van Elst, T., Leonard, A., Rosen, B., Smith, T., Low, W., Pausch, H. (2026) The diversity and evolutionary history of structural variation in European cattle genomes, *Proceedings 13th World Congress on Genetics Applied to Livestock Production*
23608: https://doi.org/10.31274/wcgalp.23608 ©2026 van Elst, et al. CC BY-NC-ND

The diversity and evolutionary history of structural variation in European cattle genomes

T. van Elst[1,*], A.S. Leonard[1], B.D. Rosen[2], T.P.L. Smith[3], W.Y. Low[4], and H. Pausch[1], on behalf of the Bovine Pangenome Consortium

[1]ETH Zürich, Zürich, Switzerland; [2]U.S. Department of Agriculture, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD, USA; [3]U.S. Department of Agriculture, Agricultural Research Service, U.S. Meat Animal Research Center, Clay Center, NE, USA; [4]Davies Livestock Research Centre, University of Adelaide, Roseworthy, SA, Australia; [*]Corresponding author: tobias.vanelst@usys.ethz.ch

Introduction

Cattle (*Bos taurus*) are one of the most economically important livestock species worldwide. The over 1,000 different breeds descend from two main lineages, taurine and indicine, which originated from independent domestication events of divergent aurochs populations (*Bos primigenius*; now extinct) in the Fertile Crescent and the Indus Valley approximately 10,000 and 8,000 years ago, respectively (Ajmone-Marsan *et al*., 2010). Because cattle lineages remained interfertile and can hybridize with other species within the genera *Bos* and *Bison*, their evolutionary history is characterized by extensive gene flow. This has been documented both among taurine and indicine lineages and between cattle and (semi-)wild bovines (Decker *et al*., 2014; Wu *et al*., 2018). For instance, genetic analyses have identified historic gene flow between European taurines and African taurines, aurochs and wisent (*Bison bonasus*) (Gautier *et al*., 2016; Park *et al*., 2015; Upadhyay *et al*., 2019).

Most insights into these processes have been obtained from analyses of single nucleotide polymorphisms (SNPs) in phylogenetic and population-genetic frameworks. Only relatively recently have studies in cattle begun to investigate the functional consequences of gene flow by identifying introgressed segments and their association with traits (e.g., Chen *et al*., 2023; Wu *et al*., 2018). However, due to technical limitations, such work has predominantly relied on short-read sequencing data and has therefore been limited in its ability to investigate structural variants (SVs; >50 bp), despite their potentially strong effects on genome evolution and gene regulation (Bickhart & Liu, 2014). Advances in long-read sequencing, the increasing availability of high-quality cattle assemblies, and the development of pangenome approaches now enable comprehensive analyses of SVs. Recent studies have demonstrated the power of cattle pangenomes to characterize SV diversity and identify introgressed SVs, particularly in Asian cattle populations (e.g., Dai *et al*., 2023). However, despite the well-documented admixture histories in European cattle lineages, the evolutionary origins of structural variation in those lineages remain largely unexplored.

This study uses a pangenome approach to characterize SVs and their evolutionary history at population scale across European cattle breeds. It traces the origins of trait-associated SVs, and assesses how gene flow with African cattle, aurochs, and wisent, along with incomplete lineage sorting and selection, have shaped the contemporary SV landscape.

Materials & Methods

Genome assemblies. We obtained 53 genome assemblies representing 27 European taurine, two African taurine and three indicine cattle breeds (mean NG50: 74.1 Mb (SD = 23.3 Mb); mean universal single-copy ortholog completeness: 97.9% (SD = 1.3%)), of which 51 were contributed by the Bovine Pangenome Consortium (BPC) and two were previously published.

Whole-genome sequencing data. Short-read whole genome sequencing data were obtained from public databases and supplemented with in-house data. This resulted in a set of 1,075 cattle samples across 80 European taurine breeds, two African taurine breeds, three indicine breeds as well as wisent, American bison (*Bison bison*), water buffalo (*Bubalus bubalis*) and the extinct aurochs. On average, each breed/taxon was represented by 12.1 (± 9.3) samples. Raw reads were trimmed with fastp v1.0.1 ("-q 15 -u 40 -g -l 15 -2"). Reads were then aligned against ARS-UCD2.0 with strobealign v0.16.1, PCR duplicates were marked with Samtools v1.22, and read depth was estimated with mosdepth 0.3.11. Short-read data were subsequently downsampled to a target depth of 30× for samples exceeding this depth. Final mean sequencing depth was 16.25x (SD = 8.0x).

Pangenome creation. Assemblies were integrated into a pangenome with Minigraph-Cactus v2.9.9, excluding sex chromosomes and using ARS-UCD2.0 as the reference backbone. Graph quality was assessed with gretl v0.1.1 and odgi v0.9.0. The graph was exported to VCF format with the deconstruct function of vg v1.66.0.

SV genotyping and small variant calling. Trimmed reads were mapped against the graph following the personalized pangenome reference approach of Sirén *et al*. (2024). Resulting GAM files were filtered with vg filter ("-r 0.9 -fu -m 1 -q 25 -D 999"), converted to CRAM format with vg surject and deduplicated with Samtools.
Variants identified in the pangenome graph were genotyped across short-read samples running pangenie v4.2.1 on filtered reads. Variants were simplified with vcfwave, and truvari collapse v5.3.0 was used to collapse SVs with ≥80% overlap and ≥50% sequence identity. Finally, SVs (≥50 bp) were extracted with BCFtools.
To call small variants (SNPs, indels < 50 bp), CRAM files were input to pangenome-aware deepvariant v1.9.0. ARS-UCD2.0 was used as the reference and the graph was supplied as the pangenome. Per-sample VCF files were merged with GLnexus v1.4.1. Subsequently, genotypes with a per sample depth smaller than 5 were masked and sites with a mean depth larger than the mean depth across sites plus three standard deviations were removed. Duplicate or closely related individuals were excluded from both data sets following Waples *et al*. (2019).

Population structure. Principal component analysis was conducted on biallelic small and structural variants using Plink v2.0.0, thinning the data based on linkage disequilibrium ("--indep-pairwise 50 5 0.5") and excluding variants and samples with more than 50% and 75% missing data, respectively. Two data sets were considered: (1) all *Bison* spp. and *Bos* spp. samples, and (2) *B. taurus taurus* and *B. primigenius* samples only.

Results & Discussion

The pangenome graph consisted of 118.5 million nodes connected by 161.9 million edges, resulting in a total graph size of 2.65 Gb. It identified 130,112 SVs, accounting for 133.5 Mb new sequence compared to the reference genome. We successfully genotyped 87,510 of these across our short read sample cohort with a missingness of 39.8% (SD_{site} = 36.3%; SD_{sample} = 8.4%). Small variant calling resulted in a set of 126,035,461 variants, with a missingness of 13.3% (SD_{site} = 7.7%; SD_{sample} = 24.9%).
PCA based on small variants and SVs recovered concordant patterns of population structure with similar proportions of variance explained (Fig. 1), revealing three well-differentiated clusters: (1) *Bos taurus indicus*, (2) B. *taurus taurus* together with European *B. primigenius*, and (3) *Bison* spp (Fig. 1A, C). Taurine cattle were separated into three clusters as well,

representing African taurine, Podolian and European taurine cattle (including aurochs; Fig. 1 B, D). Iberian cattle were positioned closer to African taurine cattle than to other breeds, supporting the previously reported gene flow among them. In addition, aurochs clustered much more closely with European taurines than with the other two groups, potentially indicative of historical gene flow as well. Overall, the concordance between small variant- and SV-based PCA indicates that the inferred SV catalogue captures evolutionary signatures similar to those observed in small-variant data.

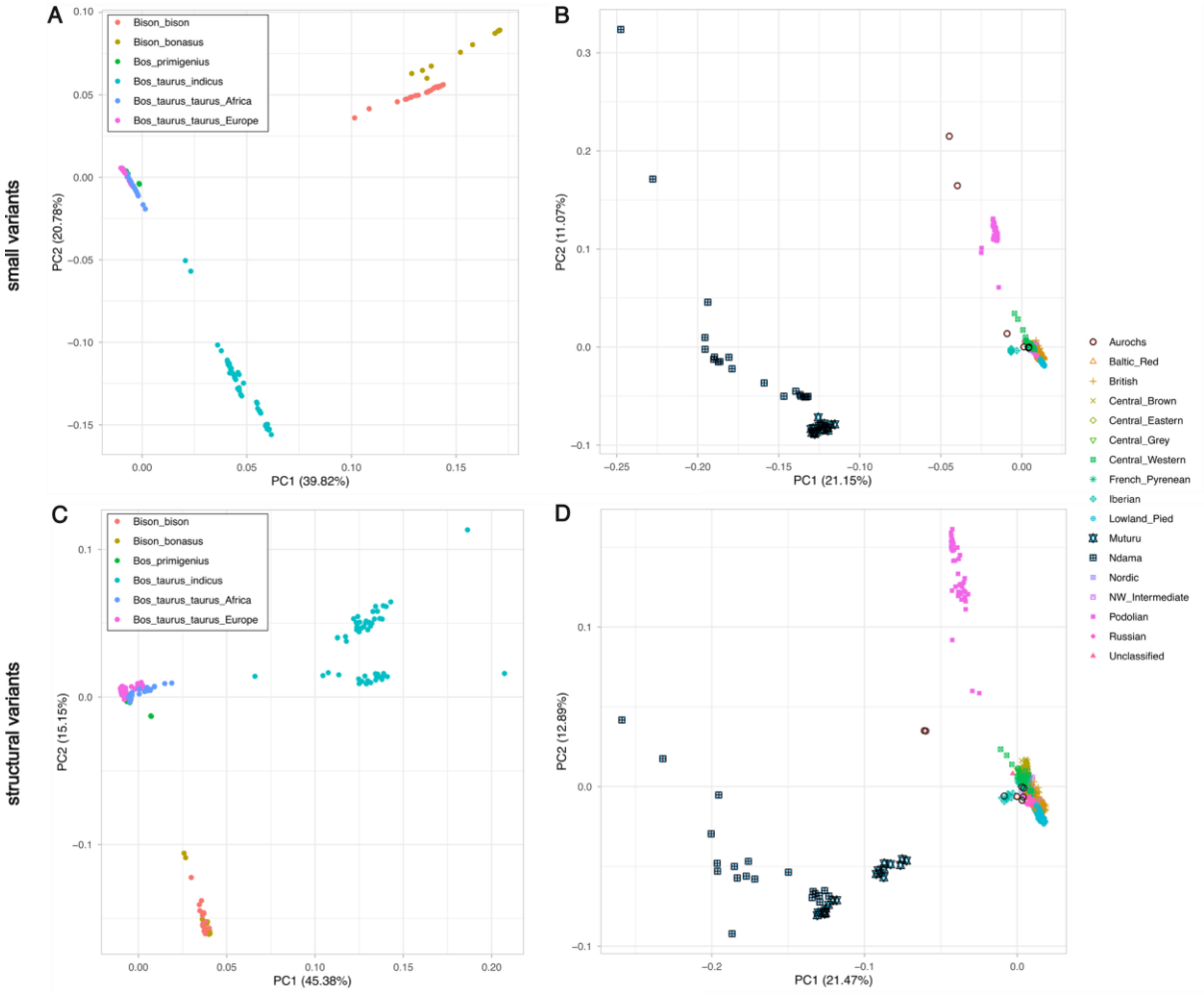

Figure 1: Principal components (PCs) inferred from small and structural variants, including *Bison* spp. and *Bos* spp. samples (panels A, C), or *B. taurus taurus* and *B. primigenius* samples (panels B, D). In panels B and D, European cattle breeds were classified according to Felius *et al*. (2011) for clarity.

Conclusions

We inferred a pangenome for European cattle breeds, enabling the characterization of SVs at population scale. This SV catalogue provides a foundation for investigating the evolutionary history of European cattle, moving beyond SNP-based analyses toward a genome-wide understanding of structural genetic variation and the evolutionary forces shaping it. The recovered SV-based population structure is largely consistent with patterns inferred from SNP data, indicating that our dataset is well suited to capture evolutionary signals in structural variation. Next steps will include more detailed classification of SVs, analyses of population

structure, comparative assessments of introgression based on SNPs and SVs, and targeted investigations of gene flow among aurochs, wisent, and domestic cattle, as well as between African taurine and southern European cattle. Additional analyses will focus on identifying introgressed structural variants, detecting signatures of selection (adaptive introgression), and evaluating the overlap between introgressed and/or selected SVs and known quantitative trait loci.

Acknowledgements

We would like to thank all members of the Bovine Pangenome Consortium for contributing genome assemblies.

References

Ajmone-Marsan P., Garcia J.F., and Lenstra J.A. (2010) Evol. Anthropol. Issues News Rev. 19:148–157. https://doi.org/10.1002/EVAN.20267

Bickhart D.M., and Liu G E. (2014) Front. Genet. 5:77142. https://doi.org/10.3389/FGENE.2014.00037/PDF

Chen N., Xia X., Hanif Q., *et al*. (2023) Nature Comm. 14:1–13. https://doi.org/10.1038/s41467-023-43626-z

Dai X., Bian P., Hu D., *et al*. (2023) Genome Res. 33:1284–1298. https://doi.org/10.1101/GR.277481.122

Decker J.E., McKay S.D., Rolf M.M., *et al*. (2014) PLOS Genet. 10:e1004254. https://doi.org/10.1371/JOURNAL.PGEN.1004254

Felius M., Koolmees P.A., Theunissen B., *et al.* (2011) Diversity 3:60–692. https://doi.org/10.3390/D3040660

Gautier M., Moazami-Goudarzi K., Levéziel H., *et al*. (2016) Mol. Biol. Evol. 33:2801–2814. https://doi.org/10.1093/MOLBEV/MSW144

Park S.D.E., Magee D.A., McGettigan P.A., *et al*. (2015) Genome Biol. 16:1–15. https://doi.org/10.1186/S13059-015-0790-2/FIGURES/6

Sirén J., Eskandar P., Ungaro M.T., *et al*. (2024) Nat. Methods 21:2017–2023. https://doi.org/10.1038/s41592-024-02407-2

Upadhyay M., Bortoluzzi C., Barbato M., *et al*. (2019) Evol. Appl. 12:951–963. https://doi.org/10.1111/EVA.12770

Waples R.K., Albrechtsen A., and Moltke I. (2019) Mol. Ecol. 28:35–48. https://doi.org/10.1111/mec.14954

Wu D.D., Ding X.D., Wang S., *et al*. (2018) Nat. Ecol. Evol. 2:1139–1145. https://doi.org/10.1038/s41559-018-0562-y