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

Matthew Hartfield[1,2], Nina Aagaard Poulsen[3], Bernt Guldbrandtsen[4,5], 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.

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

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

Abstract: Many quantitative traits are subject to 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 candidate genes for recent selection, including some relating to protein and cell regulation, the synaptic system, body growth, and immunity. We do not find strong evidence that two traits important for humans, milk–protein content and stature, have been subject to directional selection. These results inform on which genes underlie recent domestication in *B. taurus*. We propose how polygenic selection can be best investigated 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 and Di Rienzo, 2010; Pritchard et al., 2010). Many polygenic models have been formulated that account for both the response to phenotypic selection, and the maintenance of genetic variance at quantitative traits (reviewed by Sella and Barton [2019]). Among them is Fisher's infinitesimal model, which is important for its historical role in uniting population and quantitative genetics, and it's recent renaissance in the context of genome–wide association studies (Fisher, 1918; Barton and Keightley, 2002; Barton et al., 2017; Charlesworth and Edwards, 2018; Visscher and 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 and Barton, 2014; Wollstein and Stephan, 2014; Jain and Stephan, 2015, 2017a, 2017b; Stetter et al., 2018; Thornton, 2019; Hayward and Sella, 2019). Furthermore, if epistasis exists between variants, many selected alleles do not reach fixation as they eventually become deleterious (de Vladar and Barton, 2014; Jain and Stephan, 2017b). The spread of large–effect alleles may also be impeded if a faster

adaptive response can be otherwise realised through changes at many small–effect alleles (Lande, 1983; Chevin and Hospital, 2008; Pavlidis et al., 2012; Chevin, 2019). Alternatively, if the optimum shift is sufficiently large, then major–effect mutations that fix first can subsequently be replaced by small–effect variants over longer timescales (on the order of the population size; Hayward and Sella (2019)). Overall, only a small proportion of loci affected by polygenic selection are expected to fix sufficiently quickly to leave selection signatures in genomic data (Pavlidis et al., 2012; Thornton, 2019).

Due to this difficulty, earlier methods for detecting polygenic selection focused on cases where selection favours distinct phenotypes in different populations, so trait differentiation amongst populations will be greater than expected under neutral drift. Tests for this form of selection relied on comparing Q_{st} and F_{st} statistics, which respectively measured mean genetic differentiation at the trait itself and a set of neutral loci (Whitlock, 2008; Le Corre and Kremer, 2012; Savolainen et al., 2013). Yet these methods do not determine which genomic regions are subject to selection. This situation has now changed with the increased number of genome–wide association study (GWAS) data that link genotypes and phenotypes, as exemplified by the development of large cohort studies (e.g., the UK Biobank; Bycroft et al. [2018]). The release of these data have spurred a series of studies and new methods designed specifically to detect polygenic selection. These methods usually involve determining which SNPs underlying a phenotype show correlated changes in frequency (Berg and Coop, 2014; Racimo et al., 2018; Sanjak et al., 2018; Josephs et al., 2019; Berg et al., 2019; Berg et al., 2019a; Uricchio et al., 2019; Edge and Coop, 2019; Wieters et al., 2020); which sets of alleles are associated with certain environmental or climatic variations (Coop et al., 2010; Turchin et al., 2012; Robinson et al., 2015;

Yeaman et al., 2016; Exposito-Alonso et al., 2018; Zan and Carlborg, 2018; Exposito-Alonso et al., 2019; Ehrlich et al., 2020); or determining which SNPs explain a large fraction of phenotypic variance and trait heritability (Zhou et al., 2013; Yang et al., 2015; Gazal et al., 2017; Zeng et al., 2018; Schoech et al., 2019). Some of these approaches use overlapping methods.

Detecting recent polygenic selection is much harder, as long periods of time (number of generations on the order of the population size; Hayward and Sella, 2019; Thornton, 2019) may be needed to cause detectable frequency changes in weak–effect alleles. Over shorter timescales, these frequency changes are expected to be more modest and harder to detect (Stephan, 2016; Jain and Stephan, 2017a). A recent breakthrough in detecting these subtle changes was the development of the 'Singleton Density Score' (SDS), a statistic tailored to detect recent and possibly small, but coordinated allele frequency changes over many SNPs (Field et al., 2016). Recent selection at a locus favouring one particular variant will lead to a reduction in the number of singletons (i.e., variants that are only observed once) around it. The SDS detects regions that exhibit a reduction in the density of singletons, to determine candidate regions that have been subject to recent selection. Using this approach, Field et al. (2016) found correlations between SDS scores at SNPs and their associated GWAS effect sizes for several polygenic traits in the modern UK human population, including increased height, infant head circumference and fasting insulin. Their findings suggested that these traits have been subject to recent selection during the last 75 or so generations (about 2,000 years).

The SDS method is ideally suited to organisms where large amount of whole–genome data are available, along with QTL or GWAS information that link genotypes to phenotypes, and a means to correct for population stratification (as it can generate

104 spurious associations between SNPs and trait variation). The last point is important
105 as several recent studies, including the SDS analyses, reported evidence that
106 increased height is subject to polygenic selection (Turchin et al., 2012; Berg and
107 Coop, 2014; Robinson et al., 2015; Field et al., 2016; Racimo et al., 2018). However,
108 recent attempts to replicate these findings on the UK Biobank dataset have failed to
109 do so, and previous results may instead reflect unaccounted–for population structure
110 (Novembre and Barton, 2018; Barton et al., 2019; Sohail et al., 2019; Berg et al.,
111 2019; Uricchio et al., 2019; Edge and Coop, 2019).
112 Domesticated species are attractive systems for studying recent selection, as
113 selected phenotypes are often already known, and these species are subject to
114 large–scale sequencing studies. Population structure can also be controlled for by
115 focusing on specific breeds. Investigating the genetic architecture underlying rapid
116 selection in these species is also important to determine how they respond to
117 agricultural practices, and uncover selection targets that can be used to improve
118 breeding programs (Georges et al., 2018). Domestic cattle *Bos taurus* has been
119 subject to intensive genomics analyses to improve artificial selection for traits that are
120 beneficial for human use, including milk protein yield and stature (Hayes et al., 2009;
121 Meuwissen et al., 2013; Wray et al., 2019). These traits are influenced in part by an
122 individual's genome; heritability estimates of milk protein content range between 28
123 and 70% (Buitenhuis et al., 2016 and references therein), while stature estimates
124 range between 25 and 85% (Nelsen et al., 1986; Northcutt and Wilson, 1993).
125 Previous selection scans on *B. taurus* reported individual regions that were likely to
126 be subject to recent selection, some of which were close to genetic regions for
127 stature and milk protein content (Lemay et al., 2009; MacEachern et al., 2009;
128 Qanbari et al., 2010; Boitard and Rocha, 2013; Qanbari et al., 2014; Zhao et al.,

2015; Boitard et al., 2016a; Bouwman et al., 2018). However, stature and milk protein content are polygenic traits, with several genetic regions and QTLs associated with each (Lemay et al., 2009; Boitard et al., 2016a; Bouwman et al., 2018).

Here, we applied a modified version of the SDS method to whole–autosome sequencing data from 102 *B. taurus* Holstein individuals. We first determined genetic regions that have been subject to recent directional selection, and subsequently tested if evidence exists for recent selection acting on a set of regions underlying either milk protein content or stature in this breed.

Results

Methods outline

Figure 1 outlines the filtering steps applied to the 102 whole–autosome genotypes. We retained bi–allelic SNPs that had a sensible level of coverage and did not lie in putatively over–assembled regions (i.e., duplicated sections that caused many reads to assemble at a specific genetic location). Over–assembled regions are highly heterozygous with elevated coverage, and can exhibit false signatures of recent selection. We also obtained a set of singletons and filtered them to retain high quality variants where both alleles were equally well covered, to remove potentially erroneous calls.

SDS reflects the log–ratio of inferred tip lengths (and hence singleton densities) around one allele over another at a locus. Field et al. (2016) applied the statistic to polarised data where the ancestral and derived alleles were determined with high confidence. In that case, increased SDS values reflected selection favouring younger, derived SNPs over ancestral variants. However, for many species there is uncertainty around which SNPs are ancestral or derived (Keightley and

Jackson, 2018). Hence, we instead focused on the absolute value of standardized SDS statistics, which we denote asSDS. As the original SDS measurement is a log–ratio, then asSDS values reflect the relative increase of one SNP (either ancestral or derived), and hence a change in inferred tip lengths, over the other. This statistic is a broader measure of polygenic selection, as opposed to a specific test for positive selection acting on younger derived variants. Further details are available in the *Methods* section.

Estimating timescale of selection

We first determined the timescale over which we expect to detect selection in our sample using the SDS method. SDS measures the changes in singleton numbers around putatively selected SNPs, relative to background numbers in the absence of selection. As singletons arise on the tips of the underlying gene trees, the average tip length in the genealogy of sequenced samples determines the timescale over which the SDS detects a signal (Field et al., 2016). To calculate the mean tip age, we simulated gene genealogies under two scenarios. We first simulated the Holstein population demography inferred by Boitard et al. (2016b), which suggested that this population experienced a sudden decline in effective population size (N_e) since domestication, but with a present–day N_e (~793) that is much larger than that inferred from pedigree data (49; Table 2 of Sørensen et al. (2005), estimate for 1993–2003) or from temporal variation in SNP frequencies (48; Jiménez–Mena et al. 2016). Hence, we also simulated genealogies under a second model that used the Boitard et al. (2016b) demographic model but with the present–day N_e set to 49. These scenarios will be referred to as the 'High N_0' and 'Low N_0' models, respectively.

Figure 2 shows simulation results. Depending on the assumed present–day

179 N_e, the tip length in our sample of 204 alleles (i.e., assuming two per diploid
180 individual) goes back either 65 or 148 generations. Assuming 5 years per generation
181 (Boitard et al., 2016b), this time scale corresponds to between 325 and 740 years
182 ago. Since *B. taurus* domestication started around 10,000 years ago (Zeder, 2008)
183 the sample size used in this study will only capture selection acting in the very recent
184 past that is more relevant for breed formation, rather than selection during *B. taurus*
185 domestication.

186 We will focus on detecting selection signatures assuming the high N_0 model.
187 Results using the low N_0 model to calibrate scores were broadly similar. They are
188 outlined in the Supplementary Text; we will highlight when differences arise.

189

190 ## *Genome–wide asSDS*

191 Figure 3 plots asSDS values (at SNPs with minor allele frequency greater than
192 5%) across all autosomes, excluding chromosome 25 (due to an insufficient number
193 of singletons needed to obtain SDS scores after filtering). Many SNPs have elevated
194 asSDS scores (1051 SNPs at $FDR < 0.05$; 2112 for the low N_0 model). Several
195 regions contain SNPs with significantly high asSDS values (Bonferroni–corrected $P <$
196 0.05; actual $P <$ ~2.5e–8). To further investigate potential selection targets, we looked
197 for genes that either overlapped significant SNPs or lay 10kb up– or downstream of
198 them. Linkage disequilibrium (LD), as measured by r_2, decays to around 0.2 over
199 50kb in Danish Holstein breeds (Buitenhuis et al., 2016), so genes within 10kb
200 should be in LD with regions harbouring high asSDS scores. Table 1 lists these
201 genes; there is some overlap between results obtained using either a high or low N_0,
202 but more gene targets are present under the low N_0 model. Most of these genes are
203 of unknown function; the results also include unnamed genes and a snRNA. *FBXO4,*

MANBA are involved in protein regulation, while *PPM1L* is involved with cellular regulation and the activation of stress–activated protein kinases. *TRIM9* and *NRXN1* are involved in the synaptic system. *GHR* is linked to both body growth and milk yield, and has been reported in previous selection studies (Qanbari et al., 2010; Zhao et al., 2015). SNPs with significantly elevated scores are also found on chromosome 23 near the MHC region, which may reflect over–dominant selection. All Bonferroni–significant SNPs were removed from subsequent tests of recent polygenic selection. Figure S1 shows results for the low N_0 model.

Testing for polygenic selection acting on milk protein and stature

We collated asSDS scores of SNPs that either lie in genetic regions associated with milk proteins (as outlined in Lemay et al. [2009]), or those that lie close to stature QTLs (Bouwman et al., 2018). The latter were inferred from a meta–analysis of GWAS studies conducted in seven Holstein populations, but not every QTL had an effect size reported in each population. We hence investigated two overlapping consensus QTL sets, where an effect size was either reported in at least 6 of 7 populations (yielding 42 QTLs with asSDS scores associated with them), or where effect sizes were reported in at least 5 of 7 populations (58 QTLs had asSDS scores). We used a generalized linear model (GLM) to determine whether genome regions containing either milk protein genes or stature QTLs are associated with differences in asSDS scores.

Figure 4 shows the distribution of asSDS values for SNPs that lies either in milk protein genes, or close to stature QTLs, compared to the background genome–wide distribution of asSDS scores. A GLM analysis shows that while several chromosomes and allele frequency bins are significant predictors of asSDS variation,

the presence of a SNP in a milk protein gene does not explain any additional variation (effect size = 0.0212, P = 0.107; see Table S1 for the full results). SNPs near stature QTLs do not have significantly different asSDS values, irrespective of whether we use QTLs with reported effect sizes in at least 6 of 7 Holstein populations (effect = –0.210, P = 0.122; Figure 4(b), Table S2), or with effect sizes reported in at least 5 of 7 Holstein populations (effect = –0.151, P = 0.208; Figure S2 and Table S3).

Under the low N_0 model, milk protein genes have slightly elevated asSDS values (effect size = 0.0489, P = 0.000261; Figure S3, Table S4), but stature QTLs do not (Figure S3; see Tables S5, S6 for effect sizes and P–values). However, each genetic region contains several SNPs with asSDS scores that are correlated because of linkage disequilibrium (LD). To account for LD within genes, and thereby obtain a more reliable P–value associated with elevated asSDS scores, we performed a permutation test where milk–protein genes were randomly distributed along the genome. We subsequently measured the additional variance predicted by the presence or absence of these genes in the permuted datasets (see Methods for details). The observed amount of variance explained in the original data is then compared to the set of values observed for permutated data. In all cases the observed value lies within the permuted values (Figure S4). We therefore conclude that milk–protein genes as a whole do not harbour SNPs with significantly different asSDS scores compared to the rest of the genome. Permutation results were also non–significant when applied to stature QTLs (Figure S4).

Discussion

Summary of results

We have analysed an extensive *B. taurus* genomic dataset to identify signatures of recent selection, 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 the Holstein breed, our simulations suggested that the SDS method can detect very recent selection events, arising no more than approximately 740 years ago (Figure 2). A whole–genome scan for asSDS scores identified several targets of recent directional selection that overlap or lie close to protein–coding genes (Figure 3; Table 1). When the functions of these genes are known, they are involved in protein regulation, the synaptic system, and body growth. Significant values were also observed in 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 SNPs in these two groups are significantly associated with changes in asSDS values. However, asSDS values are only different in the presence of milk–protein genes when assuming a small N_0, and this difference is not significant when performing a permutation test (Figure S4). Hence, while asSDS 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

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

two traits that are important for human use, which were subject to artificial selection since domestication. One potential reason for this lack of signal is that selection on these traits was mainly driven by major–effect mutations that have already fixed in the population, with a smaller contribution from minor effect mutations. Theoretical models have shown that more major–effect QTLs are likely to fix if the population lies further from a fitness optimum (Lande, 1983; Jain and Stephan, 2017b; Thornton, 2019). Domesticated species, which experience strong directional artificial selection, could thereby fix more adaptive mutation via sweep–like processes compared to populations evolving in more stable environments (Lande, 1983; Jain and Stephan, 2017a). Furthermore, once a population has adapted to a new environment (the domestication phenotype in this case), then any remaining major–effect mutations are likely to be superseded by variants with weaker effects, which are harder to detect (Hayward and Sella, 2019). Simulations (Figure 2) suggested that SDS values obtained from our sample of 102 individuals will principally detect very recent selection related to breed formation and subsequent within–breed selection, rather than selection arising from domestication that was more likely to involve the promotion and fixation of major–effect mutations. Finally, the response to polygenic selection is weakened in smaller populations (John and Stephan, 2020), which will further hamper our ability to detect it in *B. taurus*.

Detecting polygenic selection through singleton densities is also made harder by potentially reduced tip lengths in *B. taurus*, which likely reflects successive bottlenecks due to domestication, breed formation and intense recent selection. The effective population size of many *B. taurus* breeds appears to have undergone a decline since domestication (Sørensen et al., 2005; Boitard et al., 2016b). Contracting populations produce gene genealogies with very short tip lengths

(Harpending et al., 1998). Hence it will be harder to detect differences between the tip lengths of two SNPs if the baseline tip length is already very short. A reduction in baseline singleton numbers also reduces the power to investigate asSDS values in telomeric regions. SDS values are calculated using the distance up- and downstream from a SNP to the nearest singleton, and are undefined if a certain number of samples do not harbour singletons in either direction (Field et al., 2016). SDS values are hence less likely to be defined in telomeric regions, as it is generally less feasible to observe singletons up until the end of the chromosome. This problem is exacerbated if there are few singletons overall.

The lack of a polygenic selection signal in this study also resonates with recent discussions surrounding the strength of the evidence for it in humans. Although there are larger numbers of high–quality genotypes available, recent claims of polygenic selection are likely to have been confounded by population stratification (Novembre and Barton, 2018; Barton et al., 2019; Sohail et al., 2019; Berg et al., 2019; Uricchio et al., 2019; Edge and Coop, 2019), suggesting that it is inherently difficult to detect polygenic selection from genome sequence data. One potential solution to increase power is to use recent methods to directly infer trees, and hence singleton branches, from genome data (Edge and Coop, 2019; Speidel et al., 2019). An alternative approach would be to look beyond sequence data and focus on gene networks. The recently–proposed ‘omnigenic’ model (Boyle et al., 2017; Liu et al., 2019) posits that variation in quantitative traits is principally affected by a plethora of ‘peripheral’ genes that indirectly affect them, rather than a limited set of ‘core’ genes that directly modify a trait. These numerous peripheral genes may exert their influence via regulatory effects (e.g., gene expression changes), but are also expected to be highly pleiotropic. Although fully testing the omnigenic model will require larger datasets and

326 novel experimental designs (Wray et al., 2018), there is nascent evidence that gene
327 regulation may underlie directional polygenic selection. Boitard et al. (2016a) found
328 that some adaptive signatures of *B. taurus* are located in intergenic regions;
329 regulatory changes were also proposed to guide polygenic selection in *Arabidopsis*
330 (He et al., 2016). Analyses of gene–sets associated with infection responses or
331 immunity also found evidence for polygenic selection in humans and primates (Daub
332 et al., 2013, 2017; Svardal et al., 2017). Immunity gene–sets might be exceptional
333 cases, as they are more likely to contain genes subject to very strong selection
334 (Castellano et al., 2019). Further investigations using regulatory information and a
335 broader range of gene–sets could be a promising approach to determine the impact
336 of polygenic selection.

Materials and Methods

Simulating Holstein demography

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

Genome Data Extraction

Whole genome sequencing for 102 Holstein bulls and cows were done by Illumina and BGI short read sequencing in various laboratories. Bulls were selected for sequencing had high genetic representation in the present–day Holstein population. Sequencing of close relatives was avoided. Individuals were born

between approximately 1980 and 2010. DNA was extracted from tissue, blood, or semen samples. DNA was sequenced using either BGI technology or on various Illumina platforms. Sequencing was performed using paired–end sequencing with most animals sequenced with read lengths of 100 basepairs. No raw reads were shorter than 90 basepairs. Read data were processed according to the 1,000 Bull Genomes Project (Daetwyler et al., 2014). Briefly, data were trimmed and quality filtered using Trimmomatic version 0.38 (Bolger et al., 2014). Reads were aligned to the ARS-UCD-1.2 bovine genome assembly (Rosen et al., 2018) (https://sites.ualberta.ca/~stothard/1000_bull_genomes/ARS-UCD1.2_Btau5.0.1Y.fa.gz) with the *B. taurus* Y chromosome assembly from BTau-5.0.1 added. Alignment was performed with bwa version 0.17 (Li and Durbin, 2009). Samtools (Li et al., 2009) was used for sorting and marking of PCR duplicates. Base qualities were recalibrated using Genome Analysis Toolkit (GATK; McKenna et al. [2010]) version 3.8 using a set of known variable sites (Schnabel and Chamberlain, *unpubl*). GVCF files were formed using GATK's Haplotype Caller. Genotypes were called using GATK's GenotypeGVCFs.

Initial filtering

Figure 1 outlines a schematic of the data filtering. We first used *VCFtools* (Danecek et al., 2011) to obtain a baseline list of biallelic sites containing at least one minor allele, and removed indels and sites where the genotype was unknown in any individual. For each autosome, we obtained the mean depth for each remaining site using *VCFtools*' '--site-mean-depth' option. Figure S6 shows the depth distribution for these sites after initial filtering. We fitted a Poisson distribution to these data that had the same mean (9.76) as observed in the dataset. We determined the expected

coverage range based on the 99.5% quantile range of the fitted distributions, which equalled 2 to 20. We subsequently removed sites that had mean coverage outside this range. This filtering retained 6,873,371 of 20,010,175 initial variants (all entries in each autosome VCF file, including indels), which was denoted the 'filtered' dataset.

Finding putatively over–assembled regions

Scaffolds of different genetic segments which each carry highly identical repeated regions can be 'over–assembled', where very similar chromosome regions were anchored to a single location (Chaisson et al., 2015). These over–assembled regions (OARs) manifest themselves in the assembled sequence as having high levels of heterozygosity, sequence similarity, and coverage. If not corrected for, they can be misclassified as selected sites (e.g., subject to partial sweeps or balancing selection). We used a sliding window method to identify putative OARs in the dataset. For each chromosome, in each window we calculated (i) the number of sites where the reference allele has frequency between 49% – 51%, (ii) the mean heterozygosity for each SNP (defined as the number of heterozygotes among the 102 individuals), and (iii) the mean summed allele depth. We used overlapping windows, each of size 500 SNPs with a step size of 10 SNPs. We first analysed all chromosomes to determine the distribution of values produced per window. We then re–ran the analyses, classifying windows as OARs if values for all three statistics belonged to the top 99.5% of their respective distributions, merging overlapping windows. We subtracted 1 from the start position of each region so that the leftmost boundary would also be excluded (if using '--bed-exclude' in *VCFtools*). Figure S7 shows an example where a region at the beginning of chromosome 1 was identified as an OAR. Overall, 5 OARs comprising 5,880 SNPs were identified (Table S7),

412 which were subsequently masked from the rest of the pipeline.

413

414 ## *Calculating inbreeding coefficients*

415 Inbreeding coefficients were estimated using the '--het' option of *VCFtools*,
416 which reports *F*–statistics for each chromosome per individual. Individual *F*–values
417 (Figure S5) were calculated by taking the mean over all chromosomes, weighted by
418 the chromosome size.

419

420 ## *Obtaining SDS analysis inputs*

421 ***Test SNPs:*** Focal SNPs were those with an alternate allele frequency
422 between 5% and 95%, and where each genotype was observed at least once
423 amongst all samples. 3,602,500 SNPs were retained for testing.

424 ***Singletons:*** Raw singleton data was extracted from the filtered Holstein
425 dataset using *VCFtools*' '--singleton' option. This option identified both true singletons
426 and private doubletons (i.e., where an allele is unique to an individual but present as
427 a homozygote). Only true singletons were retained for analyses. To test whether a
428 singleton had the same coverage as the non–singleton allele, we extracted the
429 sequence depth for both alleles and retained sites satisfying the following criteria.
430 First, the total allele depth was between 2 and 20 inclusive. Second, either (a) if the
431 summed depth over both alleles exceeded 5, then the binomial probability of the
432 observed allele depth exceeded 0.1; or (b) a stricter manual cut–off was applied if the
433 total allele depth equalled 5 or less. Table S8 outlines the cut–off values used;
434 554,402 of 765,822 singletons were subsequently retained.

435 ***Other parameters:*** The SDS method requires a 'singleton observability'
436 probability, to predict how likely it is that a singleton will be detected by genome

437 sequencing. Following Field et al. (2016) we used the mean depth per individual,
438 obtained using the '--depth' option in *VCFtools.* It was also necessary to state the
439 genetic boundaries between which analyses were carried out; we used a starting
440 point of 1 and end points equal to the reported size of each autosome in *B. taurus*, as
441 obtained from the ARS–UCD 1.2 genome assembly
442 (https://www.ncbi.nlm.nih.gov/genome/?term=txid9913[orgn]).
443 Raw SDS values were calculated by fitting a gamma distribution to observed
444 singleton distances, and comparing it to the expected distribution for the neutral
445 case. We used the scripts provided by Field et al. (2016)
446 (https://github.com/yairf/SDS) to generate the expected shape values for the gamma
447 distribution for both the high and low N_0 models. Finally, we used value of 10^{-7} to
448 initiate the search for a maximum value in likelihood space.
449

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

451 Out of 3,602,500 input SNPs, we retained and assigned scores to 1,983,571
452 of these. SDS scores were not assigned to a SNP if more than 5% of individuals did
453 not harbour any singletons upstream or downstream of the SNP. This cut–off tended
454 to exclude SNPs in telomeric regions. Furthermore, SDS scores were not calculated
455 for chromosome 25 as an insufficient number of singletons were present across all
456 individuals after data filtering.
457 Raw SDS scores were standardized using 18 bins, based on alternate allele
458 frequencies at the scored SNP (i.e., from 5% to 10%, from 10% to 15%, etc.). SDS
459 scores were normalised by subtracting the bin mean score from individual measures,
460 and dividing by the bin standard deviation. We subsequently took the absolute value
461 of standardized scores, which are referred to asSDS statistics. *P*–values for each

462 asSDS value were obtained from a half–normal distribution.

463 Statistical analyses were carried out in R (R Core Team, 2019). The false–
464 discovery rate (FDR) of each SNP was calculated using the 'qvalue' package (Storey
465 et al., 2019); we highlighted SNPs with an FDR of less than 0.05. Significance was
466 determined using a Bonferroni corrected *P*–value cut–off of 0.05/(1,983,571) ≈ 2.5 x
467 10^{-8}.

468

469 ## *Data sources*

470 A GTF gene annotation file for the ARS–UCD 1.2 assembly was downloaded
471 from Ensembl (available from https://www.ensembl.org/Bos_taurus/Info/Index).
472 Bedtools v2.29.0 (Quinlan and Hall, 2010) was used to obtain genetic annotations
473 10kb up– and downstream of each Bonferroni–significant SNP (overlapping windows
474 were merged).

475 A list of milk protein genes was obtained from Lemay et al. (2009), which was
476 based on proteins identified in milk in two comprehensive proteomic studies
477 (Reinhardt and Lippolis, 2006; Smolenski et al., 2007). The position of these genes in
478 the ARS–UCD 1.0.25 assembly were then determined by either locating the gene
479 name in the *B. taurus* database, or using BLAST to locate the human homologue in
480 the cattle genome. Out of 198 initial genes, new locations were obtained for 191 of
481 them. 180 were subsequently retained after removing those located on chromosome
482 25, the X chromosome, and those with unknown chromosome location.

483 Stature QTLs were obtained from Bouwman et al. (2018), which identified 164
484 QTLs in several *B. taurus* breeds, including 7 Holstein populations from different
485 countries. We initially extracted 114 QTLs for which an effect was inferred from at
486 least 5 of 7 Holstein populations. Positions were given relative to the UMD 3.1

assembly; we subsequently extracted sequence 100bp up– and downstream of the position and remapped to ARS–UCD 1.0.25. 106 QTLs were re–aligned without gaps; of the remaining 8, 4 were located close to rearrangements and discarded, while the remaining 4 were kept. After removing those on chromosome 25, 107 QTLs were retained for downstream analysis.

We analysed this full QTL set and a subset where effect sizes were reported in 6 of 7 Holstein breeds (containing 78 QTLs). Some QTLs lie at the beginning or the end of chromosomes, where asSDS scores were not available. These QTLs were not considered further; for the remaining QTLs, we identified the SNP nearest to it and assigned the asSDS value at that site to the QTL. Overall, asSDS values were assigned to 58 QTLs with effect sizes in at least 5 of 7 Holstein populations, and 42 QTLs with effect sizes in at least 6 of 7 Holstein populations.

Statistical analyses

To determine which factors explain variation in asSDS scores, we applied a generalized linear model as implemented using the glm() function in R, with a Gamma link family and inverse link function. We compared models that either included or excluded the trait of interest, of the following form:

H_0: asSDS ~ Chr + AAF

H_1: asSDS ~ Chr + AAF + Trait

'Chr' is the effect of the chromosome where the SNP resided; AAF denotes the effect of the bin of alternate allele frequency that was used to standardize raw SDS data (e.g., bin 1 denoted those with frequency between 5 and 10%). For the

512 milk protein analysis, a 'Trait' value of 1 indicated that the SNP lies within a milk
513 protein gene; for the stature QTL analyses, 1 indicates a SNP that is closest to a
514 confirmed stature QTL. Otherwise 'Trait' was set to zero. All variables are categorical.
515 Significance of the 'Trait' factor was determined by comparing the deviance of
516 models H_0 and H_1 using a likelihood ratio test (LRT). *P*–values were calculated
517 assuming that the LRT statistic was χ_{12} under H_0 .
518 To implement the permutation test for milk–protein genes, a random set of
519 non–overlapping regions were designated as associated with milk–proteins. The
520 number of regions defined equalled the actual number of milk–protein genes, and
521 each region had the same length as one of the milk–protein genes. For stature QTLs,
522 random positions were defined as being associated with stature; the number of
523 positions equalled the number of QTLs that were initially analysed, accounting for the
524 number of breeds in which a QTL effect size was reported in. The LRT was applied to
525 GLM results applied to the randomised datasets, and the deviance (a measure of
526 how much additional variation is explained by the 'Trait' term in H_1) was noted. The
527 process was repeated 1,000 times. *P*–values were calculated from the proportion of
528 deviance values that exceed the observed deviance in the actual dataset.
529
530 ***Acknowledgements.*** We would like to thank Simon Boitard for sharing his
531 results on *B. taurus* demographic inference. MH is supported by a NERC
532 Independent Research Fellowship (NE/R015686/1). NAP, BG and TB are funded by
533 a synergistic research grant from the Faculty of Science and Technology, Aarhus
534 University, Denmark. MH and TB also acknowledge financial support from the
535 European Research Council under the European Union's Seventh Framework
536 Program (FP7/20072013, ERC Grant 311341).

References

Barton NH, Etheridge AM, Véber A. 2017. The infinitesimal model: Definition, derivation, and implications. *Theor Popul Biol* **118**:50–73. doi:10.1016/j.tpb.2017.06.001

Barton NH, Hermisson J, Nordborg M. 2019. Why structure matters. *eLife* **8**:e45380. doi: 10.7554/eLife.45380

Barton NH, Keightley PD. 2002. Understanding quantitative genetic variation. *Nat Rev Genet* **3**:11–21. doi: 10.1038/nrg700

Berg JJ, Coop G. 2014. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet* **10**:e1004412. doi: 10.1371/journal.pgen.1004412

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, Coop G. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**:e39725. doi: 10.7554/eLife.39725

Berg JJ., Zhang X, Coop G. 2019. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. *bioRxiv* 167551. doi: 10.1101/167551

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. doi: 10.1534/genetics.115.181594

Boitard S, Rocha D. 2013. Detection of signatures of selective sweeps in the Blonde d'Aquitaine cattle breed. *Anim Genet* **44**:579–583. doi: 10.1111/age.12042

Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F. 2016b. Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS Genet* **12**:e1005877. doi: 10.1371/journal.pgen.1005877

562 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
563 sequence data. *Bioinformatics* **30**:2114–2120.
564 doi:10.1093/bioinformatics/btu170
565 Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel
566 FS, Sahana G, Govignon-Gion A, Boitard S, Dolezal M, Pausch H, Brøndum
567 RF, Bowman PJ, Thomsen B, Guldbrandtsen B, Lund MS, Servin B, Garrick
568 DJ, Reecy J, Vilkki J, Bagnato A, Wang M, Hoff JL, Schnabel RD, Taylor JF,
569 Vinkhuyzen AAE, Panitz F, Bendixen C, Holm L-E, Gredler B, Hozé C,
570 Boussaha M, Sanchez M-P, Rocha D, Capitan A, Tribout T, Barbat A, Croiseau
571 P, Drögemüller C, Jagannathan V, Vander Jagt C, Crowley JJ, Bieber A,
572 Purfield DC, Berry DP, Emmerling R, Götz K-U, Frischknecht M, Russ I,
573 Sölkner J, Van Tassell CP, Fries R, Stothard P, Veerkamp RF, Boichard D,
574 Goddard ME, Hayes BJ. 2018. Meta-analysis of genome-wide association
575 studies for cattle stature identifies common genes that regulate body size in
576 mammals. *Nat Genet* **50**:362–367. doi: 10.1038/s41588-018-0056-5
577 Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From
578 Polygenic to Omnigenic. *Cell* **169**:1177–1186. doi: 10.1016/j.cell.2017.05.038
579 Buitenhuis B, Poulsen NA, Gebreyesus G, Larsen LB. 2016. Estimation of genetic
580 parameters and detection of chromosomal regions affecting the major milk
581 proteins and their post translational modifications in Danish Holstein and
582 Danish Jersey cattle. *BMC Genet* **17**:114. doi:10.1186/s12863-016-0421-2
583 Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D,
584 Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean
585 G, Leslie S, Allen N, Donnelly P, Marchini J. 2018. The UK Biobank resource
586 with deep phenotyping and genomic data. *Nature* **562**:203–209. doi:

587 10.1101/250191

588 Castellano D, Uricchio LH, Munch K, Enard D. 2019. Viruses rule over adaptation in
589 conserved human proteins. *bioRxiv* 555060. doi: 10.1101/555060

590 Chaisson MJP, Wilson RK, Eichler EE. 2015. Genetic variation and the *de novo*
591 assembly of human genomes. *Nat Rev Genet* **16**:627–640. doi:
592 10.1038/nrg3933

593 Charlesworth B, Edwards AWF. 2018. A century of variance. *Significance* **15**:20–25.
594 doi: 10.1111/j.1740-9713.2018.01170.x

595 Chevin L-M. 2019. Selective Sweep at a QTL in a Randomly Fluctuating
596 Environment. *Genetics* **213**:987–1005. doi:10.1534/genetics.119.302680 doi:
597 10.1534/genetics.119.302680

598 Chevin L-M, Hospital F. 2008. Selective Sweep at a Quantitative Trait Locus in the
599 Presence of Background Genetic Variation. *Genetics* **180**:1645–1660. doi:
600 10.1534/genetics.108.093351

601 Coop G, Witonsky D, Di Rienzo A, Pritchard JK. 2010. Using Environmental
602 Correlations to Identify Loci Underlying Local Adaptation. *Genetics* **185**:1411–
603 1423. doi: 10.1534/genetics.110.114819

604 Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF,
605 Liao X, Djari A, Rodriguez SC, Grohs C, Esquerré D, Bouchez O, Rossignol
606 M-N, Klopp C, Rocha D, Fritz S, Eggen A, Bowman PJ, Coote D, Chamberlain
607 AJ, Anderson C, VanTassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B,
608 Lund MS, Veerkamp RF, Boichard DA, Fries R, Hayes BJ. 2014. Whole-
609 genome sequencing of 234 bulls facilitates mapping of monogenic and
610 complex traits in cattle. *Nat Genet* **46**:858–865. doi: 10.1038/ng.3034

611 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,

612 Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call
613 format and VCFtools. *Bioinformatics* **27**:2156–2158. doi:
614 10.1093/bioinformatics/btr330
615 Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M,
616 Excoffier L. 2013. Evidence for Polygenic Adaptation to Pathogens in the
617 Human Genome. *Mol Biol Evol* **30**:1544–1558. doi: 10.1093/molbev/mst080
618 Daub JT, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M. 2017. Detection of
619 Pathways Affected by Positive Selection in Primate Lineages Ancestral to
620 Humans. *Mol Biol Evol* **34**:1391–1402. doi: 10.1093/molbev/msx083
621 de Vladar HP, Barton N. 2014. Stability and Response of Polygenic Traits to
622 Stabilizing Selection and Mutation. *Genetics* **197**:749–767. doi:
623 10.1534/genetics.113.159111
624 Edge MD, Coop G. 2019. Reconstructing the History of Polygenic Scores Using
625 Coalescent Trees. *Genetics* **211**:235–262. doi: 10.1534/genetics.118.301687
626 Ehrlich MA, Wagner DN, Oleksiak MF, Crawford DL. 2020. Rapid polygenic selection
627 generates fine spatial structure among ecological niches in a well-mixed
628 population. *bioRxiv* 2020.03.26.009787. doi:10.1101/2020.03.26.009787
629 Exposito-Alonso M, 500 Genomes Field Experiment Team, Burbano HA, Bossdorf O,
630 Nielsen R, Weigel D. 2019. Natural selection on the *Arabidopsis thaliana*
631 genome in present and future climates. *Nature* **573**:126–129.
632 doi:10.1038/s41586-019-1520-9
633 Exposito-Alonso M, Vasseur F, Ding W, Wang G, Burbano HA, Weigel D. 2018.
634 Genomic basis and evolutionary potential for extreme drought adaptation in
635 *Arabidopsis thaliana*. *Nat Ecol Evol* **2**:352–358. doi: 10.1038/s41559-017-
636 0423-0

637 Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G,
638 Froguel P, McCarthy MI, Pritchard JK. 2016. Detection of human adaptation
639 during the past 2000 years. *Science* **354**:760–764. doi:
640 10.1126/science.aag0776
641 Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian
642 inheritance. *Trans R Soc Edinb* **52**:399–433.
643 Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, Schoech A, Bulik-
644 Sullivan B, Neale BM, Gusev A, Price AL. 2017. Linkage disequilibrium–
645 dependent architecture of human complex traits shows action of negative
646 selection. *Nat Genet* **49**:1421–1427. doi: 10.1038/ng.3954
647 Georges M, Charlier C, Hayes B. 2018. Harnessing genomic information for livestock
648 improvement. *Nat Rev Genet*. doi: 10.1038/s41576-018-0082-2
649 Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST. 1998.
650 Genetic traces of ancient demography. *Proc Natl Acad Sci* **95**:1961–1967. doi:
651 10.1073/pnas.95.4.1961
652 Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. 2009. Invited review:
653 Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci*
654 **92**:433–443. doi: 10.3168/jds.2008-1646
655 Hayward LK, Sella G. 2019. Polygenic adaptation after a sudden change in
656 environment. *bioRxiv* 792952. doi: 10.1101/792952
657 He F, Arce AL, Schmitz G, Koornneef M, Novikova P, Beyer A, de Meaux J. 2016.
658 The Footprint of Polygenic Adaptation on Stress-Responsive Cis-Regulatory
659 Divergence in the *Arabidopsis* Genus. *Mol Biol Evol* **33**:2088–2101. doi:
660 10.1093/molbev/msw096
661 Jain K, Stephan W. 2017a. Modes of Rapid Polygenic Adaptation. *Mol Biol Evol*

662 **34**:3169–3175. doi: 10.1093/molbev/msx240

663 Jain K, Stephan W. 2017b. Rapid Adaptation of a Polygenic Trait After a Sudden
664 Environmental Shift. *Genetics* **206**:389–406. doi:
665 10.1534/genetics.116.196972

666 Jain K, Stephan W. 2015. Response of Polygenic Traits Under Stabilizing Selection
667 and Mutation When Loci Have Unequal Effects. *G3* **5**:1065–1074. doi:
668 10.1534/g3.115.017970

669 Jiménez-Mena B, Tataru P, Brøndum RF, Sahana G, Guldbrandtsen B, Bataillon T.
670 2016. One size fits all? Direct evidence for the heterogeneity of genetic drift
671 throughout the genome. *Biol Lett* **12**. doi: 10.1098/rsbl.2016.0426

672 John S, Stephan W. 2020. Important role of genetic drift in rapid polygenic
673 adaptation. *Ecol Evol* **10**:1278–1287. doi:10.1002/ece3.5981

674 Josephs EB, Berg JJ, Ross-Ibarra J, Coop G. 2019. Detecting Adaptive
675 Differentiation in Structured Populations with Genomic Data and Common
676 Gardens. *Genetics* **211**:989–1004. doi: 10.1534/genetics.118.301786

677 Keightley PD, Jackson BC. 2018. Inferring the Probability of the Derived *vs.* the
678 Ancestral Allelic State at a Polymorphic Site. *Genetics* **209**:897–906.
679 doi:10.1534/genetics.118.301120

680 Kelleher J, Etheridge AM, McVean G. 2016. Efficient Coalescent Simulation and
681 Genealogical Analysis for Large Sample Sizes. *PLoS Comput Biol*
682 **12**:e1004842. doi: 10.1371/journal.pcbi.1004842

683 Lande R. 1983. The response to selection on major and minor mutations affecting a
684 metrical trait. *Heredity* **50**:47–65. doi: 10.1038/hdy.1983.6

685 Le Corre V, Kremer A. 2012. The genetic differentiation at quantitative trait loci under
686 local adaptation. *Mol Ecol* **21**:1548–1566. doi: 10.1111/j.1365-

687 294X.2012.05479.x

688 Lemay DG, Lynn DJ, Martin WF, Neville MC, Casey TM, Rincon G, Kriventseva EV,
689 Barris WC, Hinrichs AS, Molenaar AJ, Pollard KS, Maqbool NJ, Singh K,
690 Murney R, Zdobnov EM, Tellam RL, Medrano JF, German JB, Rijnkels M.
691 2009. The bovine lactation genome: insights into the evolution of mammalian
692 milk. *Genome Biol* **10**:R43. doi: 10.1186/gb-2009-10-4-r43

693 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler
694 transform. *Bioinformatics* **25**:1754–1760. doi: 10.1093/bioinformatics/btp324

695 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
696 Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The
697 Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–
698 2079. doi: 10.1093/bioinformatics/btp352

699 Liu X, Li YI, Pritchard JK. 2019. Trans Effects on Gene Expression Can Drive
700 Omnigenic Inheritance. *Cell* **177**:1022–1034.e6. doi:
701 10.1016/j.cell.2019.04.014

702 MacEachern S, Hayes B, McEwan J, Goddard M. 2009. An examination of positive
703 selection and changing effective population size in Angus and Holstein cattle
704 populations (*Bos taurus*) using a high density SNP genotyping platform and
705 the contribution of ancient polymorphism to genomic diversity in Domestic
706 cattle. *BMC Genomics* **10**:181. doi: 10.1186/1471-2164-10-181

707 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella
708 K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis
709 Toolkit: A MapReduce framework for analyzing next-generation DNA
710 sequencing data. *Genome Res* **20**:1297–1303. doi: 10.1101/gr.107524.110

711 Meuwissen T, Hayes B, Goddard M. 2013. Accelerating Improvement of Livestock

712 with Genomic Selection. *Annu Rev Anim Biosci* **1**:221–237. doi:
713 10.1146/annurev-animal-031412-103705
714 Nelsen TC, Short RE, Urick JJ, Reynolds WL. 1986. Heritabilities and Genetic
715 Correlations of Growth and Reproductive Measurements in Hereford Bulls. *J*
716 *Anim Sci* **63**:409–417. doi: 10.2527/jas1986.632409x
717 Nielsen R. 2005. Molecular Signals of Natural Selection. *Annu Rev Genet* **39**:197–
718 218. doi: 10.1146/annurev.genet.39.073003.112420
719 Nordborg M, Donnelly P. 1997. The Coalescent Process With Selfing. *Genetics*
720 **146**:1185–1195.
721 Northcutt SL, Wilson DE. 1993. Genetic parameter estimates and expected progeny
722 differences for mature size in Angus cattle. *J Anim Sci* **71**:1148–1153. doi:
723 10.2527/1993.7151148x
724 Novembre J, Barton NH. 2018. Tread Lightly Interpreting Polygenic Tests of
725 Selection. *Genetics* **208**:1351–1355. doi: 10.1534/genetics.118.300786
726 Pavlidis P, Metzler D, Stephan W. 2012. Selective Sweeps in Multilocus Models of
727 Quantitative Traits. *Genetics* **192**:225–239. doi: 10.1534/genetics.112.142547
728 Pritchard JK, Di Rienzo A. 2010. Adaptation - not by sweeps alone. *Nat Rev Genet*
729 **11**:665–667. doi: 10.1038/nrg2880
730 Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard
731 Sweeps, Soft Sweeps, and Polygenic Adaptation. *Curr Biol* **20**:R208–R215.
732 doi: 10.1016/j.cub.2009.11.055
733 Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, Nielsen R, Simianer
734 H. 2014. Classic Selective Sweeps Revealed by Massive Sequencing in
735 Cattle. *PLoS Genet* **10**:e1004148. doi: 10.1371/journal.pgen.1004148
736 Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H.

737 2010. A genome-wide scan for signatures of recent selection in Holstein cattle.
738 *Anim Genet*. doi: 10.1111/j.1365-2052.2009.02016.x
739 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing
740 genomic features. *Bioinformatics* **26**:841–842. doi:
741 10.1093/bioinformatics/btq033
742 R Core Team. 2019. R: A Language and Environment for Statistical Computing.
743 Vienna, Austria: R Foundation for Statistical Computing. http://www.R-
744 project.org
745 Racimo F, Berg JJ, Pickrell JK. 2018. Detecting Polygenic Adaptation in Admixture
746 Graphs. *Genetics* **208**:1565–1584. doi: 10.1534/genetics.117.300489
747 Reinhardt TA, Lippolis JD. 2006. Bovine Milk Fat Globule Membrane Proteome. *J*
748 *Dairy Res* **73**:406–416. doi: 10.1017/S0022029906001889
749 Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K,
750 Powell JE, Vinkhuyzen A, Berndt SI, Gustafsson S, Justice AE, Kahali B,
751 Locke AE, Pers TH, Vedantam S, Wood AR, van Rheenen W, Andreassen OA,
752 Gasparini P, Metspalu A, Berg LH van den, Veldink JH, Rivadeneira F, Werge
753 TM, Abecasis GR, Boomsma DI, Chasman DI, de Geus EJC, Frayling TM,
754 Hirschhorn JN, Hottenga JJ, Ingelsson E, Loos RJF, Magnusson PKE, Martin
755 NG, Montgomery GW, North KE, Pedersen NL, Spector TD, Speliotes EK,
756 Goddard ME, Yang J, Visscher PM. 2015. Population genetic differentiation of
757 height and body mass index across Europe. *Nat Genet* **47**: 1357-1362. doi:
758 10.1038/ng.3401
759 Rosen B, Bickhart DM, Schnabel RD, Koren S, Elsik C, Zimin AV, Dreischer C,
760 Schultheiss S, Hall R, Schroeder S, Van Tassell CP, Smith T, Medrano JF.
761 2018. Modernizing the Bovine Reference Genome Assembly. *Proc World*

Congr Genet Appl Livest Prod 11–16.

Sanjak JS, Sidorenko J, Robinson MR, Thornton KR, Visscher PM. 2018. Evidence of directional and stabilizing selection in contemporary humans. *Proc Natl Acad Sci USA* **115**:151–156. doi: 10.1073/pnas.1707227114

Savolainen O, Lascoux M, Merila J. 2013. Ecological genomics of local adaptation. *Nat Rev Genet* **14**:807–820. doi: 10.1038/nrg3522

Schoech AP, Jordan DM, Loh P-R, Gazal S, O'Connor LJ, Balick DJ, Palamara PF, Finucane HK, Sunyaev SR, Price AL. 2019. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat Commun* **10**:790. doi: 10.1038/s41467-019-08424-6

Sella G, Barton NH. 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu Rev Genomics Hum Genet* **20**:461–493. doi: 10.1146/annurev-genom-083115-022316

Smolenski G, Haines S, Kwan FY-S, Bond J, Farr V, Davis SR, Stelwagen K, Wheeler TT. 2007. Characterisation of Host Defence Proteins in Milk Using a Proteomic Approach. *J Proteome Res* **6**:207–215. doi: 10.1021/pr0603405

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N, Neale B, Mathieson I, Reich D, Sunyaev SR. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**:e39702. doi: 10.1126/science.aah5238

Sørensen AC, Sørensen MK, Berg P. 2005. Inbreeding in Danish Dairy Cattle Breeds. *J Dairy Sci* **88**:1865–1872. doi: 10.3168/jds.S0022-0302(05)72861-7

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet* **51**:1321–1329. doi:

787 10.1038/s41588-019-0484-x

788 Stephan W. 2019. Selective Sweeps. *Genetics* **211**:5–13. doi:

789 10.1534/genetics.118.301319

790 Stephan W. 2016. Signatures of positive selection: from selective sweeps at

791 individual loci to subtle allele frequency changes in polygenic adaptation. *Mol*

792 *Ecol* **25**:79–88. doi: 10.1111/mec.13288

793 Stetter MG, Thornton K, Ross-Ibarra J. 2018. Genetic architecture and selective

794 sweeps after polygenic adaptation to distant trait optima. *PLoS Genet*

795 **14**:e1007794. doi: 10.1371/journal.pgen.1007794

796 Storey JD, Bass AJ, Dabney A, Robinson D. 2019. qvalue: Q-value estimation for

797 false discovery rate control. http://github.com/StoreyLab/qvalue

798 Svardal H, Jasinska AJ, Apetrei C, Coppola G, Huang Y, Schmitt CA, Jacquelin B,

799 Ramensky V, Müller-Trutwin M, Antonio M, Weinstock G, Grobler JP, Dewar K,

800 Wilson RK, Turner TR, Warren WC, Freimer NB, Nordborg M. 2017. Ancient

801 hybridization and strong adaptation to viruses across African vervet monkey

802 populations. *Nat Genet* **49**:1705–1713. doi: 10.1038/ng.3980

803 Thornton KR. 2019. Polygenic Adaptation to an Environmental Shift: Temporal

804 Dynamics of Variation Under Gaussian Stabilizing Selection and Additive

805 Effects on a Single Trait. *Genetics* **213**:1513–1530. doi:

806 10.1534/genetics.119.302662

807 Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Genetic

808 Investigation of ANthropometric Traits (GIANT) Consortium, Hirschhorn JN.

809 2012. Evidence of widespread selection on standing variation in Europe at

810 height-associated SNPs. *Nat Genet* **44**:1015. doi: 10.1038/ng.2368

811 Uricchio LH, Kitano HC, Gusev A, Zaitlen NA. 2019. An evolutionary compass for

812 detecting signals of polygenic selection and mutational bias. *Evol Lett* **3**:69–
813 79. doi: 10.1002/evl3.97
814 Visscher PM, Goddard ME. 2019. From R.A. Fisher's 1918 Paper to GWAS a
815 Century Later. *Genetics* **211**:1125–1130. doi: 10.1534/genetics.118.301594
816 Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting Natural Selection in Genomic
817 Data. *Annu Rev Genet* **47**:97–120. doi: 10.1146/annurev-genet-111212-
818 133526
819 Whitlock MC. 2008. Evolutionary inference from Q_{ST}. *Mol Ecol* **17**:1885–1896. doi:
820 10.1111/j.1365-294X.2008.03712.x
821 Wieters B, Steige KA, He F, Koch EM, Ramos-Onsins SE, Gu H, Guo Y-L, Sunyaev
822 S, de Meaux J. 2020. Polygenic adaptation of rosette growth variation in
823 Arabidopsis thaliana populations. *bioRxiv* 2020.03.31.018341. doi:
824 10.1101/2020.03.31.018341
825 Wollstein A, Stephan W. 2014. Adaptive Fixation in Two-Locus Models of Stabilizing
826 Selection and Genetic Drift. *Genetics* **198**:685–697. doi:
827 10.1534/genetics.114.168567
828 Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. 2019. Complex Trait
829 Prediction from Genome Data: Contrasting EBV in Livestock to PRS in
830 Humans. *Genetics* **211**:1131–1141. doi: 10.1534/genetics.119.301859
831 Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. 2018. Common Disease Is
832 More Complex Than Implied by the Core Gene Omnigenic Model. *Cell*
833 **173**:1573–1580. doi: 10.1016/j.cell.2018.05.051
834 Wright S. 1951. The genetical structure of populations. *Ann Eugen* **15**:323–354.
835 Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, Robinson MR, Perry
836 JRB, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Study TLC, Esko T,

837 Milani L, Magi R, Metspalu A, Hamsten A, Magnusson PKE, Pedersen NL,
838 Ingelsson E, Soranzo N, Keller MC, Wray NR, Goddard ME, Visscher PM.
839 2015. Genetic variance estimation with imputed variants finds negligible
840 missing heritability for human height and body mass index. *Nat Genet*. **47**:
841 1114-1120. doi: 10.1038/ng.3390
842 Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, Nurkowski
843 KA, Smets P, Wang T, Gray LK, Liepe KJ, Hamann A, Holliday JA, Whitlock
844 MC, Rieseberg LH, Aitken SN. 2016. Convergent local adaptation to climate in
845 distantly related conifers. *Science* **353**:1431-1433. doi:
846 10.1126/science.aaf7812
847 Zan Y, Carlborg Ö. 2018. A Polygenic Genetic Architecture of Flowering Time in the
848 Worldwide Arabidopsis thaliana Population. *Mol Biol Evol* **36**:141–154. doi:
849 10.1093/molbev/msy203
850 Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin:
851 Origins, diffusion, and impact. *Proc Natl Acad Sci USA* **105**:11597–11604. doi:
852 10.1073/pnas.0801317105
853 Zeng J, de Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, Yap CX, Xue
854 A, Sidorenko J, McRae AF, Powell JE, Montgomery GW, Metspalu A, Esko T,
855 Gibson G, Wray NR, Visscher PM, Yang J. 2018. Signatures of negative
856 selection in the genetic architecture of human complex traits. *Nat Genet*
857 **50**:746–753. doi: 10.1038/s41588-018-0101-4
858 Zhao F, McParland S, Kearney F, Du L, Berry DP. 2015. Detection of selection
859 signatures in dairy and beef cattle using high-density genomic information.
860 *Genet Sel Evol* **47**:49. doi: 10.1186/s12711-015-0127-3
861 Zhou X, Carbonetto P, Stephens M. 2013. Polygenic Modeling with Bayesian Sparse

862 Linear Mixed Models. *PLoS Genet* **9**:e1003264. doi:

863 10.1371/journal.pgen.1003264

864

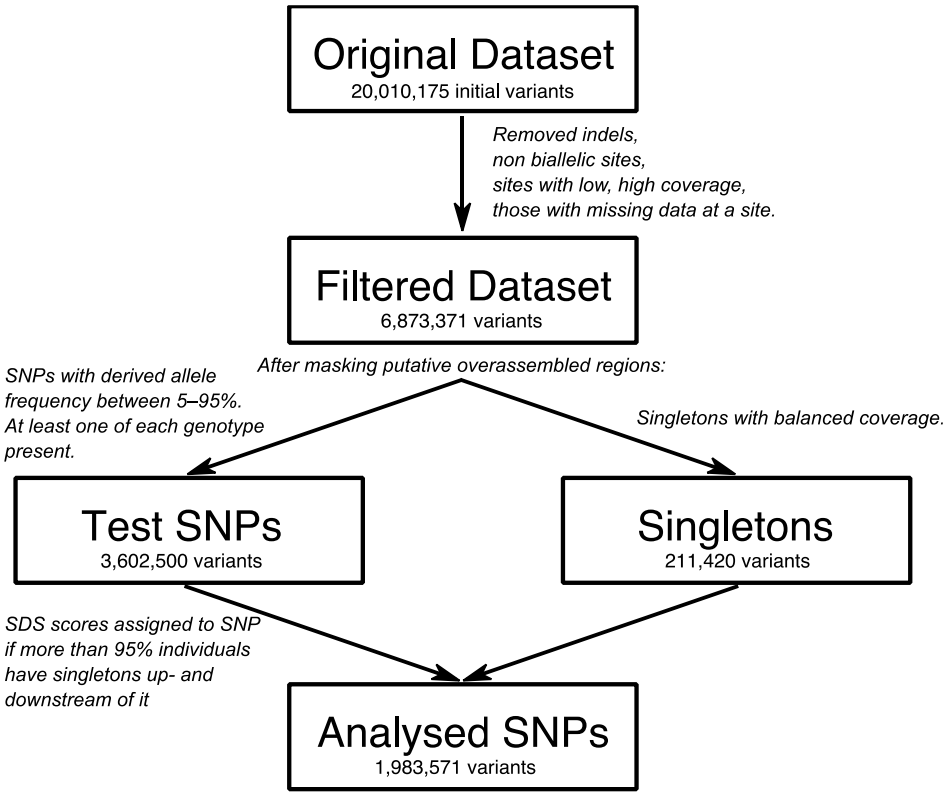
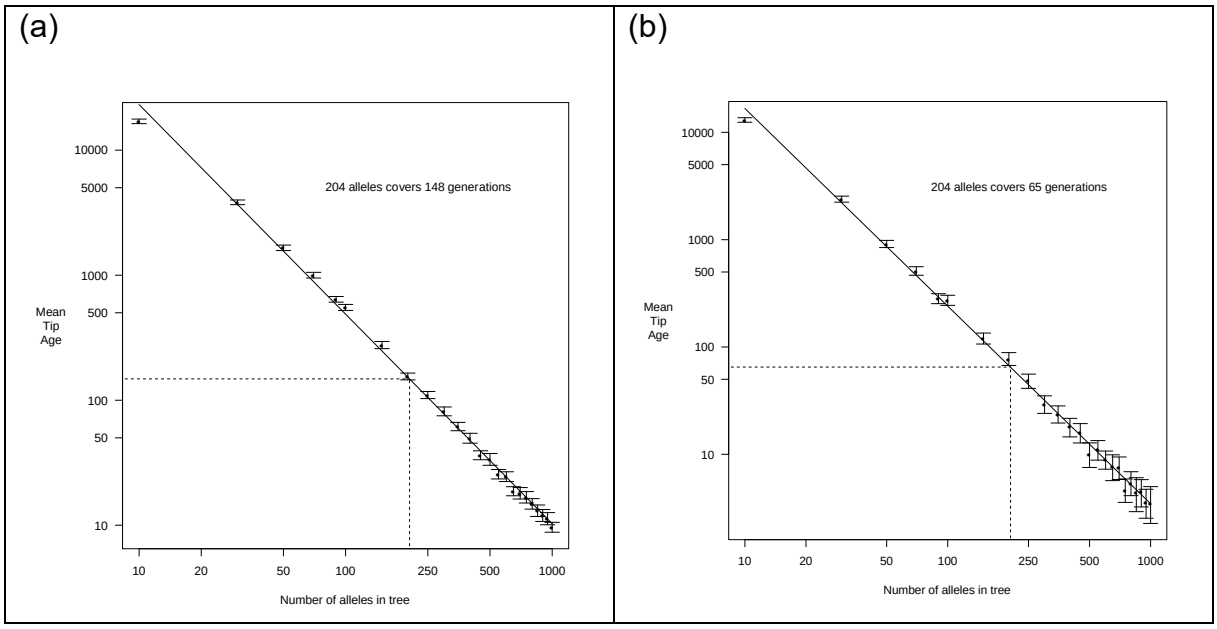
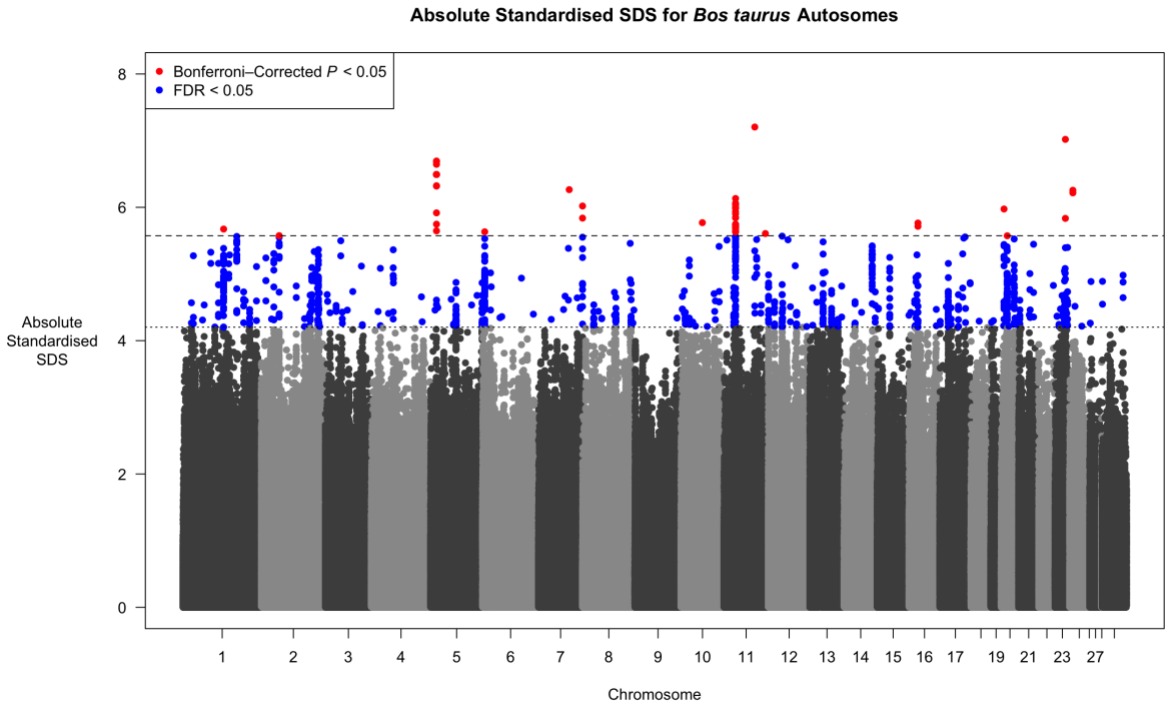


Figure 1: Schematic of data filtering.

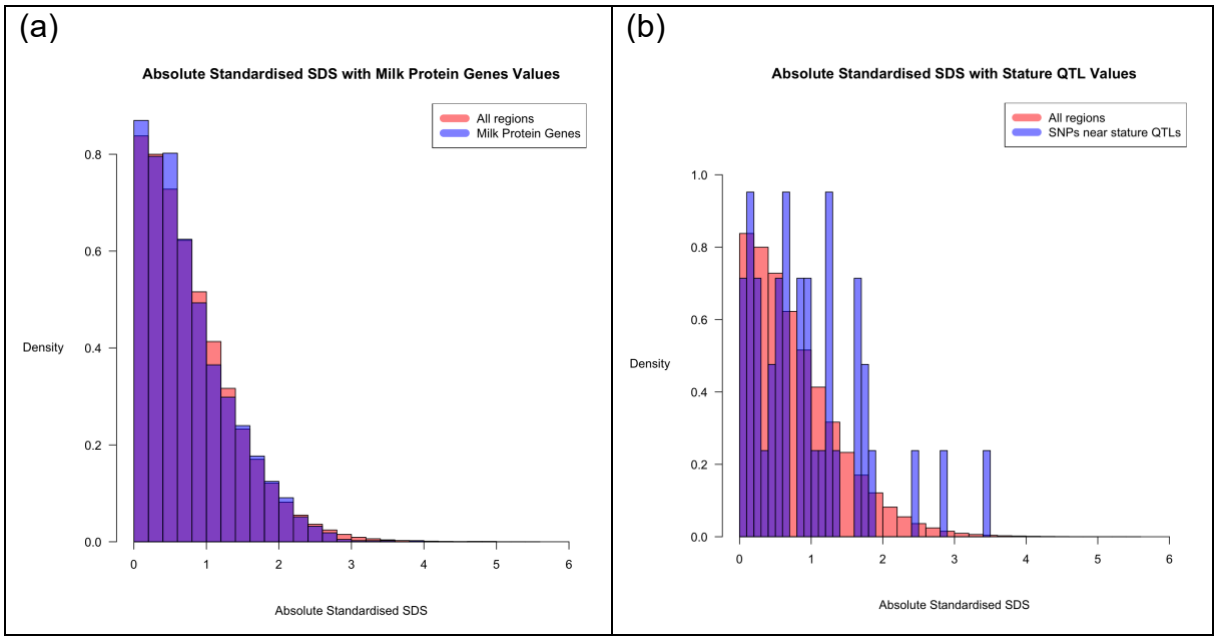


867 Figure 2: Simulated mean tip age for *B. taurus*, as a function of the number of allele
868 samples. Simulations assumed either (a) demography as inferred by Boitard et al.
869 (2016b) (the 'High N_0' model), or (b) the same but with a smaller present–day N_e of
870 49 (the 'Low N_0' model). Points are the mean values; bars show 95% confidence
871 intervals. The solid line is the best linear fit to the log of both values; dotted lines
872 show the predicted tip age for 204 alleles.



873
874

875 Figure 3: asSDS scores across *B. taurus* autosomes, as a function of the
876 chromosome. Alternating black and grey points show (non–significant) values from
877 different chromosomes. Blue points are SNPs with *FDR* < 0.05, with the cutoff
878 denoted by a horizontal dotted line. Red points are SNPs with Bonferroni–corrected
879 *P*–value < 0.05 (actual *P*–value < ~2.5e–8), with the cutoff denoted by a horizontal
880 dashed line. Figure S1 shows results for the Low N_0 model.



881 Figure 4: Histograms of asSDS, showing the background distribution over all SNPs
882 (red), compared to (a) asSDS in milk–protein genes, or (b) asSDS of the nearest
883 SNPs to stature QTLs. In (b) QTLs were obtained if effect sizes were reported in at
884 least 6 of 7 Holstein populations (as measured in Bouwman et al. (2018)). Figure S2
885 shows the distribution if using QTLs obtained with effect sizes reported in at least 5 of
886 7 Holstein populations; Figure S3 shows results for the low N_0 model.

Chromosome	Gene Name	Start Position	End Position	Gene Biotype	High, Low N_0
1	PPM1L	106405113	106727070	Protein Coding	Low
2	U6	38379710	38379816	SnRNA	Low
2	ICA1L	91143446	91177884	Protein Coding	Low
2	ADAM23	94711218	94906499	Protein Coding	Low
2	PTH2R	96667717	96752328	Protein Coding	Low
5	TMCC3	24306913	24595494	Protein Coding	High, Low
5	CEP83	24070404	24345243	Protein Coding	High, Low
6	MANBA	22062326	22189956	Protein Coding	High
7	(Unnamed)	87293323	87297625	Protein Coding	Low
8	ROR2	85905346	86141520	Protein Coding	Low
8	(Unnamed)	85959505	86086599	Protein Coding	Low
10	TRIM9	43826973	43944784	Protein Coding	High
11	NRXN1	32278324	32766620	Protein Coding	High, Low
14	GRHL2	62721044	62888891	Protein Coding	Low
17	GALNT9	44853887	44968139	Protein Coding	Low
17	RIMBP2	46406767	46715519	Protein Coding	Low
20	GHR	31868624	32178311	Protein Coding	Low
20	FBXO4	32589453	32602498	Protein Coding	Low
20	C20H5orf51	32612381	32634378	Protein Coding	Low
23	(Unnamed)	29291787	29292713	Protein Coding	High, Low
23	OR12D2	29305933	29309785	Protein Coding	High, Low
24	GAREM1	24694637	24927333	Protein Coding	High, Low

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