SUPPLEMTARY MATERALS

Fido-SNP: The first webserver for scoring the impact of single nucleotide variants in the dog genome

Emidio Capriotti[1§*], Ludovica Montanucci[2§], Giuseppe Profiti[3], Ivan Rossi[3], Diana Giannuzzi[2], Luca Aresu[4] and Piero Fariselli[2,5*]

[1] BioFolD Unit, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Via F. Selmi 3, 40126 Bologna, Italy.

[2] Department of Comparative Biomedicine and Food Science. University of Padova, Viale dell'Università, 16, 35020 Legnaro (Padova), Italy.

[3] BioDec srl. Via Calzavecchio 20, 40033 Casalecchio di Reno (Bologna), Italy.

[4] Department of Veterinary Sciences, University of Torino, Largo P. Braccini 2, 10095 Grugliasco, (Torino), Italy.

[5] Department of Medical Sciences, University of Torino, Via Santena 19, 10126, Torino, Italy.

* To whom correspondence should be addressed. Tel: +39 051 2094303; Fax: +39 051 209 4286 ; Email: E.C. (emidio.capriotti@unibo.it) and P.F. (piero.fariselli@unito.it).

§ These authors equally contributed to this work.

Evaluation measures for binary classifiers

Fido-SNP prediction output ($\bar{s}$) is rescaled around a threshold of 0.1 using the following equations.

$$\left[\begin{array}{ll} s = 5 \times \bar{s} & \bar{s} < 0.1 \\ s = \frac{5}{9} \times (\bar{s} - 0.1) + 0.5 & \bar{s} \geq 0.1 \end{array}\right. \qquad [1]$$

For each prediction, the binary classification (*Pathogenic/Benign*) is made at the output threshold (*s*). Thus, if probability of *Pathogenic* classification is *s≥0.5* the mutation is predicted to be *Pathogenic*. In all the performance measures - assuming that positives indicate *Pathogenic* and negatives indicate *Benign* - TP (true positives) are correctly predicted Pathogenic Single Nucleotide Variants (SNVs), TN (true negatives) are correctly predicted *Benign* variants, FP (false positives) *Benign* SNVs annotated as *Pathogenic*, and FN (false negatives) are *Pathogenic* variants predicted to be *Benign*.

Predictor performance was evaluated using the following metrics: true positive and negative rates (*TPR, TNR*), positive and negative predicted values (*PPV, NPV*), score and overall accuracy (*Q_2*)

$$
\begin{gathered}
Pathogenic:\ PPV = \frac{TP}{TP+FP} \quad TPR = \frac{TP}{TP+FN} \\
Benign:\ NPV = \frac{TN}{TN+FN} \quad TNR = \frac{TN}{TN+FP} \\
Q_2 = \frac{TP+TN}{TP+FP+TN+FN}
\end{gathered}
\qquad [2]
$$

We computed the Matthew's correlation coefficient *MCC* (Eq. 2) as:

$$
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad [3]
$$

We also calculated the area under the receiver operating characteristic (ROC) curve (AUC), by plotting the True Positive Rate as a function of the False Positive Rate at different probability thresholds of annotating a variant as *Pathogenic* or *Benign*. PhD-SNP^g calculates the False Discovery Rate (FDR) as a function of the returned output (s_0).

$$
Pathogenic:\ FDR(s > s_0) = \frac{FP}{FP+TP} \quad Benign:\ FDR(s < s_0) = \frac{FN}{FN+TN} \qquad [4]
$$

Supplementary Tables

Dataset	Database	# Variants	# Filtered SNVs	Task
hd-pathogenic	ClinVar (Jan 2016)	24,267	1,479	Optimization
dog-omia	OMIA (Nov 2018)	319	75	Validation
772Dogs	https://bit.ly/2KSB0LK	8,459,892	6,038,693	Validation
Lym168	PMID: 25468570	172	168	Validation
dbsnp-benign	dbSNP (build 146)	5,648,530	3,051,393	Optimization+Validation

Table S1. Composition of the data sets used for optimizing and testing Fido-SNP. Database is the resource where the variation data are collected. # Variants: number of variants initially extracted from the databases. # Filtered SNVs: number of Single Nucleotide Variants for which the PhyloP11 conservation score is available.

Supplementary Figures

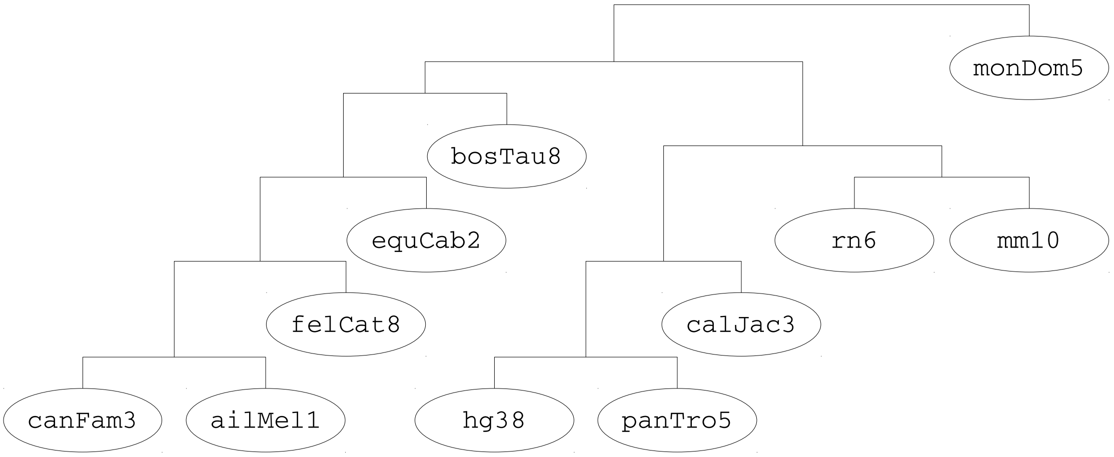

Fig S1. The phylogenetic tree used for the assembly of the pairwise alignments. The genomes of the 11 aligned species are: Dog (canFam3), Panda (ailMel1), Cow (bosTau8), Cat (felCat8) Horse (equCab2), Human (hg38), Mouse (mm10), Chimpanzee (panTro5), Rat (rn6) and Marmoset (calJac3) and Opossum (monDom5).