Figures

**Figure 1. Breed prediction and validation using breed-specific variants identified from WES data.** **A**. The heatmap shows the clustering of 680 animals (498 dogs with breed provided and 182 dogs with no breed provided), using variant allele frequency (VAF) values of the 2,783 breed-specific germline base substitution and small indel variants. For dogs with paired Normal and Tumor samples, the normal one is used; Dogs only with Normal or Tumor samples are also used. These variants were discovered with the WES dataset. The “Provided Breed” bar and the “Disease” bar respectively indicate the breed and tumor type of each dog provided by the source studies. 12 breeds with more than 10 dogs are selected for breed specific variants identification. **B**. The neighbor-joining tree reconstructed based on 2,783 breed specific variants identified from WES datasets as in Figure 1A, across 656 dogs (474 dogs with breed provided and 182 dogs with no breed provided). Only normal dogs are included in this analysis. The tree is unrooted. Tip points and color stripe are annotated based on provided breed. Tree scale is annotated as x-axis.

**Figure S1. Breed validation using breed-specific variants identified from WES data.** **A**. The heatmap shows the clustering of 498 dogs with breed provided, using variant allele frequency (VAF) values of the 2,783 breed-specific germline base substitution and small indel variants. For dogs with paired Normal and Tumor samples, the normal one is used; Dogs only with Normal or Tumor samples are also used. These variants were discovered with the WES dataset. The “Provided Breed” bar and the “Disease” bar respectively indicate the breed and tumor type of each dog provided by the source studies. 12 breeds with more than 10 dogs are selected for breed specific variants identification. **B**. The neighbor-joining tree reconstructed based on 2,783 breed specific variants identified from WES datasets as in Figure S1A, across 474 dogs with breed. Only normal dogs are included in this analysis. The tree is unrooted. Tip points and color stripe are annotated based on provided breed. Tree scale is annotated as x-axis.

**Figure 2. Neighbor-joining tree using all germline variants identified from WES data.** Shown is the neighbor-joining tree reconstructed with 321,961 germline variants identified across the whole exome from 474 dogs. Only normal samples are included in this analysis. The tree is re-rooted with Shih Tzu breed. Tip points and color stripe are annotated based on provided breed. Tree scale is annotated as x-axis.

**Figure S2. Consensus tree using all germline variants identified from WES data.** Shown is the consensus neighbor-joining tree reconstructed using 321,961 germline variants identified across the whole exome, across 474 dogs. To determine the significance of branch placement in the cladogram, the dataset is resampled 100 times by pulling a random 10% of the variants to make 100 distance matrices. The cladograms created from each of the random variant-set matrices are combined using the consense function in PHYLIP programs. Only normal samples are included in this analysis. The tree is re-rooted with Shih Tzu breed. Branch lengths are removed. Tip points and color stripe are annotated based on provided breed. Branch support values greater than 30 are annotated on branches.

**Figure 3. Breed validation using breed-specific variants identified from WGS data.** The heatmap shows the clustering of 278 dogs, using VAF values of the 6,364 breed-specific germline base substitution and small indel variants. These variants were discovered with the WGS dataset generated by Dog10K project (Meadows et al. 2023). The “Provided Breed” bar indicate the breed for each dog provided by the source study. 23 breeds with more than 10 dogs are selected for breed specific variants identification.

**Figure S3. Breed validation using breed-specific variants identified from WGS data, including all breed dogs.** The heatmap shows the clustering of 1,590 dogs, using VAF values of the 6,364 breed-specific germline base substitution and small indel variants. These variants were discovered with 278 dogs in the WGS dataset generated by Dog10K project as in Figure 3 (Meadows et al. 2023). The “Provided Breed” bar indicate the breed for each dog provided by the source study. 23 breeds with more than 10 dogs are selected for breed specific variants identification. Breed dogs other than 23 selected breeds are labeled as "Unknown/Missing".

**Figure 4. Breed prediction and validation using breed-specific variants identified from WES and WGS data.** The heatmap shows the clustering of 965 dogs (783 with breed provided, 182 without breed information), using VAF values of the 5,945 breed-specific germline base substitution and small indel variants. For dogs with paired normal and tumor samples, the normal one is used; Dogs only with normal or tumor samples are also used; Healthy dogs in Dog10K dataset are included (Meadows et al. 2023). Breed specific variants are identified from WES and WGS datasets separately, which is then concatenated. Variant filtering is applied to exclude any variant with low read depth (<10x) in more than 20% dogs. The “Provided Breed” bar and the “Disease” bar respectively indicate the breed and tumor type of each dog provided by the source studies. 12 breeds from WES as well as 23 breeds from WGS with more than 10 dogs are selected for breed specific variants identification.

**Figure S4. Breed validation using breed-specific variants identified from WES and WGS data.** The heatmap shows the clustering of 783 dogs with breed provided, using VAF values of the 5,945 breed-specific germline base substitution and small indel variants. For dogs with paired normal and tumor samples, the normal one is used; Dogs only with normal or tumor samples are also used; Healthy dogs in Dog10K dataset are included (Meadows et al. 2023). Breed specific variants are identified from WES and WGS datasets separately, which is then concatenated. Variant filtering is applied to exclude any variant with low read depth (<10x) in more than 20% dogs. The “Provided Breed” bar and the “Disease” bar respectively indicate the breed and tumor type of each dog provided by the source studies. 12 breeds from WES as well as 23 breeds from WGS with more than 10 dogs are selected for breed specific variants identification.

Tables

**Table S1. List of breed-specific variants identified from WES datasets.**

**Table S2. List of validated and/or predicted dog breeds in WES datasets.** Column “BreedCluster” indicates breed validation/prediction results inferred from VAF value clustering; Column “BreedPhylo” indicates breed validation/prediction results inferred from phylogenetic clustering, where “Others” indicates tumor-only dogs that are excluded from phylogenetic analysis; Column “BreedFinal” indicates the union set of “BreedCluster” and “BreedPhylo”, where “Others” indicates controversial breed results.

**Table S3. List of breed-specific variants of CDS regions identified from WGS datasets.**

Additional files

**Figure2\_tree.nwk.** Neighbor-joining tree file using variants identified from WES datasets.

**Figure3\_breed\_specific\_variants\_all\_wgs.txt.** List of breed-specific variants identified from WGS datasets.