Oian, A., Mancin, E., Gomez Proto, G., Rulli, E., Sartori, C., Mantovani R. (2026) Insertion–deletion variation in candidate fitness genes of a local cattle breed, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23489: https://doi.org/10.31274/wcgalp.23489 ©2026 Oian, et al. CC BY-NC-ND

Insertion–deletion variation in candidate fitness genes of a local cattle breed

A. Oian[1*], E. Mancin[1], G. Gomez Proto[1], E. Rulli[1], C. Sartori[1], and R. Mantovani[1]

[1]*University of Padova, Viale dell'Università 16, – 35020 Legnaro (PD), Italy;* [*]*Corresponding author: angelica.oian@studenti.unipd.it*

Introduction

Genetic variation plays a central role in determining fitness-related traits in livestock, including body size, fertility, immune function, and environmental adaptability. In cattle, these traits are often influenced by a combination of major genes and polygenic effects, collectively referred to as "fitness genes". Such genes contribute not only to production efficiency but also to resilience and longevity, making them critical targets for breeding programs.

Previous studies have identified a range of candidate genes associated with fitness traits. For example, PLAG1, LCORL, and NCAPG have been linked to body size and growth (Karim *et al*., 2011), PPARGC1A to metabolic efficiency and reproductive performance (Komisarek and Walendowska, 2012; Kowalewska-Łuczak and Kulig, 2013), and BoLA to immune response and disease resistance (Gowane *et al*., 2013; Rupp *et al*., 2007) . Other genes such as SOD1 and SOD2 engage in oxidative stress defence, while HSPA13, ANXA1, and TGFBR1 play roles in stress response and cellular signalling. The selection of these candidate genes allows focused investigation of loci with known or suspected functional relevance to fitness.

While single nucleotide polymorphisms (SNPs) have been extensively studied, insertions and deletions (InDels) represent an additional layer of genetic variation that can profoundly affect gene function. InDels can induce frameshift mutations, disrupt splice sites, or modify regulatory sequences, thereby potentially altering protein structure and function (Mullaney *et al*., 2010). Despite their functional importance, InDels are often underrepresented in genomic analyses.

In this study, we focused on a panel of 23 candidate genes (ANGPT1, PPARGC1A, GBE1, SOD1, HSPA13, OXCT1, PLAG1, LCORL, NCAPG, HMGA2, XKR4, FGF6, SOD2, TGFBR1, CSPP1, ANXA1, PCSK5, JMJD1C, NOD1, ITGB2, IRAK3, MZB1, and BoLA) previously implicated in growth, metabolic, immune, and reproductive traits. Using whole-genome sequencing data, we aimed to identify InDels within these genes and assess their potential functional impact, providing insights into allelic variation that may contribute to fitness-related traits in cattle.

Materials & Methods

Sample collection and library preparation. Biological material for genotyping was collected from hair samples containing a minimum of 30 follicles, provided by A.N.A.G.A. (Associazione Nazionale Allevatori Bovini di Razza Grigio Alpina), from young candidate bulls. In total, genomic data was available for 19 Alpine Grey (AG) unrelated animals. Library preparation and whole genome sequencing was conducted by Neogen corporation using Cattle Whole Genome Sequencing with 150 bp PE at an 10x estimated coverage (Neogen Corporation, Lansing, MI). After quality control and data processing, the mean coverage per sample was 7.47×.

Variant calling and quality control. Raw paired-end reads were processed using a standardized bioinformatic pipeline. Read quality was assessed with FastQC (Wingett and

Andrews, 2018), followed by adapter removal and quality trimming using fastp (Chen *et al*., 2018) (Phred≥30, minimum length 30 bp, automatic adapter detection). The first 20 bases of each read were trimmed, and post-processing quality was re-evaluated with FastQC. High-quality reads were aligned to *Bos taurus* reference genome ARS-UCD1.2 using BWA-MEM (Li, 2013). Alignments were converted to BAM, sorted, and processed with samtools (Danecek *et al*., 2021) to fix mate information, mark duplicates, and remove reads with mapping quality < 25. Alignment quality and coverage were evaluated with Qualimap (Okonechnikov *et al*., 2016). Variant calling was ran using GATK HaplotypeCaller (McKenna *et al*., 2010) in GVCF mode, followed by genotyping with CombineGVCFs and GenotypeGVCFs. Variants were filtered using GATK VariantFiltration (QD<2.0, FS>60.0, MQ<40.0) to keep high-confidence SNPs.

Variant filtering and functional annotation. Joint genotyping was performed across all samples using the GATK (v4.5.0.0) joint calling workflow. Briefly, individual GVCF files were combined using GenomicsDBImport and jointly genotyped with GenotypeGVCFs to produce a multi-sample VCF. Variant-level filtering was initially applied using VariantFiltration with standard hard-filtering thresholds (QD < 2.0, FS > 60.0, MQ < 40.0).
The resulting joint-genotyped variants were further filtered using BCFtools v1.19 (Danecek *et al*., 2021) to retain autosomal and X-chromosome InDels (≤ 50 bp). Low-confidence variants were excluded by applying a QUAL > 30 and a minimum read depth (DP) > 10. The final VCF file was compressed and indexed for downstream analyses. Functional annotation was performed using SnpEff (Cingolani *et al*., 2012), annotating variants against the Bos taurus database to predict their effects on coding and regulatory regions.

Indel extraction and genotype counting. The annotated VCF was processed in R using the vcfR package to extract InDels (Knaus and Grünwald, 2017), with particular focus on variants predicted to have high impact. For each variant, the first annotation in the ANN field was retained and used to extract the associated gene. Genotype calls were retrieved for all samples, excluding additional format information, and organized into a data frame including genomic coordinates, alleles, gene annotation, and genotypes. InDels mapping to a predefined set of candidate genes related to stress response, immunity, and metabolism were selected. For each variant, counts of homozygous reference and alternate genotypes were calculated, excluding missing data, and the resulting tables were exported as CSV files for downstream analyses.

Results & Discussion

Variant annotation summary. Whole-genome sequencing of 19 *Bos taurus* individuals yielded a total of 19,074,015 variants, including 1,521,987 insertions and 1,516,568 deletions. Functional annotation with SnpEff revealed that the vast majority of variants (~99%) were classified as MODIFIER, with only 9,762 variants (0.033%) predicted to have HIGH impact.
Focusing on InDels with HIGH impact, a total of 4,441 variants were identified. The most frequent HIGH-impact indel type was frameshift_variant (4,043), followed by intron_variant (423) and splice region variants, including splice_donor (255) and splice_acceptor (224). Less frequent categories included stop_gained, stop_lost, and start_lost variants, whereas inframe InDels and UTR-associated variants were minimal. These results indicate that only a small fraction of the total indel set is predicted to have a HIGH functional impact, with frameshift variants predominating, consistent with their potential to disrupt coding sequences.

InDels extraction and genotype counting. We investigated predicted high-impact InDels within a subset of candidate genes previously associated with fitness traits to identify variants warranting further investigation. For each variant, the number of individuals homozygous for

the reference allele and the number of individuals carrying at least one alternate allele were recorded.

Among the high-impact InDels, three genes showed allelic variation: *LCORL*, *PCSK5*, and *BoLA*. In the *LCORL* gene, encoding a transcription factor acting in body size, height, and skeletal frame regulation, a single indel at chromosome 6 (position 37,401,770) was observed in one individual homozygous for the reference allele and 18 individuals carrying at least one alternate allele. In *PCSK5*, encoding a protease in the secretory pathway converting inactive proproteins into their mature forms, one indel on chromosome 8 (position 52,270,973) was noted in four individuals homozygous for the reference allele and one individual carrying at least one alternate allele.

BoLA (Bovine Leukocyte Antigen) is a gene cluster constituting the major histocompatibility complex of cattle, essential in disease susceptibility and resistance. The *BoLA* region on chromosome 23 displayed multiple high-impact InDels with variable allele counts. Notable variants included an indel at position 27,829,895 (2 homozygous reference, 12 individuals carrying at least one alternate allele), along with additional InDels with alternate allele counts ranging from 1 to 8 individuals. This pattern suggests substantial allelic variability within the region, whereas LCORL and PCSK5 exhibited more limited variation in our dataset.

A more detailed analysis focusing on the BoLA-DRB3 gene revealed a diverse spectrum of predicted functional effects (Fig. 1). The most frequent variant type was missense_variant (n = 55), suggesting potential amino acid changes that may affect protein function. High-impact frameshift_variants (n = 14) were also observed, likely to disrupt coding sequences. Variants in untranslated regions were also detected, including 3′ UTR (n = 47) and 5′ UTR variants (n = 3), which may influence gene regulation. Rare variant types included disruptive_inframe_insertion (n = 1) and 5′ UTR premature start codon gain variants (n = 2).

Overall, these results highlight the presence of multiple potentially functional variants within BoLA-DRB3. However, as these are based on predicted effects, further validation is required to assess their actual impact on gene function and their potential association with fitness-related traits.

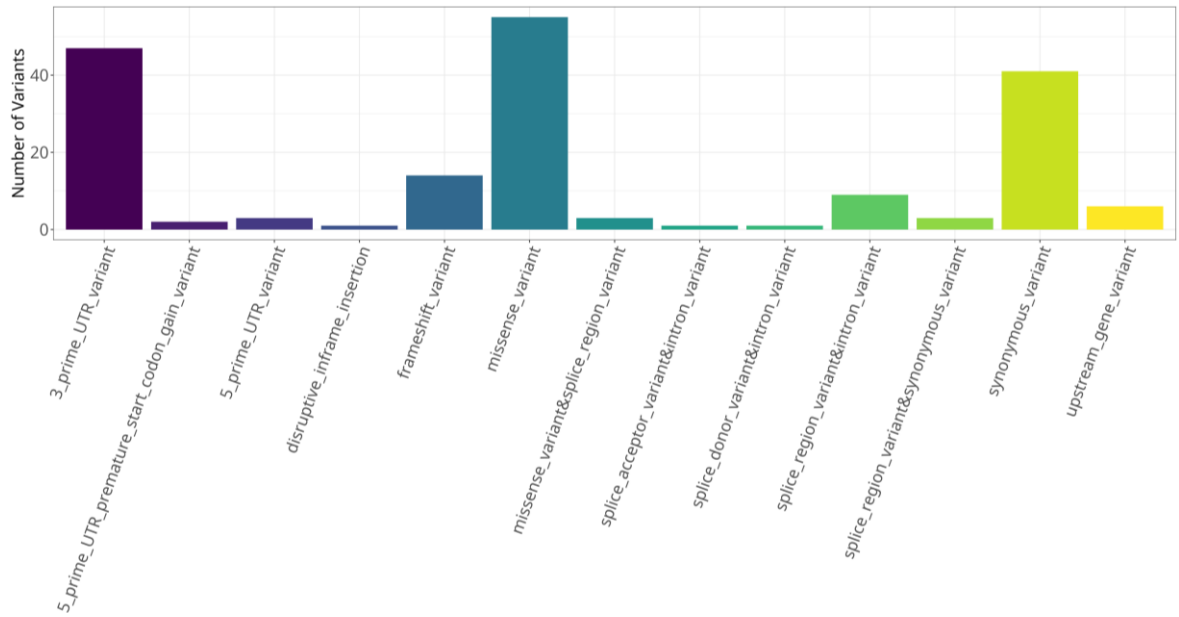

Figure 1. Distribution of variant effects in BoLA-DRB3 gene.

Conclusions

This study provides a comprehensive characterization of insertion and deletion polymorphisms within fitness-related candidate genes in Alpine Grey cattle using whole-genome sequencing data. Although InDels represented a substantial fraction of the total variant set, only a small

proportion was predicted to have high functional impact, with frameshift variants being the predominant category. This suggests that while many variants may be neutral, both coding high-impact variants and regulatory variants could exert significant biological effects. Among the 23 target candidate genes, only LCORL, PCSK5, and BoLA harbored high-impact InDels with observable allelic variation. The near fixation of the *LCORL* alternative allele suggests either strong historical selection or genetic drift, whereas the low variability in *PCSK5* indicates a more conserved genomic region. In contrast, *BoLA* exhibited marked allelic diversity and a wide range of predicted functional effects, including missense and frameshift variants, reflecting its role in immune response and adaptive variation. The diversity and functional relevance of *BoLA* InDels supports the gene key importance in immune-related fitness traits, while the identification of high-impact InDels in *LCORL* and *PCSK5* highlights loci potentially involved in growth and metabolic regulation. Overall, these findings suggest that InDels represent a potentially important, yet still underexplored, source of genetic variation in cattle. Although some variants were predicted to have functional effects, further studies integrating phenotypic data, larger sample sizes, and functional validation are needed to confirm their biological relevance. This work provides a preliminary basis for including InDels in genomic studies of fitness-related traits and for improving the understanding of resilience and adaptability in local cattle breeds.

Acknowledgements

This work received funding from the European Union's Horizon Europe Project 101136346 EUPAHW and Agritech National Research Center and European Union Next-Generation EU (PNRR, Missione 4 Componente 2, Investimento 1.4–D.D. 1032 17/06/2022, CN00000022). It reflects only Authors' views: European Union and European Commission are not responsible.

References

Chen, S., Zhou, Y., Chen, Y., *et al*. (2018) *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cingolani, P., Platts, A., Wang, L. L., *et al*. (2012) *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Danecek, P., Bonfield, J. K., Liddle, J., *et al*. (2021) *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Gowane, G. R., Sharma, A. K., Sankar, M., *et al*. (2013) *Research in Veterinary Science*, *95*(1), 156–163. https://doi.org/10.1016/j.rvsc.2013.03.001

Karim, L., Takeda, H., Lin, L., *et al*. (2011) *Nature Genetics*, *43*(5), 405–415. https://doi.org/10.1038/ng.814

Knaus, B. J., and Grünwald, N. J. (2017) *Molecular Ecology Resources*, *17*(1), 44–53. https://doi.org/10.1111/1755-0998.12549

Komisarek, J., and Walendowska, A. (2012) *Folia Biologica (Czech Republic)*, *60*(3–4), 171–174. https://doi.org/10.3409/fb60_3-4.171-174

Kowalewska-Łuczak, I., and Kulig, H. (2013) *Research in Veterinary Science*, *94*(2), 252–255. https://doi.org/10.1016/j.rvsc.2012.08.006

Li, H. (2013) http://arxiv.org/abs/1303.3997

McKenna, A., Hanna, M., Banks, E., *et al*. (2010) *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mullaney, J. M., Mills, R. E., Stephen Pittard, W., *et al*. (2010) *Human Molecular Genetics*, *19*(R2). https://doi.org/10.1093/hmg/ddq400

Neogen Corporation. 2023 Lansing, MI, USA.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016) *Bioinformatics*, *32*(2), 292–294. https://doi.org/10.1093/bioinformatics/btv566

Rupp, R., Hernandez, A., and Mallard, B. A. (2007) *Journal of Dairy Science*, *90*(2), 1029–1038. https://doi.org/10.3168/jds.S0022-0302(07)71589-8

Wingett, S. W., and Andrews, S. (2018) *F1000Research*, *7*. https://doi.org/10.12688/f1000research.15931.1