Leveraging GWAS models to map selection on complex traits in massive datasets

T.N. Rowan[1*], R.D. Schnabel[2], J.E. Decker[2]

[1] University of Tennessee Institute of Agriculture, 2506 River Drive, 37996, Knoxville, Tennessee, United States; [2] University of Missouri, 920 East Campus Drive, 65211, Columbia, Missouri, United States; [*]trowan@utk.edu

Abstract

Understanding the biology that underlies artificial selection in animal breeding programs is of significant interest. Recent and ongoing polygenic selection is difficult to map because it does not leave well-defined signatures on the genome. When populations have temporally-stratified genomic data, we can use genome-wide association (GWAS) models to map selection during the sampled period using Generation Proxy Selection Mapping (GPSM). Here, we demonstrate the power of GPSM in two large datasets from American beef cattle populations with 11,759,568 imputed SNPs. We identify ongoing polygenic selection at 294 unique loci. We observe that GPSM has high power to detect minuscule allele frequency shifts over short timescales (< 10 years). GPSM loci localized near genes, in annotated epigenetic marks, or in predicted functional genomic regions. We also observed minimal overlap in loci identified by GPSM and other traditional sweep detection methods.

Introduction

While selection occurs on phenotypes or estimated breeding values, the changes observed in a population result from changes in the frequencies of causal variants that underlie the selected trait. Occasionally, a large-effect causal variant is rapidly driven to high frequency in the population, leading to "selective sweeps" where surrounding neutral genetic variation is reduced (Smith and Haigh 1974). Studies have mapped numerous selective sweeps in livestock species caused by selection on Mendelian traits or large-effect quantitative trait loci (QTL) (Gutiérrez-Gil et al. 2015). Selective sweeps at these loci were likely essential components of domestication and breed formation. That said, it is likely that the majority of selection has been on traits that are highly polygenic (Bolormaa et al. 2014). Polygenic selection results in subtle allele frequency shifts at hundreds or thousands of causal variants of small effect, which does little to alter surrounding neutral diversity. This makes mapping polygenic selection difficult.

In contemporary livestock populations, we have accumulated large genotypic datasets containing thousands or millions of animals, including influential founder animals. We can leverage these large populations with high-density, temporally-stratified genotypes to detect small directional changes to allele frequency caused by polygenic selection. We use Generation Proxy Selection Mapping (GPSM) to map loci actively responding to polygenic selection. GPSM uses a proxy for the generations that separate an individual from the beginning of a population's pedigree as the dependent variable in a GWAS model. The goal is to detect statistically significant associations between allele frequency and an animal's generation. These associations represent directional allele frequency changes caused by selection (Decker et al. 2012; Rowan et al. 2021). Simulations in previous work showed that by using a mixed model, GPSM can effectively distinguish selection from drift regardless of effective population size, skewedness of data, family or population structure, and other confounders. We apply GPSM to datasets from two modern American beef cattle breeds, Red Angus and Simmental. Using 11,759,568 imputed genotypes and conditional joint analysis of GPSM results, we can map

putative sites underlying ongoing selection for complex traits, allowing us to understand better the biology that underlies selection for production traits.

Materials & Methods

Genotype data and imputation. We used assay genotypes from two breed associations, Red Angus (n = 46,454) and Simmental (n = 90,580), ranging in density from 25K to 770K markers. Genotypes were imputed to a set of 811K SNPs, as described in Rowan et al. 2019 (Rowan et al. 2019), followed by imputation to 43,214,290 SNPs using the 4,931 animals in Run8 of the 1000 Bulls Project (Hayes and Daetwyler 2019), phased with Eagle (v2.4) (Loh et al. 2016) as a haplotype reference. We retained only high quality variants (VQSR Tranche 90) and removed poorly imputed (internal imputation $R^2 < 0.4$), or rare (minor allele frequency < 0.01) to arrive at a set of 11,759,568 variants. We used genomic coordinates from the ARS-UCD1.2 assembly (Rosen et al. 2020) for all analyses.

Generation Proxy Selection Mapping (GPSM). We used the breeder-reported birthdate to calculate the continuous years between an animal's birth and the day we began analyzing data (19 October 2020). We used this value as the dependent variable in GPSM analyses. We used the 811K SNP sets to construct genomic relationship matrices (GRM) for each breed. We used the "--mlma" function in GCTA (v 1.92.3) (Yang et al. 2010; Yang et al. 2011) to conduct our GPSM analysis using the following model:

$$\mathbf{y}=\boldsymbol{\mu}+\mathbf{b}\mathbf{x}+\mathbf{a}+\mathbf{e} \quad (1)$$

Here, y is a vector of generation proxies, μ is the sample mean, b is a SNP's regression coefficient, on a vector of animal genotypes x, a is a random vector of polygenic terms, and e is a random error term. We performed this analysis on the full datasets of both breeds, an age-truncated Red Angus dataset (animals born < 10 years ago), and a purebred-only Simmental dataset. To refine GPSM signals, we performed a conditional and joint analysis (COJO) (Yang et al. 2012), conditioning associations on SNPs with GPSM p-values < 10^{-5}. Significant COJO SNPs were ones with conditional and genome-wide p-values < 5×10^{-8}.

Results

In Red Angus, GPSM detected 2,914 and 9,065 genome-wide significant SNPs (Bonferroni-corrected p-values < 4.29×10^{-9}) in the full and birth date-truncated datasets, respectively (**Figure 1A & 1B**). Our COJO analysis of these GPSM results identified 72 unique lead SNPs in the full dataset and 96 in the truncated dataset (COJO SNPs). Sixteen of these SNPs were identified by both Red Angus datasets. The most striking example of the truncated dataset uncovering unique GPSM associations was on BTA2 near the gene *ARHGAP15*, a major immune function gene in cattle (Noyes et al. 2011; Álvarez et al. 2016). Both GPSM analyses in Red Angus identified a known pleiotropic QTL for body weight, carcass weight, and sexual precocity on BTA14. Many significant GPSM loci had not been previously identified by other sweep mapping studies (Gutiérrez-Gil et al. 2015). The increased power in these datasets uncovered dozens of novel signatures that were not identified by previous GPSM analyses in the same populations (Rowan et al. 2021).

In the American Simmental dataset, we performed two GPSM analyses to compare and contrast selection within the purebred Simmental part of the dataset and the entire open herdbook (Animals 5-100% pedigree estimated Simmental). GPSM identified 513 genome-wide significant SNPs in the purebred dataset driven by 18 unique COJO SNPs (**Figure 1D**). The strongest GPSM signature in purebred Simmental resided on BTA5, near the genes *PMEL* and *ERBB3*. This locus is a known modulator of coat color in European Simmental populations

(Mészáros et al. 2015). Another selected locus was identified near the gene *KIT,* which contains alleles that control color-sidedness in many mammalian populations (Durkin et al. 2012). GPSM in the full Simmental dataset identified significantly more selected loci, 1,008 SNPs and 108 unique associations (**Figure 1C**). This analysis also identified strong ongoing selection at the same coat color and pattern loci and identified selection on other loci known to affect economically-important quantitative traits.

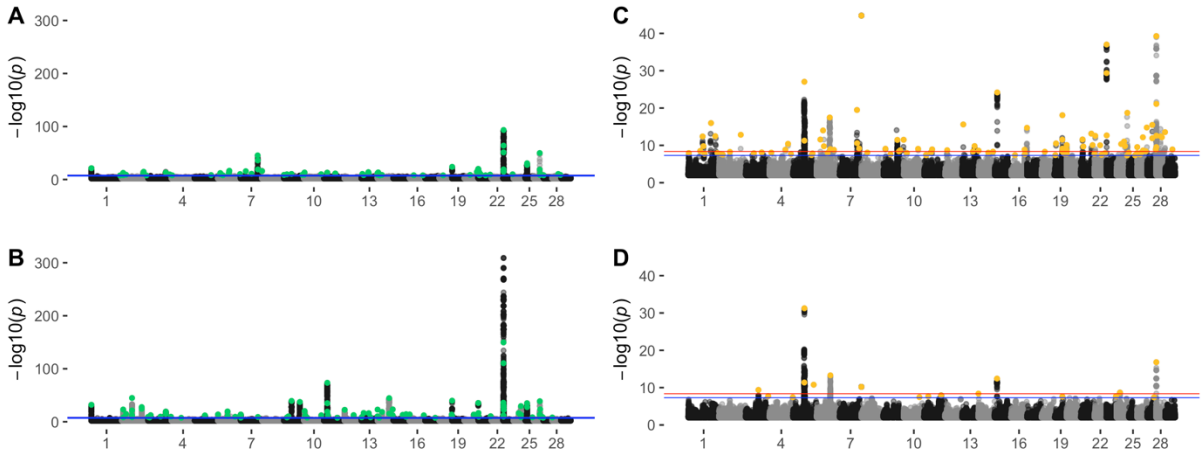

Figure 1. GPSM Manhattan plots for (A) full and (B) young animal truncated Red Angus datasets, and (C) full and (D) purebred Simmental datasets. Colored SNPs were identified as independent associations through COJO analysis. The red line represents Bonferroni significance, and the blue line represents a $5x10^{-8}$ significance threshold.

We observed only a handful of loci that were under selection in both the Red Angus and Simmental breeds. Red Angus's most significant GPSM signature resided on BTA23 (1-2 Mb), immediately upstream of the gene KHDRBS2. This area was also identified in the Simmental GPSM. However, the selected loci in the region were different, suggesting that depending on the population, selection may be acting on different regulatory variations for the same gene. Across datasets, most GPSM COJO SNPs had a clear candidate gene (**Table 1**). Many of the selected SNPs with a clear positional candidate gene were immediately upstream of the transcription start sites, suggesting possible selection on *cis*-regulatory variation.

Table 1. Significant GPSM (COJO and genome-wide $p < 5 \times 10^{-8}$) SNPs and their proximity to annotated genes.

Breed	Subset	COJO SNPs	In Gene	Near Gene (< 50kb to nearest)	Intergenic
Red Angus	Full	72	17	20	35
Red Angus	Young	96	36	24	36
Simmental	Full	108	32	36	40
Simmental	Purebred	18	6	7	3

Discussion

Our results show that GPSM, coupled with high-powered livestock datasets can detect very subtle shifts in allele frequency over short time periods, even in the absence of recorded phenotypes. Using a linear mixed model to test allelic associations with a generation, we can explicitly control the population and family structure, which often obscures selection mapping

studies. Simulations in previous studies (Rowan et al. 2021) indicate that GPSM can effectively distinguish selection from drift and is robust to genotype sampling. The combination of GPSM and sequence-density imputed genotypes lets us map this selection at an extremely high resolution.

Our results suggested that since the initial import of Simmental to America, much selection pressure within the purebred animals has made them appear more similar to Angus animals (solid black-hided). The large sample sizes of our two datasets allowed us to divide them into informative subsets to understand how selection operates at very short timescales (Red Angus "young" animals) or within a specific subset of an open herdbook breed (purebred Simmental). With large or more structured datasets, this subsetted approach to performing GPSM could be informative in understanding how changes to breeding decision-making have impacted the genome over time in a line or family-specific manner (e.g. genomic changes in terminal vs. maternal lines). Breeders might also use GPSM results as "biological priors" in Bayesian genomic prediction approaches. Further, we believe that using putatively selected loci as "functional annotations" to the genome can add important context to other mapping studies (GWAS, eQTL, etc.).

References

Alachiotis N., and Pavlidis P. (2018) Communications Biology 1: 79. doi:10.1038/s42003-018-0085-8

Bolormaa S., Pryce J.E., Reverter A., Zhang Y., Barendse, *et al.* (2014) PLoS Genetics 10: e1004198. doi:10.1371/journal.pgen.1004198

Decker J.E., Vasco D.A., McKay S.D., McClure M.C., Rolf M.M., *et al.* (2012) BMC Genomics 13: 606. doi:10.1186/1471-2164-13-606

Ferrer-Admetlla A., Liang M., Korneliussen T., and Nielsen R. (2014) Molecular Biology Evolution 31: 1275–1291. doi:10.1093/molbev/msu077

Gutiérrez-Gil B., Arranz J.J., and Wiener P. (2015) Frontiers Genetics 6: 167. doi:10.3389/fgene.2015.00167

Hayes B.J. and Daetwyler H.D. (2019) Annual Reviews Animal Bioscience 7: 89–102. doi:10.1146/annurev-animal-020518-115024

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., *et al*. (2016). Nature Genetics, 48(11), 1443–1448. https://doi.org/10.1038/ng.3679

Rowan T.N., Hoff J.L., Crum T.E., Taylor J.F., Schnabel R.D., *et al.* (2019) Genetics Selection Evolution 51: 77. doi:10.1186/s12711-019-0519-x

Rowan T.N., Durbin H.J., Seabury C.M., Schnabel R.D., and Decker J.E. (2021) PLoS Genetics 17: e1009652. doi:10.1371/journal.pgen.1009652

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G, *et al*. (2020) Gigascience 9. doi:10.1093/gigascience/giaa021

Smith J.M., Haigh J. (1974) Genetics Research 23: 23–35. doi:10.1017/S0016672300014634

Szpiech Z.A. and Hernandez R.D.(2014) Molecular Biology Evolution 31: 2824–2827. doi:10.1093/molbev/msu211

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., *et al.* (2010) Nature Genetics 2010;42: 565–569. doi:10.1038/ng.608

Yang J., Lee S.H., Goddard M.E., and Visscher P.M. (2011) American Journal of Human Genetics 88: 76–82. doi:10.1016/j.ajhg.2010.11.011

Yang J, Ferreira T, Morris AP, Medland SE, *et al.* (2012) 44: 369–75, S1–3. doi:10.1038/ng.2213