A Novel Machine Learning Method for Classification of Village Dogs

Ariella L. Gladstein, Samuel H. Vohr, Brett M. Ford, Samantha S. Hauser, Zach Lounsberry, Caleb A. Benson, Brendan J. Ritter, Stephen J. Fromm, Anthony Caragiulo, Aaron J. Sams

Embark Veterinary, Boston, MA | agladstein@embarkvet.com

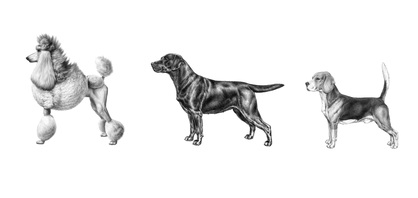
Goal

Embark has been using an IBD-based method to identify village dogs from mixed-breed and single breed dogs.

While that method is accurate, it is computationally expensive at scale.

Our aim is to develop a method that continues to accurately identify village dogs, but is also scalable.

Background

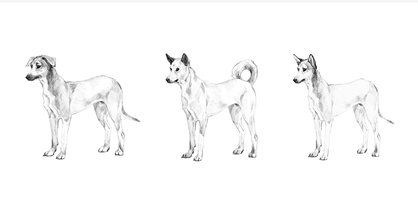


Domestic dog breeds

A group of dogs descended from common ancestors that were selected for specific traits.

Breed ancestry is commonly marked by long identical by state (IBS) segments shared between individuals of the same breed.

Fig. 1: IBS haplotypes matching breed reference panel in a mixed breed dog



Village dogs

Free-ranging, free-breeding dogs whose ancestors were indigenous to a geographic area.

Typically larger effective population size and less extensive linkage disequilibrium compared to dogs with breed ancestry, resulting in shorter IBS segments.

Fig. 2: IBS haplotypes matching breed reference panel in a village dog; black segments represent lack of large contiguous matches.

Detecting village dog ancestry is challenging due to their increased diversity and the relative underrepresentation of these populations in data sets.

Methods

Old IBD threshold method

Assumptions: Breed reference panel does not contain any village dogs

1. Match identical by descent (IBD) segments >3cM between target dog and all dogs in reference panel

   Relies on detecting IBD, which is computationally expensive at scale

2. Calculate median IBD length for each breed

3. Select breeds that likely contribute to ancestry

4. Dogs that have no IBD breed signal or not enough to explain their full ancestry are flagged as village dogs

New Random Forest method

Assumptions: Breed reference panel does not contain any village dogs and village dog reference panel only contains village dogs; Old IBD method sufficiently accurate in identifying mixed breed dogs to use in training and testing.

1. Match IBS segments between target dog and all dogs in reference panel

   >350 breeds, >25K samples, >35 countries

2. Calculate maximum likelihood estimate of ancestry mixture proportions for each breed

3. Use ancestry proportions as features in Random Forest (RF) to classify village dogs vs. not-village dog

4. Train and test RF on village dogs and not-village dogs

	Not-village dogs	Village dogs
Training	34,449	2,106
Testing	3,841	221

Results

The random forest (RF) training and evaluation run quickly, and reaches high accuracy with a relatively small training set

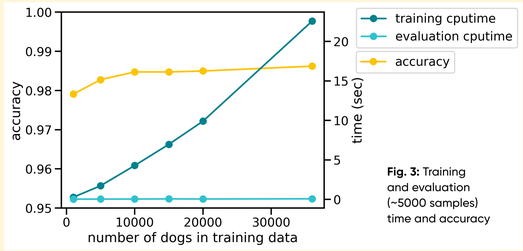


Fig. 3: Training and evaluation (~5000 samples) time and accuracy

The IBD threshold method takes ~35 sec/dog, while RF is trained once for ~20 sec and then takes <1 sec/dog to evaluate

The model accurately classifies village dogs vs. not village dogs, and we expect a small number of misclassifications

Accuracy: 0.98

	Precision	Recall
Not village dog	0.99	0.99
Village dog	0.90	0.85

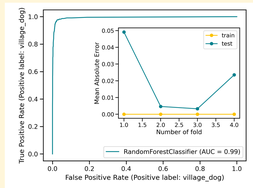


Fig. 4: ROC curve and mean absolute error of K Folds cross-validation

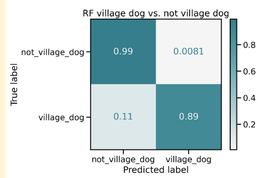


Fig. 5: Confusion matrix

The model has high concordance with the IBD-sum method when validated on customer dogs.

Concordance: 0.99

	RF not-village dog	RF village dog
IBD threshold not-village dog	128,322	514
IBD threshold village dog	426	1,416

% of dogs from countries with widespread village dog populations

	RF not-village dog	RF village dog
IBD threshold not-village dog	4%	88%
IBD threshold village dog	27%	93%

Conclusions

- Classification accuracy is high, even with a relatively small training set.
- We are primarily concerned with the misclassification of not-village dogs as village dogs.
  - We expect that only about 0.8% of not-village dogs will be misclassified as village dogs.
  - We expect the majority of identified village dogs to be true village dogs.
- The RF and IBD threshold methods have high concordance.
  - A limitation of this work is the inherent uncertainty of dog ancestry.
  - The location information provides support that the RF classified village dogs are true village dogs, but without more information on the dogs' histories we cannot be certain.
- Fast model training and evaluation means that the RF classification is scalable to a large number of samples.
- RF classification methods may be useful in other domesticated species with large free-breeding populations and purebred populations.

References

1. Shannon, L. M. et al. Genetic structure in village dogs reveals a Central Asian domestication origin. Proc National Acad Sci 112, 13639–13644 (2015).
2. Gusev, A. et al. Whole population, genome-wide mapping of hidden relatedness. Genome Res 19, 318–326 (2009).
3. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics 30, 1266–1272 (2014).