Powerful detection of polygenic selection and environmental adaptation in US beef cattle

Troy N. Rowan, Harly J. Durbin, Christopher M. Seabury, Robert D. Schnabel, Jared E. Decker*

Summary

Selection can rapidly drive evolution, especially in stressful environments. We show that novel dependent variables in mixed models with large data sets identify loci responding to polygenic selection and local adaptation. Many techniques identify mutations rapidly selected to high frequencies, based on pronounced signatures in flanking sequences. However, for complex, quantitative traits, selection does not leave these intense sweeps. Instead, in polygenic selection thousands of loci undergo small allele frequency shifts, resulting in large phenotypic changes. Directional selection and local adaptation are actively changing populations; but, identifying loci underlying polygenic or environmental selection has been difficult. However, our methods identify 207 loci responding to artificial directional selection, and hundreds more with evidence of local adaptation; with the identified genes highlighting the nervous system's central role in local adaptation. While advanced technologies increased the rate of directional selection, it has been at the expense of local adaptation, which is especially problematic in changing climates. These selection mapping approaches clarify selective forces and loci in evolutionary, model, and agricultural contexts.

Main

As climate changes, organisms either migrate, rapidly adapt, or perish. The genes and alleles that underlie adaptation have been difficult to identify, except for a handful of large-effect variants that underwent selective sweeps [1]. It is becoming increasingly apparent that for adaptation, hard sweeps are likely to be the exception, rather than the rule [2]. Polygenic selection on complex traits can cause a significant change in the mean phenotype while producing only subtle changes in allele frequencies throughout the genome [3]. Most selection mapping methods require discrete grouping of subpopulations, making the identification of selection within a largely panmictic population difficult. North American *Bos taurus* beef cattle have been exposed to strong artificial and environmental selection for more than 50 years (~10 generations) [4], making them an intriguing model for studying the impacts selection has on genomes over short time periods and across diverse environments.

Though domesticated, beef cattle are exposed to a broad spectrum of unique environments and local selection pressures, as compared to other more intensely managed livestock populations. This suggests that local adaptation and genotype-by-environment interactions play important roles in the expression of complex traits. Herein, we present two methods (**Fig. 1**), the first for detecting complex polygenic selection (Generation Proxy Selection Mapping, GPSM), and the second for identifying local adaptation (environmental GWAS, envGWAS). Both methods use genome-wide linear mixed models (LMM) incorporating novel dependent variables in large temporally and spatially dispersed datasets, while explicitly controlling for family and population structure (**Fig. 1d**). When applied to three US beef cattle populations, we identified numerous genomic regions harboring directional or environmentally selected mutations. Further, using a meta-analysis approach, we identified loci responding to region-specific selection (**Fig. 1e,f**), largely due to the erosion of local adaptation caused by gene flow among ecoregions from the use of artificial insemination sires. This study is the first step in assisting beef cattle producers to identify locally adapted individuals, which will reduce the industry's environmental footprint by increasing efficiency and resilience to stressors. Further, this study provides insight into the biology of adaptation in mammalian species.

Results

Detecting ongoing polygenic selection with Generation Proxy Selection Mapping (GPSM)

Though the first cattle single nucleotide polymorphism (SNP) genotyping assay was developed just over a decade ago[5], numerous influential males who have been deceased for 30 to 40 years have been genotyped from cryopreserved semen (**Supplementary Figure 1, Supplementary Table 1**). These types of data provide a temporal distribution of samples spanning at least ten generations for the numerically largest US beef breeds. This temporal distribution of genotypes allows us to search for allelic associations with generation number, or a proxy such as birth date, to identify loci subjected to directional selection [4]. Using a LMM, we tested for associations between an individual's generation proxy (i.e., birth date) and SNP alleles in three US cattle populations using ~830,000 SNPs. We controlled for the confounding effects of population structure, relatedness, and inbreeding by including a random effect accounting for dependency between samples using a genomic relationship matrix (GRM, **Fig. 1d**). Significant associations with birth date indicate variants undergoing frequency changes that are greater than expected due to drift over the time span represented in a sampled dataset (**Fig. 1a,c**). Our simulations show that GPSM distinguishes between selection and drift under a variety of genetic architectures, selection intensities, effective population sizes, and sampled time periods (**Fig. 1**, **Supplementary Files 1-2**).

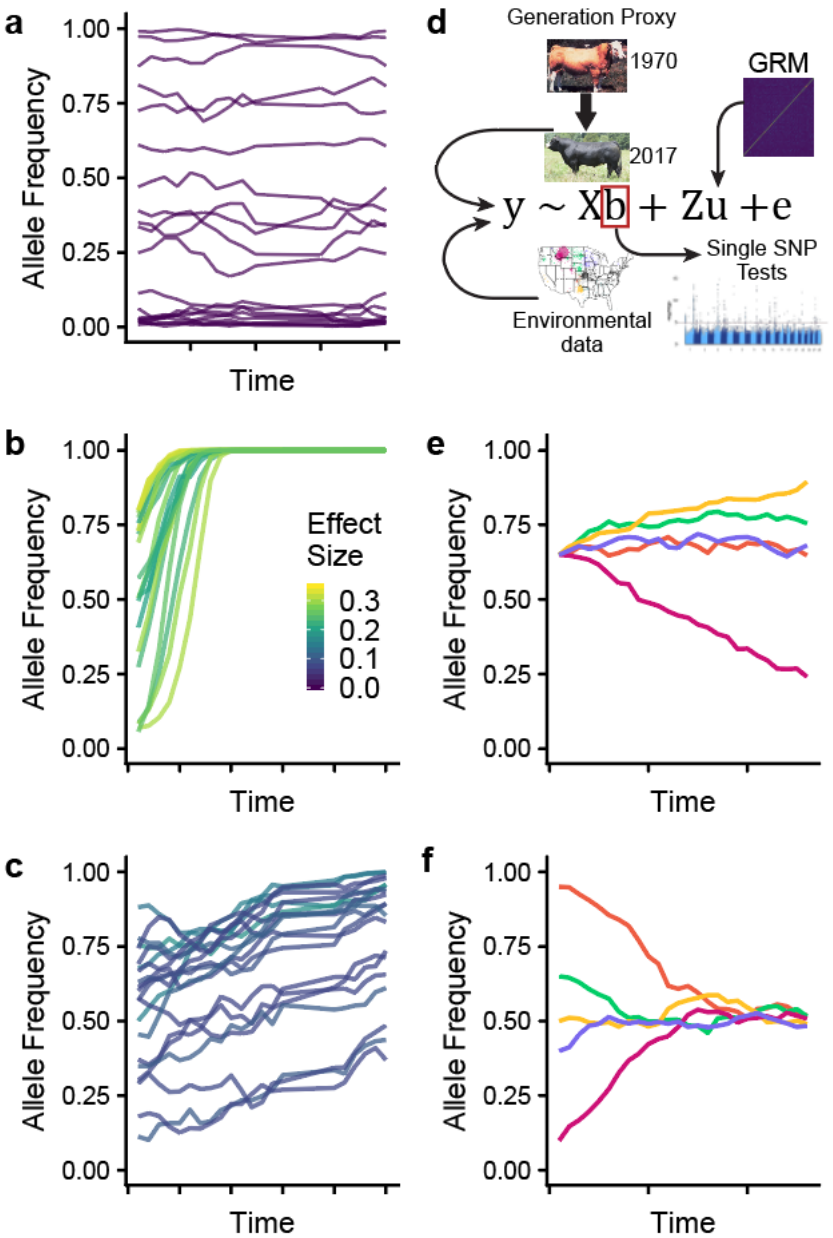

Figure 1. Simulated allele frequency trajectories and model overview. (a-c) Allele frequency trajectories for 20 SNPs colored by relative effect sizes from stochastic selection simulations. (a) Effect size = 0, representing stochastic changes in allele frequency due to genetic drift. (b) Large-effect alleles rapidly becoming fixed in the population representing selective sweeps. (c) Moderate-to-small effect size SNPs changing in frequency slowly over time, representing polygenic selection. (d) An overview of the linear mixed model approach used for Generation Proxy Selection Mapping and environmental GWAS. (e-f) A single SNP under ecoregion-specific selection. Different colors represent the trajectory of a given SNP in one of five different ecoregions. Ecoregion-specific selection can lead to allele frequencies that (e) diverge from or (f) converge to the population mean.

We used animals continuous birth date and high-density SNP genotypes for large samples of animals from three large US beef cattle populations; Red Angus (RAN; n =15,295), Simmental (SIM; n

=15,350), and Gelbvieh (GEL; n =12,031) to map loci responding to polygenic selection (**Supplementary Table 1, Supplementary Figure 1**). The LMM estimated that the proportion of variance in individuals' birth dates explained by the additive genetic effects of SNPs was large [Proportion of Variance Explained (PVE) = 0.520, 0.588, and 0.459 in RAN, SIM, and GEL, respectively], indicating that we could theoretically predict an animal's birth date from its multi-locus genotypes with an accuracy of ~70%. Obviously, birth date is not a heritable trait, but the ability to predict birth date from genotypic data is created by the time trend in allele frequencies. The PVE estimates indicate that there must be many genome-wide associations between genotype and birth date. The large amount of variance in birth date explained by the SNP genotypes persists even when the analysis is restricted to individuals born in the last 10 years (~2 generations) or 20 years (~4 generations) (**Supplementary Table 2**), demonstrating that GPSM has the power to detect subtle changes in allele frequency over extremely short periods of time. We removed the link between generation proxy and genotype by randomly permuting the animals' birth date and on reanalysis of the permuted data we observed PVE to decrease to zero (**Supplementary Table 2**).

The GPSM analyses for these three populations identified 268, 548, and 763 statistically significant SNPs (q-value < 0.1), representing at least 52, 85, and 92 genomic loci associated with birth date in RAN, SIM, and GEL, respectively (**Fig. 2a-f, Supplementary File 3**). Rapid shifts in allele frequency create highly significant GPSM signals. For example, *rs1762920* on chromosome 28 has undergone large changes in allele frequency in all three populations (**Fig. 2g)**, which in turn creates highly significant q-values (2.810×10^{-27}, 2.323×10^{-150}, 2.787×10^{-265} in RAN, SIM, and GEL, respectively). The allele frequency changes observed for this locus are extremely large compared to other significant regions, most of which have only small to moderate changes in allele frequency over the last ~10 generations. When we regressed allele frequency (0, 0.5, or 1.0 representing *AA*, *AB*, and *BB* genotypes per individual) on birth date, the average allele frequency changes per generation (ΔAF) for significant GPSM associations were 0.017, 0.024, and 0.022 for RAN, SIM, and GEL, respectively (**Supplementary Table 3**). In the analyses of each dataset, GPSM identified significant SNPs with $\Delta AF < 1.1\times10^{-4}$. The generally small allele frequency changes detected by GPSM are consistent with the magnitude of allele frequency changes expected for traits with polygenic trait architectures [3]. Consequently, we suspect that many of these loci would go undetected when using other selection mapping methods.

We performed a genomic REML analysis to identify how much of the variation in birth date was explained by various classes of GPSM SNPs. We built three GRMs using different SNP sets: One set with GPSM genome-wide significant SNPs (q < 0.1), the second with an equivalent number of the next most suggestive GPSM SNPs outside of loci (> 1 Mb from a q < 0.1 significant SNP), and the third an equivalent number of moderate minor allele frequency (MAF) (MAF > 0.15), non-significant SNPs (p > 0.5) intended to represent loci randomly drifting in the population. For each population, we observed that nearly all of the variation in birth date was explained by the significant and suggestive GRMs. While genome-wide significant loci explain the majority of genetic variance associated with birth year, an equivalent number of suggestive, but not significant SNPs have only slightly smaller PVEs (**Table 1**). We suspect that these SNPs are undergoing directional allele frequency changes too small to detect at genome-wide significance, even in this large dataset. We suspect that future sample size increases will provide power to detect these signatures of polygenic selection at a genome-wide significance level. Regardless of the number of SNPs used in the drift GRM, the variance associated with drift was consistently minimal (**Table 1**).

Table 1. Variation in birth date explained by three classes of SNPs. The PVE estimates (standard error in parentheses) from a Genomic Restricted Maximum Likelihood (GREML) variance component analysis of birth date using three GRMs created from: 1) genome-wide significant SNPs ($q < 0.1$), 2) an equivalent number of the next most significant SNPs outside of genome-wide significant associated regions, and 3) an equivalent number of non-significant SNPs ($p>0.5$) randomly sampled from genomic regions that did not harbor genome-wide significant associations.

Population	Genome-wide significant SNPs*	Suggestive significant SNPs*	Other SNPs*	Total
Red Angus	0.187 (0.026)	0.148 (0.015)	0.031 (0.004)	0.366 (0.023)
Simmental	0.239 (0.020)	0.194 (0.014)	0.037 (0.004)	0.470 (0.016)
Gelbvieh	0.225 (0.017)	0.193 (0.013)	0.008 (0.003)	0.426 (0.018)

*Contained 268, 548, and 763 SNPs for Red Angus, Simmental, and Gelbvieh, respectively

As proof-of-concept, GPSM identified known targets of selection. In Simmental, we identified significant associations at three Mendelian loci that explain the major differences in appearance between early imported European Simmental and modern US Simmental (**Fig. 2h)**. These loci: *POLLED* (absence of horns [6]), *ERBB3/PMEL* (European Simmental cream colour [7]), and *KIT* (piebald coat colouration [8]) have not appreciably changed in allele frequency since 1995, making their GPSM signature significant, but less so than other loci actively changing in frequency.

In addition to these three known Mendelian loci, we detected numerous novel targets of selection within and across the populations. While the majority of the genomic regions detected as being under selection were population-specific (79.8%, 79.8%, and 77.2% of the significant regions in RAN, SIM, and GEL, respectively), we identified seven loci that are under selection in all three populations, and fifteen more under selection in two (**Supplementary Table 4**). Of the regions identified in multiple populations, many correspond to genes with predicted production-related functions in cattle (*DACH1*-Growth, *LRP12*-Growth, *MYBPH*-Muscle Growth, *RHOU*-Carcass Weight, *BIRC5*-Feed Intake). However, GPSM did not identify any of the well-established large-effect growth loci (i.e., *PLAG1*, *LCORL*). Growth phenotypes (e.g., birth, weaning, and yearling weights) are known to be under strong selection in all three populations [9], but antagonistic pleiotropic effects such as increased calving difficulty prevent directional selection from changing frequencies at these large-effect loci. We also identified immune function genes under polygenic selection (*ARHGAP15*, *ADORA1*, *CSF2RA*). While immune function has not been directly artificially selected in cattle, healthy cattle perform better than their sick counterparts [10]. Many of the selection signatures that were identified in at least two of the populations have no known functions or phenotype associations in cattle, highlighting the ability of GPSM to identify important loci under polygenic selection without the need for any phenotype data.

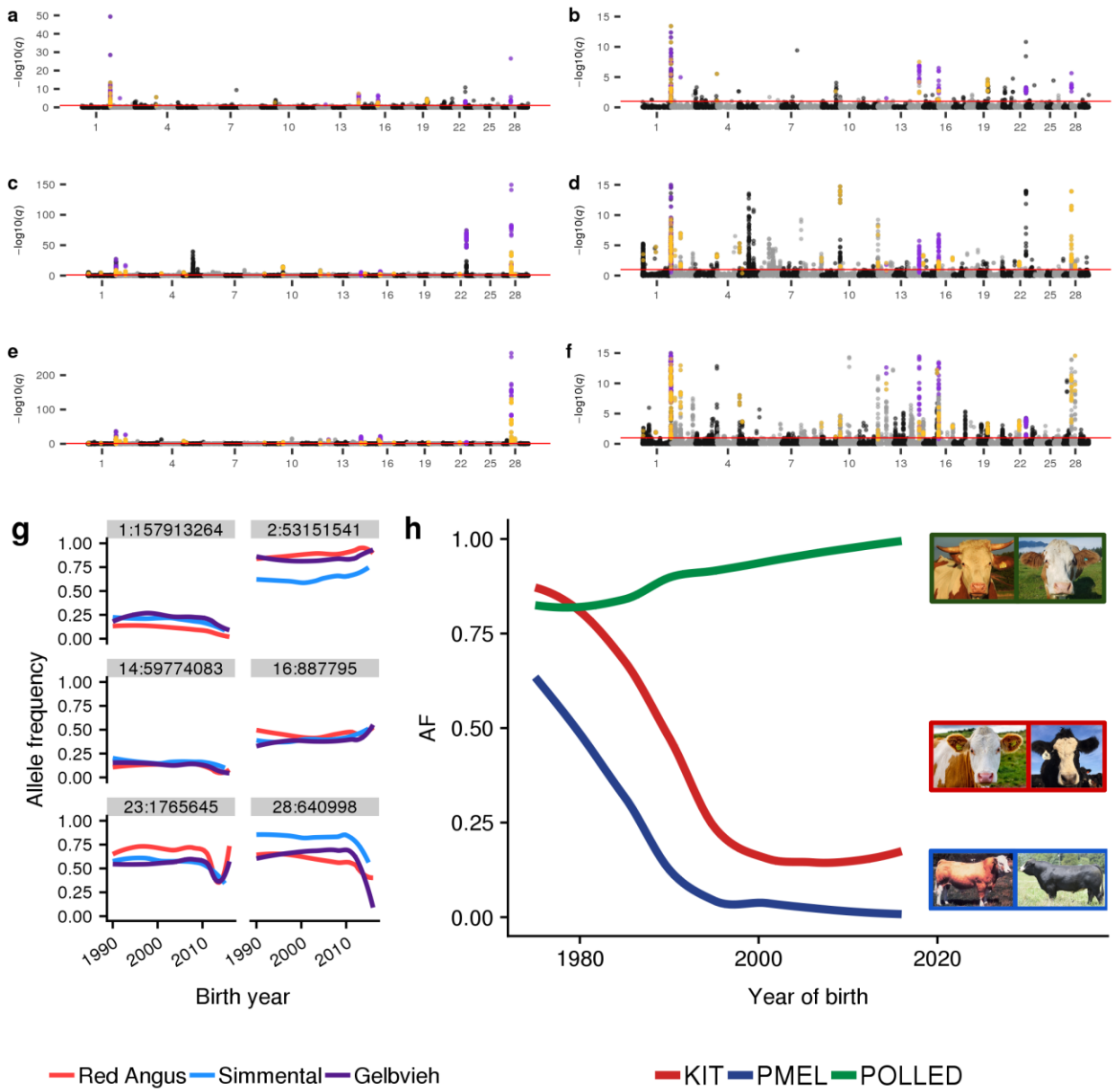

Figure 2. Generation Proxy Selection Mapping identifies signals of polygenic selection in three major U.S. cattle populations. Full and truncated ($-\log_{10}(q) < 15$) Manhattan plots for GPSM analysis of Red Angus (a & b), Simmental (c & d), and Gelbvieh (e & f). Purple points indicate SNPs significant in all three population-specific GPSM analyses and orange points indicate SNPs significant in two. (g) Smoothed allele frequency histories for the six most significant loci identified as being under selection in all three datasets. (h) Allele frequency histories for three known Mendelian loci that control differences in visual appearance between introduced European and modern US Simmental cattle.

Biological processes and pathways enriched in genes located proximal to GPSM SNP associations point to selection on drivers of production efficiency and on population-specific characteristics (**Supplementary File 4**). In each population, we identified numerous biological processes involved in cell cycle control, which are directly involved in determining muscle growth rate, as being under selection [11].

In Red Angus and Gelbvieh we identified multiple cancer pathways as being under selection. This likely represents further evidence of selection on cell cycle regulation and growth rather than on any cancer related phenotypes [12]. Red Angus cattle are known to be highly fertile with exceptional maternal characteristics [13]. We identified the "ovarian steroidogenesis" pathway as being under selection, a likely driver of hormone balance and female fertility [14]. We also identify numerous other processes involved in the production and metabolism of hormones. Hormone metabolism is a central regulator of growth in cattle [15], but could also represent selection for increased female fertility in Red Angus. Further, Tissue Set Enrichment Analyses (TSEA) of Red Angus GPSM candidate genes showed suggestive expression differences ($p < 0.1$) in multiple human reproductive tissues (**Supplementary Files 5-6**). Enrichments in these tissues did not exist in TSEA of Simmental or Gelbvieh GPSM gene sets, suggesting explicit within-population selection on fertility. Gelbvieh cattle are known for their rapid growth rate and carcass yield. Selection on these phenotypes likely drives the identification of the six biological processes identified which relate to muscle development and function in the Gelbvieh GPSM gene set. Consequently, this gene set is significantly enriched for expression in human skeletal muscle (**Supplementary Files 5-6**). A complete list of genomic regions under population-specific selection and their associated candidate genes is in **Supplementary File 4**.

Detecting environmental adaptation using envGWAS

Using an equivalent form of model to GPSM, but with continuous environmental variables (30 year normals for temperature, precipitation, and elevation) or statistically-derived discrete ecoregions as the dependent variable allows us to identify environmental adaptive loci that have been subjected to artificial and, perhaps in this context more importantly, natural selection [16]. We refer to this method as environmental GWAS (envGWAS). envGWAS extends the theory of the Bayenv approach of Coop et al. (2010) [17] which searches for allele frequency correlations along environmental gradients to identify potentially adaptive loci. However, our approach scales to biobank-sized datasets. We used *K*-means clustering with 30-year normal values for temperature, precipitation, and elevation to partition the United States into 9 discrete ecoregions (**Fig. 3a**). These ecoregions are largely consistent with those represented in previously-published maps from the environmetrics and atmospheric science literature [18], and reflect well-known differences in cattle production environments. The resulting ecoregions capture not only combinations of environmental variables, but associated differences in forage type, local pathogens, and ecoregion-wide management differences to which animals are exposed. The three studied populations are not universally present in all ecoregions (**Fig. 3b, Supplementary Figures 3b & 4b, Supplementary Table 5**) and since the development of these US populations in the late 1960s and early 1970s, registered seedstock animals from these populations have a small footprint in desert regions with extreme temperatures and low rainfall.

Although environmental variables and ecoregions are not inherited, the estimated PVE measures the extent to which genome-wide genotypes change in frequency across the environments in which the animals were born and lived. The PVE explained by SNPs ranged from 0.586 to 0.691 for temperature, 0.526 to 0.677 for precipitation, and 0.585 to 0.644 for elevation (**Supplementary Table 6**). In Red Angus, PVE for ecoregion membership ranged from 0.463 for the Arid Prairie to 0.673 for the Fescue Belt (**Supplementary Table 7**). We observe similar environmental PVE in both Simmental and Gelbvieh datasets. These measures suggest that genetic associations exist along both continuous environmental gradients and within discrete ecoregions. Despite this genetic signal, principal component analysis (PCA)

does not suggest that ecoregion-driven population structure exists in any of the populations (**Supplementary Figure 5**). Permutation tests that shuffled environmental dependent variables, removing the relationship between the environment and the animal's genotype, resulted in all PVEs being reduced to ~ 0, strongly suggesting that the detected associations between genotype and environment were not spurious. An additional permutation test that permuted animals' zip codes, such that all animals from a given zip code were assigned the same "new" zip code from a potentially different ecoregion provided similar results, indicating that bias due to sampling at certain zip codes was not producing envGWAS signals. From 10 rounds of permutation, there were no SNP associations with p-values $< 1\times10^{-5}$. Consequently, we used this empirically-derived p-value threshold to determine SNP significance in all of the envGWAS analyses, which is also in agreement with the significance threshold used by Wellcome Trust Case Control Consortium [19].

Discrete ecoregion envGWAS

In Red Angus, we identified 54 variants defining 18 genomic loci significantly associated with membership of an ecoregion in the discrete multivariate envGWAS analysis (**Fig. 3c**). Of these loci, only two overlapped with loci identified in the continuous envGWAS analyses, suggesting that using alternative definitions of environment in envGWAS may detect different sources of adaptation. Of the 18 significant loci, 17 were within or near (< 100 kb) candidate genes (**Supplementary Table 8, Supplementary File 7**), many of which have potentially adaptive functions. For example, envGWAS identified SNPs immediately (22.13 kb) upstream of *CUX1* (Cut Like Homeobox 1) gene on chromosome 25. *CUX1* controls hair coat phenotypes in mice [20]. Alleles within *CUX1* can be used to differentiate between breeds of goats raised for meat versus those raised for fiber [21]. The role of *CUX1* in hair coat phenotypes makes it a strong adaptive candidate in environments where animals are exposed to heat, cold, or toxic ergot alkaloids from fescue stress [22]. Other candidate genes identified by envGWAS have previously been identified as targets of selection between breeds of cattle (*MAGI2*, *CENPP*), or in other species (*DIRC1*-humans, *GORASP2*-fish, *ADRB1*-dogs) (**Supplementary Table 8**). Adaptive signatures shared between cattle and other species may point to shared biological processes that drive environmental adaptation. We also identified four adaptive candidate genes known to possess immune functions (*RASGEF1B*, *SPN*, *ZMYND8*, *LOC100298064*/*HAVCR1*). The envGWAS identified variants within or near immune function genes under ongoing selection in all three populations, thereby suggesting that genetic adaptations conferring resistance or tolerance to local pathogens and immune stressors may be as important as adaptations to abiotic stressors like heat or cold stress.

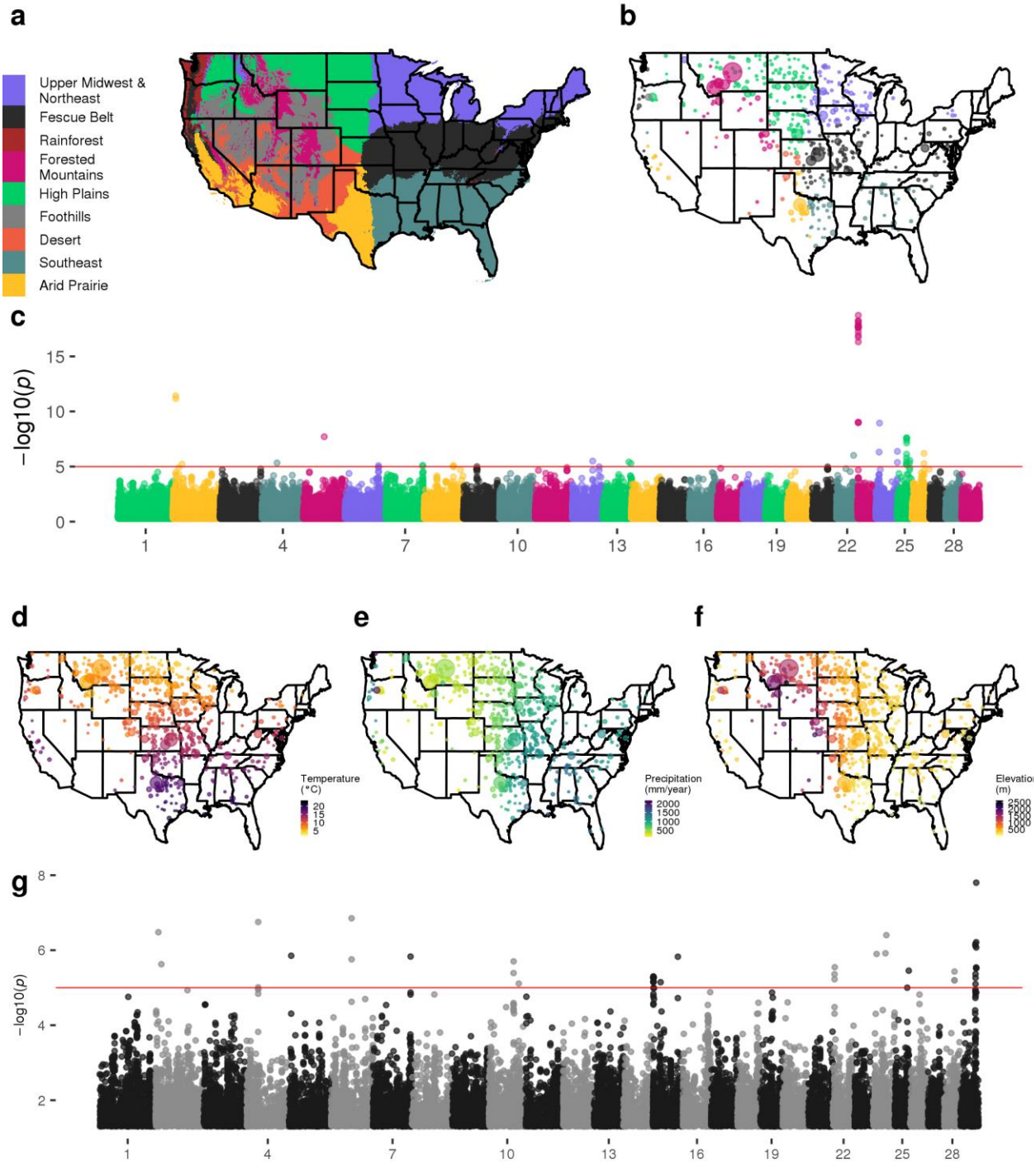

Figure 3. Manhattan plots for discrete and continuous envGWAS in Red Angus cattle. (a) Nine continental US ecoregions defined by *K*-means clustering of 30-year normal temperatures, precipitations, and elevations. (b) Locations of sampled Red Angus animals coloured by breeder's ecoregion and sized by the number of animals at that location. (c) Multivariate discrete envGWAS (case-control for six regions with > 600 animals). Locations of sampled Red Angus animals coloured by (d) 30-year normal temperature, (e) 30-year normal precipitation, and (f) elevation. (g) Multivariate continuous envGWAS with

temperature, precipitation, and elevation as dependent variables. For all Manhattan plots the red line indicates the empirically-derived p-value significance threshold from permutation analysis ($p < 1 \times 10^{-5}$).

In Simmental, we identified 11 loci tagged by 39 variants significantly associated with membership of an ecoregion in the multivariate envGWAS analysis (**Supplementary Figure 3**). In Gelbvieh, 66 variants identified 33 local adaptation loci (**Supplementary Figure 4**). In the analyses of all three datasets, we identified a common local adaptation signature on chromosome 23 (peak SNP *rs1023574*). Multivariate analyses in all three populations identified alleles at this SNP to be significantly associated with one or more ecoregions ($q = 1.24 \times 10^{-13}$, 3.15×10^{-12}, 4.82×10^{-5} in RAN, SIM, and GEL, respectively). In all three datasets, we identified *rs1023574* as a univariate envGWAS association with membership of the Forested Mountains ecoregion. However, the most significant univariate association in Red Angus was with the Arid Prairie region which was excluded from both the Simmental and Gelbvieh analyses due to low within-region sample size. In the multivariate analysis for Red Angus, the associated locus spanned 18 SNPs from (1,708,914 to 1,780,836 bp) and contained the pseudogene *LOC782044*. The nearest annotated gene, *KHDRBS2* (KH RNA Binding Domain Containing, Signal Transduction Associated 2) has previously been identified by other adaptation studies in cattle, sheep, and pigs [23–25]. This variant was not significantly associated with any continuous environmental variable in Red Angus. However, *rs1023574* was significantly associated with temperature, elevation, and humidity variables in Simmental. The *KHDRBS2* locus was preferentially introgressed between *Bos taurus* and domestic yak [26]. Further, this locus shows an abnormal allele frequency trajectory (**Fig. 4c**), indicating that it may be a target of balancing selection.

Continuous environmental variable envGWAS

Using continuous temperature, precipitation, and elevation data as quantitative dependent variables in a multivariate envGWAS analysis of Red Angus animals, we identified 46 significantly associated SNPs (**Fig. 3g**). These SNPs tag 17 loci, many of which are within 100 kb of strong candidate genes. Univariate envGWAS identified 23, 17, and 10 variants associated with temperature, precipitation, and elevation, respectively (**Supplementary Figure 6**). The most significant multivariate association in Red Angus is located on chromosome 29 within *BBS1* (Bardet-Biedl syndrome 1), which is involved in energy homeostasis [27]. *BBS1* mutant knock-in mice show irregularities in photoreceptors and olfactory sensory cilia [28] functions that are likely important to an individual's ability to sense its local environment. This region was not significantly associated in any of the univariate analyses of environmental variables, and was not identified in any of the discrete ecoregion envGWAS. Of the candidate genes identified in this Red Angus analysis, 9 have previously been implicated in adaptive functions in humans or cattle (*DIRC1*, *ABCB1*, *TBC1D1*, *AP5M1*, *GRIA4*, *LRRC4C*, *RBMS*3, *GADL1*, *ADCYAP1*, *CUX1*, and *PLA2G12B*) (**Supplementary Table 9**). Significant SNPs and their corresponding candidate genes for all three datasets are reported in **Supplementary File 7**.

While we found few candidate genes to overlap between populations, we identified multiple shared biological pathways and processes (**Supplementary File 8**) derived from lists of envGWAS candidate genes. Pathways in common between populations were driven by largely different gene sets. Across all populations, we identified the "axon guidance" pathway, and numerous gene ontology (GO) terms related to axon development and guidance as under region-specific selection. Ai et al. (2015) suggested that axon development and migration in the central nervous system is essential for the maintenance of homeostatic temperatures by modulating heat loss or production [29]. Further, the direction and organization of axons is

an essential component of the olfactory system which is frequently implicated in environmental adaptation through the recognition of local environmental cues [30]. In addition to axonal development, a host of other neural signaling pathways were identified in multiple populations. A genome-wide association study for gene-by-environment interactions with production traits in Simmental cattle by Braz et al. (2019) identified a similar set of enriched pathways [31]. These common neural signaling pathways identified by envGWAS are regulators of stress response, temperature homeostasis, and vasoconstriction [32]. We identified other shared pathways involved in the control of vasodilation and vasoconstriction (relaxin signaling, renin secretion, and insulin secretion). Vasodilation and vasoconstriction are essential to physiological temperature control in cattle and other species [33]. The ability to mount a physiological response to temperature stress has a direct impact on cattle performance, making vasodilation a prime candidate for environment-specific selection. Pathways and processes identified by envGWAS signals are reported in **Supplementary File 8.**

To further explore the biology underlying adaptive signatures, we performed Tissue Set Enrichment Analysis of our envGWAS candidate gene lists. These analyses, using expression data from humans and worms (*C. elegans*), identified brain and nerve tissues as the lone tissues where envGWAS candidate genes show significantly enriched expression (**Supplementary Files 9-12**). Tissue-specific expression in the brain further supports our observed enrichment of local adaptation pathways involved in neural signaling and development.

Identifying loci undergoing region-specific selection with GPSM ecoregion meta-analysis

envGWAS detects allelic associations with continuous and discrete environmental variables, but does not address whether selection is towards increased local adaptation, or whether local adaptation is being eroded by the exchange of germplasm between ecoregions via artificial insemination. We used the spatiotemporal stratification of genotyped animals to identify loci undergoing ecoregion-specific selection. We performed GPSM within each sufficiently genotyped ecoregion and identified variants with high effect size heterogeneity (Cochran's Q statistic) between ecoregions. Variants with significant heterogeneity across regions that were also significant in at least one within-region GPSM analysis imply ecoregion-specific allele frequency change. These changes could have been due either to selection for local adaptation (**Fig. 1e**), or locally different allele frequencies moving towards the population mean (**Fig. 1f**). We identified 59, 38, and 46 significant SNPs in Red Angus, Simmental, and Gelbvieh, respectively undergoing ecoregion-specific selection. These represent 15, 21, and 26 genomic loci (> 1 Mb to nearest next significant SNP) (**Fig. 4a**). In most cases, these variants have an effect (posterior probability of an effect: m-value > 0.9) in only one or two ecoregions (**Fig. 4b**). Further, nearly all represent the decay of ecoregion-specific allele frequencies towards the population mean (**Fig. 4c**) as opposed to on-going directional selection for ecoregion specific beneficial adaptations (**Supplementary File 8**).

Despite the apparent ongoing decay of local adaptation, this meta-analysis of ecoregion-specific GPSM identified several interesting candidate genes for environmental adaptation. A significant locus on chromosome 1 at ~73.8 Mb (lead SNP *rs254372*) lies within *OPA1* (OPA1 mitochondrial dynamin like GTPase), which was under selection in the Fescue Belt ecoregion. *OPA1* has been implicated in the regulation of circadian rhythm in mice [34], and is known to regulate metabolic and cardiac adaptations [35] through mitochondrial interactions. We also identified variants within *ADAMTS16* (ADAM Metallopeptidase With Thrombospondin Type 1 Motif 16), which regulates blood pressure in mice [36] and has previously been identified in other adaptation studies [37]. These genes are of particular interest, because the primary symptoms of fescue toxicosis are due to vasoconstriction caused by ergot alkaloids synthesized by endophytes in fescue [38]. Adaptive alleles at these loci are being driven in frequency towards the population mean allele frequency (**Fig. 4c**), which is typically a low minor allele frequency (**Supplementary Figures 9-11**).

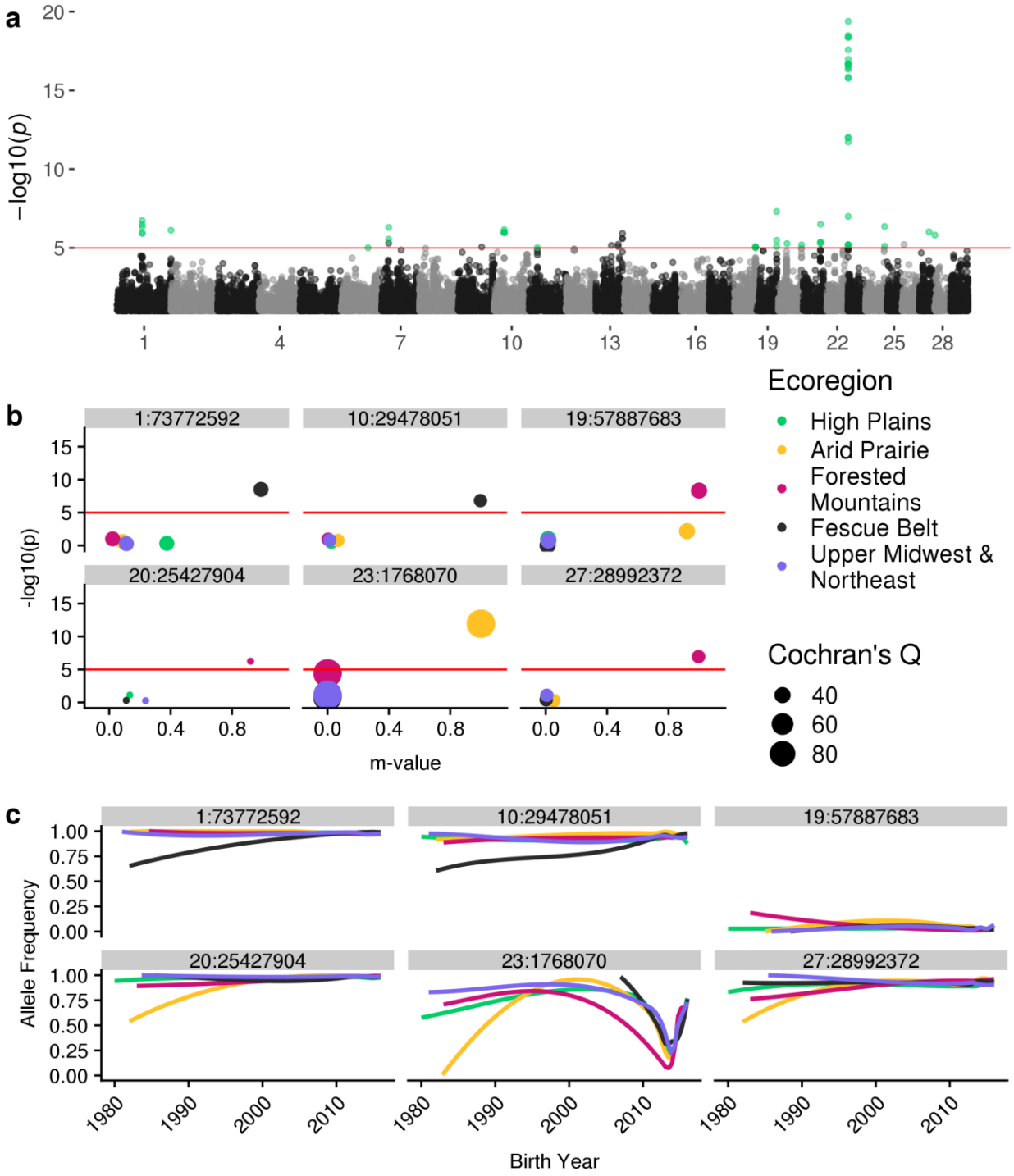

Figure 4. Meta-analysis of within-ecoregion GPSM for Red Angus cattle. A) Manhattan plot of per-variant Cochran's Q p-values. Points coloured green had significant Cochran's Q ($p < 1 \times 10^{-5}$) and were significant in at least one within-region GPSM analysis ($p < 1 \times 10^{-5}$). B) Ecoregion effect plots for lead SNPs from six loci from (A). Points are coloured by ecoregion and are sized based on Cochran's Q value. C) Ecoregion-specific allele frequency histories for SNPs from (B), coloured by ecoregion.

DISCUSSION

We leveraged large commercially-generated genomic datasets from three major US beef cattle populations to map polygenic selection and environmental adaptation using novel GWAS applications [39]. Using temporally-stratified genotype data we detected very small selection-driven changes in allele frequency throughout the genome. This is consistent with expectations of polygenic selection acting on a large number of variants with individual small effects. Which phenotypes are being selected and driving the allele frequency changes at particular loci is not definitively known. However, GPSM is agnostic to the selected phenotypes, and identifies important loci changing in frequency due to selection without the need to measure potentially difficult or expensive phenotypes. Further, GPSM differentiates between selection and drift while accounting for confounding effects such as population structure, relatedness, and inbreeding. With the availability of large samples our analytical frameworks solve the long-standing population genetics problem of identifying the loci subjected to polygenic selection.

Future studies exploring the effects of selection from the context of complex trait networks could explain how hundreds or thousands of selected genes act together to shape genomic diversity under directional selection. Candidate genes identified by GPSM identify pathways and processes involved in production efficiency (growth, digestion, muscle development, and fertility). In addition to a small number of loci, for which function is known, we identify hundreds of novel signatures of ongoing selection.

The envGWAS identified 174, 125, and 130 SNPs associated with both continuous or discrete environmental factors in Red Angus, Simmental, and Gelbvieh, respectively. Identified candidate genes have functions related to environmental adaptation. Using these environmentally-associated candidate genes we identified an enrichment of pathways and tissues involved in neural development and signaling. These envGWAS associations emphasize the role that the nervous system plays in recognizing and responding to environmental stress in mammals. In addition to neural pathways, we observe significantly enriched expression of envGWAS genes in the brain tissues of humans, mice, and worms. Other pathways associated with environmental adaptation reveal the importance of mechanisms involved in regulating vasoconstriction and vasodilation, both of which are essential for responses to heat, cold, altitude, and toxic fescue stressors in cattle.

The statistical power and wide geographical distribution of the cattle comprising these data highlights that the utilized approaches can be leveraged to understand the genomic basis of adaptation in many other studies and species. The small allele frequency differences identified by envGWAS are consistent with a polygenic model of local adaptation, likely driven by small changes in gene expression [40]. Further, envGWAS identifies candidate genes (i.e. *KHDRBS2*) and pathways previously implicated as domestication-related [26]. This suggests that these genes are under natural and balancing selection to cope with environmental stress, and not specifically part of the domestication process. Further, because different genes in the same pathways were detected in the analyses of the different populations, we hypothesize that these pathways influence local adaptation in many mammals and should be studied in other ecological systems. This knowledge will become increasingly valuable as species attempt to adjust to a changing climate.

Artificial insemination in cattle has allowed the ubiquitous use of males which have been found to be superior when progeny performance has been averaged across US environments. Our results suggest that environmental associations are widespread in cattle populations, but that the widespread use of artificial insemination has caused US cattle populations to lose ecoregion-specific adaptive variants. We identified 16, 21, and 30 loci undergoing ecoregion-specific selection in Red Angus, Simmental, and Gelbvieh, respectively. In almost every case, selection has driven allele frequencies within an ecoregion back towards the population mean allele frequency (**Fig. 1f and Fig. 4c**). In three independent datasets, we identified a

single shared environmentally-associated locus near the gene *KHDRBS2*. This locus has been identified as introgressed in yak, and exhibits an irregular allele frequency trajectory which suggests that it may be subject to balancing selection [41]. Though we identified only a single common envGWAS locus, we observed significant overlap in the pathways regulated by candidate genes within the associated loci. This reveals that adaptive networks are complex and that adaptation can be influenced by selection on functional variants within combinations of genes from these networks. As we work to breed more environmentally-adapted cattle, there will be a need for selection tools that incorporate genotype-by-environment interactions to ensure that cattle become increasingly locally adapted.

Methods

Genotype Data:

SNP assays for three populations of genotyped *Bos taurus* beef cattle ranging in density from ~25K SNPs to ~770K SNPs were imputed to a common set of 830K SNPs using the large multi-breed imputation reference panel described by Rowan et al. 2019 [42]. Genomic coordinates for each SNP were from the ARS-UCD1.2 reference genome [43]. Genotype filtering for quality control was performed in PLINK (v1.9)[44], reference-based phasing was performed with Eagle (v2.4) [45], and imputation with Minimac3 (v2.0.1) [46]. Following imputation, all three datasets contained 836,118 autosomal SNP variants. All downstream analyses used only variants with minor allele frequencies > 0.01.

Upon filtering, we performed a principal component analysis for each population in PLINK. This was to assess if there were discrete subpopulations within the populations and if there were patterns of structure related to ecoregions.

Generation Proxy Selection Mapping (GPSM):

To identify alleles that had changed in frequency over time, we fit a univariate genome-wide linear mixed model (LMM) using GEMMA (Version 0.98.1) [47]. Here, we used the model:

EQUATION 1:

$$y = Xg + Zu + e$$
$$u \sim N(0, G\sigma_a{}^2)$$
$$e \sim N(0, \sigma_e{}^2 I)$$

where **y** is an individual's generation proxy, in our case birth date, and **X** was an incidence matrix that related SNPs to birth dates within each individual and **g** was the estimated effect size for each SNP. Birth date was calculated from an animal's age in years calculated on April 5, 2017. We control for confounding population structure, relatedness, and inbreeding with a polygenic term **u** that uses a standardized genomic relationship matrix (GRM) **G** [48] and we estimated $\sigma_a{}^2$ *and* $\sigma_e{}^2$ using restricted maximum likelihood estimation. Here, continuous age served as a proxy for generation number from the beginning of the pedigree. Other than the tested SNP effects, no fixed effects other than the overall mean were included in the model. We tested each SNP for an association with continuous birth date. We converted p-values to FDR corrected q-values and used a significance threshold of q < 0.1. We performed additional negative-control analyses in each dataset by permuting the date of birth associated with each animal's genotypes to

ensure that the detected GPSM signals were likely to be true positives. Permutation was performed ten times for each population. To visualize the allele frequency history of loci undergoing the strongest selection, we fit a loess and simple linear regressions for date of birth and allele frequencies scored as 0, 0.5 or 1.0 within each individual using R [49]. Results were visualized using ggplot2 [50].

Birth date variance component analysis

To estimate the amount of variation in birth date explained by GPSM significant SNPs, we performed multi-GRM GREML analyses for birth date in GCTA (v1.92.4) [51]. We built separate GRMs using genome-wide significant markers and all remaining makers outside of significant GPSM loci (> 1 Mb from significant GPSM SNPs to control for markers physically linked to significant GPSM SNPs). To further partition the variance in birth date explained by subsets of SNPs, we performed a GREML analysis using three GRMs created with genome-wide significant ($p < 1 \times 10^{-5}$) SNPs, an equal number of the next most significant SNPs, and an equal number of unassociated ($p > 0.5$) markers with minor allele frequencies > 0.15, to match the allele frequencies of significant SNPs. These three GRM were each constructed using 268, 548, and 763 SNPs for Red Angus, Simmental, and Gelbvieh, respectively.

Environmental Data:

Thirty-year normals (1981-2010) for mean temperature ((average daily high (°C) + average daily low (°C)/2), precipitation (mm/year), and elevation (m above sea level) for each 4 km^2 of the continental US were extracted from the PRISM Climate Dataset [52], and used as continuous dependent variables in envGWAS analysis. Optimal *K*-means clustering of these three variables grouped each 4 km^2 of the continental US into 9 distinct ecoregions. Using the reported breeder zip code for each individual, we linked continuous environmental variables to animals and partitioned them into discrete environmental cohorts for downstream analysis. For ecoregion assignments, latitude and longitude were rounded to the nearest 0.1 degrees. As a result, some zip codes were assigned to multiple ecoregions. Animals from these zip codes were excluded from the discrete region envGWAS but remained in analyses that used continuous measures as dependent variables.

Environmental Genome-wide Association Studies (envGWAS):

To identify loci segregating at different frequencies within discrete ecoregions or along continuous climate gradients, we used longitudinal environmental data for the zip codes attached to our study individuals as dependent variables in univariate and multivariate genome-wide LMMs implemented in GEMMA (Version 0.98.1). We fit three univariate envGWAS models that used 30-year normal temperature, precipitation, and elevation data as dependent variables. These used an identical model to **EQUATION 1**, but used environmental values as the dependent variable (**y**) instead of birth date. We also fit a combined multivariate model using all three environmental variables to increase power. To identify loci associated with entire climates as opposed to only continuous variables, we fit univariate and multivariate case-control envGWAS analyses using an individual's region assignment described in the "Environmental Data" section as binary phenotypes. Proportion of variation explained (PVE), phenotypic correlations, and genetic correlations were estimated for continuous environmental variables and discrete environmental regions using GEMMA's implementation of REML.

To ensure that envGWAS signals were not driven by spurious associations, we performed two separate permutation analyses. In the first, we randomly permuted the environmental variables and regions

associated with an individual prior to performing each envGWAS analysis, detaching the relationship between an individual's genotype and their environment. In the second, to ensure that envGWAS signals were not driven by the over-sampling of individuals at particular zip codes, we permuted the environmental variables associated with each zip code prior to envGWAS analysis. These two types of permutation analyses were performed for each dataset and for each type of univariate and multivariate envGWAS analysis. We determined significance using a permutation-derived p-value cutoff ($p < 1 \times 10^{-5}$) [53].

GPSM meta-analyses:

To identify variants undergoing ecoregion-specific allele frequency changes, we performed GPSM analyses within each region with more than 600 individuals. The SNP significance testing effects and standard errors from each of the within-region GPSM analyses were combined into a single meta-analysis for each population using METASOFT (v2.0.1) [54]. We identified loci with high heterogeneity in allele effect size, suggesting region-specific selection. An m-value indicating the posterior-probability of a locus having an effect in a particular ecoregion was calculated for each of these loci [55].

Gene set and tissue set enrichment analysis:

Using the NCBI annotations for the ARS-UCD1.2 *Bos taurus* reference assembly, we located proximal candidate genes near significant SNPs from each of our analyses. We generated two candidate gene lists each from significant GPSM and envGWAS SNPs. Lists contained all annotated genes within 10 kb or 100 kb from significant SNPs. We consolidated significant SNPs from all envGWAS analyses to generate a single candidate gene list for each breed. Using these candidate gene lists, we performed gene ontology (GO) and KEGG pathway enrichment analysis using ClueGO (v2.5.5) [56] implemented in Cytoscape (v3.7.2) [57]. We identified pathways and GO terms where at least two members of our candidate gene list comprised at least 1.5% of the term's total genes. We applied a Benjamini-Hochberg multiple-testing correction to reported p-values and GO terms with FDR corrected p-values < 0.1 were considered significant.

Using the above gene sets, we performed three separate Tissue Set Enrichment Analyses (TSEA) using existing databases of human, mouse, and worm gene expression data. We searched for enriched gene expression with data from the Human Protein Atlas [58] and Mouse ENCODE [59] using the Tissue Enrich tool (v1.0.7) [60]. Additionally, we performed another Tissue Set Enrichment Analysis using GTEx data [61] and a targeted Brain Tissue Set Enrichment Analysis in the pSI R package (v1.1) [62]. Finally, we used Ortholist2 [63] to identify *C. elegans* genes orthologous with members of our envGWAS and GPSM gene lists. We then queried these lists in WormBase's Tissue Enrichment Analysis tool [64,65] to identify specific tissues and neurons with enriched expression in *C. elegans*. We used each tool's respective multiple-testing correction to determine significance. We deemed an enrichment in a tissue "suggestive" when its p-value was < 0.1.

Data availability

The raw data used in this research are available from the Red Angus Association of America, American Simmental Association, and American Gelbvieh Association under a Data Use Agreement, but are not publicly available. Derived data (analytical results) are however available as supplementary files associated with this publication.

Ethics approval

Because phenotypic records and tissue samples were collected as part of routine livestock production practices, an ACUC exemption was granted (University of Missouri ACUC Protocol Number EX 8596).

Acknowledgements
This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2016-68004-24827 from the USDA National Institute of Food and Agriculture.
We appreciate comments from Wes Warren, Jeremy Taylor, and William Lamberson while writing this manuscript.

Author Contributions:
JED and TNR conceptualized and designed the research. RDS managed data transfer, storage, and retrieval. TNR performed the majority of the analyses, including imputation, simulations, association analyses and enrichment analyses, with assistance from HJD. TNR and JED interpreted results, with assistance from RDS and CMS. TNR and JED wrote the initial version of the manuscript, which was edited by HJD and CMS.

The authors declare no competing interests

Additional Information:

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Jared Decker, DeckerJE@missouri.edu.

Reprints and permissions information is available at www.nature.com/reprints.

References

1. Gutiérrez-Gil, B., Arranz, J. J. & Wiener, P. An interpretive review of selective sweep studies in Bos taurus cattle populations: identification of unique and shared selection signals across breeds. *Front. Genet.* **6**, 167 (2015).
2. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
3. Höllinger, I., Pennings, P. S. & Hermisson, J. Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS Genet.* **15**, e1008035 (2019).
4. Decker, J. E. *et al.* A novel analytical method, Birth Date Selection Mapping, detects response of the Angus (Bos taurus) genome to selection on complex traits. *BMC Genomics* **13**, 606 (2012).
5. Matukumalli, L. K. *et al.* Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**, e5350 (2009).
6. Wiedemar, N. *et al.* Independent polled mutations leading to complex gene expression differences in cattle. *PLoS One* **9**, e93435 (2014).
7. Mészáros, G., Petautschnig, E., Schwarzenbacher, H. & Sölkner, J. Genomic regions influencing coat color saturation and facial markings in Fleckvieh cattle. *Anim. Genet.* **46**, 65–68 (2015).
8. Fontanesi, L., Tazzoli, M., Russo, V. & Beever, J. Genetic heterogeneity at the bovine KIT gene in cattle breeds carrying different putative alleles at the spotting locus. *Anim. Genet.* **41**, 295–303 (2010).
9. Kuehn, L. A. & Thallman, R. M. Across-Breed EPD Tables For The Year 2016 Adjusted To Breed Differences For Birth Year Of 2014. (2016).
10. Weber, K. L. *et al.* Identification of Gene Networks for Residual Feed Intake in Angus Cattle Using Genomic Prediction and RNA-seq. *PLoS One* **11**, e0152274 (2016).
11. Guo, B. *et al.* Transcriptome analysis of cattle muscle identifies potential markers for skeletal muscle growth rate and major cell types. *BMC Genomics* **16**, 177 (2015).
12. Rolf, M. M. *et al.* Genome-wide association analysis for feed efficiency in Angus cattle. *Anim.*

Genet. **43**, 367–374 (2012).

13. Breeds - Red Angus. *The Cattle Site* https://www.thecattlesite.com/breeds/beef/99/red-angus/.
14. Gareis, N. C. *et al.* Impaired insulin signaling pathways affect ovarian steroidogenesis in cows with COD. *Anim. Reprod. Sci.* **192**, 298–312 (2018).
15. Davis, S. L., Hossner, K. L. & Ohlson, D. L. Endocrine Regulation of Growth in Ruminants. in *Manipulation of Growth in Farm Animals: A Seminar in the CEC Programme of Coordination of Research on Beef Production, held in Brussels December 13–14, 1982* (eds. Roche, J. F. & O'Callaghan, D.) 151–178 (Springer Netherlands, 1984).
16. Hill, W. G. Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics* **196**, 1–16 (2014).
17. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**, 1411–1423 (2010).
18. Sathiaraj, D., Huang, X. & Chen, J. Predicting climate types for the Continental United States using unsupervised clustering techniques. *Environmetrics* vol. 30 e2524 (2019).
19. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
20. Sansregret, L. & Nepveu, A. The multiple roles of CUX1: insights from mouse models and cell-based assays. *Gene* **412**, 84–94 (2008).
21. Bertolini, F. *et al.* Signatures of selection and environmental adaptation across the goat genome post-domestication. *Genet. Sel. Evol.* **50**, 57 (2018).
22. Aiken, G. E., Klotz, J. L., Looper, M. L., Tabler, S. F. & Schrick, F. N. Disrupted hair follicle activity in cattle grazing endophyte-infected tall fescue in the summer insulates core body temperatures1. *The Professional Animal Scientist* **27**, 336–343 (2011).
23. León, C. D., De León, C., Manrique, C., Martínez, R. & Rocha, J. F. Research Article Genomic association study for adaptability traits in four Colombian cattle breeds. *Genetics and Molecular Research* vol. 18 (2019).

24. Guo, J. *et al.* Whole-genome sequencing reveals selection signatures associated with important traits in six goat breeds. *Sci. Rep.* **8**, 10405 (2018).

25. Gurgul, A. *et al.* A genome-wide detection of selection signatures in conserved and commercial pig breeds maintained in Poland. *BMC Genet.* **19**, 95 (2018).

26. Medugorac, I. *et al.* Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks. *Nat. Genet.* **49**, 470–475 (2017).

27. Guo, D.-F. *et al.* The BBSome Controls Energy Homeostasis by Mediating the Transport of the Leptin Receptor to the Plasma Membrane. *PLoS Genet.* **12**, e1005890 (2016).

28. Davis, R. E. *et al.* A knockin mouse model of the Bardet–Biedl syndrome 1 M390R mutation has cilia defects, ventriculomegaly, retinopathy, and obesity. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19422–19427 (2007).

29. Ai, H. *et al.* Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* **47**, 217–225 (2015).

30. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3**, e90 (2007).

31. Braz, C. U., Rowan, T. N., Schnabel, R. D. & Decker, J. E. Extensive genome-wide association analyses identify genotype-by-environment interactions of growth traits in Simmental cattle. doi:10.1101/2020.01.09.900902.

32. Morrison, S. F. Central control of body temperature. *F1000Res.* **5**, (2016).

33. Garner, J. B. *et al.* Genomic Selection Improves Heat Tolerance in Dairy Cattle. *Sci. Rep.* **6**, 34114 (2016).

34. Davies, V. J. *et al.* Opa1 deficiency in a mouse model of autosomal dominant optic atrophy impairs mitochondrial morphology, optic nerve structure and visual function. *Hum. Mol. Genet.* **16**, 1307–1318 (2007).

35. Patten, D. A. *et al.* OPA1-dependent cristae modulation is essential for cellular adaptation to metabolic demand. *EMBO J.* **33**, 2676–2691 (2014).

36. Gopalakrishnan, K. *et al.* Targeted disruption of Adamts16 gene in a rat genetic model of hypertension. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20555–20559 (2012).

37. Dong, K. *et al.* Genomic scan reveals loci under altitude adaptation in Tibetan and Dahe pigs. *PLoS One* **9**, e110520 (2014).

38. Strickland, J. R., Aiken, G. E., Spiers, D. E., Fletcher, L. R. & Oliver, J. W. Physiological Basis of Fescue Toxicosis. in *Tall Fescue for the Twenty-first Century* (eds. Fribourg, H. A., Hannaway, D. B. & West, C. P.) 203–227 (American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, 2009).

39. Decker, J. E. Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power. *PLoS Genet.* **11**, e1005621 (2015).

40. Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Res.* **23**, 1089–1096 (2013).

41. Castric, V., Bechsgaard, J., Schierup, M. H. & Vekemans, X. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet.* **4**, e1000168 (2008).

42. Rowan, T. N. *et al.* A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. *Genet. Sel. Evol.* **51**, 77 (2019).

43. Rosen, B. D. *et al.* De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**, (2020).

44. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

45. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443 (2016).

46. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

47. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).

48. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).

49. R Core Team, R. & Others. R: A language and environment for statistical computing. (2013).

50. Wickham, H. ggplot2. *Wiley Interdisciplinary Reviews: Computational* (2011).

51. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

52. PRISM Climate Group. PRISM 30-year Normal Climate Data.

53. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).

54. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

55. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet.* **8**, e1002555 (2012).

56. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).

57. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

58. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

59. Mouse ENCODE Consortium *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* **13**, 418 (2012).

60. Jain, A. & Tuteja, G. TissueEnrich: Tissue-specific gene enrichment analysis. *Bioinformatics* **35**, 1966–1967 (2019).

61. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

62. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression

analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.* **34**, 1420–1431 (2014).

63. Kim, W., Underwood, R. S., Greenwald, I. & Shaye, D. D. OrthoList 2: A New Comparative Genomic Analysis of Human and Caenorhabditis elegans Genes. *Genetics* **210**, 445–461 (2018).
64. Angeles-Albores, D., N Lee, R. Y., Chan, J. & Sternberg, P. W. Tissue enrichment analysis for C. elegans genomics. *BMC Bioinformatics* **17**, 366 (2016).
65. Angeles-Albores, D., Lee, R. Y. N., Chan, J. & Sternberg, P. W. Two new functions in the WormBase Enrichment Suite. microPublication Biology: https://doi. org/10.17912. (2018).