

Machine Learning–Driven Discovery of Host Genetic Factors for Paratuberculosis in Goats Within the One Health Framework

Yalçın Yaman

yalcingaman@gmail.com

Siirt University

Ahmet ESER

Siirt University

Devran Coşkun

Siirt University

Ramazan Aymaz

General Directorate of Agricultural Research and Policies

Yiğit Emir Kişi

General Directorate of Agricultural Research and Policies

Murat Keleş

General Directorate of Agricultural Research and Policies

Serdar Yağcı

General Directorate of Agricultural Research and Policies

Özgül Gülaydın

Siirt University

Serkan Süleyman Şengül

General Directorate of Agricultural Research and Policies

Kıvanç İrak

Siirt University

Memiş Bolacalı

Ahi Evran University

Research Article

Keywords: Paratuberculosis, Johne's disease, Machine learning–based GWAS, Host genetic resistance, Caprine genomics, One Health

Posted Date: January 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8537890/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Veterinary Research on April 7th, 2026. See the published version at <https://doi.org/10.1186/s12917-026-05430-x>.

Machine Learning–Driven Discovery of Host Genetic Factors for Paratuberculosis in Goats Within the One Health Framework

Yalçın YAMAN¹, Ahmet ESER¹, Devran COŞKUN¹, Ramazan AYMAZ², Yiğit Emir KİŞİ¹, Murat KELEŞ², Serdar YAĞCI³, Özgül GÜLAYDIN¹, Serkan Süleyman ŞENGÜL², Kıvanç IRAK¹, Memiş BOLACALI⁴

¹*Faculty of Veterinary Medicine, Siirt University, TR-56100, Siirt – TÜRKİYE*

²*Department of Breeding Techniques, Sheep Breeding Research Institute, TR-10200 Bandırma, Balıkesir – TÜRKİYE*

³*General Directorate of Agricultural Research and Policies, Department of Livestock and Aquatic Products, TR-06000 Ankara – TÜRKİYE*

⁴*Department of Biostatistics and Medical Informatics, Faculty of Medicine, Kırşehir Ahi Evran University, TR-40100, Kırşehir – TÜRKİYE*

*** Correspondance:**

Yalçın YAMAN: yalcinyaman@gmail.com
yalcin.yaman@siirt.edu.tr

Abstract

Paratuberculosis, caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP), remains a persistent One Health concern due to its slow clinical course, environmental resilience, wide circulation in ruminant systems, and unresolved zoonotic implications. To characterise MAP exposure across Türkiye's goat populations, we conducted a nationwide genomic survey encompassing seven breeds from 36 farms in 11 provinces. High-density SNP genotyping combined with mutual-information-based feature preselection retained informative, non-redundant loci capturing both linear and nonlinear components of disease architecture.

Nine complementary machine-learning models were applied to identify host genetic factors underlying MAP infection, and an ensemble importance framework resolved 31 FDR-controlled SNPs consistently associated with MAP status. Functional annotation implicated immune processes including cytokine–receptor signalling, antigen presentation, glycan-mediated T-cell regulation, and NF- κ B-linked inflammation. Concordance with mixed linear models and genome-wide McNemar tests suggested that both additive and non-additive genetic effects shape the observed signal. These reproducible, albeit preliminary, markers outline a genomic foundation for breeding MAP-resilient goats and point to opportunities for reducing pathogen shedding at its source within a broader One Health strategy.

Keywords: Paratuberculosis, Johne's disease, Machine learning-based GWAS, Host genetic resistance, Caprine genomics, One Health

Introduction

Mycobacterium avium subsp. *paratuberculosis* (MAP), the causative agent of paratuberculosis (Johne's disease) in ruminants, constitutes a critical One Health concern with direct implications for animal productivity, human health, and environmental safety. Infected hosts often remain asymptomatic for prolonged periods while persistently shedding MAP in milk, faeces, and pasture environments, enabling quiet but extensive contamination of farms and facilitating entry of the pathogen into the human food chain. This covert circulation contributes to substantial economic losses—particularly in regions characterised by dense ruminant populations and suboptimal control programmes—and heightens zoonotic concern amid growing interest in potential associations with chronic inflammatory disease in humans. Its environmental persistence, transboundary dissemination, and systematic under-diagnosis collectively pose escalating threats to public health, food security, and sustainable livestock

systems, underscoring the need for integrated One Health strategies that unite animal surveillance, human disease monitoring, and environmental stewardship [1].

MAP pathogenesis reflects a dynamic interplay between host immunity and bacterial virulence factors, with the intestinal mucosa as the principal site of infection and granulomatous inflammation as the hallmark lesion. Clinically, affected ruminants exhibit progressive diarrhoea, weight loss, and eventual cachexia [2]. Diagnosis is hindered by an extended asymptomatic phase and the limited sensitivity of current assays—serology, faecal culture, and PCR—particularly in subclinical carriers [3]. Beyond the veterinary context, public health relevance is amplified by reported associations with Crohn’s disease [4] and autoimmune conditions including type 1 diabetes, rheumatoid arthritis, Hashimoto’s thyroiditis, and multiple sclerosis, potentially mediated through molecular mimicry [1, 5].

Globally, MAP ranks second only to coccidiosis among food-animal diseases in terms of projected human health burden [6]. Asymptomatic shedding throughout the life of infected animals accelerates its global spread [7]. The organism’s spore-like phenotype enables survival for months in barns, slurry, and surface water, and even persistence through pasteurisation, thereby contaminating bulk milk, dairy products, and infant formula [5]. Transmission through insects has also been documented [8].

The disease exhibits substantial geographic variation in prevalence; a comprehensive review across 48 countries revealed that in approximately half of the countries with available data, more than 20% of herds and flocks were infected with MAP [9]. Regional studies demonstrate marked heterogeneity, with true individual animal seroprevalence reaching 8.4% in sheep and 25.2% in goats in southern Spain, where herd-level seropositivity was detected in 66.3% of sheep herds and 90.0% of goat herds [10]. In Sicily, a large-scale serological survey revealed an overall apparent herd-level prevalence of 71.8% in sheep and 60.8% in goat farms, with animal-level prevalence of 4.5% and 5.1% in sheep and goats, respectively [11]. Systematic reviews from Latin America and the Caribbean reported animal-level prevalence of 16.0% in sheep and 4.3% in goats [12]. Studies from Inner Mongolia, China documented seroprevalence of 18.48% in sheep [13]. Notably, the disease remains particularly underreported in developing countries, especially Africa, where it is considered a neglected disease despite its significant impact on rural communities dependent on small ruminant production [14].

In Türkiye, sheep seroprevalence ranges from 8.3% [15] to 48% [16]. A comprehensive survey of 2,257 animals from seven native and four composite breeds demonstrated marked breed effects, with rates ranging from 0% in Çine Çaparı sheep to 29.4% in Sakız, yielding a population mean of 7.3% and herd-level infection exceeding 90% [17]. Caprine paratuberculosis studies are scarce in Türkiye; a single field survey in Burdur Province reported 24% seropositivity in goats using ELISA [16].

Taken together, MAP is a globally pervasive and environmentally resilient pathogen that undermines livestock productivity, compromises food safety, and presents a growing public health risk. Its insidious subclinical course, continual shedding, and capacity for environmental persistence necessitate coordinated One Health solutions that span the animal–human–environment interface. In the absence of effective treatment and given the uncertain influence of the only licensed vaccine on shedding dynamics, attention has shifted toward long-term control strategies, particularly the identification of genetic determinants of resistance and susceptibility as a route to improving host resilience.

Host genetic resistance to MAP reflects a complex architecture shaped by innate recognition, cytokine regulation, and antigen presentation. In cattle, candidate-gene studies indicate that polymorphisms within pattern-recognition receptors and cytokine-modulating pathways—including *TLR1*, *TLR2*, *TLR4*, and *IL10RA*—alter the balance between inflammatory activation and immune tolerance [18, 19]. Genome-wide studies extend this view, highlighting additional loci linked to inflammasome signalling and interferon-mediated defence [20] and illustrating the polygenic nature of resistance.

In sheep, early evidence implicated *SLC11A1* and MHC-linked regions in modulating susceptibility [21]. Subsequent studies identified functional substitutions within *TLR1*, *TLR2*, and *TLR4* that modify innate recognition of mycobacterial ligands [22]. Genome-wide signals in *PCP4* and *CD109* [23] further suggest that both innate sensing and downstream immunoregulatory processes are integral to disease outcome.

Studies from Türkiye refine these insights by demonstrating that several globally proposed candidates show no association with MAP serostatus in native sheep populations [24, 25]. Conversely, a structurally significant amino-acid replacement within *TLR2* confers clear protection [17]. The absence of this variant in Turkish Saanen goats implies species-specific genetic routes to resistance.

Caprine research remains limited, yet available findings consistently point to antigen-presentation and phagosomal-transport pathways. Variation in *MHC II-DRB* alleles [26] and microsatellite-linked polymorphisms of *SLC11A1* [27] highlight roles for T-cell priming and macrophage microbicidal activity. Preliminary genome-wide signals [28] further support a widely distributed genetic architecture in which resistance emerges from coordinated modulation of innate sensing, intracellular control, and adaptive immunity.

To extend current knowledge, we conducted the largest caprine MAP serosurvey undertaken in Türkiye. A total of 3,372 goats were sampled across four geographic regions—Marmara, Mediterranean, Central Anatolia, and Southeastern Anatolia—spanning 11 provinces and seven widely farmed native breeds (Hairy goat, Honamlı, Angora, Kilis, Turkish Saanen, Maltız, and Damascus). All animals were screened using an indirect ELISA, from which the case and control groups were derived. To minimise confounding, a matched-pair panel of 240 MAP-positive goats and 240 MAP-negative controls ($n = 480$) was constructed through manual matching for farm, herd, breed, age, and sex, thereby controlling for geographic, climatic, and management effects, as well as breed-, age-, and sex-related influences in downstream genetic analyses.

Parallel advances in analytic methodology are reshaping efforts to characterise host genetic resilience to MAP. Conventional GWAS approaches often struggle to detect small-effect variants, fail to capture nonlinear or non-additive relationships within the genome, and have limited capacity to account for higher-order gene–environment interactions [29, 30]. Machine learning (ML) has therefore emerged as a complementary tool, capable of enhanced predictive performance and more effective prioritisation of candidate SNPs associated with susceptibility or resilience [31, 32].

To address these gaps and establish a rigorous foundation for genotype–phenotype inference, we employed a high-resolution genotyping platform (Illumina Goat 64K BeadChip) coupled with stringent quality control and a model-independent Mutual Information (MI) approach to reduce dimensionality while preserving variants informative across linear and non-linear regimes. Building on this foundation, we developed a multi-algorithm machine-learning framework that integrates complementary modelling paradigms—including Linear/Generalised Linear Models (Logistic Regression L2, Elastic Net, Linear SVM), Probabilistic/Generative Models (Naive Bayes), Tree-Based Ensemble Learners (Random Forest, Extra Trees), and Gradient Boosting Algorithms (Gradient Boosting, XGBoost, LightGBM)—to capture a broad

spectrum of genetic signals. Model performance was systematically evaluated using a suite of discrimination and calibration metrics across training, validation, and independent test sets, with overfitting tightly controlled through predefined divergence thresholds.

To identify robust molecular predictors, feature importance was extracted using model-appropriate strategies and consolidated into a unified SNP ranking via weighted averaging, Borda count, and Kemeny–Young aggregation, incorporating persistence, rank stability, and bootstrap-based confidence. All analytical steps were rigorously structured to eliminate data leakage and ensure reproducibility. Together, this integrated framework provides a comprehensive and methodologically resilient approach for resolving the genomic determinants of MAP resistance, offering a principled bridge between statistical learning and biological interpretation.

Materials and Methods

Sampling Strategy and Study Populations

To capture the genetic and serological diversity of goats raised under Türkiye’s heterogeneous husbandry systems, sampling was conducted across 36 farms located in 11 provinces representing four major geographical regions. This design ensured comprehensive coverage of the national goat population. From each of the seven constituent breeds, approximately 450–500 animals were selected. Only goats aged two years or older were included to ensure adequate time for seroconversion following potential MAP exposure. For serological testing, blood was collected from the jugular vein into gel tubes, and for genetic testing, into EDTA-containing tubes. Serum samples were separated immediately. Both whole blood in EDTA tubes and serum samples were placed in a –20 °C deep freezer as soon as possible and stored until laboratory analyses were performed. Authorized permissions were obtained for all sampling procedures. After blood collection, all animals continued their routine management under standard husbandry conditions at their respective farms, and no animals were euthanized at any stage of the study. The distribution of sampled animals is summarised in Table 4.

Table 4: The sampling distribution across breeds, provinces, and geographic regions.

Breeds	<i>n</i>	Male	Female	Farm	Location (Provinces)	Geographic region
Hairy goat	500	17	483	6	Antalya, Burdur, Kahramanmaraş, Mersin	Mediterranean
Honamlı goat	464	11	453	6	Antalya, Burdur	
Damascus	470	22	448	6	Hatay	
Angora goat	486	22	464	6	Ankara, Siirt	Central Anatolia and Southeastern Anatolia
Kilis goat	499	10	489	4	Gaziantep, Kilis	Southeastern Anatolia
Turkish Saanen	475	6	469	4	Çanakkale	Marmara
Maltız goat	478	8	470	4	Kırklareli	Marmara
Total	3.372	96	3.276	36	11	4

Serological Screening

Serological screening for paratuberculosis was performed on 3,372 goat samples using the Indirect Enzyme-Linked Immunosorbent Assay (indirect ELISA) with commercial kits (IDvet, ID Screen Paratuberculosis Indirect Screening Test, Grabels, France), following the manufacturer's protocol. Optical density (OD) was measured at 450 nm, and Sample-to-Positive percentage (SP%) values were calculated as:

$$SP\% = \frac{OD_{\text{sample}} - \overline{OD}_{\text{negative control}}}{\overline{OD}_{\text{positive control}} - \overline{OD}_{\text{negative control}}} \times 100$$

Tests were validated when the ratio of mean negative to mean positive control SP% was $\leq 1/5$. Samples with $SP\% \leq 60\%$ were considered negative, $\geq 70\%$ positive, and 60–70% borderline. Of the total, ELISA failed in three samples, while the remaining 3,369 were successfully screened for paratuberculosis.

Construction of the Case–Control Matched-Pairs Panel

To enable genotype–phenotype association analyses with maximal control over environmental and demographic confounding, a panel of 240 matched case–control pairs was constructed. Matching criteria included farm, herd, breed, age, and sex, such that each MAP-positive

individual was paired with a MAP-negative individual raised under identical conditions. Prior to matching, all seropositive and seronegative samples were ranked based on their SP% values. Strict thresholds were applied to maximise contrast: SP% ≥ 100 for inclusion as positive and SP% ≤ 10 for inclusion as negative. Samples with intermediate or borderline SP% values were excluded to avoid phenotype misclassification. The distribution of matched pairs by breed, herd, and age class is provided in Table 5.

Table 5. Distribution of the matched-pairs panel across breeds, ages, and herds

Breeds	Number of herds	Ages									Total pairs
		2	3	4	5	6	7	8	9	10	
Hairy goat	4	0	3	9	7	7	5	1	0	0	32
Honamlı	4	1	10	14	4	0	1	3	0	0	33
Angora	3	2	2	8	3	2	0	2	0	0	19
Kilis	4	2	17	15	5	1	0	0	0	0	40
Türkish Saanen	3	12	4	12	7	4	0	0	0	0	39
Maltız	4	1	5	17	10	3	0	3	0	1	40
Damascus	6	5	13	8	6	5	0	0	0	0	37
Overall	28	23	54	83	42	22	6	9	0	1	240

Genotyping, Quality Control, and Genotype Imputation

Genotyping was conducted on the Illumina GoatSNP 64K BeadChip, yielding an overall genotyping success rate of 0.999204. Quality control (QC) procedures were performed in PLINK v1.9. Sex chromosomes and SNPs lacking defined genomic positions (“chromosome 0”) were removed, followed by exclusion of duplicated SNP identifiers. SNP- and sample-level QC thresholds included SNP call rate $>95\%$, individual call rate $>90\%$, minor allele frequency (MAF) $>5\%$, and Hardy–Weinberg equilibrium p-value $>1 \times 10^{-5}$. Linkage disequilibrium (LD) pruning was performed at $r^2 < 0.2$ to minimise redundancy. Missing genotypes were imputed using Beagle v5.0. After QC, the dataset consisted of 44,375 SNPs and 474 individuals (237 case–control pairs).

Feature Preselection via Mutual Information

Rationale and Theoretical Basis

Mutual Information (MI) was employed as a model-agnostic, filter-based feature selection strategy to reduce genomic dimensionality while preserving informative variation relevant to phenotype prediction. Because MI quantifies both linear and nonlinear statistical dependencies between SNP genotypes and binary outcomes, it provides an advantage over correlation-based filters in complex genomic landscapes [33, 34]. Moreover, its invariance to monotonic transformations makes MI particularly robust to heterogeneous SNP encoding schemes and non-Gaussian phenotype distributions. Importantly, MI also mitigates redundancy introduced by linkage disequilibrium (LD) by prioritizing tag SNPs with maximal discriminatory capacity [33, 35, 36].

Implementation within Nested Cross-Validation

Feature preselection was performed using the `mutual_info_classif` function from `scikit-learn`, retaining the top ~1% of genome-wide SNPs (500 total). MI computation was embedded within a rigorously structured nested cross-validation (CV) design consisting of 5 outer folds and 4 inner folds, repeated across 10 independent random seeds (50 total test evaluations). All hyperparameter optimisation—including the number of neighbours used in the k-nearest-neighbour MI estimator and the size of the feature subset—was restricted strictly to inner folds to ensure proper separation between model selection and performance estimation.

A k-nearest-neighbour MI estimator was used to capture model-free dependencies between discrete genotype states and the binary phenotype. For each outer fold, optimal hyperparameters derived from the inner loop were used to recompute MI scores on the complete outer-training portion. Final MI rankings were computed by globally aggregating fold-wise scores across all outer folds and seeds using min–max normalisation.

Preservation of Matched-Pair Integrity

All data partitioning procedures respected the structure of case–control matched pairs. Samples sharing a pair identifier were never separated across folds, ensuring that no MAP-positive individual was ever evaluated independently of the MAP-negative partner originating from the same farm, breed, age group, and sex. This design prevented information leakage, preserved biological dependency structures, and allowed explicit control over potential confounders influencing MAP dynamics. Pair-based constraints were enforced identically across inner and

outer folds, guaranteeing that hyperparameter optimisation, feature selection, and final model evaluation all operated under equivalent biological matching conditions.

Feature Aggregation and Stability Estimation

Global MI profiles were derived by normalising and consolidating feature scores across all 50 independent test evaluations. Feature-selection stability was quantified through two complementary measures: (i) selection frequency across outer folds and seeds, and (ii) seed-consistency metrics capturing the reproducibility of SNP prioritisation under different random initialisations. Together, these metrics provided a rigorous assessment of signal robustness and ensured that downstream modelling relied on features consistently supported across resampling iterations.

Machine Learning Framework for SNP Discovery

Model Description

To interrogate complementary dimensions of genotype–phenotype architecture, we implemented a diverse ensemble of nine machine learning algorithms: Random Forest, Gradient Boosting, Logistic Regression with L2 regularisation (Ridge), Elastic Net, Linear Support Vector Machine, Naive Bayes, XGBoost, LightGBM, and Extra Trees. This heterogeneous model suite spans penalised linear inference, multivariate regularisation, and non-linear interaction modelling, thereby offering a unified analytical landscape capable of detecting both additive and higher-order genetic effects.

Population Structure Modelling and Covariate Processing

Population structure was quantified using a PCA restricted to the pre-selected SNP panel, thereby preventing information leakage from genome-wide variation. All covariate handling—including demographic variables and PCs—was performed strictly within the nested cross-validation framework to preserve complete isolation between training and test partitions. Standardisation parameters were estimated exclusively from each training fold and subsequently applied to the corresponding test fold without recalibration. To avoid confounding and ensure stable effect estimation, demographic covariates and principal components were identified, validated for alignment, and standardised independently within every training split. The resulting scaled covariates were then horizontally merged with the normalised SNP

features, producing fold-specific augmented matrices that jointly captured genomic structure and non-genomic predictors. This rigorous preprocessing pipeline ensured that inferred associations reflected genuine genetic contributions rather than artefacts arising from information leakage or improperly adjusted confounders.

Machine Learning Pipeline

Hyperparameter Optimisation

Hyperparameter tuning was performed using stratified 5-fold inner cross-validation nested within each outer fold. Optuna's Tree-Parzen Estimator (TPE) algorithm guided the search across up to 50 trials per model per fold, while Hyperband pruning terminated underperforming configurations to accelerate convergence. Balanced accuracy served as the primary optimisation objective. Search spaces encompassed model-specific parameters including tree depth, number of estimators, minimum samples per split, and regularisation strengths (C), L1/L2 ratios.

Nested Cross-Validation and Leakage Prevention

A fully nested design comprising 5 outer folds $\times$ 5 seeds (25 independent evaluations) ensured unbiased performance estimation. Case-control matched pairs were preserved in all splits, maintaining biological dependencies while preventing inadvertent leakage from farm, breed, age, or sex structure. Inner folds were used exclusively for hyperparameter optimisation; outer folds were reserved for independent testing. All preprocessing—standardisation, covariate alignment, PCA application—occurred independently within each fold. No information from any test partition influenced training operations.

Performance metrics—including AUC, balanced accuracy, MCC, F1-score, recall, and specificity—were computed for each of the nine machine-learning models across 25 independent test datasets, and the aggregated results and confidence intervals summarised model-level stability.

Feature Importance Estimation

Feature importance was extracted using model-native criteria: Gini importance or gain for tree-based models, standardised coefficients for linear models, and permutation importance for models lacking intrinsic interpretability. Each metric was normalised to the [0,1] interval to

enable cross-model comparability. Importances were computed exclusively within the validation folds of the inner CV to avoid leakage from outer test sets. The resulting fold-specific matrices (features $\times$ 25 evaluations) were aggregated to yield mean effect sizes and stability indices, including variance, Spearman correlations, and Jaccard similarity across folds.

Statistical Validation and Quality Control

Multiple validation layers were embedded throughout the pipeline:

- (i) train–test independence was verified using automated overlap detection.
- (ii) overfitting was monitored via performance gaps exceeding 0.05 AUC or 0.07 balanced accuracy.
- (iii) calibration quality was evaluated through Brier scores and calibration slopes within the acceptable 0.9–1.1 range.
- (iv) consensus feature rankings underwent post-hoc permutation testing (10,000 iterations) with Benjamini–Hochberg correction ($q < 0.05$).

Global Leakage and Overfitting Safeguards

Pipeline integrity was ensured through stringent safeguards:

- (i) strict isolation of training and test folds,
- (ii) fold-specific scaling and covariate handling,
- (iii) nested cross-validation for unbiased hyperparameter selection,
- (iv) automated flagging of models exhibiting excessive train–test divergence, and
- (v) calibration-based reliability assessments.

SNP Discovery Methodology

Multi-Method Consensus Framework

To derive a stable and generalisable set of predictive variants, we constructed a multi-method consensus framework integrating three complementary aggregation strategies—weighted model-based scoring, Borda count ranking, and an approximate Kemeny–Young consensus. By combining direct importance estimation with rank-oriented and pairwise-preference methods, this framework reduces the idiosyncrasies of individual models and yields robust cross-model feature prioritisation.

339 ***Performance-Driven and Diversity-Aware Model Weighting***

340 *Composite Performance Metrics*

341 Each model’s contribution began with a composite validation score that blended three elements:
342 discriminatory power (captured through AUC), sensitivity to class imbalance (balanced
343 accuracy), and precision–recall efficiency (F1 score). These components were combined in
344 fixed proportions, with AUC receiving the highest weight, followed by balanced accuracy and
345 then F1.

346 To account for performance reliability, we penalized models whose AUC values fluctuated
347 substantially across cross-validation folds. Models exhibiting large variance had their
348 contribution proportionally reduced. The product of the composite score and this stability
349 penalty defined a model’s base weight.

350 *Diversity Adjustment*

351 To ensure that the ensemble did not overweight highly similar models, we quantified
352 redundancy by computing pairwise correlations between each model’s feature-importance
353 rankings. A diversity score was then assigned to each model based on how weakly correlated it
354 was with the remainder of the ensemble. Models providing more independent information
355 received higher scores.

356 Final model weights combined the base (performance-driven) weight with the diversity
357 component, using a heavier emphasis on the former. These weights were subsequently
358 normalised so that their total sum equalled one.

359 ***Rank Aggregation Approaches***

360 *Weighted Importance Estimation*

361 For each feature, we averaged the normalised importance values provided by all models,
362 weighting each model’s contribution according to its final ensemble weight. This produced the
363 primary “weighted consensus” importance estimate.

364 *Borda Count Consensus*

To incorporate ordering stability, each feature received a Borda score reflecting its relative position in each model. These scores were aggregated across models using the ensemble weights and normalised to a common scale.

Kemeny–Young Pairwise-Optimal Consensus

We also employed an approximate Kemeny–Young procedure to quantify how consistently one feature was preferred over another across models. Summing these weighted pairwise preferences produced a global Kemeny score per feature. Due to computational constraints, Kemeny scoring was restricted to the top 100 features identified by the weighted consensus method.

Final Composite Ranking

The three ranking strategies were integrated as:

$$\mathbf{final}(\mathbf{f}) = 0.40 (\mathbf{weighted}) + 0.30 (\mathbf{Borda}) + 0.30 (\mathbf{Kemeny}),$$

balancing direct score magnitude, rank stability, and pairwise-optimal ordering. This yielded a robust, cross-model consensus list suitable for downstream genomic interpretation.

Permutation-Based Statistical Validation and Persistency Threshold

False-discovery control was performed using permutation-derived p-values based on logistic-regression AUCs evaluated under 5-fold cross-validation, benchmarked against null distributions generated via Westfall–Young minP permutations (10,000 iterations applied to the top 100 SNPs). Multiple testing correction followed the Benjamini–Hochberg procedure ($q < 0.05$). All permutation procedures were executed strictly on the outer test folds to prevent information leakage. SNPs surpassing the Benjamini–Hochberg threshold were subsequently filtered using a persistency cutoff of >0.25 , a stability metric that aggregates the model-specific recurrence of high-importance rankings across outer-fold splits (25 folds in total) into a single consensus score.

Genomic Annotation, Gene Enrichment, and Pathway Analysis

Genomic annotation of significant SNPs was performed using the ARS1 reference assembly through the Ensembl platform, Genome Data Viewer, and Genome Browser. Genes located within ± 200 kb of each significant SNP were subjected to functional enrichment and pathway

analysis using ShinyGO database, incorporating all available databases for the *Capra hircus* genome. A false discovery rate of 0.05 was applied to identify significantly enriched terms. All visualisations and graphical outputs were generated using ShinyGO.

Results

Serological Survey Results

Serological screening across seven indigenous goat breeds—based on 3,369 successfully tested animals from 36 herds spanning four geographic regions (Figure 1)—revealed a heterogeneous yet broadly coherent pattern of MAP exposure. The mean individual seroprevalence was 13.95%, with breed-level estimates ranging from 8.6% to 25.41%. At the herd level, a prevalence of 95.55% highlighted the deep and widespread establishment of MAP in Türkiye's goat production systems. Although absolute seropositivity differed among breeds, the near-universal presence of positive animals across herds indicated that MAP circulation is extensive and persistent.

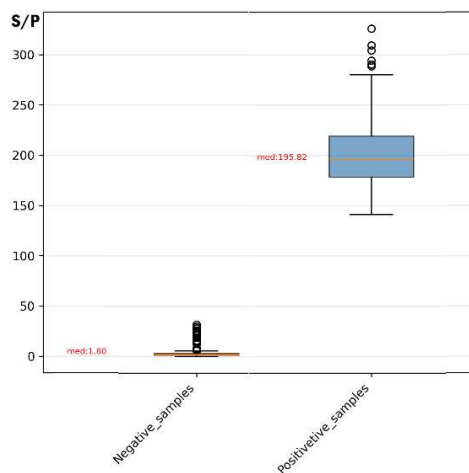
Figure 1. Sampling map spanning 11 provinces and 4 geographic regions



Across the 240 matched pairs—constructed according to herd, breed, age, and sex and subsequently used to generate the dataset for genetic analyses—the seronegative group

exhibited a mean Sample-to-Positive (SP) value of 3.5%, whereas the seropositive group showed a markedly higher mean of 202.0 (Figure 2). This clear separation between positive and negative animals provided a reliable foundation for downstream analyses, ensuring consistency and interpretive confidence.

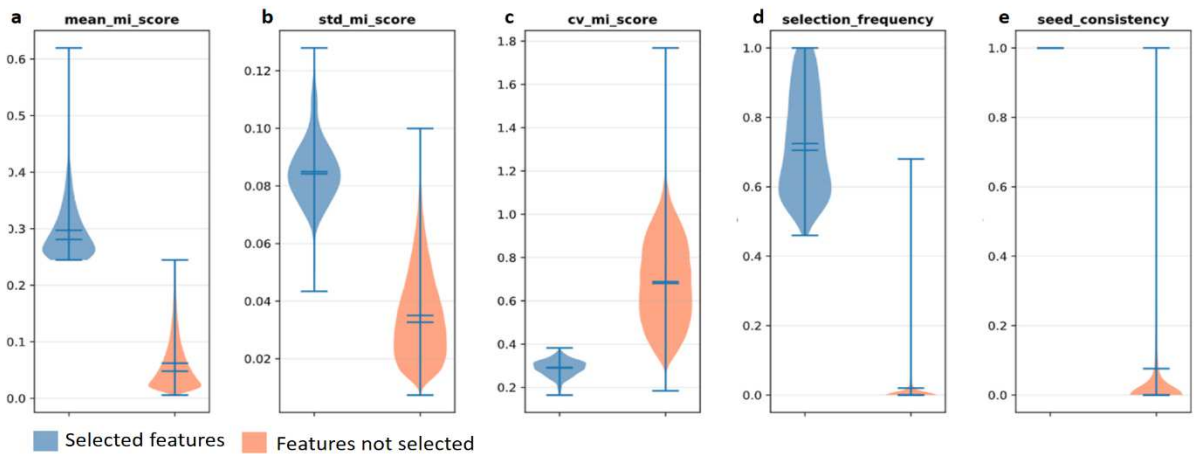
Figure 2. Box plot illustrating the SP% distribution between positive and negative groups in the matched-pairs panel.



Discriminative Performance of Preselected Features

Comparative analysis between Top 500 SNPs and remaining variants demonstrates exceptional discriminative capacity across all quality metrics. Selected SNPs about exhibit four-fold elevated mean MI scores (median ~ 0.28 vs ~ 0.05) (Figure 3a) with superior cross-fold stability (Figure 3b), as evidenced by lower coefficient of variation values (~ 0.30 vs ~ 0.7) (Figure 3c). The selection frequency distribution reveals that Top 500 SNPs span 0.4-1.0 (Figure 3d) indicating consistent prioritization across cross-validation iterations, while non-selected variants cluster near zero (< 0.05). Perfect seed consistency (1.0) for selected features contrasts sharply with near-zero reproducibility (~ 0.05) for remaining variants (Figure 3e), confirming deterministic convergence to biologically relevant markers independent of stochastic initialization. Notably, the coefficient of variation (Figure 3c) displays a bimodal pattern where Top 500 SNPs concentrate at lower CV values (~ 0.30), signifying superior stability, while the remaining variants exhibit higher relative variability (~ 0.60), reflecting their inconsistent and likely spurious signals across different data partitions. These metrics establish clear separation between feature spaces, validating the Top 500 SNPs as a stable, information-rich subset for predictive modeling.

Figure 3. Distribution of MI metrics for the Top 500 SNPs selected by the Mutual Information Algorithm versus all remaining excluded SNPs



(a) *mean_mi_score*: Mean Mutual Information across resampling iterations, indicating overall informativeness of each SNP.

(b) *std_mi_score*: Standard deviation of MI values, reflecting the stability of each SNP's information contribution across iterations.

(c) *cv_mi_score*: Coefficient of variation of MI scores, capturing relative dispersion and highlighting features with unstable or noise-driven signal.

(d) *selection_frequency*: Proportion of resampling runs in which each SNP was selected, representing robustness and repeatability of selection.

(e) *seed_consistency*: Consistency of SNP selection across different random seeds, quantifying reproducibility of feature importance under stochastic variation.

Performance Evaluation Across Multiple Machine Learning Algorithms

Table 1 demonstrates strong concordance across nine classification architectures, with test-set AUC values tightly constrained to 0.982–0.992 (1.0% spread), indicating convergence on a shared decision boundary. Naive Bayes achieves the highest discriminative performance (AUC = 0.992) while also exhibiting the smallest train–test divergence ($\Delta\text{AUC} = 0.006$), consistent with an optimal bias–variance balance.

Test-set balanced accuracy ranges from 0.947 to 0.971, peaking with Random Forest (0.971) and reaching its minimum with SVM Linear (0.947). Naive Bayes (0.969), LightGBM (0.967), Gradient Boosting (0.967), and XGBoost (0.967) form a narrow 0.2% performance cluster, indicating functional equivalence. Naive Bayes again shows the smallest train–test gap ($\Delta =$

0.007), whereas Elastic Net displays the largest decline ($\Delta = 0.044$), suggesting sensitivity to regularisation or insufficient model complexity.

F1-scores closely mirror balanced accuracy, spanning 0.947–0.971 on the test set. Random Forest leads (0.971), followed by Naive Bayes (0.969) and Gradient Boosting (0.968). Train–test F1 divergences range from 0.007 to 0.044, with Naive Bayes and XGBoost showing exceptional stability, while Elastic Net, SVM Linear, and Logistic Regression L2 exhibit larger performance drops, consistent with the known limitations of linear models in capturing complex non-linear feature interactions.

The Matthews Correlation Coefficient (MCC), a class-imbalance–invariant metric integrating all confusion matrix components, ranges from 0.895 to 0.944 on the test set. Random Forest (0.944) and Naive Bayes (0.940) show the strongest agreement between predictions and ground truth, followed by Gradient Boosting, XGBoost, and LightGBM (all 0.936). Logistic Regression L2, Elastic Net, and SVM Linear occupy the lower boundary (0.900, 0.900, and 0.895), indicating reduced capacity to optimise all classification quadrants concurrently.

Train–test MCC degradation spans 0.013–0.087, representing a 6.7-fold variation in generalisation stability. Naive Bayes, XGBoost, and Random Forest maintain MCC loss below 5%, whereas Elastic Net, SVM Linear, and Logistic Regression L2 approach or exceed 6.5%. Gradient Boosting and Extra Trees exhibit intermediate degradation, reflecting moderate capacity–complexity trade-offs.

Test-set precision varies from 0.949 to 0.970, with Random Forest and Naive Bayes achieving the highest positive predictive value (0.970), corresponding to minimal false-alarm rates. Gradient Boosting shows comparatively lower precision (0.956), consistent with a bias toward sensitivity. Precision train–test gaps range from 0.004 to 0.043, with Naive Bayes showing exceptional stability and Gradient Boosting the greatest volatility, likely due to amplification of training-set idiosyncrasies inherent to sequential error correction.

Recall (sensitivity) spans a wider range than other performance metrics (0.947–0.983), peaking with XGBoost (0.983), followed by Gradient Boosting (0.982) and LightGBM (0.980), thereby identifying gradient-based ensembles as optimal when false-negative costs dominate. Elastic Net (0.952) and SVM Linear (0.947) exhibit the lowest recall, consistent with the limited representational capacity of linear decision boundaries for complex class separation. Train–test

recall degradation remains modest across models ($\Delta = 0.005$ – 0.045), with XGBoost ($\Delta = 0.005$) and Naive Bayes ($\Delta = 0.009$) showing exceptional generalization, whereas Elastic Net ($\Delta = 0.045$) and Logistic Regression L2 ($\Delta = 0.038$) demonstrate more pronounced decline.

Recall and specificity exhibit balanced performance across classifiers, spanning 0.947–0.983 and 0.947–0.968, respectively, with no pronounced trade-offs between true positive and true negative rates. Gradient Boosting, XGBoost, and LightGBM attain the highest recall (0.982, 0.983, and 0.980), favouring these ensembles for recall-critical settings, whereas Naive Bayes achieves maximal specificity (0.968), highlighting its strength in minimizing false positives.

Mean prediction confidence displays a marked bimodal distribution (0.661–0.997). Naive Bayes yields near-maximal confidence (0.997) with negligible train–test degradation, consistent with well-calibrated posterior probabilities under conditional independence assumptions. Gradient Boosting similarly shows high confidence (0.988) with minimal volatility. In contrast, Random Forest exhibits systematic underconfidence (0.661), representing a 33.7% deficit relative to Naive Bayes, attributable to probability regression induced by ensemble averaging. Gradient-based ensembles combine high confidence with strong discrimination, with XGBoost (0.952) and LightGBM (0.974) offering robust probability estimates alongside accurate classification.

Confidence variability ranges from 0.017 to 0.119, with Naive Bayes showing minimal dispersion (0.017) and Elastic Net (0.117) and SVM Linear (0.119) exhibiting approximately sevenfold greater heterogeneity, reflecting instability of probability estimates under linear decision boundaries.

Brier scores span 0.029–0.137, with Naive Bayes achieving optimal calibration (0.029), representing a 79% improvement over Random Forest (0.137). LightGBM (0.030), Gradient Boosting (0.031), and XGBoost (0.030) form a tightly clustered elite group, confirming the consistent calibration strength of gradient-based ensembles. Linear models show intermediate and progressively degrading calibration (SVM Linear 0.050, Logistic Regression L2 0.106, Elastic Net 0.117), while Extra Trees occupies an intermediate position (0.054), indicating that gains in discrimination do not necessarily translate into probabilistic accuracy.

521 **Table 1.** Comparative performance metrics of machine learning classifiers on training and test sets

Model	Random Forest		Gradient Boosting		Logistic Regression L2		Elastic Net		SVM Linear		Naive Bayes		XGBoost		Light GBM		Extra Trees	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
AUC	1.000	0.991	1.000	0.989	0.998	0.982	1.000	0.986	0.998	0.982	0.999	0.992	0.998	0.987	1.000	0.991	1.000	0.988
Balanced Accuracy	0.995	0.971	0.999	0.967	0.983	0.949	0.993	0.949	0.982	0.947	0.976	0.969	0.981	0.967	0.997	0.967	0.996	0.963
F1	0.996	0.971	0.999	0.968	0.983	0.949	0.993	0.949	0.982	0.947	0.976	0.969	0.982	0.968	0.997	0.967	0.996	0.964
MMC	0.991	0.944	0.999	0.936	0.966	0.900	0.987	0.900	0.964	0.895	0.952	0.940	0.963	0.936	0.993	0.936	0.993	0.929
Precision	0.993	0.970	0.999	0.956	0.979	0.952	0.990	0.949	0.980	0.949	0.973	0.970	0.975	0.955	0.996	0.957	0.995	0.955
Recall	0.998	0.975	1.000	0.982	0.986	0.948	0.997	0.952	0.984	0.947	0.979	0.970	0.988	0.983	0.998	0.980	0.998	0.975
Specificity	0.993	0.967	0.999	0.952	0.979	0.950	0.990	0.947	0.980	0.947	0.973	0.968	0.975	0.950	0.996	0.953	0.995	0.952
Confidence mean	0.746	0.661	0.993	0.988	0.762	0.726	0.868	0.817	0.931	0.894	0.997	0.997	0.956	0.952	0.981	0.974	0.928	0.864
Confidence Std	0.043	0.043	0.013	0.034	0.071	0.092	0.069	0.117	0.093	0.119	0.028	0.017	0.056	0.058	0.018	0.035	0.057	0.080
Confidence min	0.032	0.043	0.376	0.293	0.037	0.020	0.093	0.117	0.017	0.053	0.115	0.395	0.053	0.163	0.360	0.285	0.067	0.081
Confidence max	0.331	0.243	0.497	0.497	0.417	0.379	0.485	0.117	0.500	0.499	0.500	0.500	0.483	0.480	0.491	0.490	0.475	0.443
Brier score	0.076	0.137	0.001	0.031	0.078	0.106	0.031	0.117	0.019	0.050	0.021	0.029	0.015	0.030	0.004	0.030	0.017	0.054

Overfitting control (Obtained by subtracting the test metric from the corresponding train metric.)

AUC	0.009	0.011	0.016	0.013	0.016	0.006	0.011	0.009	0.012
Balanced Accuracy	0.025	0.033	0.034	0.044	0.035	0.007	0.015	0.030	0.033
F1	0.024	0.032	0.034	0.044	0.035	0.007	0.014	0.029	0.032
MMC	0.047	0.063	0.066	0.087	0.068	0.013	0.027	0.058	0.064
Precision	0.024	0.043	0.028	0.042	0.031	0.004	0.021	0.038	0.040
Recall	0.023	0.018	0.038	0.045	0.037	0.009	0.005	0.018	0.023
Specificity	0.027	0.047	0.029	0.044	0.033	0.005	0.025	0.042	0.043

522

523 *AUC (Area Under Curve): Model's discrimination power; ability to distinguish between positive and negative classes (0-1 range, higher=better)*

524 *Balanced Accuracy: Mean of sensitivity and specificity, providing equal weight to both classes*

525 *F1: Harmonic mean of Precision and Recall; overall performance measure balancing both metrics*

526 *MCC (Matthews Correlation): Balanced classification quality using all confusion matrix elements*
527 *Precision: Proportion of positive predictions that are actually correct; measures exactness*
528 *Recall (sensitivity): Proportion of actual positives correctly identified; measures completeness*
529 *Specificity: Proportion of actual negatives correctly identified; true negative rate*
530 *Confidence Mean: Average prediction confidence score across all samples*
531 *Confidence Std: Standard deviation of confidence scores; measures prediction uncertainty variation*
532 *Confidence Min: Lowest confidence score among all predictions*
533 *Confidence Max: Highest confidence score among all predictions*
534 *Brier Score: Mean squared difference between predicted probabilities and actual outcomes (0-1, lower=better)*

535

536

537 ***Train-Test Performance Differences Across 25 Folds***

538 The composite overfitting metric integrates multiple performance gaps into a single quantitative
539 index of generalisation stability. For each fold, discrepancies between training and test
540 performance are calculated for AUC, balanced accuracy, and F1 score, capturing
541 complementary dimensions of potential overfitting. These gaps are then combined using a
542 weighted scheme that assigns the greatest emphasis to AUC, followed by balanced accuracy
543 and F1, producing a unified score that reflects the overall extent of performance inflation on
544 the training data. During analysis, any fold exceeding the weighted overfitting threshold of 0.05
545 was excluded from further evaluation. The resulting composite measure is interpreted using
546 predefined thresholds that distinguish acceptable generalisation from moderate or pronounced
547 overfitting, enabling a clear and systematic assessment of model stability across folds (Figure
548 4).

549

550

551

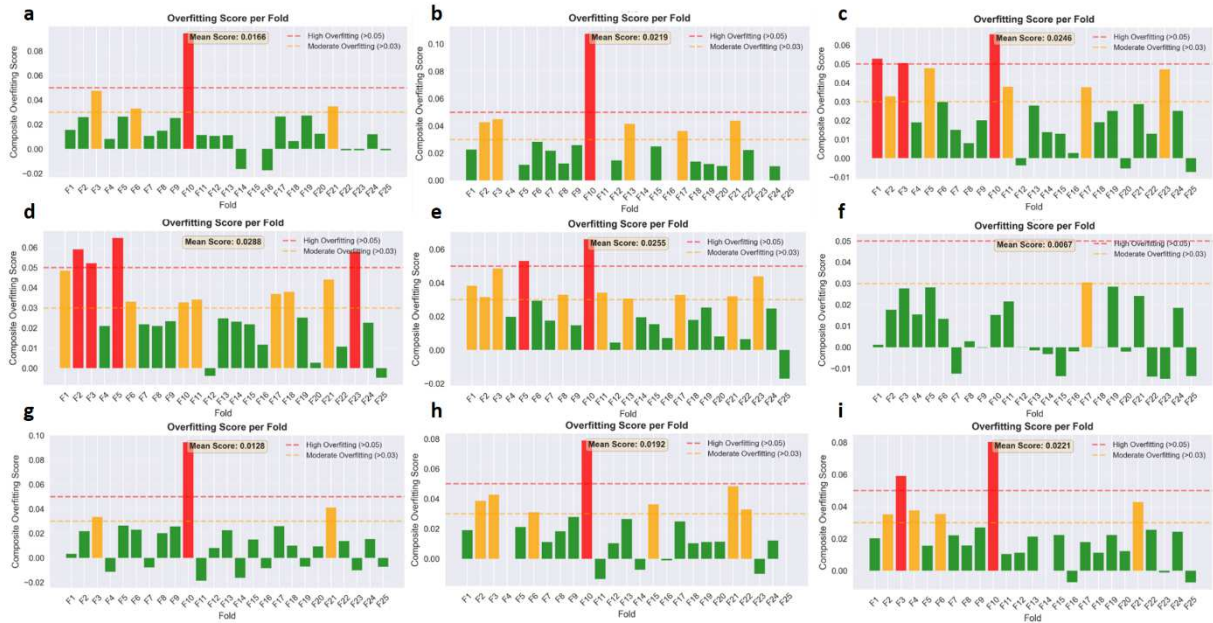
552

553

554

555

Figure 4. Variability of train–test metrics across 25-fold cross-validation



a, Random Forest; b, Gradient Boosting; c, Logistic Regression L2; d, Elastic Net; e, SVM Linear; f, Naive Bayes; g, XGBoost; h, Light GBM; i, Extra Trees

Cross-Model Feature Importance Concordance Analysis

To quantify cross-algorithmic consistency in genomic feature prioritization, we constructed a comprehensive feature importance correlation matrix spanning all nine classification architectures. For each model, fold-averaged feature importance scores were computed, yielding a single importance value per SNP. Pairwise linear agreement was assessed via Pearson's correlation coefficient applied to aligned SNP-level importance vectors, with only shared SNPs retained and missing values excluded. The resulting symmetric 9×9 matrix with unity diagonal elements is visualized as a heatmap (Figure 5).

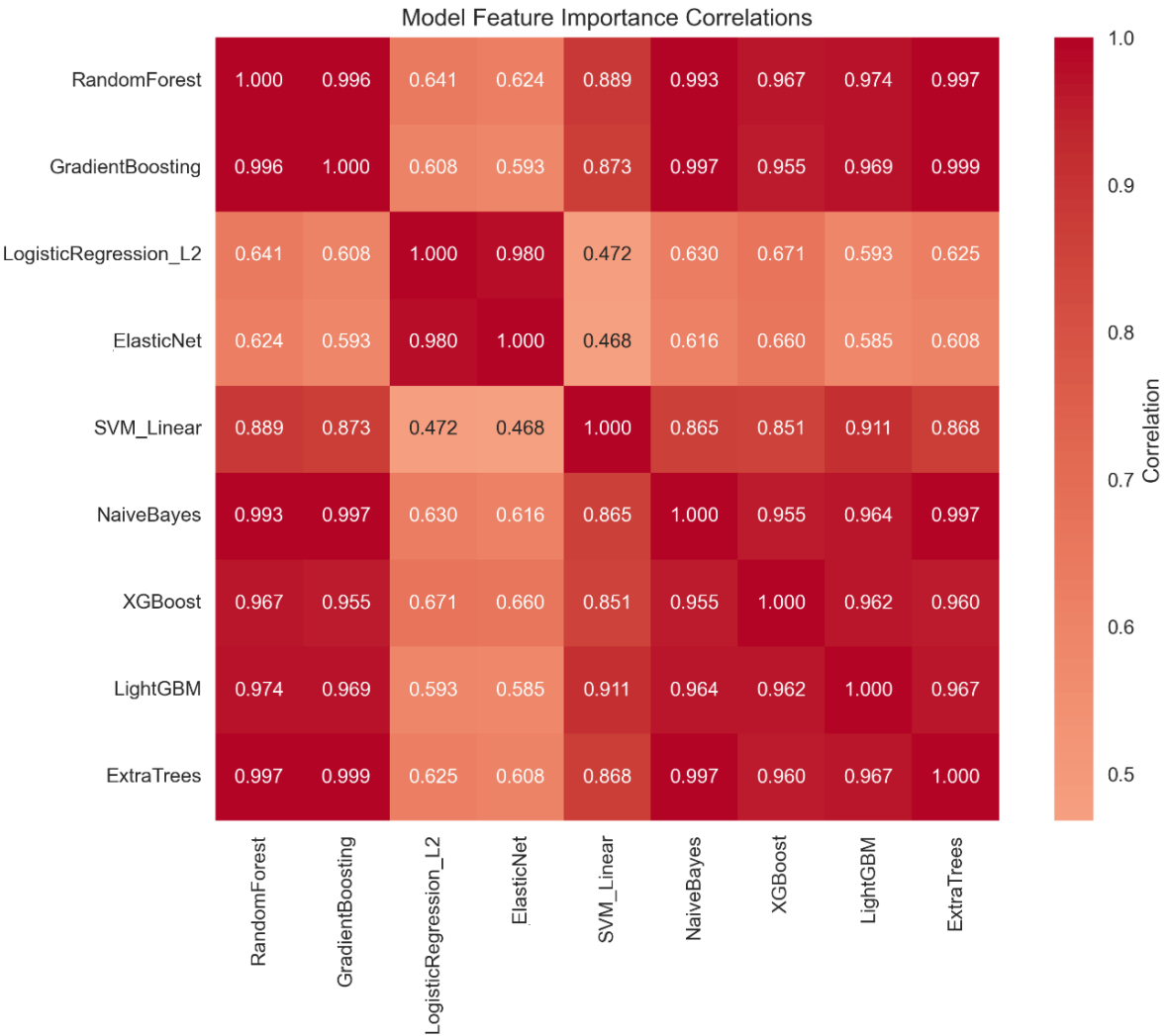
The correlation structure reveals pronounced architectural clustering. Tree-based ensemble methods—Random Forest, Gradient Boosting, Extra Trees— and Naive Bayes demonstrate exceptionally high concordance ($r = 0.993$ – 0.999), with Gradient Boosting–Extra Trees achieving near-perfect agreement ($r = 0.999$). This convergence indicates that despite divergent algorithmic mechanisms, these methods identify substantively identical genomic signal structures.

Among linear models, Logistic Regression L2 and Elastic Net demonstrate strong mutual agreement ($r = 0.980$), indicating substantial concordance in feature attribution despite their different regularization strategies. However, both exhibit markedly reduced concordance with SVM Linear ($r = 0.468\text{--}0.472$), suggesting that margin-based optimization diverges fundamentally from likelihood-based linear frameworks in feature attribution patterns. This internal heterogeneity within the linear model family underscores that algorithmic paradigm—margin maximization versus likelihood optimization—exerts greater influence on feature importance rankings than regularization architecture alone.

On the other hand, linear architectures—Logistic Regression L2, Elastic Net, and SVM Linear—demonstrate substantially diminished concordance with tree-based methods ($r = 0.468\text{--}0.671$). Elastic Net manifests the most pronounced divergence ($r = 0.468\text{--}0.660$), potentially reflecting dual-penalty regularization's aggressive coefficient shrinkage for correlated features. This systematic linear-nonlinear discordance stems from fundamental representational constraints: linear models assign global importance weights invariant to feature interactions, whereas tree-based methods capture interaction-dependent feature salience through recursive partitioning.

Naive Bayes occupies a distinctive position within the correlation landscape: it aligns exceptionally well with tree-based ensembles ($r = 0.993\text{--}0.997$ with Random Forest, Gradient Boosting, and Extra Trees), yet shows only moderate concordance with linear models ($r = 0.616\text{--}0.865$ with Elastic Net and Linear SVM). This dual concordance pattern positions Naive Bayes as effectively bridging probabilistic and decision-boundary paradigms through its conditional independence architecture. Its near-perfect alignment with nonlinear methods suggests that probabilistic feature attribution naturally captures interaction structures, while its sustained—albeit attenuated—linear model agreement reflects shared reliance on feature-level marginal associations.

Figure 5. Cross-model feature importance correlation matrix



Significant Markers Identified Through Ensemble Machine Learning

Using an ensemble machine learning strategy that integrated nine models and employed extensive permutation testing (n=10,000), a total of 31 genomic loci were identified as having high predictive importance, each exceeding a precision threshold of 0.25. These SNPs showed consistent performance across multiple ranking approaches, with composite scores between 0.794 and 1.000, indicating stable and reproducible associations. Their chromosomal distribution was uneven, with notable clustering on chromosomes 2, 6, 11, and 21, suggesting potential functional concentration in these regions (Table 2).

Among these loci, snp4242-scaffold1131-545695 on chromosome 6 stood out, achieving maximum normalized values across both Borda and Kemeny rankings (1.000) and exhibiting high rank stability (0.96), making it the most influential marker. Additional variants on

chromosome 6—such as snp18252-scaffold1850-148547 and snp36198-scaffold433-398052 —also ranked highly, with weighted normalized scores above 0.85 and strong persistence values (0.40–0.76), further highlighting the relevance of this chromosome. Chromosome 11 similarly contained several high-impact loci, including snp42545-scaffold564-283910 and snp17821-scaffold185-652809, both of which showed composite scores over 0.93 and substantial rank stability.

Combined with rank stability coefficients generally above 0.70, these results support the view that the observed associations are biologically meaningful rather than random. A subset of loci on chromosomes 2, 4, and 15 showed strong weighted normalized scores but moderate persistence levels, suggesting potential context-dependent or epistatic contributions that merit further investigation. Overall, the distributed yet functionally aligned set of loci provides a solid basis for follow-up mechanistic research and potential translational applications.

Table 2. Genomic variants with high predictive importance identified by ensemble machine learning

Feature	CHR	Pos	variant	Weighted norm	Borda score norm	Kemeny score norm	Composite score norm	Persistence	Rank stability
snp14117-scaffold1564-138906	1	23951701	G	0.547	0.803	0.706	0.882	0.36	0.84
snp22402-scaffold2210-358741	1	146000000	A	0.571	0.776	0.685	0.860	0.36	0.76
snp7408-scaffold127-1272890	2	27102072	G	0.588	0.803	0.721	0.894	0.52	0.78
snp5726-scaffold1200-178160	2	60706561	G	0.606	0.715	0.624	0.800	0.32	0.65
snp5772-scaffold1203-488300	2	67188067	A	0.647	0.862	0.788	0.823	0.48	0.85
snp28267-scaffold301-1618884	2	108000000	A	0.583	0.842	0.758	0.794	0.36	0.87
snp58409-scaffold948-634796	2	134000000	G	0.729	0.837	0.770	0.813	0.28	0.74
snp24990-scaffold257-835886	3	55323900	A	0.588	0.724	0.627	0.805	0.36	0.66
snp31149-scaffold345-718909	4	63112927	G	0.773	0.877	0.825	0.859	0.36	0.82
snp31527-scaffold350-360780	4	82149468	G	0.749	0.862	0.818	0.846	0.52	0.77
snp47327-scaffold665-1059656	5	34657160	G	0.723	0.778	0.728	0.900	0.28	0.70
snp51038-scaffold740-716213	5	89531145	A	0.627	0.891	0.832	0.851	0.4	0.91
snp18252-scaffold1850-148547	6	1756463	G	0.914	0.896	0.884	0.907	0.4	0.84
snp4242-scaffold1131-545695	6	3701899	G	0.878	1.000	1.000	1.000	0.92	0.96
snp36198-scaffold433-398052	6	82628596	G	0.859	0.910	0.904	0.915	0.76	0.89
snp19388-scaffold196-254025	7	107000000	A	0.820	0.941	0.909	0.929	0.56	0.88
snp9078-scaffold133-416791	10	19118447	C	0.641	0.790	0.702	0.885	0.4	0.70
snp42545-scaffold564-283910	11	1534514	A	0.809	0.955	0.932	0.943	0.48	0.90
snp41247-scaffold536-90392	11	7102955	G	0.766	0.918	0.876	0.898	0.28	0.89
snp28637-scaffold309-203183	11	35431796	G	0.542	0.727	0.643	0.807	0.36	0.77
snp17821-scaffold185-652809	11	73210757	G	0.800	0.945	0.925	0.934	0.4	0.88
snp42713-scaffold568-1111566	12	18563769	G	1.000	0.908	0.891	0.926	0.44	0.84
snp23736-scaffold24-170008	13	23886395	A	0.510	0.745	0.655	0.818	0.52	0.76
snp58744-scaffold959-252718	13	43364413	C	0.764	0.890	0.839	0.870	0.4	0.85
snp1942-scaffold1054-281717	14	78601781	A	0.797	0.961	0.936	0.947	0.6	0.91
snp40457-scaffold516-2400865	15	4706995	A	0.984	0.792	0.784	0.824	0.32	0.71
snp42323-scaffold557-1353363	18	5395057	A	0.837	0.838	0.817	0.843	0.4	0.78
snp57371-scaffold913-1533576	20	8211993	A	0.875	0.828	0.799	0.835	0.28	0.74
snp6499-scaffold1230-873345	21	32528800	G	0.831	0.904	0.871	0.897	0.36	0.83
snp36676-scaffold444-459060	21	39010838	A	0.909	0.864	0.843	0.875	0.64	0.79
snp25274-scaffold2598-218661	21	58428665	A	0.647	0.881	0.805	0.839	0.36	0.87

Weighted score: SNP's consensus importance derived from model-weighted averaging of normalized importances.

637 *Borda score: Rank-based consensus score computed via weighted Borda count across all model rankings.*

638 *Kemeny score: Pairwise-rank aggregation score reflecting the SNP's position under the Kemeny-Young rule.*

639 *Composite score: Integrated ensemble metric combining weighted, Borda and Kemeny scores (40 %, 30 %, 30*
640 *%) to produce the final consensus importance value for each SNP.*

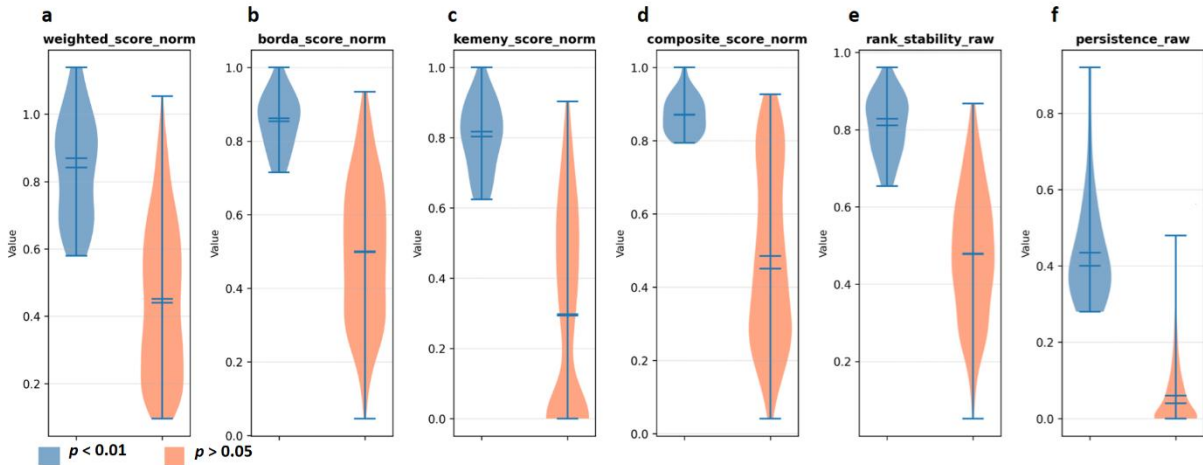
641 *Persistency: a stability metric that aggregates model-specific consistency of high importance rank across outer-*
642 *fold splits into a single consensus score for each SNP.*

643 *Rank stability: normalized rank consistency across models, where lower variance in feature ranking yields*
644 *higher stability for the consensus SNP list.*

645

646 The ensemble-derived metrics shown in Figure 6 provide a consistent visual distinction between
647 SNPs classified as significant (blue; FDR $q < 0.05$) and non-significant (orange; FDR $q > 0.05$),
648 with the normalised (“norm”) scores generally concentrated near the upper range for the
649 significant group. In panel **a**, the normalised weighted score exhibits clear separation, with
650 significant SNPs clustered tightly at high values, while non-significant SNPs are more dispersed
651 and substantially lower. Panel **b** displays a similar pattern in the normalised Borda score, where
652 significant SNPs show minimal variance and consistently high values, reflecting strong
653 agreement across models. Panel **c** shows an even more pronounced contrast in the normalised
654 Kemeny score, as non-significant SNPs exhibit wide variability and frequent near-zero values,
655 suggesting inconsistent model support. In panel **d**, the normalised composite score effectively
656 synthesises these indicators and maintains the strong separation between groups. Panel **e**
657 demonstrates that raw rank stability is high and concentrated for significant SNPs—indicating
658 stable ordering across folds—whereas non-significant SNPs exhibit much broader spread.
659 Finally, panel **f** shows that raw persistence, representing the proportion of folds in which a SNP
660 is selected, declines sharply for non-significant SNPs, while significant SNPs appear
661 persistently across folds. Together, these panels present a coherent and objective multimetric
662 pattern that differentiates SNPs with consistent model support from those without.

Figure 6. Multimetric ensemble profiling of significant and non-significant SNPs

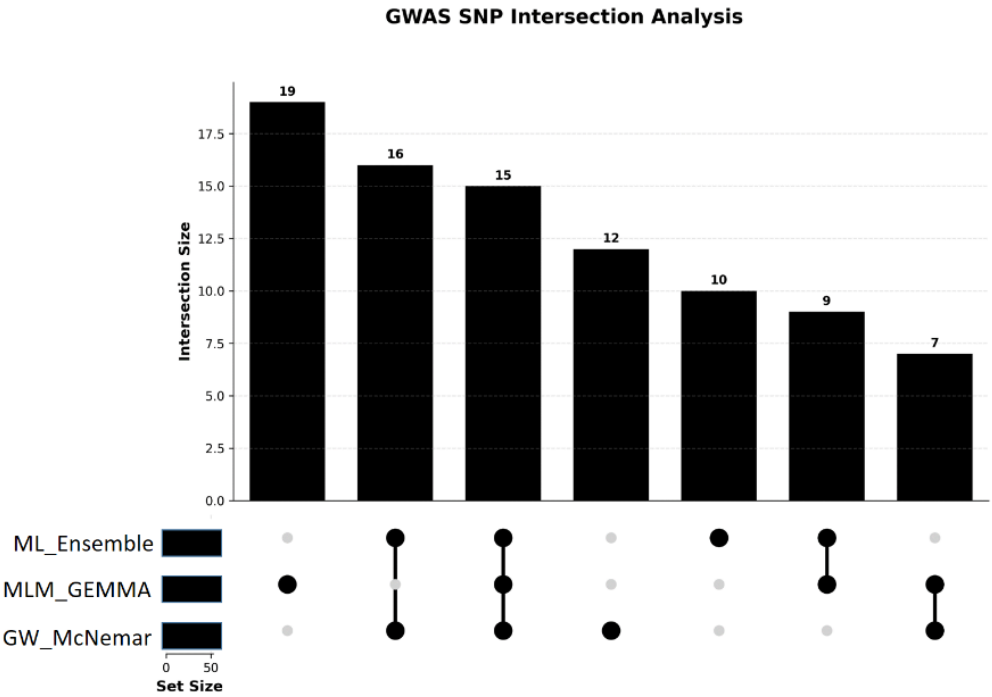


Cross-model Concordance and Benchmarking of Association Signals

Finally, the ensemble results derived from the nine ML models were benchmarked against two additional approaches: a classical mixed linear model (MLM) implemented in GEMMA and a genome-wide McNemar test performed on the matched pairs using a custom Python script. No SNP preselection was applied in either of these models. As in the ensemble framework, the first ten principal components were included as covariates in the MLM to account for population structure. The genome-wide McNemar test was conducted exclusively on the matched pairs to ensure proper control of confounding.

To evaluate concordance across methods, we quantified the overlap among the top 50 SNPs identified by each approach. The analysis revealed that 21 SNPs were shared between the ML ensemble and the MLM, 31 between the ML ensemble and the genome-wide McNemar test, and 15 SNPs were jointly supported by all three methods (Figure 7).

Figure 7. Cross-model overlap structure of benchmark association signals



Gene Functional Annotation, GO-Term Enrichment, And Pathway Analysis

Across the genome, 31 significant SNPs were identified on 16 chromosomes, with the vast majority located in intergenic regions (Table 3). Only four variants fell within intronic sequences: *CACNA2D1* on chromosome 4, *RGS6* on chromosome 10, *BUB1* on chromosome 11, and *ENSCHIG00000014388* on chromosome 15. The remaining 27 SNPs were intergenic but frequently positioned within ± 200 kb of biologically relevant genes. Notably, several mapped near immune-related loci, including interferon-stimulated genes (*IFI44*, *IFI44L*, *IFI27L2*) and members of the interleukin receptor cluster (*IL1R1*, *IL1RL2*, *IL18R1*, *IL18RAP*).

Additional associations were observed adjacent to a dense serpin gene cluster on chromosome 21 (*SERPINA1*, *SERPINA6*, *SERPINA10*, *SERPINA11*, *SERPINA12*, *SERPINA14*), highlighting a region involved in regulating innate immunity, inflammation, hormone transport, coagulation balance, metabolic homeostasis, and reproductive immunomodulation.

The predominance of intergenic variants suggests that regulatory elements—rather than alterations in coding sequence—likely drive the observed associations. It is also plausible that these SNPs are in high linkage disequilibrium with functional coding variants in nearby genes, thereby serving as proxies for causal mutations.

702 **Table 3.** Functional Annotation and Pathway Analysis

Feature	Chr	Pos	Located genes within ± 200 KB range	Location
snp14117-scaffold1564-138906	1	23951701	<i>na</i>	intergenic
snp22402-scaffold2210-358741	1	146238603	<i>S100B</i> ; <i>PRMT2</i> ; <i>DIP2A</i>	intergenic
snp7408-scaffold127-1272890	2	27102072	<i>na</i>	intergenic
snp5726-scaffold1200-178160	2	60706561	<i>na</i>	intergenic
snp5772-scaffold1203-488300	2	67188067	<i>na</i>	intergenic
snp28267-scaffold301-1618884	2	107809284	<i>B3GALT1</i>	intergenic
snp58409-scaffold948-634796	2	133639821	<i>ENSCHIG00000025950</i>	intergenic
snp24990-scaffold257-835886	3	55323900	<i>ADGRL4</i> ; <i>IFI44</i> ; <i>IFI44L</i>	intergenic
snp31149-scaffold345-718909	4	63112927	<i>LRRN3</i>	intergenic
snp31527-scaffold350-360780	4	82149468	<i>CACNA2D1</i>	intron
snp47327-scaffold665-1059656	5	34657160	<i>PLEKHA5</i>	intergenic
snp51038-scaffold740-716213	5	89531145	<i>PLEKHA5</i>	intergenic
snp18252-scaffold1850-148547	6	1756463	<i>ENSCHIG00000013370</i>	intergenic
snp4242-scaffold1131-545695	6	3701899	<i>CCNA2</i> ; <i>BBS7</i> ; <i>EXOSC9</i>	intergenic
snp36198-scaffold433-398052	6	82628596	<i>na</i>	intergenic
snp19388-scaffold196-254025	7	107193370	<i>RUFY</i> ; <i>CBY3</i> ; <i>HNRNP1</i> ; <i>CANX</i> ; <i>MAML1</i> ; <i>LTC4S</i> ; <i>MGAT4B</i> ; <i>MRNIP</i>	intergenic
snp9078-scaffold133-416791	10	19118447	<i>RGS6</i>	intron
snp42545-scaffold564-283910	11	1534514	<i>BUB1</i>	intron
snp41247-scaffold536-90392	11	7102955	<i>IL1R1</i> ; <i>IL1RL2</i> ; <i>IL18R1</i> ; <i>IL18RAP</i> ; <i>SLC9A4</i> ; <i>SLC9A2</i> ; <i>MSDF9</i> ; <i>TMEM182</i>	intergenic
snp28637-scaffold309-203183	11	35431796	<i>na</i>	intergenic
snp17821-scaffold185-652809	11	73210757	<i>KIF3C</i> ; <i>GAREM2</i> ; <i>RAB10</i> ; <i>ASXL2</i>	intergenic
snp42713-scaffold568-1111566	12	18563769	<i>ENSCHIG00000011884</i>	intergenic
snp23736-scaffold24-170008	13	23886395	<i>OTUD1</i> ; <i>C10orf67</i>	intergenic
snp58744-scaffold959-252718	13	43364413	<i>AKR1E2</i> ; <i>ENSCHIG00000021551</i> ; <i>ENSCHIG00000015988</i> ; <i>ENSCHIG00000025612</i>	intergenic
snp1942-scaffold1054-281717	14	78601781	<i>KCNK9</i> ; <i>TRAPPC9</i>	intergenic
snp40457-scaffold516-2400865	15	4706995	<i>ENSCHIG00000014388</i>	intron
snp42323-scaffold557-1353363	18	5395057	<i>MON1B</i>	intergenic
snp57371-scaffold913-1533576	20	8211993	<i>UTP15</i> ; <i>BTF3</i> ; <i>ANKRA2</i> ; <i>FOXD1</i>	intergenic
snp6499-scaffold1230-873345	21	32528800	<i>LINGO1</i> ; <i>IMP3</i> ; <i>CIMAP1C</i> ; <i>SNX33</i> ; <i>CSPG4</i> ; <i>SNUPN</i> ; <i>PTPN9</i>	intergenic
snp36676-scaffold444-459060	21	39010838	<i>na</i>	intergenic
snp25274-scaffold2598-218661	21	58428665	<i>IFI27L2</i> ; <i>PPP4R4</i> ; <i>SERPINA1</i> ; <i>SERPINA6</i> ; <i>SERPINA10</i> ; <i>SERPINA11</i> ; <i>SERPINA12</i> ; <i>SERPINA14</i>	intergenic

703

704 Functional enrichment analyses reinforced this inference: candidate genes showed significant

705 overrepresentation of immune pathways, particularly those governing cytokine signalling and

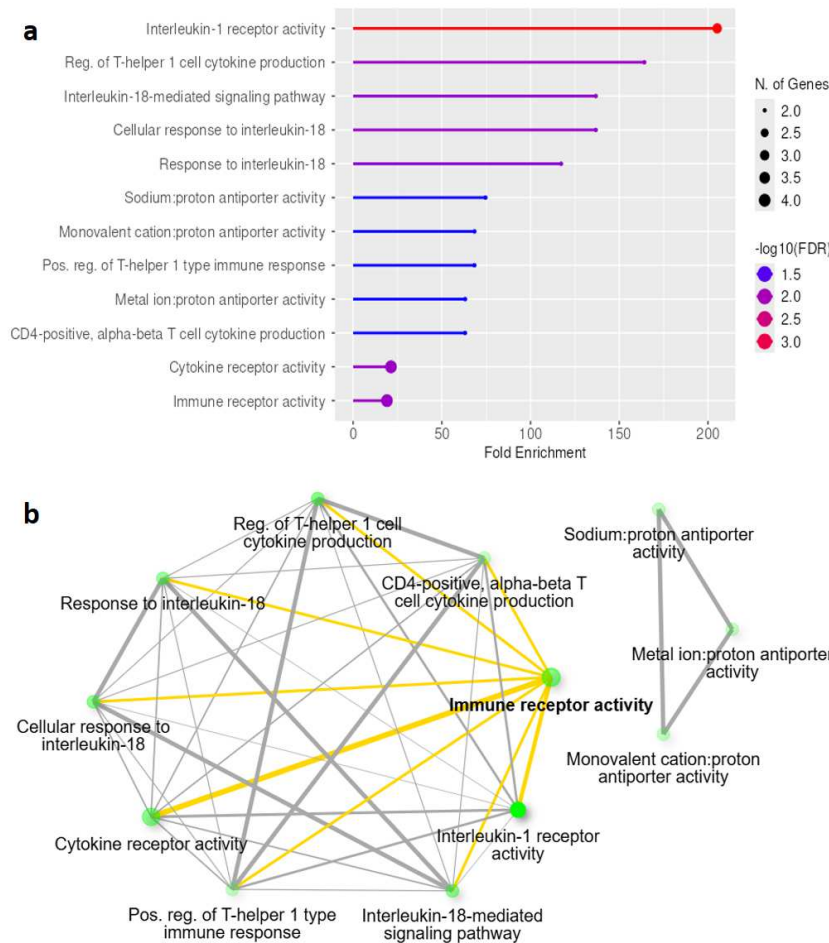
706 T-helper cell responses. Interleukin-1 receptor activity exhibited the strongest enrichment,

707 followed by Th1-associated cytokine production and IL-18-mediated signalling cascades. The

708 coordinated enrichment of IL-1 and IL-18 pathways underscores the centrality of pro-

inflammatory cytokine responses in modulating MAP susceptibility. More broadly, the enrichment of cytokine and immune receptor-related pathways suggests that host genetic variation shaping receptor-mediated immune recognition plays a key role in determining differential vulnerability to MAP infection (Figure 8a, b).

Figure 8. GO-Term and Pathway Enrichment Mapping of Immune Signalling Processes



Discussion

Model Performances

Benchmarking nine heterogeneous classifiers reveals a convergent pattern of discrimination: despite their different inductive assumptions, the models achieve similar AUC values and balanced sensitivity–specificity profiles, indicating that a reproducible, phenotype-associated signal is visible across modelling paradigms. Comparable discrimination, however, does not imply that the genetic host architecture is dominated by additive effects; rather, non-additive

and interaction effects may also contribute. Tree-based ensembles and sparsity-oriented linear methods emphasise different SNP subsets, and the most coherent set of variants emerges when these complementary importance patterns are combined under permutation-based FDR control. Calibration analyses show that almost identical ranking performance can occur alongside very different probability outputs, underscoring the need to assess both ranking ability and probabilistic accuracy. Across model types, train–test gaps remain narrow, consistent with a favourable signal-to-noise structure that limits simple memorisation. Cross-model importance correlations reveal additional structure: tree ensembles converge on a shared locus hierarchy, whereas linear models diverge, meaning that concordant signals represent the most stable effects while discordant ones illuminate alternative inferential perspectives. Partial agreement with mixed linear models and McNemar analyses shows that the ensemble captures both classically additive and proportion-based signals, while also pointing to additional loci that may reflect non-linear or interaction-driven effects. Together, these observations support an integrative ensemble strategy for characterising a multidimensional genomic architecture.

Immunogenetic Architecture Underlying Innate and Adaptive Immune Regulation

A diverse array of immune-regulatory genes orchestrates host defense, shaping antiviral resistance, inflammatory homeostasis, antigen processing, and immunopathology. The genes examined here converge on several core immunological modules—including interferon-stimulated antiviral programs, cytokine receptor signaling, immune checkpoint control, phagosome and autophagy pathways, glycan-mediated T-cell regulation, and NF- κ B-dependent inflammation—revealing a multilayered network governing both innate and adaptive immunity. In this context, we further highlight the subset of genes identified through our Machine Learning-based GWAS annotation pipeline that are already known to participate in immune system regulation, providing an integrated synthesis of ML-driven discovery with established immunobiological function.

1. Interferon-Stimulated Genes and Antiviral Innate Immunity

Interferon-inducible loci such as IFI27L2, IFI44, and IFI44L encode antiviral effectors that shape cell-intrinsic immunity. IFI27L2 localizes to mitochondria where it promotes apoptosis and restricts replication of neurotropic viruses including West Nile virus, operating as a cell type-specific antiviral determinant [37, 38]. IFI44 and IFI44L suppress IRF3 and NF- κ B signaling by FKBP5-dependent inhibition of IKK kinases, functioning as negative feedback

regulators that fine-tune type I/III IFN responses [39, 40]. Their strong induction across infection contexts—including influenza, RSV, HIV and hepatitis viruses—illustrates their conserved role in antiviral defense [41]. In bacterial immunity, IFI44L constrains early *Mycobacterium tuberculosis* growth, while epigenetic dysregulation of IFI44L and elevated IFI44 expression contributes to autoimmunity [42]. Collectively, these ISGs integrate pathogen sensing with inflammatory resolution, balancing antiviral potency and immunopathology.

2. Cytokine Receptors Governing Innate–Adaptive Crosstalk

The IL18R1–IL18RAP–IL1R1–IL1RL2 receptor cluster constitutes a central hub coordinating early inflammatory signaling. IL1R1 drives neutrophil recruitment and acute inflammation through MyD88-dependent NF- κ B activation [43]. The IL-18R1/IL-18RAP complex synergizes with IL-12 to elicit robust IFN- γ production by NK and Th1 cells, promoting type-1 immunity against intracellular pathogens [44]. Genetic evidence links IL18RAP variants to Crohn’s disease, ulcerative colitis, and leprosy, where alleles reduce receptor surface expression and weaken NF- κ B/MAPK responses in monocyte-derived macrophages [45, 46]. IL1RL2 (IL-36R) activates epithelial-driven inflammation, particularly in fungal infection and psoriasis [47]. Together, this receptor cluster modulates the magnitude and quality of innate recruitment, Th1/Th17 polarization, epithelial immunity, and chronic inflammatory susceptibility.

3. Immune Checkpoint and Immune-Evasion Regulators

Multiple genes in the dataset directly influence immune escape, T-cell suppression, and immune checkpoint dynamics. IMP3 (IGF2BP3) represses stress-induced NKG2D ligands ULBP2 and MICB, thereby dampening NK-cell surveillance and facilitating tumor immune evasion [48]. Additional IMP family members enhance immune checkpoint expression and limit interferon signaling, contributing to immunotherapy resistance [49]. HNRNPH1, a key RNA-binding protein, stabilizes NEAT1v2 to promote IL-8 transcription and controls the translation of JAK2 and PD-L1/PD-L2, integrating innate inflammation with adaptive immune suppression [50, 51]. Its transcriptional repression by AhR upregulates PD-L1 and IDO1, further strengthening immunosuppressive programs [52]. Additionally, ADGRL4, enriched in endothelial cells, fosters an immunosuppressive tumor microenvironment through M2-like macrophage polarization and PD-L1 upregulation, while impairing T-cell infiltration [53, 54]. These pathways collectively highlight the tight coupling between stromal signaling, NK/T-cell activation thresholds, and immune evasion.

4. *NF-κB Pathway Regulators and Inflammasome-Linked Immunity*

Innate immunity relies heavily on NF-κB-mediated transcriptional activation. OTUD1, a deubiquitinase, acts as a central negative regulator by targeting MAVS, TRAF3/6, and IRF3 for inactivation, thereby constraining type I IFN production. OTUD1 deficiency markedly enhances antiviral and antifungal immunity [55–57]. PPP4R4, through PP4C complexes, also suppresses TBK1 phosphorylation to downregulate antiviral signaling, reflecting an evolutionarily conserved mechanism for preventing excessive inflammation [58]. TRAPPC9 (NIBP) enhances NF-κB activation downstream of TLR signaling and is associated with heightened inflammatory cytokine production in macrophage-driven infectious mastitis [59, 60]. S100B, functioning as a danger-associated molecular pattern (DAMP), triggers MyD88/TRIF-dependent NF-κB activation and amplifies cytokine secretion, integrating metabolic stress with pathogen-driven inflammation [61]. These networked regulators calibrate the threshold between protective inflammation and harmful immunopathology.

Together these genes define a multi-tiered immunogenetic framework integrating antiviral interferon responses, cytokine receptor signaling, immune checkpoint regulation, vesicular antimicrobial pathways, and glycan-mediated T-cell control. Their coordinated activity illustrates how genetic variation shapes susceptibility to infection, chronic inflammation, immune evasion, and therapeutic responsiveness—providing a scalable foundation for precision immunology and host-directed interventions.

Pathway Analysis

Pathway analysis provides a mechanistic perspective on how SNP-level associations converge on coordinated biological processes, enabling the identification of signalling networks that may underlie differential susceptibility to infectious diseases. By integrating multi-model GWAS outputs with functional annotation, this analysis highlights immune pathways whose perturbation can alter host defence, cytokine signalling dynamics, and the interplay between innate and adaptive immunity.

Interleukin-1 receptor (IL-1R) signalling constitutes one of the core pathways of innate immune activation and plays a central role in host defence against infectious diseases. IL-1R family members—particularly IL-1R1 and IL-1R2—mediate the actions of the pro-inflammatory cytokines IL-1α and IL-1β, triggering neutrophil recruitment, fever induction, and acute-phase

protein synthesis [62]. Through MyD88-dependent signalling, IL-1R activation drives NF- κ B and MAPK pathways, thereby enhancing antimicrobial peptide production and phagocyte function [63]. Impaired IL-1R activity increases susceptibility to bacterial, viral, and fungal pathogens and delays pathogen clearance [64]. Beyond innate immunity, IL-1R signalling provides essential cues for adaptive immune activation by modulating T-cell differentiation and antibody responses [65].

Regulation of Th1 cytokine production represents a crucial axis of cellular immunity and is predominantly orchestrated through the reciprocal interplay between IL-12 and interferon- γ (IFN- γ). IL-12 derived from antigen-presenting cells activates STAT4, initiating the differentiation of naïve CD4⁺ T cells toward a Th1 phenotype and inducing the expression of the transcription factor T-bet. This IL-12–STAT4 signalling cascade forms a critical bridge between innate and adaptive immunity, as pathogen-derived stimuli stimulate IL-12 secretion, which in turn promotes robust IFN- γ production by natural killer and T cells [66].

The IL-18–mediated signalling pathway is another fundamental regulator of both innate and adaptive immune responses. IL-18 binds to a heterodimeric receptor composed of IL-18R α and IL-18R β , initiating MyD88-dependent signalling through its TIR domain [67]. As a member of the IL-1 cytokine family, IL-18 relays danger signals to immune cells that cannot directly recognise microbial antigens, thereby strengthening innate immune activation and linking it to downstream adaptive responses [68]. During infectious disease, IL-18 plays a particularly pronounced role. For example, during influenza A virus infection, NLRP3 inflammasome activation converts pro-IL-18 into its mature form, promoting effective adaptive immune responses [69].

IL-18 functions synergistically with IL-12 to activate natural killer cells and protect the host against infections and malignancies. Mouse models lacking IL-18 show impaired NK-cell activity and heightened infection susceptibility. Clearance of intracellular bacteria, fungi, protozoa, and viruses requires host-derived interferon- γ , for which IL-18 is indispensable. In *Mycobacterium avium* infection, IL-18 mediates protective immunity by inducing IFN- γ , while in *Leishmania major* infection, co-administration of IL-12 and IL-18 confers resistance to reinfection. IL-18 also enhances cytotoxic CD8⁺ T-cell activation and supports antiviral defence, as demonstrated by impaired viral clearance in IL-18-deficient mice. The proteolytic conversion of pro-IL-18 to its active form by caspase-1 following inflammasome activation

amplifies innate immune responses to microbial antigens and danger signals, thereby providing a mechanistic link between early immune activation and subsequent adaptive immunity [68].

The regulation of Th1-type immune responses forms a central determinant of host protection, as the balance between Th1 and Th2 differentiation dictates susceptibility to bacterial, parasitic, and viral pathogens. IL-12 directs Th1 polarization, whereas IL-4 promotes Th2 commitment, establishing a cytokine environment that determines adaptive immune trajectories. During early responses to intracellular pathogens, macrophage-derived IL-12 acts as the initiating cue by stimulating natural killer cells to produce IFN- γ . This initial IFN- γ burst enhances macrophage activation and induces further IL-12 secretion, forming a self-reinforcing loop that stabilises Th1 programming. At the transcriptional level, IL-12 signalling activates the STAT family—especially STAT4—which serves as a principal regulator of Th1 lineage commitment. Disruption of this axis severely compromises Th1 immunity, highlighting its non-redundant role in antiviral and antibacterial defence [70].

Cytokine receptor activity, defined as the binding of cytokines to their specific cell-surface receptors and the initiation of downstream signal transduction, is essential for regulating innate immunity through the JAK-STAT pathway [71]. Upon pathogen recognition by pattern-recognition receptors, inflammatory cytokines such as IL-1, IL-6, TNF- α , and IFN- γ bind their cognate receptors to activate JAK proteins, trigger STAT phosphorylation, and induce transcription of antimicrobial effector genes [72]. Interleukin receptors are particularly critical in bacterial infection; for example, IL-27 receptor signalling drives excessive immunosuppression during neonatal Gram-negative sepsis [73].

Immune receptors provide the molecular foundation for pathogen detection and immune activation. Key receptor families—including Toll-like receptors (TLRs) and nucleotide-binding oligomerization domain (NOD)-like receptors (NLRs)—identify pathogen-associated and damage-associated molecular patterns [74, 75]. Their activation initiates signalling cascades that guide both innate and adaptive immune responses, ensuring coordinated host defence against invading microorganisms [76].

In conclusion, our findings suggest a genomic landscape in which MAP susceptibility in Türkiye's goat populations presents as a biologically coherent and methodologically consistent signal. The agreement observed across diverse machine-learning frameworks—each operating under different inductive assumptions—indicates that the detected structure of the trait is not

strongly tied to any single modelling approach. This convergence, together with stable performance on unseen data, lends support to the reliability of the patterns identified and illustrates how modern learning methods can complement classical tools in exploring complex genetic architectures. If similar signals are observed in broader datasets, they may help refine genomic prediction efforts in small ruminants and inform selection strategies aimed at reducing MAP prevalence and supporting long-term herd resilience. Additionally, by contributing to marker-assisted selection efforts that limit pathogen circulation at its animal-source origin, such genomic insights may align with broader One Health objectives, promoting improvements in livestock management, environmental health, and, indirectly, human well-being.

Acknowledgements

This project was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK), and we extend our sincere gratitude to TÜBİTAK for their support. The corresponding author also expresses profound gratitude to Gazi Mustafa Kemal ATATÜRK, the founding leader of modern Türkiye.

Author contributions

Y.Y. conceptualized and designed the study, performed the GWAS analysis, and wrote the manuscript. Y.Y., R.A., M.K., Ö.G., and K.I. conducted the laboratory work. S.Y. organized the field work. Y.Y., A.E., D.C., Y.E.K., S.S.Ş., and M.B. conducted the field work and data collection.

Funding

This research project was funded by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) (Project No: TOVAG-222O107).

Data availability

The data used in this study have been deposited in the figshare.com database under <https://figshare.com/s/6936b8fd55d0f15ae47c>

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the guidelines of the Local Ethics Committee for Animal Experiments and with an experimental protocol approved by the “Ethics Committee for the Use of Animals in Research and Experimentation” at the Sheep Breeding and Research Institute, Türkiye (Approval No: 066/14.02.2023). Informed consent was obtained from the Ministry of Agriculture and Forestry, Siirt Provincial Directorate of Agriculture (Approval No: E-64380313-325.01-8848896/08.02.2023) prior to the study. The authors also complied with the ARRIVE guidelines.

Consent to Publish declaration

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References:

1. Dow CT, Alvarez BL. Mycobacterium paratuberculosis zoonosis is a One Health emergency. *EcoHealth*. 2022;19:164–74. <https://doi.org/10.1007/s10393-022-01602-x>.
2. Garcia AB, Shalloo L. Invited review: The economic impact and control of paratuberculosis in cattle. *J Dairy Sci*. 2015;98:5019–39. <https://doi.org/10.3168/jds.2014-9241>.
3. Maraudam V, Mohana B, Praveen K, Raj D. Paratuberculosis: Diagnostic Methods and their Constraints. *J Vet Sci Technol*. 2015;6. <https://doi.org/10.4172/2157-7579.1000259>.
4. Sechi LA, Scanu AM, Mollicotti P, Cannas S, Mura M, Dettori G, et al. Detection and Isolation of Mycobacterium avium Subspecies paratuberculosis from Intestinal Mucosal Biopsies of Patients with and without Crohn's Disease in Sardinia. *Off J Am Coll Gastroenterol ACG*. 2005;100:1529.
5. Garvey M. Mycobacterium avium subspecies paratuberculosis: A possible causative agent in human morbidity and risk to public health safety. *Open Vet J*. 2018;8:172–81. <https://doi.org/10.4314/ovj.v8i2.10>.
6. O'Brien D, Scudamore J, Charlier J, Delavergne M. DISCONTTOOLS: a database to identify research gaps on vaccines, pharmaceuticals and diagnostics for the control of infectious diseases of animals. *BMC Vet Res*. 2017;13:1. <https://doi.org/10.1186/s12917-016-0931-1>.
7. McGovern SP, Purfield DC, Ring SC, Carthy TR, Graham DA, Berry DP. Candidate genes associated with the heritable humoral response to Mycobacterium avium ssp. paratuberculosis in dairy cows have factors in common with gastrointestinal diseases in humans. *J Dairy Sci*. 2019;102:4249–63. <https://doi.org/10.3168/jds.2018-15906>.
8. Biet F, Boschioli ML, Thorel MF, Guilloteau LA. Zoonotic aspects of Mycobacterium bovis and Mycobacterium avium-intracellulare complex (MAC). *Vet Res*. 2005;36:411–36. <https://doi.org/10.1051/vetres:2005001>.
9. Whittington R, Donat K, Weber MF, Kelton D, Nielsen SS, Eisenberg S, et al. Control of paratuberculosis: who, why and how. A review of 48 countries. *BMC Vet Res*. 2019;15:198. <https://doi.org/10.1186/s12917-019-1943-4>.
10. Jiménez-Martín D, García-Bocanegra I, Rialde MA, Fernández-Molera V, Jiménez-Ruiz S, Isla J, et al. Epidemiology of paratuberculosis in sheep and goats in southern Spain. *Prev Vet Med*. 2022;202:105637. <https://doi.org/10.1016/j.prevetmed.2022.105637>.
11. Di Marco Lo Presti V, Ippolito D, Migliore S, Tolone M, Mignacca SA, Marino AMF, et al. Large-scale serological survey on Mycobacterium avium subsp. paratuberculosis infection in sheep and goat herds in Sicily, Southern Italy. *Front Vet Sci*. 2024;11. <https://doi.org/10.3389/fvets.2024.1334036>.
12. Fernández-Silva JA, Correa-Valencia NM, Ramírez NF. Systematic review of the prevalence of paratuberculosis in cattle, sheep, and goats in Latin America and the Caribbean. *Trop Anim Health Prod*. 2014;46:1321–40. <https://doi.org/10.1007/s11250-014-0656-8>.

975 13. Yu Y, Zhang S, Xu G, Xu D, Zheng H, Li B, et al. Identification of *Mycobacterium avium*
976 subspecies paratuberculosis in sheep farms in Bayannaoer, Inner Mongolia, China (short
977 communication). *BMC Vet Res.* 2022;18:281. <https://doi.org/10.1186/s12917-022-03293-6>.

978 14. Okuni JB, Hansen S, Eltom KH, Eltayeb E, Amanzada A, Omega JA, et al.
979 Paratuberculosis: A Potential Zoonosis and a Neglected Disease in Africa. *Microorganisms.*
980 2020;8:1007. <https://doi.org/10.3390/microorganisms8071007>.

981 15. Buyuk F, Celebi O, Akca D, Otlu S, Tazegul E, Gulmez A, et al. Estimated apparent and
982 true prevalences of paratuberculosis in sheep herds of the Kars Region in Northeastern
983 Turkey. *Veterinární Medicína.* 2014;59:331–5. <https://doi.org/10.17221/7619-VETMED>.

984 16. Celik A, Turutoglu H. Seroprevalence of paratuberculosis in cattle, sheep and goats in
985 Burdur, Southwestern Turkey. 2017;72:30–6.

986 17. Yaman Y, Aymaz R, Keleş M, Bay V, Ün C, Heaton MP. Association of TLR2 haplotypes
987 encoding Q650 with reduced susceptibility to ovine Johne's disease in Turkish sheep. *Sci*
988 *Rep.* 2021;11:7088. <https://doi.org/10.1038/s41598-021-86605-4>.

989 18. Cinar MU, Hizlisoy H, Akyüz B, Arslan K, Aksel EG, Gümüşsoy KS. Polymorphisms in
990 toll-like receptor (TLR) 1, 4, 9 and SLC11A1 genes and their association with
991 paratuberculosis susceptibility in Holstein and indigenous crossbred cattle in Turkey. *J Genet.*
992 2018;97:1147–54. <https://doi.org/10.1007/s12041-018-1008-7>.

993 19. Koets A, Santema W, Mertens H, Oostenrijk D, Keesstra M, Overdijk M, et al.
994 Susceptibility to paratuberculosis infection in cattle is associated with single nucleotide
995 polymorphisms in Toll-like receptor 2 which modulate immune responses against
996 *Mycobacterium avium* subspecies paratuberculosis. *Prev Vet Med.* 2010;93:305–15.
997 <https://doi.org/10.1016/j.prevetmed.2009.11.008>.

998 20. Mallikarjunappa S, Sargolzaei M, Brito LF, Meade KG, Karrow NA, Pant SD. Short
999 communication: Uncovering quantitative trait loci associated with resistance to
1000 *Mycobacterium avium* ssp. paratuberculosis infection in Holstein cattle using a high-density
1001 single nucleotide polymorphism panel. *J Dairy Sci.* 2018;101:7280–6.
1002 <https://doi.org/10.3168/jds.2018-14388>.

1003 21. Reddacliff LA, Beh K, McGregor H, Whittington RJ. A preliminary study of possible
1004 genetic influences on the susceptibility of sheep to Johne's disease. *Aust Vet J.* 2005;83:435–
1005 41. <https://doi.org/10.1111/j.1751-0813.2005.tb13087.x>.

1006 22. Bhide MR, Mucha R, Mikula I, Kisova L, Skrabana R, Novak M, et al. Novel mutations
1007 in TLR genes cause hyporesponsiveness to *Mycobacterium avium* subsp. paratuberculosis
1008 infection. *BMC Genet.* 2009;10:21. <https://doi.org/10.1186/1471-2156-10-21>.

1009 23. Moioli B, D'Andrea S, De Grossi L, Sezzi E, De Sanctis B, Catillo G, et al. Genomic scan
1010 for identifying candidate genes for paratuberculosis resistance in sheep. *Anim Prod Sci.*
1011 2016;56:1046–55.

1012 24. Yaman Y. Association of toll-like receptor 4 (TLR 4) gene exon 3 variants with serostatus
1013 of the ovine Johne's disease (paratuberculosis) in Turkish sheep. *Turk J Vet Anim Sci.*
1014 2020;44:542–7. <https://doi.org/10.3906/vet-1909-98>.

- 1015 25. Yaman Y, Aymaz R, Keleş M, Bay V, Hatipoğlu E, Kaptan C, et al. Evaluation of CD109,
1016 PCP4 and SEMA3D genes for their association with Ovine Johne's disease in Turkish sheep.
1017 Anim Biotechnol. 2021;32:519–25. <https://doi.org/10.1080/10495398.2021.1887880>.
- 1018 26. Singh PK, Singh SV, Singh MK, Saxena VK, Horin P, Singh AV, et al. Effect of genetic
1019 variation in the MHC Class II DRB region on resistance and susceptibility to Johne's disease
1020 in endangered Indian Jamunapari goats. Int J Immunogenet. 2012;39:314–20.
1021 <https://doi.org/10.1111/j.1744-313X.2012.01092.x>.
- 1022 27. Abraham A, Naicy T, Raghavan KC, Siju J, Aravindakshan T. Evaluation of the
1023 association of SLC11A1 gene polymorphism with incidence of paratuberculosis in goats. J
1024 Genet. 2017;96:641–6. <https://doi.org/10.1007/s12041-017-0820-9>.
- 1025 28. Cecchi F, Russo C, Iamartino D, Galiero A, Turchi B, Fratini F, et al. Identification of
1026 candidate genes for paratuberculosis resistance in the native Italian Garfagnina goat breed.
1027 Trop Anim Health Prod. 2017;49:1135–42. <https://doi.org/10.1007/s11250-017-1306-8>.
- 1028 29. Yoosefzadeh-Najafabadi M, Eskandari M, Torabi S, Torkamaneh D, Tulpan D, Rajcan I.
1029 Machine-Learning-Based Genome-Wide Association Studies for Uncovering QTL
1030 Underlying Soybean Yield and Its Components. Int J Mol Sci. 2022;23:5538.
1031 <https://doi.org/10.3390/ijms23105538>.
- 1032 30. Yoosefzadeh-Najafabadi M, Torabi S, Tulpan D, Rajcan I, Eskandari M. Application of
1033 SVR-Mediated GWAS for Identification of Durable Genetic Regions Associated with
1034 Soybean Seed Quality Traits. Plants Basel Switz. 2023;12:2659.
1035 <https://doi.org/10.3390/plants12142659>.
- 1036 31. De Velasco Oriol J, Vallejo EE, Estrada K, Taméz Peña JG, Disease Neuroimaging
1037 Initiative TA. Benchmarking machine learning models for late-onset alzheimer's disease
1038 prediction from genomic data. BMC Bioinformatics. 2019;20:709.
1039 <https://doi.org/10.1186/s12859-019-3158-x>.
- 1040 32. Gudhe NR, Hartikainen JM, Tengstrom M, Pylkas K, Winqvist R, Kosma V-M, et al.
1041 GenoGraph: an Interpretable Graph Contrastive Learning Approach for Identifying Breast
1042 Cancer Risk Variants. IEEE Trans Comput Biol Bioinforma. 2025;PP.
1043 <https://doi.org/10.1109/TCBBIO.2025.3617088>.
- 1044 33. Elmas A, Ou Yang T-H, Wang X, Anastassiou D. Discovering Genome-Wide Tag SNPs
1045 Based on the Mutual Information of the Variants. PloS One. 2016;11:e0167994.
1046 <https://doi.org/10.1371/journal.pone.0167994>.
- 1047 34. Vergara JR, Estévez PA. A review of feature selection methods based on mutual
1048 information. Neural Comput Appl. 2014;24:175–86. <https://doi.org/10.1007/s00521-013-1368-0>.
- 1050 35. Behravan H, Hartikainen JM, Tengström M, Pylkäs K, Winqvist R, Kosma V-M, et al.
1051 Machine learning identifies interacting genetic variants contributing to breast cancer risk: A
1052 case study in Finnish cases and controls. Sci Rep. 2018;8:13149.
1053 <https://doi.org/10.1038/s41598-018-31573-5>.
- 1054 36. Ruigrok M, Xue B, Catanach A, Zhang M, Jesson L, Davy M, et al. The Relative Power
1055 of Structural Genomic Variation versus SNPs in Explaining the Quantitative Trait Growth in

1056 the Marine Teleost *Chrysophrys auratus*. *Genes*. 2022;13:1129.
1057 <https://doi.org/10.3390/genes13071129>.

1058 37. Lucas TM, Richner JM, Diamond MS. The Interferon-Stimulated Gene Ifi2712a Restricts
1059 West Nile Virus Infection and Pathogenesis in a Cell-Type- and Region-Specific Manner. *J*
1060 *Virol*. 2015;90:2600–15. <https://doi.org/10.1128/JVI.02463-15>.

1061 38. Villamayor L, López-García D, Rivero V, Martínez-Sobrido L, Nogales A, DeDiego ML.
1062 The IFN-stimulated gene IFI27 counteracts innate immune responses after viral infections by
1063 interfering with RIG-I signaling. *Front Microbiol*. 2023;14:1176177.
1064 <https://doi.org/10.3389/fmicb.2023.1176177>.

1065 39. DeDiego ML, Martinez-Sobrido L, Topham DJ. Novel Functions of IFI44L as a Feedback
1066 Regulator of Host Antiviral Responses. *J Virol*. 2019;93:e01159-19.
1067 <https://doi.org/10.1128/JVI.01159-19>.

1068 40. DeDiego ML, Nogales A, Martinez-Sobrido L, Topham DJ. Interferon-Induced Protein 44
1069 Interacts with Cellular FK506-Binding Protein 5, Negatively Regulates Host Antiviral
1070 Responses, and Supports Virus Replication. *mBio*. 2019;10:e01839-19.
1071 <https://doi.org/10.1128/mBio.01839-19>.

1072 41. Busse DC, Habgood-Coote D, Clare S, Brandt C, Bassano I, Kaforou M, et al. Interferon-
1073 Induced Protein 44 and Interferon-Induced Protein 44-Like Restrict Replication of
1074 Respiratory Syncytial Virus. *J Virol*. 2020;94:e00297-20. [https://doi.org/10.1128/JVI.00297-](https://doi.org/10.1128/JVI.00297-20)
1075 20.

1076 42. Zhao M, Zhou Y, Zhu B, Wan M, Jiang T, Tan Q, et al. IFI44L promoter methylation as a
1077 blood biomarker for systemic lupus erythematosus. *Ann Rheum Dis*. 2016;75:1998–2006.
1078 <https://doi.org/10.1136/annrheumdis-2015-208410>.

1079 43. Mizgerd JP, Lupa MM, Kogan MS, Warren HB, Kobzik L, Topulos GP. Nuclear factor-
1080 kappaB p50 limits inflammation and prevents lung injury during *Escherichia coli* pneumonia.
1081 *Am J Respir Crit Care Med*. 2003;168:810–7. <https://doi.org/10.1164/rccm.200303-412OC>.

1082 44. Nakanishi K, Yoshimoto T, Tsutsui H, Okamura H. Interleukin-18 regulates both Th1 and
1083 Th2 responses. *Annu Rev Immunol*. 2001;19:423–74.
1084 <https://doi.org/10.1146/annurev.immunol.19.1.423>.

1085 45. Liu H, Irwanto A, Tian H, Fu X, Yu Y, Yu G, et al. Identification of IL18RAP/IL18R1 and
1086 IL12B as Leprosy Risk Genes Demonstrates Shared Pathogenesis between Inflammation and
1087 Infectious Diseases. *Am J Hum Genet*. 2012;91:935–41.
1088 <https://doi.org/10.1016/j.ajhg.2012.09.010>.

1089 46. Zhernakova A, Festen EM, Franke L, Trynka G, Diemen CC van, Monsuur AJ, et al.
1090 Genetic Analysis of Innate Immunity in Crohn's Disease and Ulcerative Colitis Identifies Two
1091 Susceptibility Loci Harboring CARD9 and IL18RAP. *Am J Hum Genet*. 2008;82:1202–10.
1092 <https://doi.org/10.1016/j.ajhg.2008.03.016>.

1093 47. Mahil SK, Catapano M, Di Meglio P, Dand N, Ahlfors H, Carr IM, et al. An analysis of
1094 IL-36 signature genes and individuals with IL1RL2 knockout mutations validates IL-36 as a
1095 psoriasis therapeutic target. *Sci Transl Med*. 2017;9:aan2514.
1096 <https://doi.org/10.1126/scitranslmed.aan2514>.

1097 48. Schmiedel D, Tai J, Yamin R, Berhani O, Bauman Y, Mandelboim O. The RNA binding
1098 protein IMP3 facilitates tumor immune escape by downregulating the stress-induced ligands
1099 ULPB2 and MICB. *eLife*. 2016;5:e13426. <https://doi.org/10.7554/eLife.13426>.

1100 49. Elcheva IA, Gowda CP, Bogush D, Gornostaeva S, Fakhardo A, Sheth N, et al. IGF2BP
1101 family of RNA-binding proteins regulate innate and adaptive immune responses in cancer
1102 cells and tumor microenvironment. *Front Immunol*. 2023;14:1224516.
1103 <https://doi.org/10.3389/fimmu.2023.1224516>.

1104 50. Mineo M, Lyons SM, Zdioruk M, von Spreckelsen N, Ferrer-Luna R, Ito H, et al. Tumor
1105 Interferon Signaling Is Regulated by a lncRNA INCR1 Transcribed from the PD-L1 Locus.
1106 *Mol Cell*. 2020;78:1207-1223.e8. <https://doi.org/10.1016/j.molcel.2020.05.015>.

1107 51. Tanu T, Taniue K, Imamura K, Onoguchi-Mizutani R, Han H, Jensen TH, et al.
1108 hnRNPH1-MTR4 complex-mediated regulation of NEAT1v2 stability is critical for IL8
1109 expression. *RNA Biol*. 2021;18:537–47. <https://doi.org/10.1080/15476286.2021.1971439>.

1110 52. Lara B, Snyder M, Fimbres J, Yang E, Song G, Duggineni VK, et al. The AhR regulates
1111 IFN-induced immune checkpoints in lung cancer cells through HNRNPH1, an RNA-binding
1112 protein, and INCR1, a novel long non-coding RNA. *J Biol Chem*. 2025;301:110316.
1113 <https://doi.org/10.1016/j.jbc.2025.110316>.

1114 53. Huang H, Georganaki M, Conze LL, Laviña B, van Hooren L, Vemuri K, et al. ELTD1
1115 deletion reduces vascular abnormality and improves T-cell recruitment after PD-1 blockade in
1116 glioma. *Neuro-Oncol*. 2022;24:398–411. <https://doi.org/10.1093/neuonc/noab181>.

1117 54. Sheldon H, Bridges E, Silva I, Masiero M, Favara DM, Wang D, et al. ADGRL4/ELTD1
1118 Expression in Breast Cancer Cells Induces Vascular Normalization and Immune Suppression.
1119 *Mol Cancer Res MCR*. 2021;19:1957–69. <https://doi.org/10.1158/1541-7786.MCR-21-0171>.

1120 55. Chen X, Zhang H, Wang X, Shao Z, Li Y, Zhao G, et al. OTUD1 Regulates Antifungal
1121 Innate Immunity through Deubiquitination of CARD9. *J Immunol Baltim Md 1950*.
1122 2021;206:1832–43. <https://doi.org/10.4049/jimmunol.2001253>.

1123 56. Zhang L, Liu J, Qian L, Feng Q, Wang X, Yuan Y, et al. Induction of OTUD1 by RNA
1124 viruses potently inhibits innate immune responses by promoting degradation of the
1125 MAVS/TRAF3/TRAF6 signalosome. *PLoS Pathog*. 2018;14:e1007067.
1126 <https://doi.org/10.1371/journal.ppat.1007067>.

1127 57. Zhang Z, Wang D, Wang P, Zhao Y, You F. OTUD1 Negatively Regulates Type I IFN
1128 Induction by Disrupting Noncanonical Ubiquitination of IRF3. *J Immunol Baltim Md 1950*.
1129 2020;204:1904–18. <https://doi.org/10.4049/jimmunol.1900305>.

1130 58. Zhan Z, Cao H, Xie X, Yang L, Zhang P, Chen Y, et al. Phosphatase PP4 Negatively
1131 Regulates Type I IFN Production and Antiviral Innate Immunity by Dephosphorylating and
1132 Deactivating TBK1. *J Immunol Baltim Md 1950*. 2015;195:3849–57.
1133 <https://doi.org/10.4049/jimmunol.1403083>.

1134 59. Khan MZ, Khan A, Xiao J, Ma J, Ma Y, Chen T, et al. Overview of Research
1135 Development on the Role of NF-κB Signaling in Mastitis. *Anim Open Access J MDPI*.
1136 2020;10:1625. <https://doi.org/10.3390/ani10091625>.

1137 60. Khan MZ, Dari G, Khan A, Yu Y. Genetic polymorphisms of TRAPPC9 and CD4 genes
1138 and their association with milk production and mastitis resistance phenotypic traits in Chinese
1139 Holstein. *Front Vet Sci.* 2022;9:1008497. <https://doi.org/10.3389/fvets.2022.1008497>.

1140 61. Sorci G, Giovannini G, Riuzzi F, Bonifazi P, Zelante T, Zagarella S, et al. The danger
1141 signal S100B integrates pathogen- and danger-sensing pathways to restrain inflammation.
1142 *PLoS Pathog.* 2011;7:e1001315. <https://doi.org/10.1371/journal.ppat.1001315>.

1143 62. Dinarello CA. Overview of the IL-1 family in innate inflammation and acquired
1144 immunity. *Immunol Rev.* 2018;281:8–27. <https://doi.org/10.1111/imr.12621>.

1145 63. Garlanda C, Dinarello CA, Mantovani A. The interleukin-1 family: back to the future.
1146 *Immunity.* 2013;39:1003–18. <https://doi.org/10.1016/j.immuni.2013.11.010>.

1147 64. Netea MG, Simon A, Veerdonk F van de, Kullberg B-J, Meer JWMV der, Joosten LAB.
1148 IL-1 β Processing in Host Defense: Beyond the Inflammasomes. *PLOS Pathog.*
1149 2010;6:e1000661. <https://doi.org/10.1371/journal.ppat.1000661>.

1150 65. Ben-Sasson SZ, Hogg A, Hu-Li J, Wingfield P, Chen X, Crank M, et al. IL-1 enhances
1151 expansion, effector function, tissue localization, and memory response of antigen-specific
1152 CD8 T cells. *J Exp Med.* 2013;210:491–502. <https://doi.org/10.1084/jem.20122006>.

1153 66. Trinchieri G. Interleukin-12 and the regulation of innate resistance and adaptive immunity.
1154 *Nat Rev Immunol.* 2003;3:133–46. <https://doi.org/10.1038/nri1001>.

1155 67. Adachi O, Kawai T, Takeda K, Matsumoto M, Tsutsui H, Sakagami M, et al. Targeted
1156 disruption of the MyD88 gene results in loss of IL-1- and IL-18-mediated function. *Immunity.*
1157 1998;9:143–50. [https://doi.org/10.1016/s1074-7613\(00\)80596-8](https://doi.org/10.1016/s1074-7613(00)80596-8).

1158 68. Ihim SA, Abubakar SD, Zian Z, Sasaki T, Saffarioun M, Maleknia S, et al. Interleukin-18
1159 cytokine in immunity, inflammation, and autoimmunity: Biological role in induction,
1160 regulation, and treatment. *Front Immunol.* 2022;13:919973.
1161 <https://doi.org/10.3389/fimmu.2022.919973>.

1162 69. Ichinohe T, Lee HK, Ogura Y, Flavell R, Iwasaki A. Inflammasome recognition of
1163 influenza virus is essential for adaptive immune responses. *J Exp Med.* 2009;206:79–87.
1164 <https://doi.org/10.1084/jem.20081667>.

1165 70. Taki S, Sato T, Ogasawara K, Fukuda T, Sato M, Hida S, et al. Multistage regulation of
1166 Th1-type immune responses by the transcription factor IRF-1. *Immunity.* 1997;6:673–9.
1167 [https://doi.org/10.1016/s1074-7613\(00\)80443-4](https://doi.org/10.1016/s1074-7613(00)80443-4).

1168 71. Morris R, Kershaw NJ, Babon JJ. The molecular details of cytokine signaling via the
1169 JAK/STAT pathway. *Protein Sci Publ Protein Soc.* 2018;27:1984–2009.
1170 <https://doi.org/10.1002/pro.3519>.

1171 72. Murray PJ, Allen JE, Biswas SK, Fisher EA, Gilroy DW, Goerdts S, et al. Macrophage
1172 activation and polarization: nomenclature and experimental guidelines. *Immunity.*
1173 2014;41:14–20. <https://doi.org/10.1016/j.immuni.2014.06.008>.

1174 73. Seman BG, Vance JK, Rawson TW, Witt MR, Huckaby AB, Povroznik JM, et al. Elevated
1175 Levels of Interleukin-27 in Early Life Compromise Protective Immunity in a Mouse Model of

1176 Gram-Negative Neonatal Sepsis. *Infect Immun*. 2020;88:e00828-19.
 1177 <https://doi.org/10.1128/IAI.00828-19>.

1178 74. Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. *Cell*.
 1179 2006;124:783–801. <https://doi.org/10.1016/j.cell.2006.02.015>.

1180 75. Franchi L, Muñoz-Planillo R, Núñez G. Sensing and reacting to microbes through the
 1181 inflammasomes. *Nat Immunol*. 2012;13:325–32. <https://doi.org/10.1038/ni.2231>.

1182 76. Negishi H, Matsuki K, Endo N, Sarashina H, Miki S, Matsuda A, et al. Beneficial innate
 1183 signaling interference for antibacterial responses by a Toll-like receptor–mediated
 1184 enhancement of the MKP-IRF3 axis. *Proc Natl Acad Sci*. 2013;110:19884–9.
 1185 <https://doi.org/10.1073/pnas.1320145110>.

1186