Alonso, M., 500 Genomes Field Experiment Team, Burbano, H.A., Bossdorf, O., Nielsen, R. & Weigel, D. (2019) Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature*, **573**, 126–129.

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H.A. & Weigel, D. (2018) Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*. *Nat. Ecol. Evol.*, **2**, 352–358.

Exposito-Alonso, M., Wilton, P. & Nielsen, R. (2020) Non-additive polygenic models improve predictions of fitness traits in three eukaryote model species. *bioRxiv*, 2020.07.14.194407.

Fagny, M. & Austerlitz, F. (2021) Polygenic Adaptation: Integrating Population Genetics and Gene Regulatory Networks. *Trends Genet.*, **37**, 631–638.

Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., *et al.* (2016) Detection of human adaptation during the past 2000 years. *Science*, **354**, 760–764.

Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, **52**, 399–433.

Fuhrmann, N., Prakash, C. & Kaiser, T.S. (2021) Polygenic adaptation from standing genetic variation allows rapid ecotype formation. *bioRxiv*, 2021.04.16.440113.

Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., *et al.* (2017) Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.*, **49**, 1421–1427.

Georges, M., Charlier, C. & Hayes, B. (2018) Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.*

Gramlich, S., Liu, X., Favre, A., Alex Buerkle, C. & Karrenberg, S. (2021) A polygenic architecture with conditionally neutral effects underlies ecological differentiation in *Silene*. *bioRxiv*, 2021.07.06.451304.

Haile-Mariam, M., Nieuwhof, G.J., Beard, K.T., Konstatinov, K.V. & Hayes, B.J. (2013) Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations: Implication of genomic heritability for genomic evaluation. *J. Anim. Breed. Genet.*, **130**, 20–31.

Harland, C., Coppieters, W., Mullaart, E., Charlier, C. & Georges, M. (2018) *Rate of de novo mutation in dairy cattle and potential impact of reproductive technologies*. Proc. World Congr. Genet. Appl. to Livest. Prod.

Harpending, H.C., Batzer, M.A., Gurven, M., Jorde, L.B., Rogers, A.R. & Sherry, S.T. (1998) Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA*, **95**, 1961–1967.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.*, **92**, 433–443.

Hayward, L.K. & Sella, G. (2019) Polygenic adaptation after a sudden change in environment. *bioRxiv*, 792952.

He, F., Arce, A.L., Schmitz, G., Koornneef, M., Novikova, P., Beyer, A., *et al.* (2016) The Footprint of Polygenic Adaptation on Stress-Responsive Cis-Regulatory Divergence in the *Arabidopsis* Genus. *Mol. Biol. Evol.*, **33**, 2088–2101.

Jain, K. & Stephan, W. (2015) Response of Polygenic Traits Under Stabilizing Selection and Mutation When Loci Have Unequal Effects. *G3*, **5**, 1065–1074.

759 Jain, K. & Stephan, W. (2017a) Modes of Rapid Polygenic Adaptation. *Mol. Biol.*
760 *Evol.*, **34**, 3169–3175.
761 Jain, K. & Stephan, W. (2017b) Rapid Adaptation of a Polygenic Trait After a Sudden
762 Environmental Shift. *Genetics*, **206**, 389–406.
763 Jiménez-Mena, B., Tataru, P., Brøndum, R.F., Sahana, G., Guldbrandtsen, B. &
764 Bataillon, T. (2016) One size fits all? Direct evidence for the heterogeneity of genetic
765 drift throughout the genome. *Biol. Lett.*, **12**.
766 John, S. & Stephan, W. (2020) Important role of genetic drift in rapid polygenic
767 adaptation. *Ecol. Evol.*, **10**, 1278–1287.
768 Josephs, E.B., Berg, J.J., Ross-Ibarra, J. & Coop, G. (2019) Detecting Adaptive
769 Differentiation in Structured Populations with Genomic Data and Common Gardens.
770 *Genetics*, **211**, 989–1004.
771 Kelleher, J., Etheridge, A.M. & McVean, G. (2016) Efficient Coalescent Simulation
772 and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.*, **12**,
773 e1004842.
774 Kreiner, J.M., Tranel, P.J., Weigel, D., Stinchcombe, J.R. & Wright, S.I. (2020) The
775 genetic architecture and genomic context of glyphosate resistance in *Amaranthus*
776 *tuberculatus*. *bioRxiv*, 2020.08.19.257972.
777 Lande, R. (1983) The response to selection on major and minor mutations affecting a
778 metrical trait. *Heredity*, **50**, 47–65.
779 Le Corre, V. & Kremer, A. (2012) The genetic differentiation at quantitative trait loci
780 under local adaptation. *Mol. Ecol.*, **21**, 1548–1566.
781 Lemay, D.G., Lynn, D.J., Martin, W.F., Neville, M.C., Casey, T.M., Rincon, G., *et al.*
782 (2009) The bovine lactation genome: insights into the evolution of mammalian milk.

783 *Genome Biol.*, **10**, R43.

784 Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows–

785 Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

786 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009) The

787 Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

788 Liu, X., Li, Y.I. & Pritchard, J.K. (2019) Trans Effects on Gene Expression Can Drive

789 Omnigenic Inheritance. *Cell*, **177**, 1022-1034.e6.

790 MacEachern, S., Hayes, B., McEwan, J. & Goddard, M. (2009) An examination of

791 positive selection and changing effective population size in Angus and Holstein cattle

792 populations (*Bos taurus*) using a high density SNP genotyping platform and the

793 contribution of ancient polymorphism to genomic diversity in Domestic cattle. *BMC*

794 *Genom.*, **10**, 181.

795 MacLachlan, I.R., McDonald, T.K., Lind, B.M., Rieseberg, L.H., Yeaman, S. & Aitken,

796 S.N. (2021) Genome-wide shifts in climate-related variation underpin responses to

797 selective breeding in a widespread conifer. *Proc. Natl. Acad. Sci. USA*, **118**,

798 e2016900118.

799 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., *et*

800 *al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-

801 generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

802 Meuwissen, T., Hayes, B. & Goddard, M. (2013) Accelerating Improvement of

803 Livestock with Genomic Selection. *Annu. Rev. Anim. Biosci.*, **1**, 221–237.

804 Nielsen, R. (2005) Molecular Signals of Natural Selection. *Annu. Rev. Genet.*, **39**,

805 197–218.

806 Nordborg, M. & Donnelly, P. (1997) The Coalescent Process With Selfing. *Genetics*,

807 **146**, 1185–1195.

808 Novembre, J. & Barton, N.H. (2018) Tread Lightly Interpreting Polygenic Tests of

809 Selection. *Genetics*, **208**, 1351–1355.

810 Pavlidis, P., Metzler, D. & Stephan, W. (2012) Selective Sweeps in Multilocus Models

811 of Quantitative Traits. *Genetics*, **192**, 225–239.

812 Peng, B. & Kimmel, M. (2005) simuPOP: a forward-time population genetics

813 simulation environment. *Bioinformatics*, **21**, 3686–3687.

814 Pritchard, J.K. & Di Rienzo, A. (2010) Adaptation - not by sweeps alone. *Nat. Rev.*

815 *Genet.*, **11**, 665–667.

816 Pritchard, J.K., Pickrell, J.K. & Coop, G. (2010) The Genetics of Human Adaptation:

817 Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Curr. Biol.*, **20**, R208–R215.

818 Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T.M., Fries, R., *et al.* (2014)

819 Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. *PLoS Genet.*,

820 **10**, e1004148.

821 Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R., *et al.*

822 (2010) A genome-wide scan for signatures of recent selection in Holstein cattle.

823 *Anim. Genet.*

824 Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing

825 genomic features. *Bioinformatics*, **26**, 841–842.

826 R Core Team. (2021) *R: A Language and Environment for Statistical Computing*. R

827 Foundation for Statistical Computing, Vienna, Austria.

828 Racimo, F., Berg, J.J. & Pickrell, J.K. (2018) Detecting Polygenic Adaptation in

829 Admixture Graphs. *Genetics*, **208**, 1565–1584.

830 Robinson, M.R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T.,

Shakhbazov, K., *et al.* (2015) Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.*, **47**, 1357.

Rosen, B., Bickhart, D.M., Schnabel, R.D., Koren, S., Elsik, C., Zimin, A.V., *et al.* (2018) Modernizing the Bovine Reference Genome Assembly. *Proc. World Congr. Genet. Appl. Livest. Prod.*, 11–16.

Rowan, T.N., Durbin, H.J., Seabury, C.M., Schnabel, R.D. & Decker, J.E. (2021) Powerful detection of polygenic selection and evidence of environmental adaptation in US beef cattle. *PLoS Genet.*, **17**, e1009652.

Sanjak, J.S., Sidorenko, J., Robinson, M.R., Thornton, K.R. & Visscher, P.M. (2018) Evidence of directional and stabilizing selection in contemporary humans. *Proc. Natl. Acad. Sci. USA*, **115**, 151–156.

Savolainen, O., Lascoux, M. & Merila, J. (2013) Ecological genomics of local adaptation. *Nat. Rev. Genet.*, **14**, 807–820.

Schoech, A.P., Jordan, D.M., Loh, P.-R., Gazal, S., O'Connor, L.J., Balick, D.J., *et al.* (2019) Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.*, **10**, 790.

Sella, G. & Barton, N.H. (2019) Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu. Rev. Genom. Hum. Genet.*, **20**, 461–493.

Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., *et al.* (2019) Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, **8**, e39702.

Sørensen, A.C., Sørensen, M.K. & Berg, P. (2005) Inbreeding in Danish Dairy Cattle Breeds. *J. Dairy Sci.*, **88**, 1865–1872.

855 Soyeurt, H., Gillon, A., Vanderick, S., Mayeres, P., Bertozzi, C. & Gengler, N. (2007)
856 Estimation of Heritability and Genetic Correlations for the Major Fatty Acids in Bovine
857 Milk. *J. Dairy Sci.*, **90**, 4435–4442.
858 Speidel, L., Forest, M., Shi, S. & Myers, S.R. (2019) A method for genome-wide
859 genealogy estimation for thousands of samples. *Nat Genet.*, **51**, 1321–1329.
860 Stephan, W. (2016) Signatures of positive selection: from selective sweeps at
861 individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.*,
862 **25**, 79–88.
863 Stephan, W. (2019) Selective Sweeps. *Genetics*, **211**, 5–13.
864 Stern, A.J., Speidel, L., Zaitlen, N.A. & Nielsen, R. (2021) Disentangling selection on
865 genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum.*
866 *Genet.*, **108**, 219–239.
867 Stetter, M.G., Thornton, K. & Ross-Ibarra, J. (2018) Genetic architecture and
868 selective sweeps after polygenic adaptation to distant trait optima. *PLoS Genet.*, **14**,
869 e1007794.
870 Storey, J.D., Bass, A.J., Dabney, A. & Robinson, D. (2021) qvalue: Q-value
871 estimation for false discovery rate control [WWW Document]. URL
872 http://github.com/StoreyLab/qvalue [accessed on 2021].
873 Svardal, H., Jasinska, A.J., Apetrei, C., Coppola, G., Huang, Y., Schmitt, C.A., *et al.*
874 (2017) Ancient hybridization and strong adaptation to viruses across African vervet
875 monkey populations. *Nat. Genet.*, **49**, 1705–1713.
876 Teshima, K.M. & Innan, H. (2009) mbs: modifying Hudson's ms software to generate
877 samples of DNA sequences with a biallelic site under selection. *BMC Bioinform.*, **10**,
878 166.

Thornton, K.R. (2019) Polygenic Adaptation to an Environmental Shift: Temporal Dynamics of Variation Under Gaussian Stabilizing Selection and Additive Effects on a Single Trait. *Genetics*, **213**, 1513–1530.

Turchin, M.C., Chiang, C.W., Palmer, C.D., Sankararaman, S., Reich, D., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, *et al.* (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet.*, **44**, 1015–1019.

Uricchio, L.H., Kitano, H.C., Gusev, A. & Zaitlen, N.A. (2019) An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol. Lett.*, **3**, 69–79.

Visscher, P.M. & Goddard, M.E. (2019) From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics*, **211**, 1125–1130.

Vitti, J.J., Grossman, S.R. & Sabeti, P.C. (2013) Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.*, **47**, 97–120.

Vladar, H.P. de & Barton, N. (2014) Stability and Response of Polygenic Traits to Stabilizing Selection and Mutation. *Genetics*, **197**, 749–767.

Whitlock, M.C. (2008) Evolutionary inference from F_{ST}. *Mol. Ecol.*, **17**, 1885–1896.

Wieters, B., Steige, K.A., He, F., Koch, E.M., Ramos-Onsins, S.E., Gu, H., *et al.* (2021) Polygenic adaptation of rosette growth in Arabidopsis thaliana. *PLoS Genet.*, **17**, e1008748.

Wollstein, A. & Stephan, W. (2014) Adaptive Fixation in Two-Locus Models of Stabilizing Selection and Genetic Drift. *Genetics*, **198**, 685–697.

Wray, N.R., Kemper, K.E., Hayes, B.J., Goddard, M.E. & Visscher, P.M. (2019) Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in

Humans. *Genetics*, **211**, 1131–1141.

Wray, N.R., Wijmenga, C., Sullivan, P.F., Yang, J. & Visscher, P.M. (2018) Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell*, **173**, 1573–1580.

Wright, S. (1951) The genetical structure of populations. *Ann. Eugen.*, **15**, 323–354.

Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J.M., Zhang, Y., Tokarska, M., *et al.* (2018) Pervasive introgression facilitated domestication and adaptation in the *Bos* species complex. *Nat. Ecol. Evol.*, **2**, 1139–1145.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., *et al.* (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.*, **47**, 1114–1120.

Yeaman, S., Hodgins, K.A., Lotterhos, K.E., Suren, H., Nadeau, S., Degner, J.C., *et al.* (2016) Convergent local adaptation to climate in distantly related conifers. *Science*, **353**, 1431.

Zan, Y. & Carlborg, Ö. (2018) A Polygenic Genetic Architecture of Flowering Time in the Worldwide Arabidopsis thaliana Population. *Mol. Biol. Evol.*, **36**, 141–154.

Zeder, M.A. (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. USA*, **105**, 11597–11604.

Zeng, J., Vlaming, R. de, Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., *et al.* (2018) Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.*, **50**, 746–753.

Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L.R., Wu, Y., Wang, H., *et al.* (2021) Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat. Commun.*, **12**, 1164.

Zhang, W., Reeves, G.R. & Tautz, D. (2021) Testing Implications of the Omnigenic Model for the Genetic Analysis of Loci Identified through Genome-wide Association. *Curr. Biol.*, **31**, 1092-1098.e6.

Zhao, F., McParland, S., Kearney, F., Du, L. & Berry, D.P. (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genet. Sel. Evol.*, **47**, 49.

Zhou, X., Carbonetto, P. & Stephens, M. (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.*, **9**, e1003264.

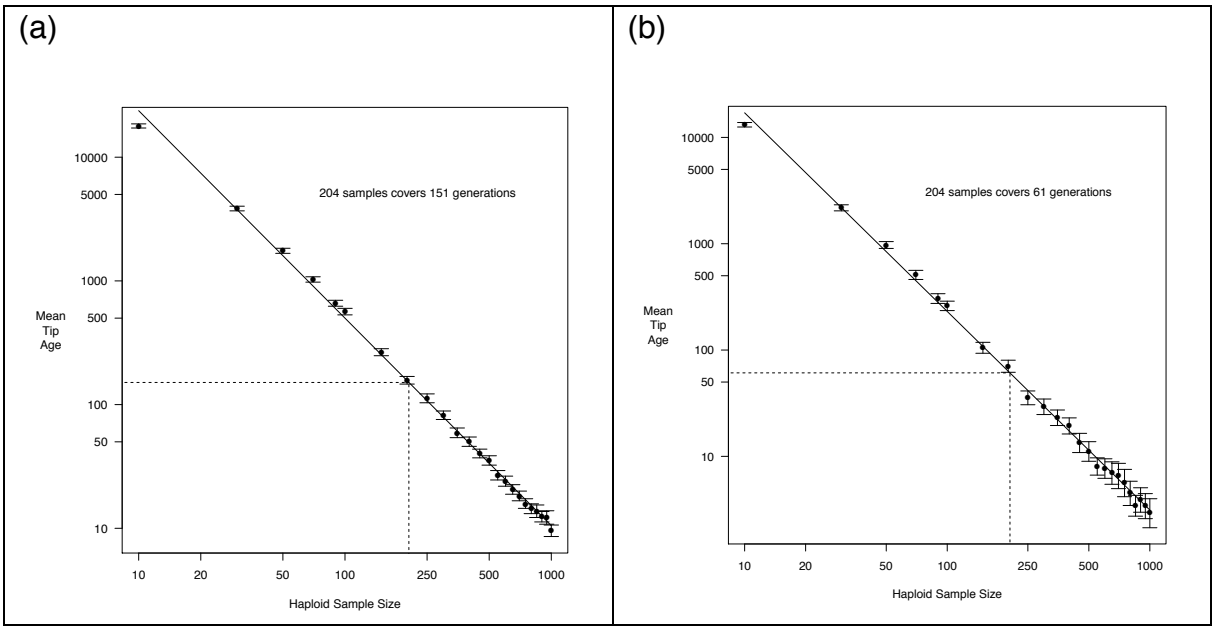


Figure 1: Simulated mean tip age for *B. taurus*, as a function of the number of haploid samples. Simulations assumed either (a) demography as inferred by Boitard et al. (2016b) (the 'High N_0' model), or (b) the same but with a smaller present–day N_e of 49 (the 'Low N_0' model). Points are the mean values; bars show 95% confidence intervals. The solid line is the best linear fit to the log of both values; dotted lines show the predicted tip age for 204 alleles.

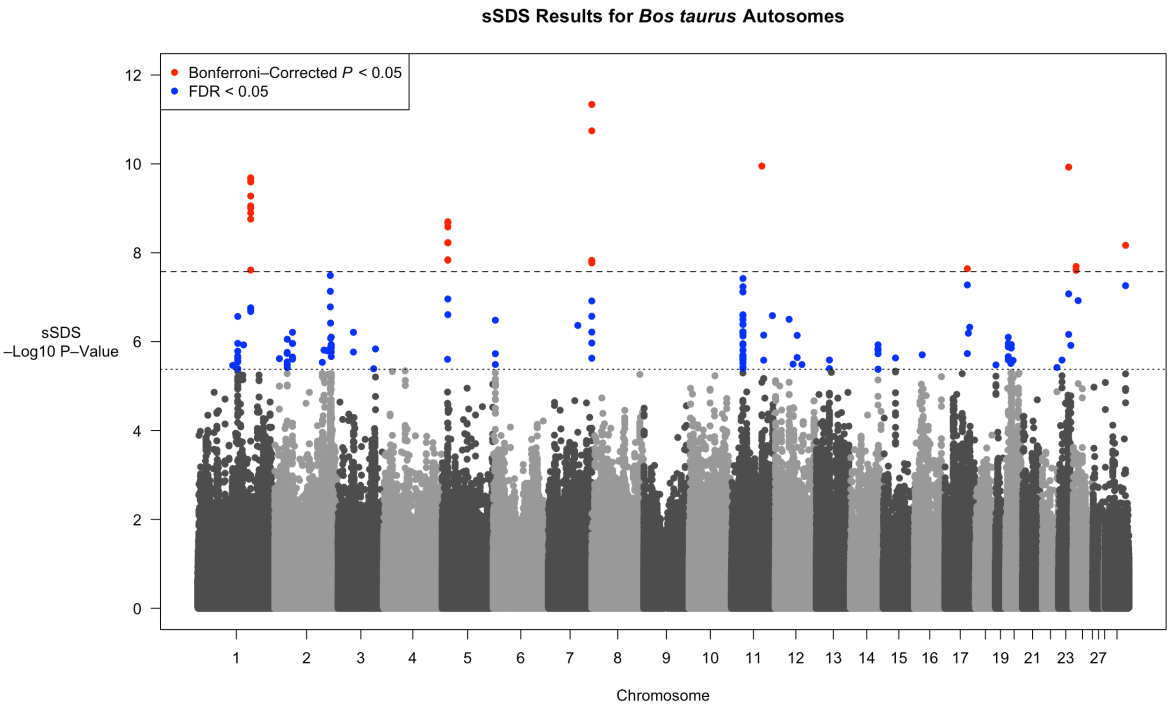
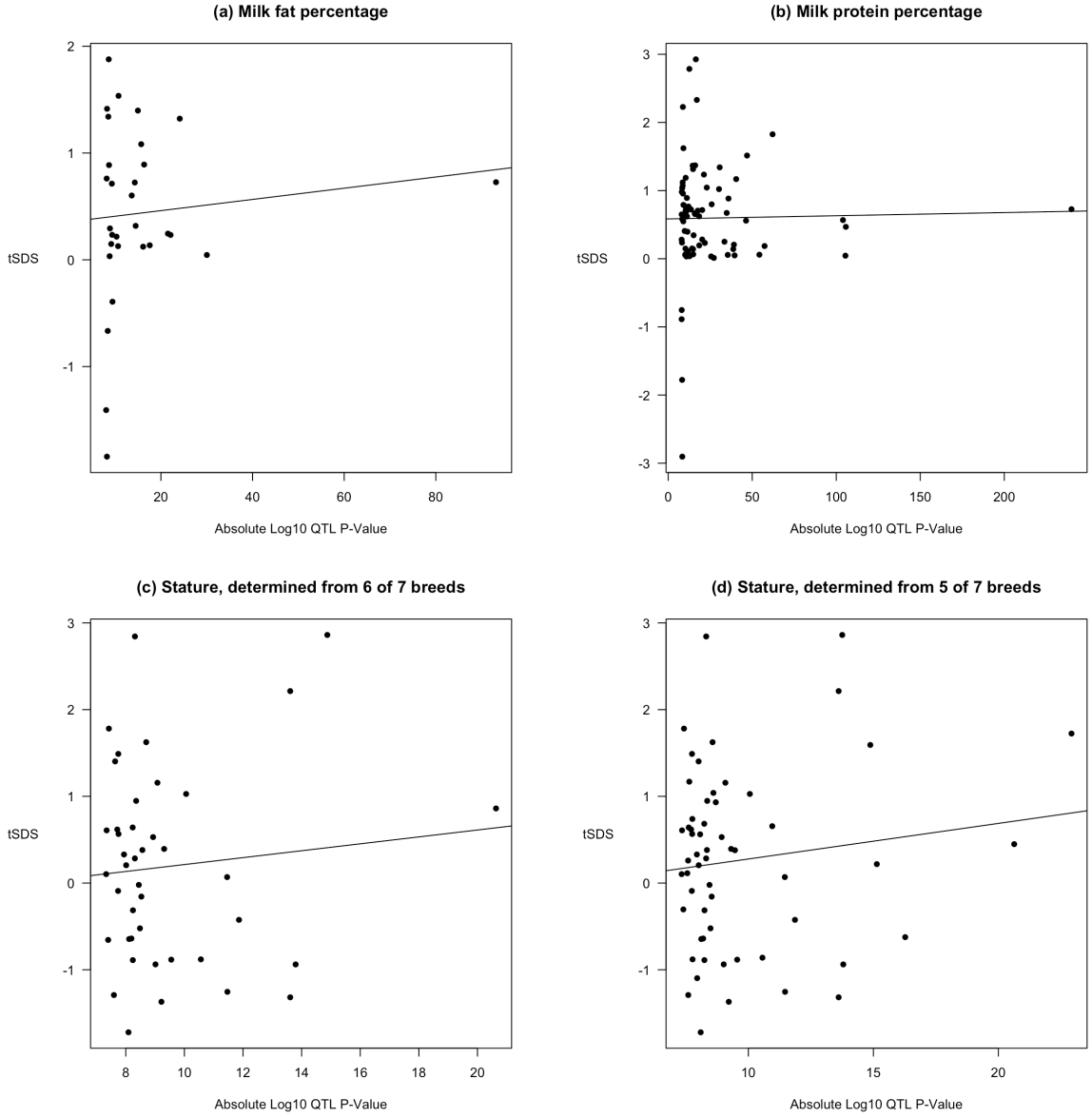


Figure 2: *P*–values of sSDS scores across *B. taurus* autosomes, as plotted on a negative Log_{10} scale, as a function of the chromosome. Alternating black and grey points show (non–significant) values from different chromosomes. Blue points are SNPs with *FDR* < 0.05, with the cut-off denoted by a horizontal dotted line. Red points are SNPs with Bonferroni–corrected *P*–value < 0.05 (actual *P*–value < ~2.7 x 10^{-8}), with the cut-off denoted by a horizontal dashed line. Figure S1 shows results for the Low N_0 model.



952

953 Figure 3: Relationship between tSDS scores near milk or stature QTLs, as noted in

954 the subheadings, and the absolute log *P*-value of QTLs. Lines show a linear model

955 regression fit. Figure S3 shows results assuming a low N_0 model.

956

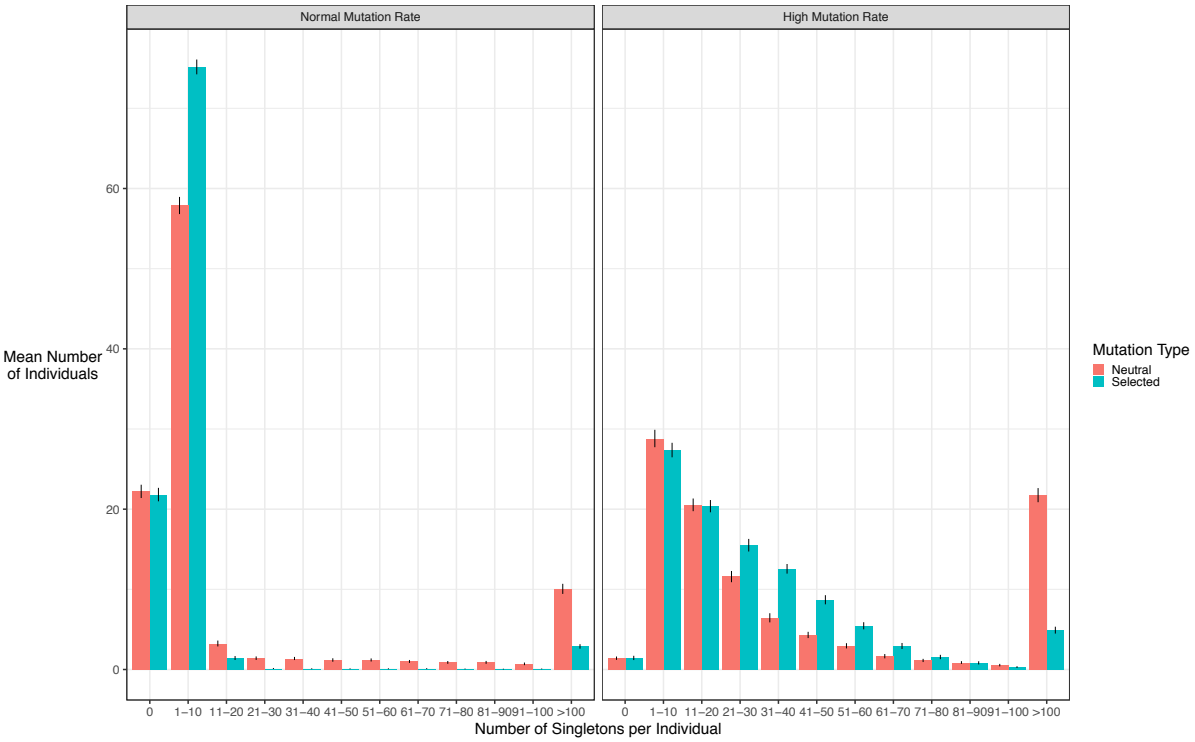


Figure 4: Mean distribution of singleton numbers per individual for each simulation, wither assuming a standard mutation rate (left) or a 10-fold higher mutation rate (right). Bars represent 95% confidence intervals.

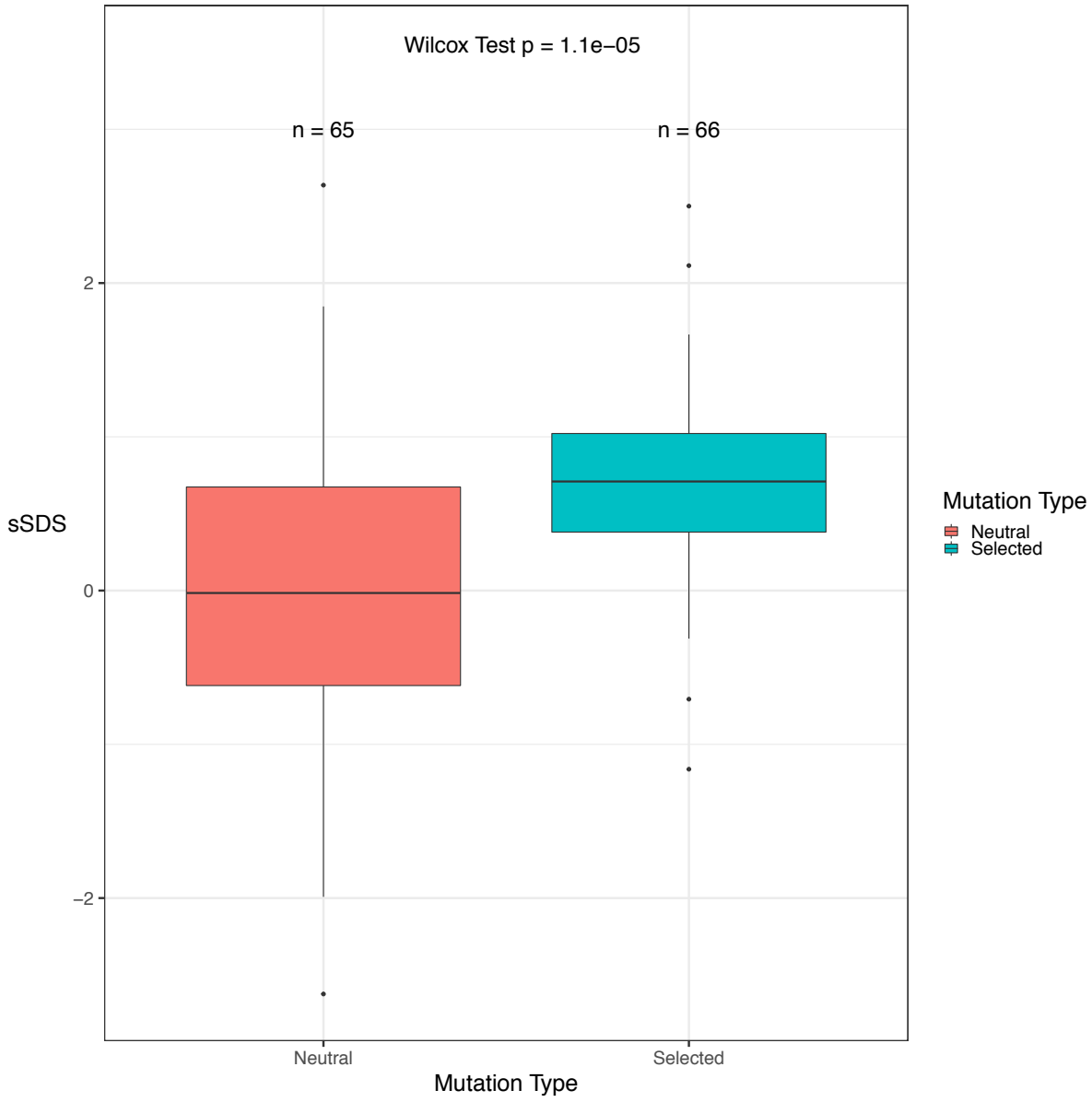


Figure 5: Distribution of simulated sSDS scores assuming a high mutation rate, for the neutral and selected cases. Numbers above each box plot denotes how many simulations produced SDS scores and were included in the plot.

Chromosome	Gene Name	Start Position	End Position	Gene Biotype	High, Low N_0
1	PPM1L	106405113	106727070	Protein Coding	High, Low
5	TMCC3	24306913	24595494	Protein Coding	High, Low
5	CEP83	24070404	24345243	Protein Coding	High, Low
17	U6	43381106	43381209	snRNA	Low
17	CTSO	43364999	43381605	Protein Coding	Low
17	TDO2	43386894	43403747	Protein Coding	High, Low
23	OR12D2H	29291787	29292713	Protein Coding	High, Low
23	OR12D2E	29305933	29309785	Protein Coding	High, Low
24	GAREM1	24694637	24927333	Protein Coding	High, Low
29	NTM	34576918	34994005	Protein Coding	High, Low

Table 1: Genes that overlap or lie close to Bonferroni–significant sSDS regions. The 'High, Low N_0' column specifies which genes are close to significant SNPs for each N_0 model.