

Supplementary Information for:

Cross-Species Analysis of Genetic Architecture and Polygenic Risk Scores for Non-Contact ACL Rupture in Dogs and Humans

Mehdi Momen^{1*}, Hannah K. Kearney¹, Margaret M. Patterson¹, Susannah J. Sample¹, Zijie Zhao², Qiongshi Lu², Guilherme J.M. Rosa³, Peter Muir^{1*}

¹Department of Surgical Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

²Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

³Department of Animal and Dairy Sciences, College of Agriculture and Life Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

This PDF file includes:

Supplementary
Methods

Supplementary
Results

Supplementary References and
Discussion

Table S1

Figures S1 to S13

Supplementary Methods

Anterior cruciate ligament (ACL) rupture narrow sense heritability estimation

Restricted Maximum Likelihood (REML) mixed model. We considered the non-contact ACL rupture control/case phenotype as a 0/1 variable and used the REML procedure for a quantitative trait to calculate the heritability of the binary phenotype on the observed scale and then converted the estimate to the liability scale. We performed the REML mixed model analysis using GCTA (1). The analysis involved a total of 1,006 individual dogs, with 440 Labrador Retriever cases and 566 controls. Heritability may be calculated based on either the unobserved liability scale (denoted h_l^2) or the observed binary disease phenotype (denoted h_o^2). Lee et al. (2) proposed an extended derivation to account for the fact that, in a case-control study, cases tend to be over-represented compared to the population prevalence, arriving at the following expression:

$$h_l^2 = h_o^2 \frac{K(1-K)K(1-K)}{z^2 P(1-P)}$$

Here, h_o^2 is observed single nucleotide polymorphism (SNP) heritability, K is population disease prevalence, P is the proportion of cases in the sample and z is the height of the standard normal probability density function at the truncation threshold. A population prevalence of 5.79% in the Labrador Retriever (3) was used for this calculation.

Probit Bayesian linear mixed model. A probit Bayesian linear mixed model (LMM) was used to estimate genetic variance components and obtain the posterior distribution of non-contact ACL rupture heritability. Since dogs were phenotyped as case or control, this follows a binomial distribution as we modeled ACL rupture using a normally distributed latent variable (liability score). Therefore, we consider the probability of ACL rupture for the j -th outcome on the i -th dog (p_{ij}), can be modeled as:

$$\text{logit}(p_{ij}) = \log \left[\frac{p_{ij}}{1-p_{ij}} \right] = \mu + X_{ij}\beta_{ij} + g_i + e_{ij}$$

Where μ , is the total population mean which is a constant common to all dogs; X_{ij} can be dedicated to a set of recorded nongenetic or systematic effects and here we considered a sex effect (1 for male and 2 for female). β_{ij} is the unknown vector of effects associated with a sex effect; g_i is the additive genetic contribution to ACL rupture status for the i -th dog and assumed to be sampled from a multivariate normal density with $g|\sigma_g^2 \sim N(0, G\sigma_g^2)$, so it was a vector of liability values for all 1,006 dogs, and G being the genomic relationship matrix (GRM). For GRM calculation, we used a standard approach (4).

$$G = \frac{WW'}{2 \sum_{k=1}^m p_j(1-p_j)}$$

Where m refers to the number of SNPs, $W(n \times m)$ is a matrix containing all additive contributions from m SNPs of n animals, coded as genotypes 0 (AA), 1 (AB), and 2 (BB). In this model the residuals are assumed to follow a multivariate normal distribution with mean zero and identity variance (I) i.e., $e \sim N(0, I\sigma_e^2)$. Notably, in a binary or ordinal latent model, the residual variance may not be identifiable. A solution to overcome this complexity is to fix one of the variance components to a known constant C (for example $\sigma_e^2=C$). The model was fitted using the MCMCglmm function in R. MCMC sampling was employed to explore the posterior distribution of model parameters. The MCMC algorithm ran for 1 million iterations, with a burn-in period of 50,000 iterations and a thinning rate of 2 to ensure efficient sampling and convergence. Incorporating prior knowledge, we defined informative priors for the model parameters. The residual variance (e) was assigned a fixed value of 1, addressing identifiability concerns associated with residual variance. For the genetic effects (g), we assumed a weakly informative prior, allowing flexibility in the genetic covariance estimation. We set the variance of the random genetic effect to 1, expressing a prior expectation of moderate genetic variability. The degrees of freedom (ν) for the inverse-Wishart prior were set to 1,000, indicating limited prior information about the genetic covariance pattern. Additionally, no prior bias was imposed on the prior mean of the random genetic effect ($\alpha.\mu = 0$), while the variance of the prior distribution was set to 1 ($\alpha.V=1$).

$$h_{probit}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2 + 1}$$

Regional heritability mapping (RHM) of ACL rupture

The genome was divided into regions (30 single nucleotide polymorphisms (SNPs) using sequential distinct windows of genotyped markers to assess the contribution from each region to non-contact ACL rupture variance (5). A logistic linear random effects model was applied to binary case/control phenotypes. Two random effects were considered which included the regional genomic polygenic and whole genome polygenic matrices using the *gaston* R package. The following mixed model was used:

$$y = X\beta + Zg + Qr + e$$

Where y is the vector of case-control cruciate ligament rupture status, X is an incidence matrix for fixed effects (sex and overall mean), β is the regression coefficient for fixed effects, Z and Q are the design matrices for the whole (without the window) and the regional random effects, respectively. The distributions and covariance structures of g and r were $g \sim N(0, G\sigma_g^2)$ and $r \sim N(0, G_r\sigma_r^2)$, respectively. The residual term followed $e \sim N(0, I\sigma_e^2)$. The regional heritabilities to each genomic window segment are:

$$h_{RHM}^2 = \frac{\sigma_r^2}{\sigma_g^2 + \sigma_r^2 + \sigma_e^2}$$

Here G is the matrix of additive genomic relationship (4) was given by the following:

$$G = \frac{WW'}{2 \sum_{j=1}^p p_j q_j}$$

Where $W(N \times p)$ is the additive incidence matrix of SNPs, p_j and q_j were allele frequencies of the j -th marker. G_r is a matrix like G , but using a subset of the matrix W .

Dissecting the genetic architecture of non-contact ACL rupture and evidence of selection using BayesR and BayesS

In this approach, all SNP associations were treated as random effects drawn from a normal distribution to enable an unbiased estimate of variance explained by the SNPs. This is a useful approach for studying the genetic architecture of common complex diseases (6). This approach can be further tailored to GWAS of complex phenotypes by treating SNP effects (β) as drawn from a mixture of normal distributions corresponding to differing SNP effect sizes, including a distribution for SNPs with zero effect. BayesR (6) is one such implementation that models SNP effects using a mixture of four normal distributions each with zero mean but with variance ranging from zero to 1% of total genetic variance, which more accurately models the effect size distribution expected from a complex phenotype. The relative variance for each mixture component is:

$$p(\beta_j | \pi, \sigma_g^2) = \pi_1 \times N(0, 0 \times \sigma_g^2) + \pi_2 \times N(0, 10^{-4} \times \sigma_g^2) + \pi_3 \times N(0, 10^{-3} \times \sigma_g^2) + \pi_4 \times N(0, 10^{-2} \times \sigma_g^2)$$

In the above equation, the mixture proportions are constrained to sum to unity, and it is assumed that the mixing proportions are drawn from a Dirichlet distribution with parameters (1,1,1,1). This prior is purposefully ambiguous to have little influence on the final estimates of the mixing proportions, which are primarily determined by the data (7), and σ_g^2 is the total genetic variance. We used 100,000 Markov Chain Monte Carlo (MCMC) samples with a burn in of 20,000 samples for the analysis. The posterior mean $\pm$ standard deviation SNP number in each mixture class was then determined.

Given that our previous studies suggest non-contact ACL rupture is a polygenic trait, theoretically, selection pressure is spread over many small effect variants making it difficult to detect selection signals. However, it is possible to infer evidence of natural or artificial selection from the relationship between effect size and MAF for all genome-wide variants. We also investigated signatures of negative selection in the genetic architecture of ACL rupture using the Bayesian mixed linear model BayesS (8). This model simultaneously estimates SNP-based heritability, polygenicity (proportion of SNPs with non-zero effects), and the relationship between SNP effect size and MAF. The approach modeling the relationship between SNP effect size and MAF uses the following mixture distribution as a prior for each SNP effect:

$$\beta_j \sim N \left(0, [2p_j(1-p_j)]^s \sigma_\beta^2 \right) \gamma + \varphi (1-\gamma)$$

Where β_j is the SNP effect or allelic substitution effect of the SNP j , p_j is the MAF of the SNP, σ_β^2 is the variance of SNP effects under a neutral model, φ is a point mass at zero, and γ is the proportion of SNPs with nonzero effects as an indicator for disease polygenicity.

$$[2p_j(1-p_j)]^S \sigma_\beta^2$$

is the variance of non-zero SNP effects. The model assumes that S follows a normal distribution as $S \sim N(0, \sigma_S^2)$. If there is selection, the effect size can be positive ($S > 0$) or negatively ($S < 0$) related to MAF. When $S = 0$, the effect size is independent of MAF (neutral model) and there is no selection.

Polygenic risk score prediction of ACL rupture risk

Association analysis using mixed logistic regression. Using the non-contact ACL rupture phenotypes, a logistic regression model was used to estimate the genetic variant effects. The following model was considered for each SNP:

$$\text{logit}(P[Y = 1 | \mathbf{X}, \mathbf{G}, \boldsymbol{\omega}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\omega}$$

Where $\mathbf{Y}$ was an n -dimensional vector composed of binary elements, representing zeroes and ones. The matrix $\mathbf{X}$, dimension $n \times p$, contains covariates, encompassing a column of ones that serves as the intercept and a column as the sex (male or female) of the dogs. Additionally, $\mathbf{G}$ represents a vector denoting genotypes, typically encoded as values 0, 1, or 2. Furthermore, $\boldsymbol{\omega}$ stands as a random vector that adheres to a multivariate normal distribution $MVN(\mathbf{0}, \boldsymbol{\tau K})$, where $\mathbf{0}$ signifies the mean vector and $\boldsymbol{\tau K}$ characterizes the covariance genomic relationship matrix. The AIREML algorithm was employed to estimate the model parameters. A score test was performed, independently of the value of each SNPs to obtain the P -values (Fig. S10).

SNP selection for polygenic risk score prediction. In dogs, we selected 1% (1,420 SNPs), 2% (2,841 SNPs), 3% (4,262 SNPs), and 4% (5,682 SNPs) of top SNPs according to the lowest P -value for predictions and evaluated the best predictive SNPs subset. The top 2,841 SNPs (2%) with the strongest association with ACL rupture yielded the highest performance and were used for the final risk prediction (Fig. S11). In all models, we included sex as the only covariate that can potentially influence prediction of ACL rupture risk.

Model evaluation using coefficient of determination in dogs

Machine learning prediction models. We used four different machine learning models: least absolute shrinkage and selection operator (LASSO), elastic net regression (EN), weighted subspace random forest (RF), and support vector machine (SVM). The LASSO approach (9) was used for efficient feature selection based on the assumption of linear dependency between input features and output values. In a general form, the lasso estimator uses the ℓ_1 penalized least squares criterion to obtain a sparse solution to the following optimization problem:

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Where, $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_i^n (y_i - x_i^T \boldsymbol{\beta})^2$, is the ℓ_2 -norm (quadratic) loss function (i.e. residual sum of squares), x_i^T is the i -th row of $\mathbf{X}$, and the

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$$

is the ℓ_1 -norm penalty on $\boldsymbol{\beta}$, which induces sparsity in the solution, and $\lambda \geq 0$ is a tuning parameter. The ℓ_1 penalty enables the lasso to simultaneously regularize the least squares fit and shrinks some components of $\hat{\boldsymbol{\beta}}_{Lasso}$ to zero for some suitably chosen λ . EN adds both the ℓ_1 and ℓ_2 penalties to the regression's loss function is a hybrid of ridge regression and lasso regularization as follows:

$$\hat{\boldsymbol{\beta}}_{EN} = \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|^2$$

The quadratic penalty term makes the loss function strongly convex, and it therefore has a unique minimum:

$$\|\beta\|^2 = \sum_{j=1}^p \beta_j^2$$

The glmnet function from the “glmnet” R-package were used for fitting LASSO model. The cv.glmnet() function in this package was used to determine optimum values for λ using a cross validation procedure. We used a weighted RF model because the weighted form of the RF model can achieve high accuracy in classifying high dimensional data (10), such as datasets with thousands of SNPs. Because such data typically contain many uninformative features for classification, random sampling often doesn't include informative features in selected subspaces. So, we used the Weighted Subspace Random Forest algorithm (wsrf) (11), using wsrf R package (12). The SVM algorithm (13) aims at identifying a function that has a maximum deviation epsilon (ϵ) from the observed values (y) and has a maximum margin for the set of prediction variables (S). To fit this model the e1071 R package was used. The power of the SVM resides in a particular component known as the kernel. In this study, we used a simple linear kernel. These kernels transform our genotypic data points to pass a linear hyperplane and thus classify our data.

Bayesian regression prediction models. Bayes B, Bayes C, Bayesian Lasso (BL), and Bayesian ridge regression (BRR) models were fitted and compared in terms of their prediction accuracy. For each sample we assumed that there is a genomic variable predictor i.e.:

$$G = \{g_{ij}\} \text{ with } i = 1, \dots, n, j = 1, \dots, p_g$$

The phenotypic vector $y = \{y_i\}$ defined as two possible values including $y_i = 0$ for phenotype-negative controls and $y_i = 1$ for non-contact ACL rupture cases for the y individual. A probit link function as $P(y_i = 1|G_i) = \Phi(\eta_i)$ was used to estimate the model parameters. Where, Φ is a standard normal cumulative distribution function and η_i is a linear predictor that follows the function given below:

$$\eta_i = \mu + \sum_{j=1}^{p_g} g_{ij}\beta_j$$

In the above equation, μ is an intercept or general mean, g_{ij} is the genotype of the i -th individual at the j -th marker, and β_j is the j -th marker effect. The probit link implemented used a latent normally distributed variable $l_i = \eta_i + \varepsilon_i$ and a measurement model $y_i = 0$ if $l_i < \gamma$, and 1 otherwise, where γ is a threshold parameter; ε_i is an independent normal model residual with mean zero and with variance set equal to one. A standard Bayesian linear model was used for prediction was as follows:

$$p(\theta_g|y, \omega_g) \propto p(y|\theta_g) p(\theta_g|\omega_g)$$

where $p(\theta_g|y, \omega_g)$ is the conditional posterior density of the genomic parameters ($\theta_g = \{\mu, \sigma_e^2, \beta\}$), including the residual variance (σ_e^2), which was assigned a scaled-inverse χ^2 prior density, μ was assigned a flat prior density, and the marker effects (β) were assigned independent and identically distributed informative priors, depending on the model, and ω_g is the genomic hyperparameter for indexing the prior density of marker effects in the different regression models.

BB and BC both have an extra parameter ‘ π ’ which is the probability of the marker effect be equal to zero. This parameter is treated as unknown and assigned a Beta prior $\pi \sim \text{beta}(p_0, \pi_0)$, with $p_0 > 0$ and $\pi_0 \in [0, 1]$ (14). The BB model assumes the prior for marker effects follows a normal mixture distribution given by:

$$\beta_i|\pi \sim (1 - \pi)N(0, \sigma_{\beta_i}^2) + \pi N(0, \sigma_{\beta_i}^2 = 0)$$

Here, the $\sigma_{\beta_i}^2$ denotes that each SNP has its own variance with prior distribution given by:

$$\sigma_{\beta_i}^2|v_\beta, S_\beta \sim v_\beta S_\beta \chi_{v_\beta}^{-2}$$

In BC, the prior distribution for marker effects is also given by a normal mixture distribution:

$$\beta_i|\pi \sim (1 - \pi)N(0, \sigma_\beta^2) + \pi N(0, \sigma_\beta^2 = 0)$$

but this model assumes the same genetic variance for all markers with prior distribution given by:

$$\sigma_{\beta}^2 | v_{\beta}, S_{\beta} \sim v_{\beta} S_{\beta} \chi_{v_{\beta}}^{-2}$$

which is like BRR model's assumptions.

In Bayesian LASSO, the regression parameter β_j is assumed to follow a double exponential (DE) prior distribution regression (15). In this context,

$$\beta_i | \tau_i, \sigma_e^2 \sim N(0, \sigma_{\beta_i}^2 = \tau_i^2 \sigma_e^2)$$

where $\tau_i^2 | \lambda^2 \sim \text{Exp}(\lambda^2)$, the shrinkage factor λ is further assigned with a hyper prior of a gamma distribution ($\lambda^2 | s, r$), and so it can be estimated as other model parameters. Under this approach, the marginal prior distribution for the marker effect is given by $\beta_i | \lambda \sim \text{Double-Exp.}(0, \lambda)$. This double-exponential distribution presents higher mass at zero, but it does not necessarily set coefficients exactly to zero. The shape (s) and rate (r) parameters of the Gamma prior was specified to $s = 1.1$ and

$$r = \frac{(s-1)}{2 \times (1-R^2)/R^2 \times MS_x} \quad (14),$$

where MS_x represents the sum of the variances of genotype values of each SNP, and $R^2 = 0.5$.

With BRR (16), the model assumes the same genetic variance for all markers, i.e., $\beta_i \sim N(0, \sigma_{\beta}^2)$. The prior distribution for marker genetic variance is the following scaled inverted chi-squared distribution,

$$\sigma_{\beta}^2 | v_{\beta} S_{\beta} \sim v_{\beta} S_{\beta} \chi_{v_{\beta}}^{-2}$$

with hyper-parameters v_{β} (degrees of freedom) and S_{β} (scale parameter).

In this study we used the BGLR package (14) and R (12) to fit all Bayesian regression models. Analyses were carried out using a Monte Carlo Markov Chain algorithm with a total of 1,000,000 iterations, and burn-in of 20,000 iterations. Global convergence was checked by visual inspection of trace plots.

Coefficients of determination (R^2). In this study we obtained the coefficient of determination, R^2 , of models using two different formulas (17). The first one was obtained by fitting a linear regression for the observed binary phenotypes (0 and 1), on the predicted probability of case-control status, was written as:

$$R_o^2 = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i^N (y_i - \bar{y})^2} = \frac{\text{var}(\hat{b}_{obs} g_i)}{\hat{K}(1 - \hat{K})}$$

Where, $\text{var}(\hat{b}_{obs} g_i)$ is the variance due to the explanatory variable (genetic variance) on the observed probability scale and $\hat{K}$ is the mean proportion of cases in the sample. The second one was based on the Nagelkerke's scale method (18), which is a modified version of the R^2 proposed by Cox and Snell (19) and can be approximated as:

$$R_{C\&S}^2 \cong 1 - \frac{\sum_i^N \frac{(y_i - \hat{y})^2}{N}}{\frac{\sum_i^N (y_i - \bar{y})^2}{N}} = 1 - \frac{\sigma_e^2}{K(1 - K)}$$

Nagelkerke (1991) corrected Cox and Snell's R^2 by scaling it by the maximum attainable value:

$$R_N^2 = \frac{R_{C\&S}^2}{R_{max}^2}$$

Where:

$$R_{max}^2 = 1 - K^{2K} \cdot (1 - K)^{2(1-K)}$$

and K, population prevalence of non-contact ACL rupture.

R^2 on the liability scale was computed according to Lee et. al (2012) for ascertained case-control studies (17):

$$R_t^2 = \frac{R_o^2 C}{1 + R_o^2 \theta C}$$

C and θ are correcting factors for ascertainment, K sample disease prevalence, t : the threshold on the normal distribution truncating the proportion of disease prevalence K , $m = \frac{z}{K}$ mean liability for cases and $z = f(t)$ the value of probability density function of standard normal distribution at t . Represents the adjusted coefficient of determination on the liability scale, reflecting the proportion of variance explained by the model while accounting for ascertainment biases.

We also computed the F1 score = $(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$ as the harmonic mean of precision and recall of a classification model. Recall (also known as Sensitivity or True Positive Rate) quantifies the ratio of true positive predictions to the total actual positives ($TP / (TP + FN)$). Precision (also known as Positive Predictive Value) quantifies the ratio of true positive predictions to the total positive predictions made. Additionally, another metric calculated was Accuracy, which is the ratio of correctly predicted observations to the total observations $((TP + TN) / (TP + TN + FP + FN))$.

Evaluation of GWAS summary-based model for human non-contact ACL rupture

To evaluate the performance of non-contact ACL rupture prediction models based on GWAS summary statistics, we employed the recently developed novel algorithm PUMAS (20). The approach is a comprehensive methodology that combines subsampling of GWAS association statistics, evaluating PRS model performance using summary data, and applying model-tuning strategies to optimize the performance of polygenic risk score prediction models. The approach assumes that in the context of the methodology's foundation:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \text{ where } \mathbf{X} = (X_1, \dots, X_p)$$

signifies the random vector of p SNPs, $\beta = (\beta_1, \dots, \beta_p)^T$ forms a p -dimensional vector that fixed SNP effect sizes and ε pertains to the error term, which follows a normal distribution with a zero mean and $\mathbf{y}$ is the observed phenotypes of N independent individuals. The dimension for $\mathbf{y}$ and $\mathbf{X}$ is in $N \times 1$ and $N \times p$, respectively, and assumes both $\mathbf{y}$ and the variables X_j 's have been centered. Furthermore, the summary association statistics for GWAS are derived through marginal linear regressions. With having only the GWAS association summary statistics, we can generate marginal summary statistics for the training and validating datasets from summary statistics of the full dataset. With this description the algorithm follows these steps:

Subsampling GWAS association statistics. Defining the distribution of summary statistics based on the central limit theorem and randomly sampling a subset of $N - n$ individuals as the training set

$$\mathbf{X}^{(tr)T} \mathbf{y}^{(tr)} \text{ and } \mathbf{X}^{(v)T} \mathbf{y}^{(v)}:$$

$$\mathbf{X}^T \mathbf{y} \sim N(N E(\mathbf{X}^T \mathbf{y}), N \text{var}(\mathbf{X}^T \mathbf{y}))$$

$$\mathbf{X}^{(tr)T} \mathbf{y}^{(tr)} \sim N((N - n) E(\mathbf{X}^T \mathbf{y}), (N - n) \text{var}(\mathbf{X}^T \mathbf{y}))$$

$$\mathbf{X}^{(v)T} \mathbf{y}^{(v)} \sim N(n E(\mathbf{X}^T \mathbf{y}), n \text{var}(\mathbf{X}^T \mathbf{y}))$$

Approximation of covariance matrix of summary statistics $\mathbf{X}^T \mathbf{y}$, for $j = 1, \dots, p$ and $i \neq j$ we have:

$$\hat{\Sigma}_{ji} = \hat{\beta}_j \hat{\beta}_i \hat{\sigma}_j^2 \hat{\sigma}_i^2$$

Obtaining the validating or testing summary statistics by:

$$\mathbf{X}^{(v)T} \mathbf{y}^{(v)} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^{(tr)T} \mathbf{y}^{(tr)}$$

So, subsampled GWAS summary statistics for the training set can be estimated by:

$$\hat{\beta}_j^{(tr)} = [(N - n) \hat{\sigma}_j^2]^{-1} \mathbf{X}^{(tr)T} \mathbf{y}^{(tr)} \text{ and } SE(\hat{\beta}_j^{(tr)}) = \sqrt{\frac{N}{N - n}} SE(\hat{\beta}_j)$$

Evaluate model performance using GWAS summary data and calculate R^2 , a commonly used metric to quantify PRS predictive performance.

$$R^2 \approx \frac{(\frac{1}{n} \sum_{j=1}^p \omega_j X_j^{(v)T} y^{(v)})^2}{N \max_j [SE(\hat{\beta}_j)^2] \sum_{j=1}^p \omega_j^2 \hat{\sigma}_j^2}$$

Where $\mathbf{W}^T = (w_1, \dots, w_p)$ is the weight for SNPs in PRS. In a traditional PRS, marginal regression coefficients from GWAS are used as the weight values, i.e., $\mathbf{w} = \hat{\beta}$.

Analysis of biological processes and pathway enrichment

The pathway enrichment analyses of non-contact ACL rupture associated genes were performed using Metascape (<https://metascape.org/>), which leverages over forty independent knowledge bases, to investigate biological processes and pathways relevant to the disease in humans and dogs (21). The analyses included two curated gene lists, comprising 66 non-contact ACL rupture-related genes in humans and 54 in dogs, which were used to map identifiers to known genes and proteins in its database. Metascape method applies a mature complex identification algorithm, called the Molecular Complex Detection (MCODE) method (22), which automatically extracts protein complexes embedded within a large network. Taking advantage of Metascape's functional enrichment analysis capabilities, the most significantly enriched ontology terms were combined to annotate putative biological roles for each MCODE complex (21). The cumulative hypergeometric distribution was used to calculate the P -values and corrected for multiple testing by the Benjamini-Hochberg method to identify all ontology terms that contain a statistically significant greater number of genes in common with an input list than expected by chance.

Supplementary Results and Discussion

The enrichment analysis showed that the most significant pathway is extracellular matrix organization (R-HSA-1474244) in both humans and dog (Figs. S7-S9). Additionally, several key biological processes and signaling pathways crucial for tissue integrity, healing, and response to injury were highlighted in the analysis. Beside the extracellular matrix organization, other important elements included collagen synthesis and elastic fiber formation, which are essential for maintaining the structural integrity of the ACL. Critical signaling pathways such as PI3K-Akt-mTOR, focal adhesion, and VEGF play vital roles in cellular responses to stress, survival, proliferation, and angiogenesis, especially under hypoxic conditions commonly seen in injured tissues. Furthermore, pathways related to growth factor responses, cellular development, and differentiation indicate the complex regulation needed for effective ACL ligament homeostasis and repair. Other aspects like axon guidance and the response to peptides also suggest broader regulatory networks influencing pain perception and healing dynamics. Understanding these interconnected pathways offers insights into potential therapeutic targets, which could help enhance ACL repair and reduce the risk of long-term complications like osteoarthritis.

Of the 16 shared risk genes (Fig. 5) it is worth noting that these are novel risk genes regarding the pathogenesis of ACL rupture. The function of these genes is summarized below.

1. **STK32B**. Encodes a protein kinase involved in transferring phosphate molecules to serine and threonine residues. Mutations or deletions in this gene can cause skeletal dysplasia. In the context of ACL rupture, *STK32B* might affect the structural integrity of skeletal muscle and connective tissues due to its role in phosphorylation processes essential for cellular functions and tissue stability.
2. **ARID1B**. Alterations in this gene can lead to low muscle tone and other health issues. About 80% of individuals with *ARID1B*-related syndrome exhibit low muscle tone, which may influence muscle strength and joint stability. This could increase susceptibility to ACL injuries due to weakened muscle support.
3. **LOXHD1**. Associated with mechanotransduction defects, which can affect how muscles respond to mechanical stress and strain. Disruptions in mechanotransduction could impair muscle adaptation to mechanical forces, potentially contributing to ACL rupture susceptibility.
4. **RSBN1L**. May be involved in the maintenance of muscle mass, as it is expressed in skeletal muscle tissue. Proper maintenance of muscle mass is crucial for supporting joint stability and function, which can influence the risk of ACL rupture.
5. **NIBAN2**. Plays a role in cell survival under stress conditions, potentially impacting muscle tissues through stress response pathways. Enhanced stress response might influence muscle resilience and recovery, indirectly affecting risk of ACL rupture.
6. **TEX9**. Involved in energy-related processes within skeletal muscle cells. Efficient energy metabolism is essential for muscle contraction and overall muscle health, which can impact muscle strength and joint stability, influencing ACL rupture risk.
7. **GRIN2A**. Modulates excitatory neurotransmission, critical for muscle contraction. Proper neurotransmission is vital for effective muscle function and coordination, which can affect muscle support and stability around the ACL.
8. **MBTPS1**. Regulates lipid metabolism, stress responses, and the unfolded protein response (UPR) in the endoplasmic reticulum. Disruptions in these processes can affect muscle cell function and stress responses, potentially impacting muscle health and joint stability.
9. **HSDL1**. Involved in osteoclast function, osteoblast differentiation, and bone formation. These processes are critical for skeletal development and bone health. Disruptions can affect bone density and strength, indirectly influencing the risk of ACL injuries.
10. **DNAAF1**. Affects the assembly of dynein arms and regulates cilia and microvilli. Since cilia and microvilli are involved in cellular signaling and mechanotransduction, their dysfunction could impact muscle cell function and stability, potentially contributing to risk of ACL rupture.
11. **TAF1C**. A transcription factor involved in the TFIID complex, essential for transcription of protein-coding genes. Its role in skeletal muscle development and function is still under investigation, but disruptions could impact muscle growth and repair, influencing risk of ACL rupture.
12. **PTPRM**. A protein tyrosine phosphatase involved in cell signaling that regulates various cellular processes including growth and differentiation. Its role in skeletal muscle might impact muscle development and joint stability, potentially affecting risk of ACL rupture.

13. **RAB31**. Plays a role in *TGF- β* -induced actin remodeling and cytoskeleton rearrangement. Proper cytoskeleton dynamics are crucial for muscle cell function and stability, and disruptions could impact muscle homeostasis and neuromuscular control.
14. **PTPRT**. A non-receptor-type protein tyrosine phosphatase that interacts with insulin receptors and substrates. Its role in cellular signaling could affect muscle cell function and repair mechanisms, which may influence risk of ACL rupture.
15. **DSCAM**. Involved in cell adhesion processes. Proper cell adhesion is important for maintaining tissue integrity and stability, including in muscle and connective tissues. Disruptions could affect muscle and joint stability, potentially impacting risk of ACL rupture.
16. **RNF165**. Also known as *Ark2C*, is an E3 ubiquitin ligase that enhances bone morphogenetic protein signaling. Efficient bone morphogenetic protein signaling mediated by *RNF165* might contribute to the maintenance of ligament strength and resilience. Deficiencies or dysregulation in this signaling could lead to weaker ligaments and an increased susceptibility to ACL rupture.

Table S1. The estimated Area Under the Receiver Operating Characteristic Curve (AUC), Coefficients of determination (R^2), Accuracy (ACC) and F1 Score for non-contact ACL rupture polygenic risk score (PRS) prediction.

Model	Metric	LAB→LAB	MA→ROT	LAB→ROT
BB	AUC	0.696±0.035	0.702±0.052	0.532
	R^2	0.329±0.117	0.408±0.121	0.210
	ACC	0.632±0.055	0.633±0.066	0.574
	F ₁ Score	0.474±0.031	0.415±0.045	0.307
BC	AUC	0.694±0.037	0.648±0.088	0.634
	R^2	0.329±0.116	0.313±0.131	0.220
	ACC	0.630±0.058	0.578±0.118	0.490
	F ₁ Score	0.473±0.032	0.388±0.052	0.352
BL	AUC	0.684±0.056	0.679±0.040	0.656
	R^2	0.305±0.142	0.347±0.169	0.190
	ACC	0.623±0.057	0.618±0.074	0.556
	F ₁ Score	0.472±0.045	0.418±0.032	0.368
BRR	AUC	0.683±0.062	0.718±0.051	0.529
	R^2	0.328±0.117	0.411±0.099	0.214
	ACC	0.629±0.059	0.667±0.071	0.639
	F ₁ Score	0.470±0.036	0.395±0.041	0.303
EN	AUC	0.635±0.073	0.579±0.092	0.721
	R^2	0.289±0.166	0.295±0.120	0.267
	ACC	0.617±0.069	0.624±0.078	0.361
	F ₁ Score	0.454±0.051	0.336±0.037	0.236
LASSO	AUC	0.615±0.088	0.587±0.067	0.586
	R^2	0.290±0.180	0.314±0.121	0.246
	ACC	0.595±0.078	0.658±0.051	0.454
	F ₁ Score	0.467±0.044	0.332±0.073	0.274
RF	AUC	0.59±0.0479	0.626±0.092	0.600
	R^2	0.259±0.134	0.338±0.136	0.217
	ACC	0.552±0.073	0.591±0.066	0.435
	F ₁ Score	0.487±0.055	0.314±0.038	0.270
SVM	AUC	0.674±0.056	0.66±0.0649	0.665
	R^2	0.311±0.119	0.323±0.118	0.236
	ACC	0.621±0.047	0.384±0.052	0.676
	F ₁ Score	0.464±0.028	0.341±0.059	0.353

Notes: Prediction in a within-breed scenario, a single ancestry population (LAB→LAB), a multi-ancestry population (MA→Rot), and an across ancestry (breed) population scenario (LAB→Rot) were examined. LAB – Labrador Retriever, ROT – Rottweiler. LAB→LAB: PRS prediction in a single ancestry population.

LAB→ROT: PRS prediction in a multi-ancestry population between two distinct breeds of dog. MA (LAB+ROT) → ROT prediction in a multi-ancestry population with admixing of the training set used for prediction. Mean±SD are reported. The models referenced include BayesB (BB), BayesC (BC), Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Elastic Net (EN), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), and Support Vector Machine (SVM).

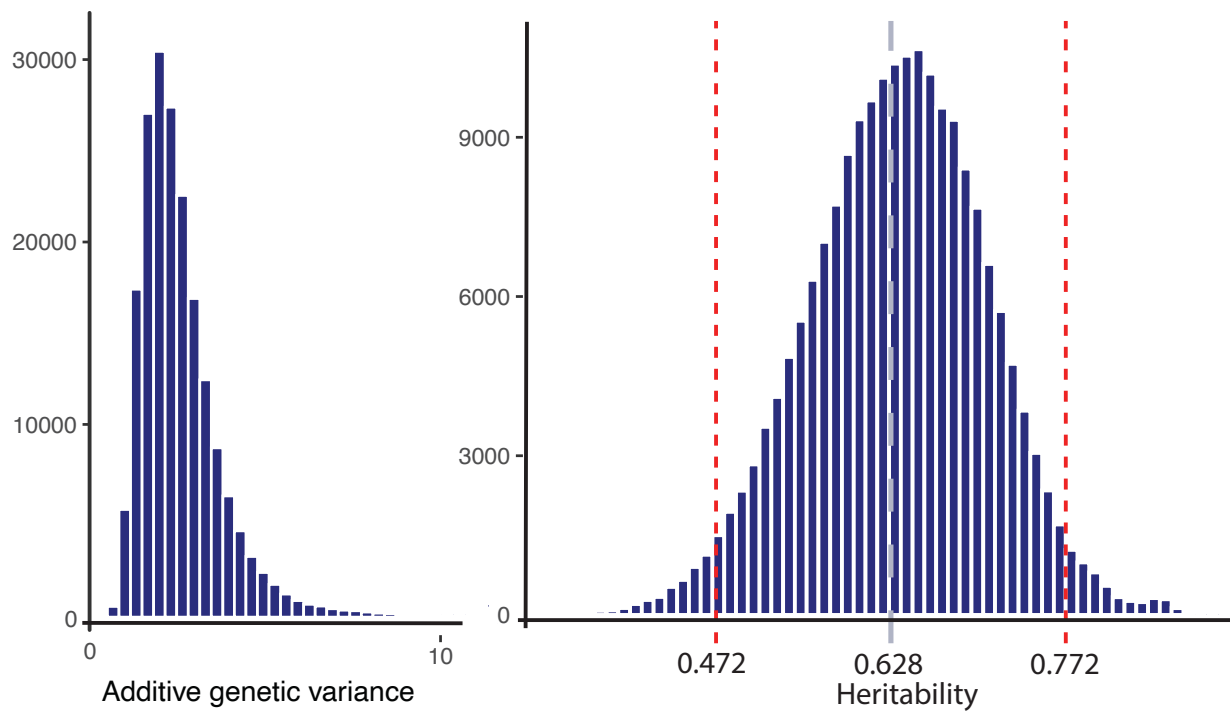


Fig. S1. Posterior densities for additive genetic variance component and heritability of non-contact ACL rupture in the Labrador Retriever. The dashed grey line denotes the mean (SD) of the heritability distribution. The heritability estimate was 0.63 ± 0.08 on the liability scale. The red dashed lines represent the highest posterior density (HPD) interval of 0.471-0.772 at threshold of 0.95%.

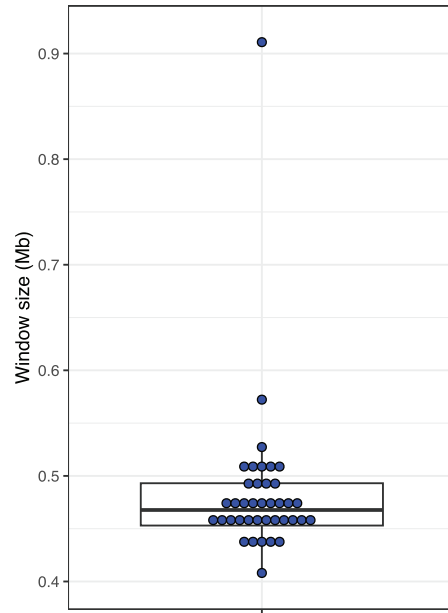


Fig. S2. The boxplot represents distribution of average of regional windows size according to Mb for each chromosome. Each point represents chromosome length/number of 30 single nucleotide polymorphism (SNP) windows.



Fig. S3. Regional heritability analysis of non-contact ACL rupture in dogs, with bar colors indicating the strength of window heritability. The x-axis shows the number of chromosomes, and each tile represents a window of 30 SNPs.

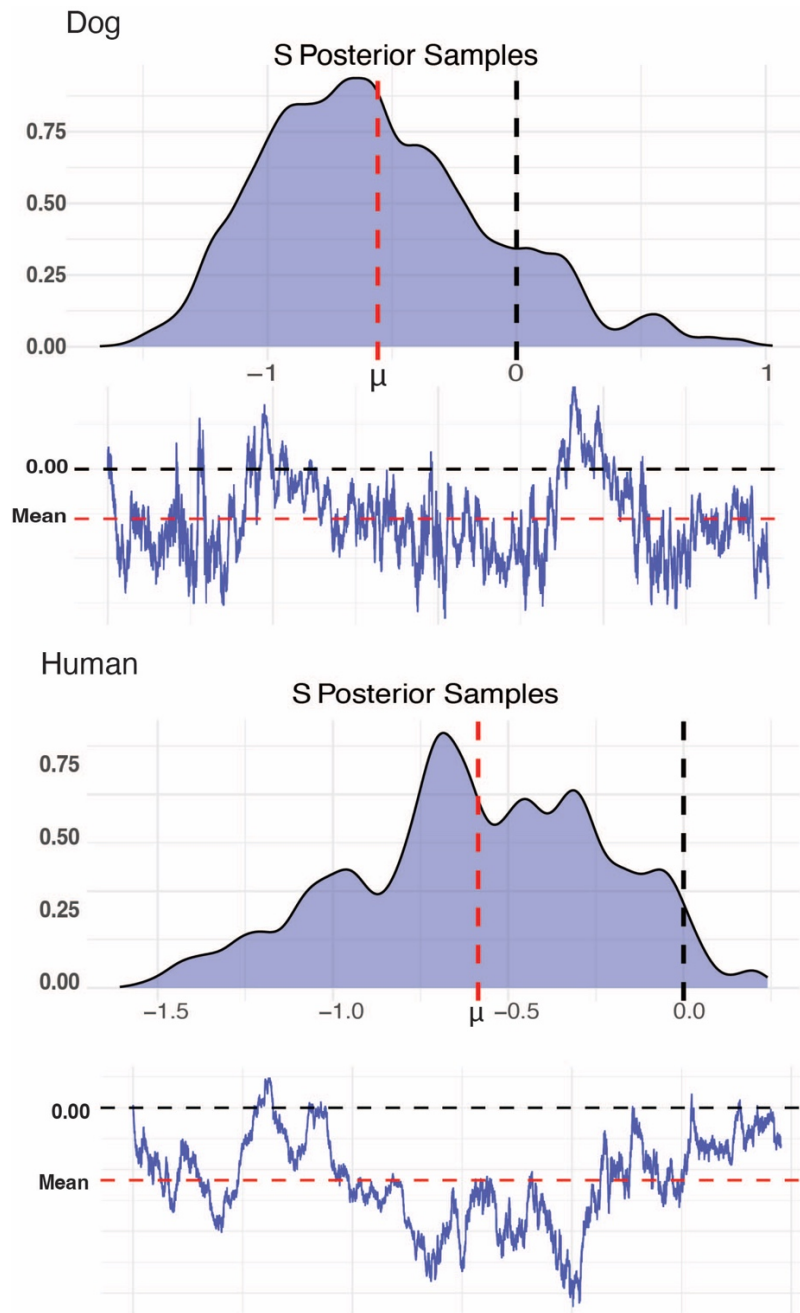


Fig. S4. *Dog.* The hyperparameter S from the BayesS regression model showed a mean of -0.558 and a median of -0.613 and a mode of -0.64, indicating a central tendency towards negative values. About 87% of the posterior samples were negative, suggesting a strong negative tendency. The 95% HPD interval ranges from -1.363 to 0.314, with most of the credible interval falling in the negative range. *Human.* Posterior analysis of the hyperparameter S from the sBayesS regression model revealed a strong tendency towards negative values. The mean (-0.59) and median (-0.59) were both negative and the mode was -0.686. ~97% of the samples fall in the negative range, indicating a predominant negative skew. The standard deviation was 0.360, reflecting moderate variability. The 95% HPD interval ranges from -1.285 to 0.054, confirming that most of the credible interval is negative. The red dashed line represents mean of posterior samples, and the black dashed line represent 0 values as separating positive and negative spaces.

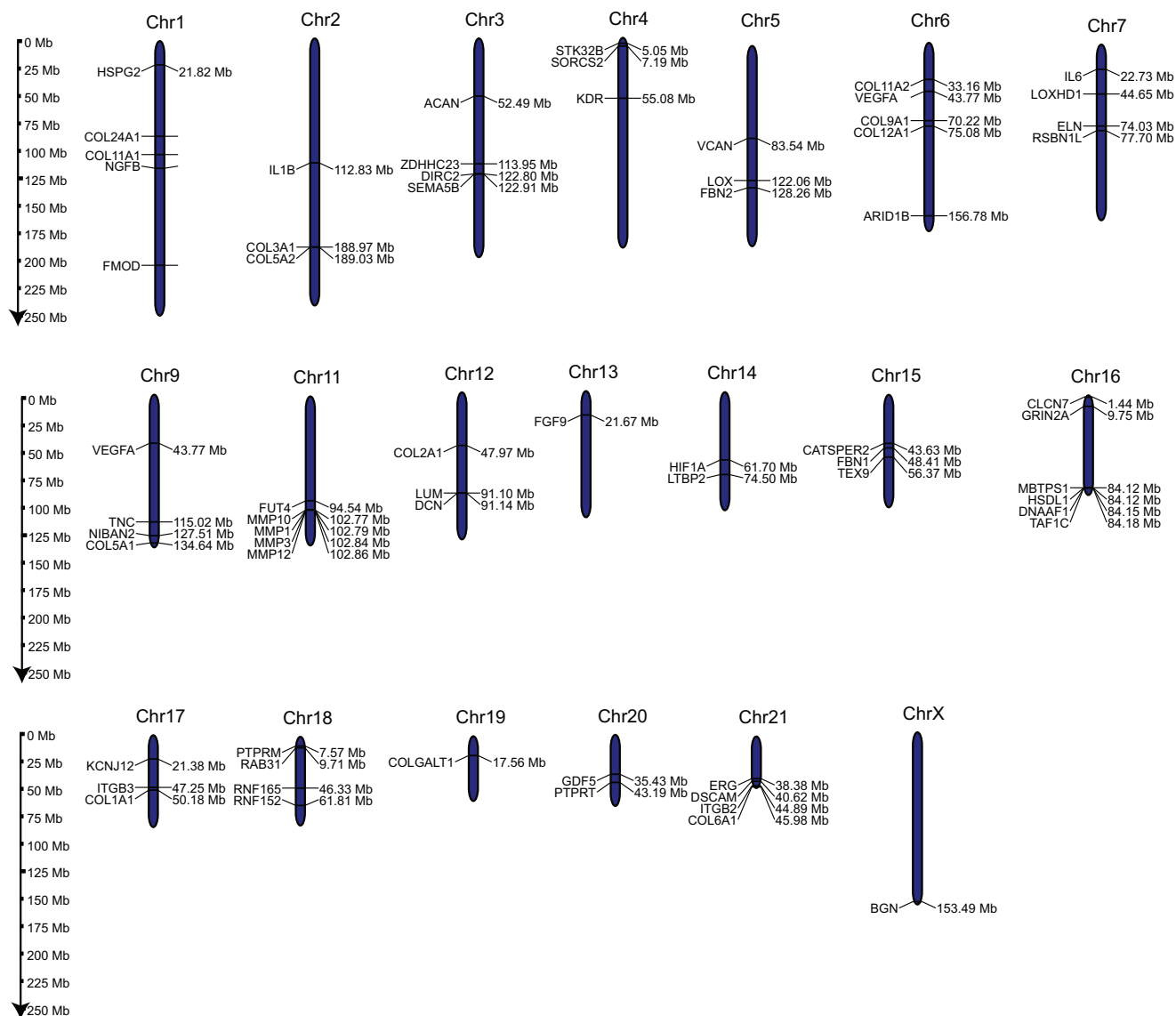


Fig. S5. Comprehensive list all reported genes related to non-contact ACL rupture in human beings.

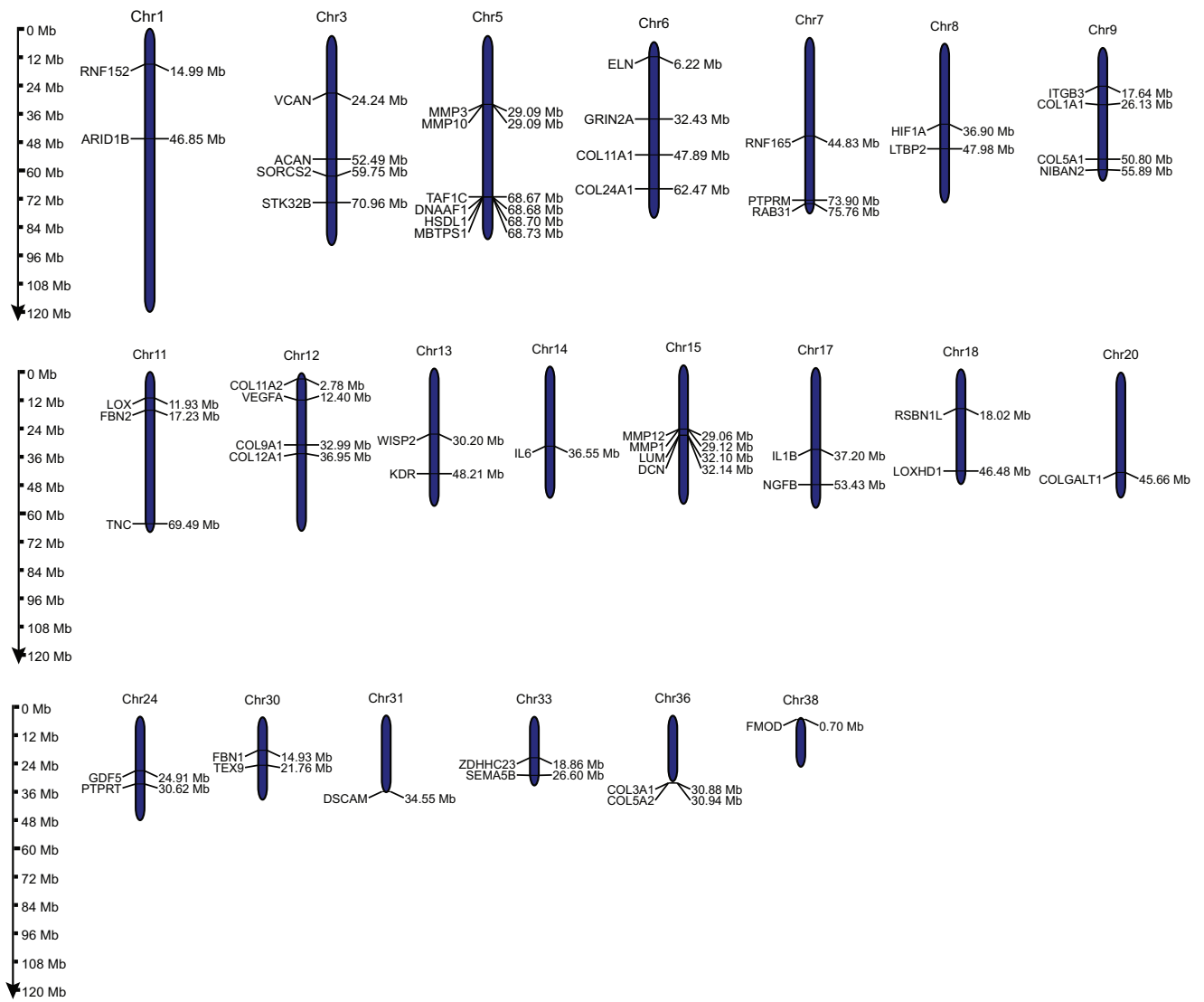


Fig. S6. Comprehensive list all reported genes related to non-contact ACL rupture in dogs.

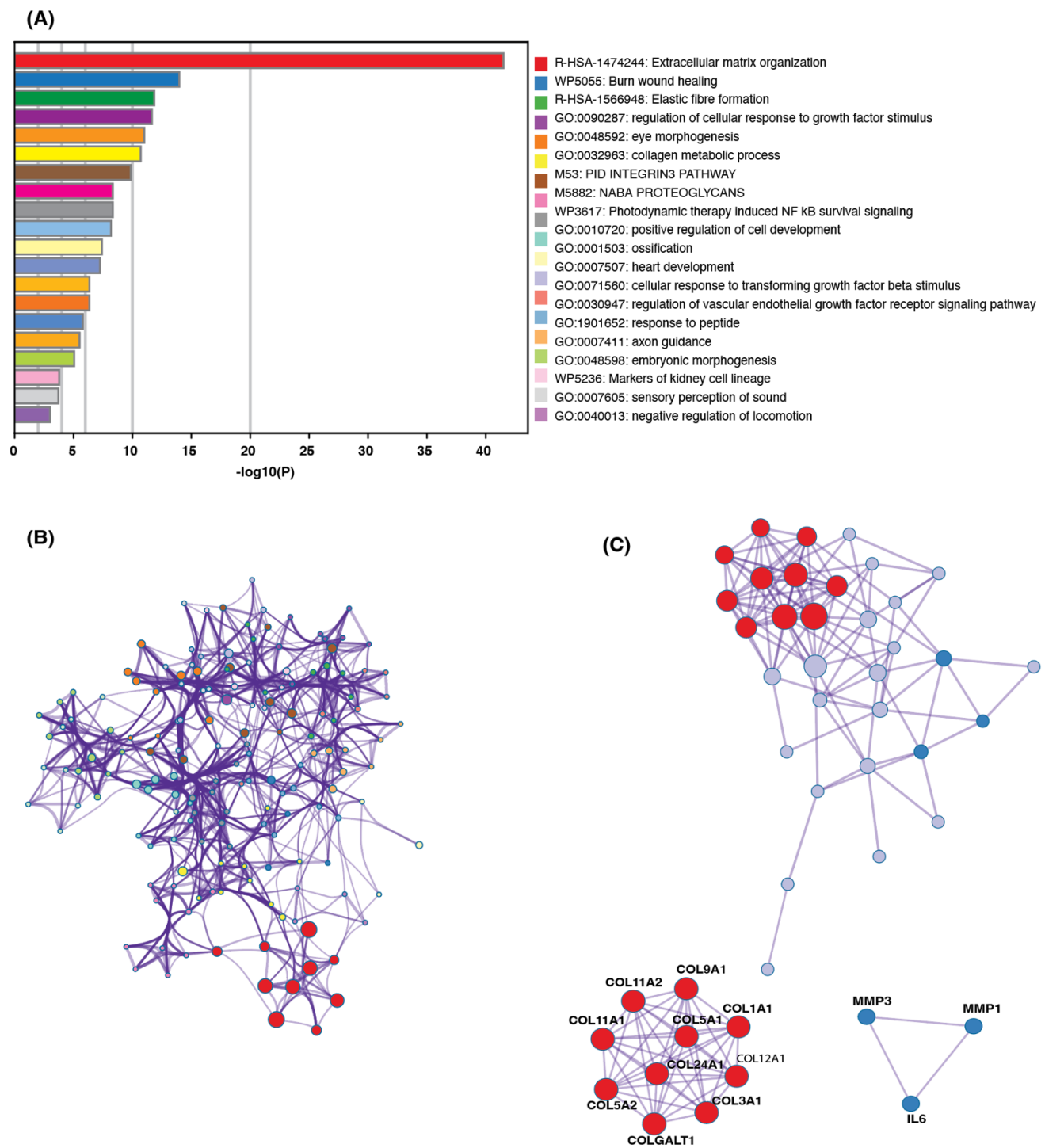


Fig. S7. Pathway network of enriched terms for dog non-contact ACL rupture. (A) Metascape bar graph for viewing the top 20 biological pathways with significantly enriched GO terms and pathways, using a discrete color scale to represent statistical significance, which had a significance level of $P < 0.01$. (B) Network of all enriched terms. This enrichment network shows the intra-cluster and inter-cluster similarities of enriched terms, up to ten terms per cluster. Cluster annotations are shown in color code from (A). (C) Protein-protein interaction (PPI) enrichment network and identified MCODE components colored by their identities. The MCODE algorithm subclustered the network into two subclusters (red color: MCODE 1, blue color: MCODE 2).

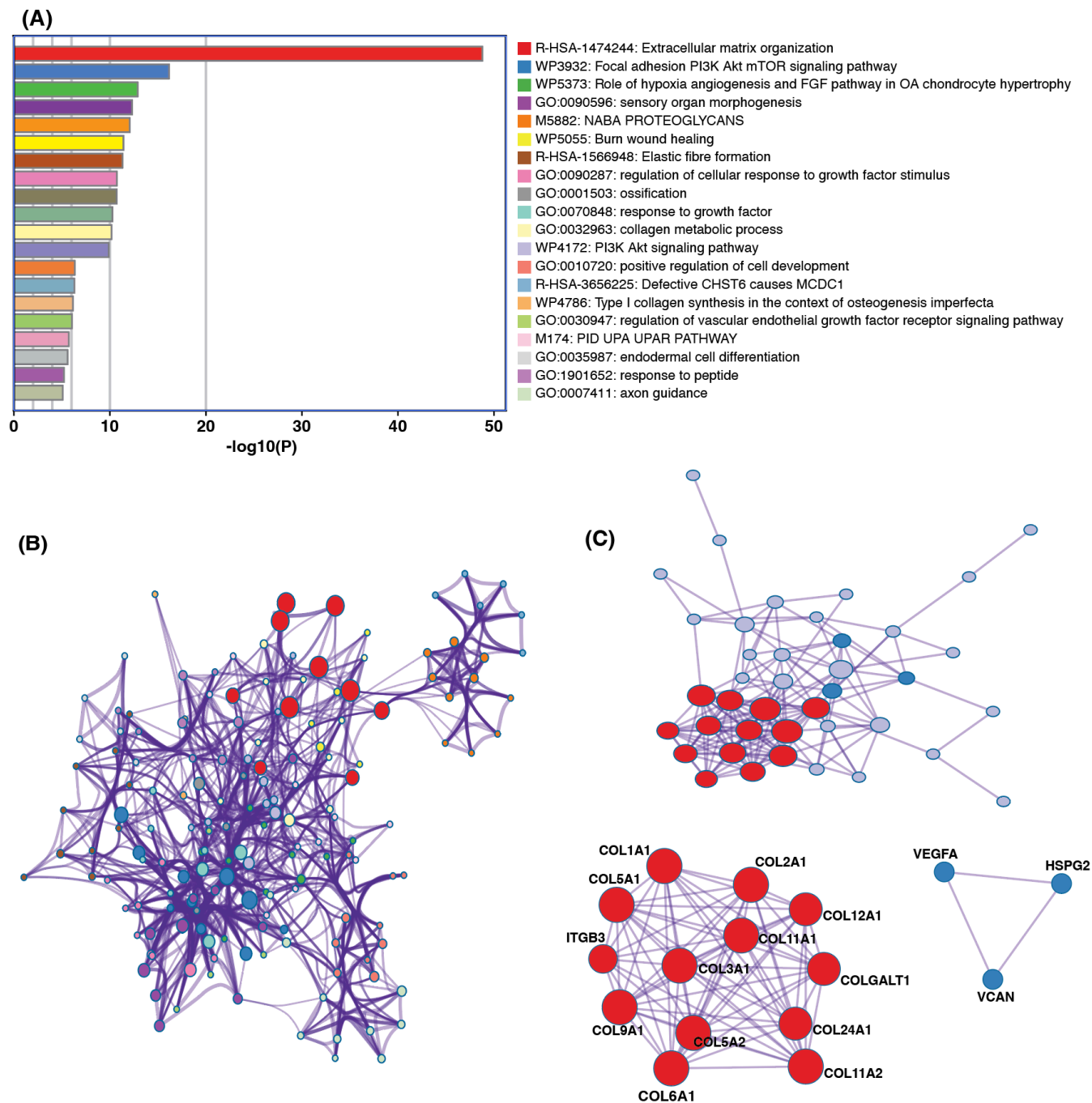
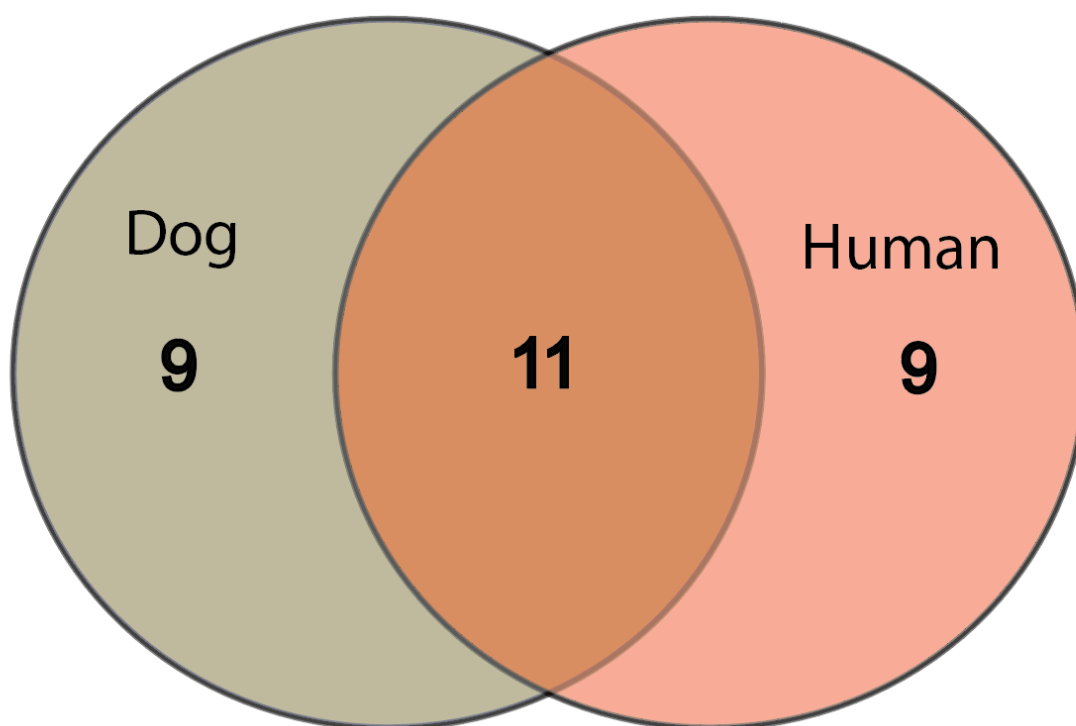


Fig. S8. Pathway network of enriched terms for human non-contact ACL rupture. (A) Metascape bar graph for viewing the top 20 biological pathways with significantly enriched GO terms and pathways, using a discrete color scale to represent statistical significance which had a significance level of $P < 0.01$. (B) Network of all enriched terms. This enrichment network shows the intra-cluster and inter-cluster similarities of enriched terms, up to ten terms per cluster. Cluster annotations are shown in color code from (A). (C) Protein-protein interaction (PPI) enrichment network and identified MCODE components colored by their identities. The MCODE algorithm subclustered the network into two subclusters (red color: MCODE 1, blue color: MCODE 2).



Groups	Count	Terms
Common Dog-Human	10	GO:0001503 WP5055 GO:0090287 GO:0007411 GO:0030947 GO:0032963 GO:0010720 GO:1901652 R-HSA-1566948 M5882 R-HSA-1474244
Dog	9	WP5373 GO:0035987 M174 WP3932 R-HSA-3656225 WP4786 WP4172 GO:0090596 GO:0070848
Human	9	M53 GO:0048592 GO:0071560 WP3617 GO:0040013 GO:0007507 GO:0048598 WP5236 GO:0007605

Fig. S9. Venn diagram illustrating the number of shared biological pathways and go terms between dogs and humans from the top 20 terms.

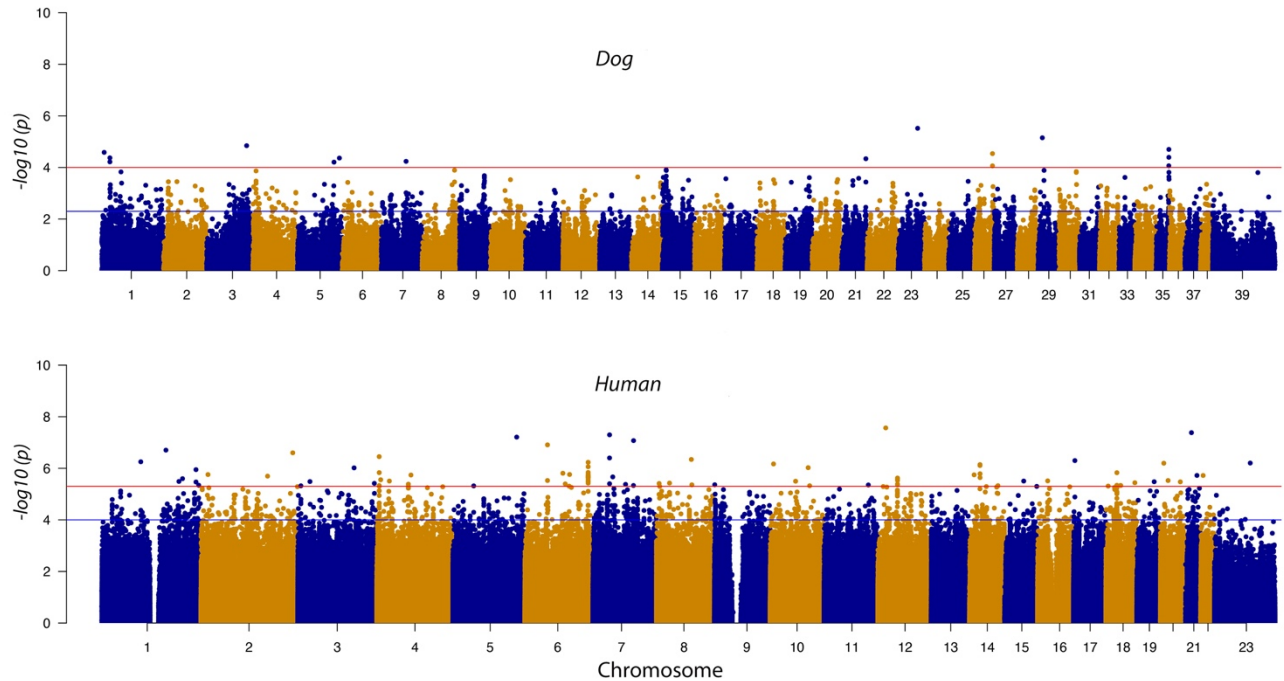


Fig. S10. Manhattan plot of the results from the GWAS analysis of non-contact anterior cruciate ligament (ACL) rupture in the Labrador Retriever and in human beings. The y axis represents $-\log_{10} P$ -values for association of with ACL rupture. The horizontal blue and red lines represent the thresholds used for genome-wide significance at $-\log_{10}(5e-03)$ and $-\log_{10}(1e-04)$ for dogs, and $-\log_{10}(1e-04)$ and $-\log_{10}(5e-6)$ for humans. In dogs, the less stringent threshold was determined through GWAS permutation testing using randomized phenotypes to account for high linkage disequilibrium. We established genome-wide significance thresholds using 95% confidence intervals derived from the empirical distribution of P -values obtained under the null hypothesis of no association (23,24). To construct this distribution, we performed 500 random permutations of the phenotypes and reran the GWAS each time.

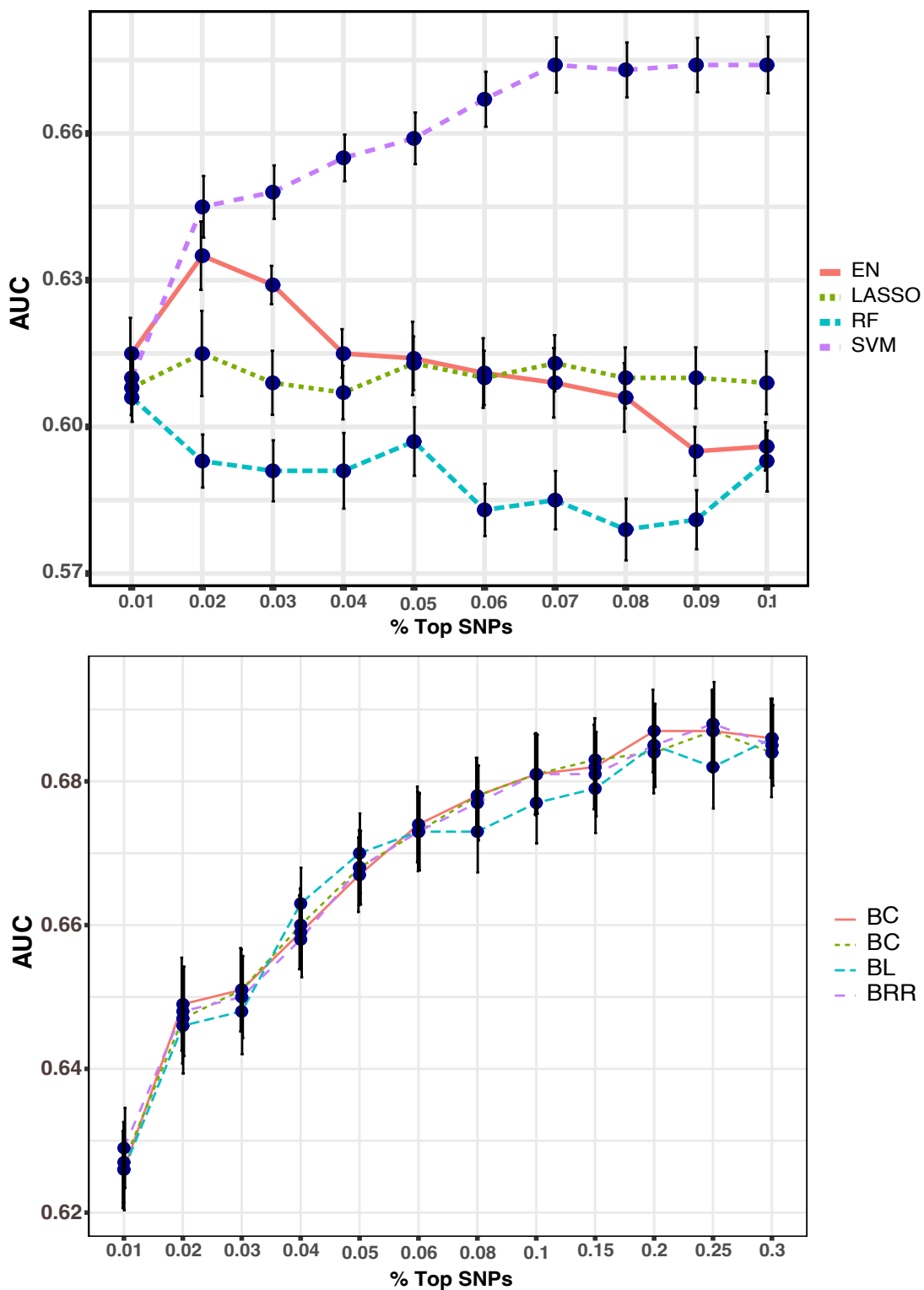


Fig. S11. The overall predictive performance (AUC mean $\pm$ standard deviation) of all machine learning and Bayesian models across different percentage of top SNPs for prediction of dog non-contact ACL rupture risk.

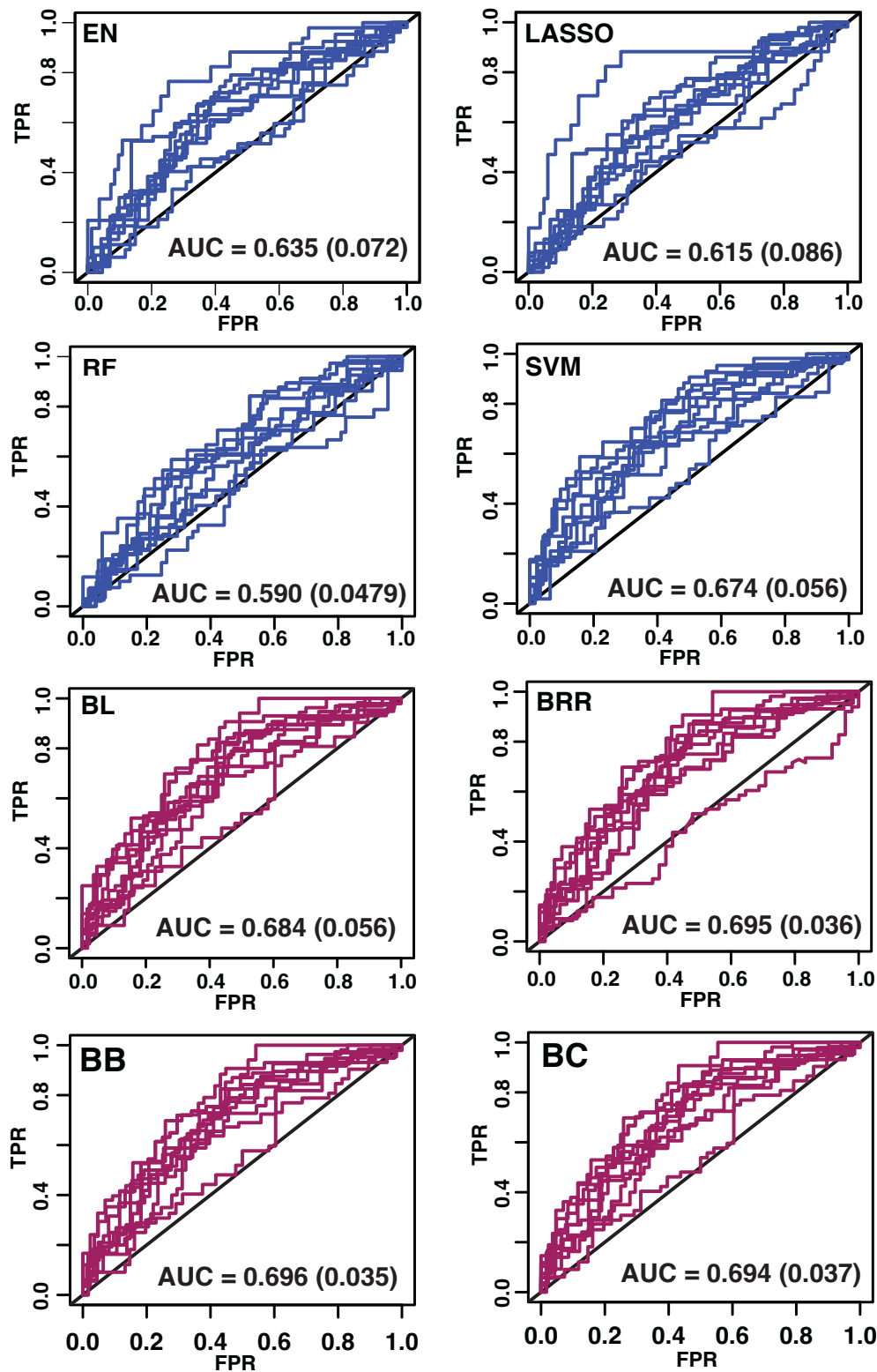


Fig. S12. The estimated Area Under the Receiver Operating Characteristic Curve (AUC) for non-contact ACL rupture polygenic risk score (PRS) prediction in a within-breed or single ancestry population (LAB→LAB). Machine learning models (blue color) involved Elastic Net (EN), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), and Support Vector Machine (SVM) and Bayesian regression models (red color) include BayesB (BB), BayesC (BC), Bayesian Lasso (BL), Bayesian Ridge Regression (BRR). The value in each plots shows the average of AUC values (standard deviation) over all 10 folds.

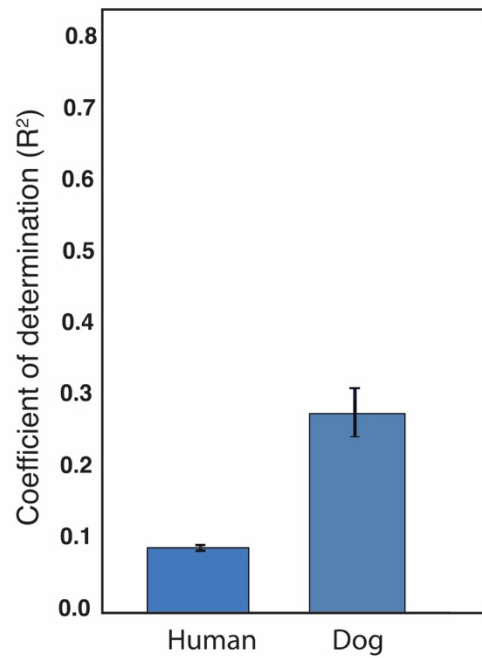


Fig. S13. Polygenic risk score (PRS) prediction of non-contact ACL rupture risk in dogs and humans. PRS predictions in humans used public GWAS summary statistics and sBLUP, Lassosum and sBayesS models. In the dog, ensemble predictions on the continuous liability scale using four machine learning (Elastic Net, Least Absolute Shrinkage and Selection Operator, subspace weighted Random Forest, Support Vector Machine and four Bayesian regression models (BayesB, BayesC, Bayesian Lasso, Bayesian Ridge Regression) and were then transformed to R^2 values for comparison. Results represent mean $\pm$ standard error from five-fold cross validation. Despite similar heritability, the coefficient of determination for human non-contact ACL rupture was lower, likely because of the different methodology used for human summary statistic data from biobank phenotypes and use of less accurate individual phenotypes from biobank data.

Supplementary References

1. J. Yang, S.H. Lee, M.E. Goddard, P.M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).
2. S.H. Lee, N.R. Wray, M.E. Goddard, P.M. Visscher, Estimating missing heritability for disease from genome-wide association studies. *Am. J. Human. Genet.* **88**, 294-305 (2011).
3. T.H. Witsberger, J.A. Villamil, L.G. Schultz, A.W. Hahn, J.L. Cook, Prevalence of and risk factors for hip dysplasia and cranial cruciate ligament deficiency in dogs. *J. Am. Vet. Med. Assoc.* **232**, 1818-1824 (2008).
4. P.M. VanRaden, Efficient methods to compute genomic predictions. *J. Dairy. Sci.* **91**, 4414-4423 (2008).
5. Y. Nagamine *et al.*, Localising loci underlying complex trait variation using regional genomic relationship mapping. *PLoS. One.* **7**, e46501; 10.1371/journal.pone.0046501 (2012).
6. G. Moser *et al.*, Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS. Genet.* **11**, e1004969; 10.1371/journal.pgen.1004969 (2015).
7. M.E. Goddard *et al.*, Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. Biol. Sci.* **283**, 20160569; 10.1098/rspb.2016.0569 (2016).
8. J. Zeng *et al.*, Signatures of negative selection in the genetic architecture of human complex traits. *Nature. Genet.* **50**, 746-753 (2018).
9. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B. Stat. Methodol.* **58**, 267-288 (1996).
10. R.E. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer, A comparison of decision tree ensemble creation techniques. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **29**, 173-180 (2007).
11. B. Xu, *et al.*, Classifying very high-dimensional data with random forests built from small subspaces. *International Journal of Data Warehousing and Mining* **8**, 44-63 (2012).
12. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, A. (n.d.). *R Core Team* (2021). <https://www.r-project.org/>.
13. C. Cortes, V. Vapnik, Support-vector networks. *Mach. Learn.* **20**, 273-297 (1995).
14. P. Pérez, G. de Los Campos, Genome-wide regression and prediction with the BGLR statistical package. *Genetics.* **198**, 483-495 (2014).
15. T. Park, G. Casella, The bayesian lasso. *J. Am. Stat. Assoc.* **103**, 681-686 (2008).
16. T.H.E. Meuwissen, B.J. Hayes, M. Goddard, Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* **157**, 1819-1829 (2001).
17. S.H. Lee, M.E. Goddard, N.R. Wray, P.M. Visscher. A better coefficient of determination for genetic profile analysis. *Genet. Epid.* **36**, 214-224, (2012).
18. N.J.D. Nagelkerke, A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691-692 (1991).
19. D.R. Cox, E.J. Snell. The analysis of binary data. London: Chapman and Hall (1989).
20. Z. Zhao, *et al.*, PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. *Genome. Biol.* **22**, 257; 10.1186/s13059-021-02479-9 (2021).
21. Y. Zhou, *et al.*, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Commun.* **10**, 1523; 0.1038/s41467-019-09234-6 (2019).
22. G.D. Bader, C.W.V. Hogue, An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 1-27 (2003).
23. M. Momen, *et al.*, Genetic architecture and polygenic risk score prediction of degenerative suspensory ligament desmitis (DSLSD) in the Peruvian Horse. *Front Genet* **14**, 1201628; 10.3389/fgene.2023.1201628 (2023).
24. E.K Karlsson, *et al.*, Genome-wide analyses implicate 33 loci in heritable dog osteosarcoma, including regulatory variants near CDKN2A/B. *Genome Biol* **14**, R132; 10.1186/gb-2013-14-12-r132 (2013).