Dog Aging Project | Genomic Data User Guide

2022-04-13

Genetic Data Release

The Dog Aging Project is a nationwide, longitudinal effort to study all dynamics of aging and health span in tens of thousands of companion dogs. As an open, community science initiative, dog-owning participants are an integral part of the project, and all the data that we collect will be shared with scientists around the world. As part of this initiative, we will perform low-coverage DNA sequencing in 10,000 dogs and launch analyses that delineate the genetic and environmental components of aging, longevity, health disorders, and age-related cognitive decline.

Overview

The following package contains genetic data derived from whole genome, low-coverage DNA sequencing and genotype imputation. Derived data, including ancestry and inbreeding, were also generated for each dog.

Sampling

Following cohort enrollment, a saliva sample kit is assigned to the dog and mailed to the participant's address by GBF, Inc., High Point, NC. The participant samples the dog using a DNA Genotek Performagene swab, and returns the sample kit to the mail.

Library preparation and sequencing

The sample kit arrives at Neogen Corporation - GeneSeek Operations (Lincoln, NE) for library preparation. Prepared libraries undergo short read, low-coverage Illumina DNA sequencing at their platform.

Alignment and imputation

Raw sequencing read data is imported by Illumina BaseSpace to the Gencove Platform for alignment and statistical genotype imputation using software *loimpute* by Gencove, Inc. (New York, NY). Not all sequencing runs reach the quality assurance and control metrics needed for genotype imputation. Sequencing reads are aligned to the CanFam3.1 reference genome assembly (NCBI accession GCF_000000145.2). Genotype imputation is performed using *loimpute* and the dog low-pass v2.0 [0.1x-6x] panel of reference haplotypes, consisting of 34,191,821 single nucleotide polymorphisms and 11,943,064 insertions / deletions and representing 540 dogs of known breed ancestry distributed among ~133 breeds, 28 dogs of mixed breed ancestry, 12 dogs of unknown ancestry, 62 worldwide indigenous or village dogs, 33 wolves, and 1 coyote.

Curation and assignment

Successful data deliverables migrate to a Dog Aging Project workspace on Terra.bio for curation and assignment to study IDs. For sequencing runs that do not reach quality needed for genotype imputation, only the raw sequencing reads (FASTQ files) are migrated. In the case of multiple samples per study ID, only genotyping data from the sequencing run with the higher effective genome coverage are assigned to the dog.

Workspace

Data model

The Terra data model table `dna` provides sample information and genomic reports for each dog included in the genotyping data release.

Columns:

- **dog_id** (entity): study ID for the dog
- **swab_id**: swab ID for sample and sequencing run
- **bioproject**: accession for NCBI BioProject
- **biosample**: accession for NCBI BioSample
- **sex**: sex, confirmed by chromosome X coverage (see *Sex confirmation*)
- **coi**: coefficient of inbreeding, estimated from runs of homozygosity (see *Coefficients of inbreeding*)
- **size**: genomic size prediction score, ranging from 0 (under cm tall) to 4 (over cm tall) (see *Size predictions*)

Data files

The Terra workspace bucket includes the following files:

- Genotyping data
  - *PLINK1* bfile data set: `DogAgingProject_GeneticData_CuratedRelease_2021.bed` `DogAgingProject_GeneticD` `DogAgingProject_GeneticData_CuratedRelease_2021.fam`
- Genomic reports
  - Sample information, sex confirmation, coefficients of inbreeding, and genomic size prediction scores: `DogAgingProject_GeneticData_CuratedRelease_2021.tsv`
  - Principal components of genetic variance (dog x principal component): `DogAgingProject_GeneticData_CuratedRe`
  - Pairwise kinship (dog x dog): `DogAgingProject_GeneticData_CuratedRelease_2021.kinship.tsv`
  - Global ancestry (dog x population): `DogAgingProject_GeneticData_CuratedRelease_2021.ancestry.tsv`

Raw sequencing data

Raw sequencing read data (FASTQ files) are deposited to the NCBI Sequence Read Archive under BioProject PRJNA800779. Sample IDs correspond to swab IDs (`swab_id`), not study IDs (`dog_id`).

Genotyping data

Genotyping data (imputed variant calls) are released as a *PLINK1* bfile set containing filtered SNP genotypes. Sample IDs correspond to study IDs (`dog_id`).

Genomic reports

The genomic reports are data derived and inferred from genetic data that are neither raw sequencing nor genotyping data. Sample IDs correspond to study IDs (`dog_id`).

Methods

Sex confirmation

We inferred the sex of each dog from aligned sequencing read data (BAM files) using *SAMtools* to measure the ratio of X chromosome coverage to autosomal coverage. Ratios over or equal to 0.7 were inferred as female, under 0.7 inferred as male. We compared genetically-inferred sex to owner-reported sex given in the Health and Life Experience Survey (HLES). No sex discrepancies exist for dogs included in the data release.

Confirmed sexes for each dog are reported as `sex` in file: `DogAgingProject_GeneticData_CuratedRelease_2021.tsv`

Filtering

Each variant call file (VCF) was filtered for calls of genotype probability above 70% (max(GP) > 0.7) using *BCFtools*. Sample IDs were converted from swab IDs to study IDs and genotyping data from multiple samples were merged in batches using *BCFtools*. Each merged VCF was converted to a *PLINK2* pfile data set (.pgen / .pvar / .psam) containing all variant calls (insertions, deletions, single nucleotide polymorphisms, multiallelic variants). Then, only biallelic single nucleotide polymorphisms (SNPs) were selected (35,875,299 SNPs total) and converted to a *PLINK1* bfile data set (.bed / .bim / .fam). After filtering for a minimum minor allele frequency of 1% and minimum genotype rate of 95%, the final data set included 12,161,842 SNPs. Confirmed sexes were encoded into this final *PLINK1* data set.

Coefficients of inbreeding

We estimated coefficients of inbreeding as autozygosity, or the proportion of genome covered by runs of homozygosity (ROH). We scanned for ROH using *PLINK1* across the 35,875,898 unfiltered biallelic SNP genotypes of genotype probability >70% with the following settings: minimum run length of 500kb (`--homozyg-kb 500`) and minimum SNP count of 100 SNPs (`--homozyg-snp 100`), at a density of 1kb per SNP (`--homozyg-density 1`), with no two SNPs more than 500kb apart (`--homozyg-gap 500`), and only 1 heterozygous genotype tolerated per window (`--homozyg-window-het 1`), performing scans without LD-based pruning on chromosomes 1-38 (`--chr 1-38`). All other settings were *PLINK1* defaults. We then calculated the autosomal ROH-estimated coefficient of inbreeding (FROH, or CoI) from the total ROH segment length divided by the total SNP-covered autosomal length (2,203,765,000 bases) used for ROH detection.

Coefficients of inbreeding are reported as `coi` in file: `DogAgingProject_GeneticData_CuratedRelease_2021.tsv`

Genomic size prediction

Genomic size predictions are reported as `size` in file: `DogAgingProject_GeneticData_CuratedRelease_2021.tsv`

Global ancestry

We selected publicly available genotype data from 109 modern breeds with at least 4 dogs per breed, 3 regional village dog populations (4 Nigerian village dogs, 5 Vietnamese village dogs, 55 Chinese village dogs), and 2 wolf populations (19 North American wolves and 25 Eurasian wolves) (see *Populations* for full list of population labels and counts). We used *PLINK2* to identify ancestry-informative markers. We selected 4,267,732 biallelic single nucleotide polymorphisms with <10% missing genotypes, and calculated the Wright's F-statistics using Hudson method for (1) each dog breed versus all other purebred dogs; (2) all village dogs versus all other purebred dogs; (3) each regional village dog population; (4) all wolves versus all other dogs; (5) North American wolves versus Eurasian wolves. We selected 1,569,037 SNPs with FST > 0.5 across all comparisons, and performed LD-based pruning in 250kb windows for r2 > 0.2 to extract 115,427 markers for global ancestry inference.

We merged genotype data for these biallelic SNPs from query samples with genotype data from reference samples, then performed global ancestry inference using *ADMIXTURE* in supervised mode (random seed: 43) to infer ancestry from these populations. We report only admixture proportions over 1% for each dog.

The global ancestry inferred for each dog are reported in file: `DogAgingProject_GeneticData_CuratedRelease_2021.ancest`

Principal components of genetic variance

We extracted the top 10 principal components from the variance-standardized relationship matrix generated using *PLINK2* to estimate broad population structure. These top 10 PCs explain a cumulative 8.8% of the total variance (313.625 / 3546.08), calculated from summing the diagonal of the relationship matrix.

The principal component weights for each dog are reported in file: `DogAgingProject_GeneticData_CuratedRelease_2021.e`

Pairwise kinship

We applied the *PLINK2* implementation of the KING-robust estimator to measure kinship k between pairs of dogs from mixed populations. Relationships were either unrelated, and removed from the resulting table, or labeled as related ($k > 0$), second-degree ($k >= 0.125$), or first-degree ($k >= 0.25$). We apply a cutoff $k >= 0.35$ to identify duplicate samples. No data from duplicate samples are included in this release.

The relatedness for each pair of dogs are reported in file: `DogAgingProject_GeneticData_CuratedRelease_2021.kinship.t`

Software

The following versions of software were used when referenced:

- *SAMtools*: SAMtools v1.11
- *BCFtools*: BCFtools v1.8
- *PLINK1*: PLINK v.1.9 Stable (19 Oct 2020)
- *PLINK2*: PLINK v.2.00 Development (02 Mar 2021)
- *ADMIXTURE*: ADMIXTURE v.1.3.0

Populations

The full list of populations and sample sizes used for ancestry inference are as follows:

Dog Breeds

- `afghan_hound` (13)
- `airedale_terrier` (17)
- `akita` (14)
- `alaskan_malamute` (16)
- `american_cocker_spaniel` (21)
- `american_pit_bull_terrier` (30)
- `australian_cattle_dog` (19)
- `australian_shepherd` (32)
- `basenji` (15)
- `basset_hound` (17)
- `beagle` (30)
- `bearded_collie` (13)
- `belgian_groenendael` (7)
- `belgian_malinois` (15)
- `belgian_tervuren` (23)
- `berger_picard` (5)
- `bernese_mountain_dog` (24)
- `bichon_frise` (15)
- `bloodhound` (16)
- `border_collie` (31)
- `border_terrier` (17)
- `borzoi` (13)
- `boston_terrier` (18)
- `bouvier_des_flandres` (4)
- `boxer` (29)
- `brittany` (16)
- `bull_terrier` (13)
- `bullmastiff` (23)
- `cairn_terrier` (14)
- `cavalier_king_charles_spaniel` (20)
- `chesapeake_bay_retriever` (16)
- `chihuahua` (20)
- `chinese_crested` (17)
- `chinook` (14)
- `chow_chow` (15)
- `collie` (15)
- `dachshund` (32)
- `dalmatian` (16)
- `doberman_pinscher` (22)
- `english_bulldog` (20)
- `english_cocker_spaniel` (14)
- `english_setter` (15)
- `english_shepherd` (16)
- `english_springer_spaniel` (16)
- `entlebucher` (13)
- `finnish_spitz` (13)
- `french_bulldog` (17)
- `german_shepherd_dog` (33)
- `german_shorthaired_pointer` (17)
- `golden_retriever` (88)
- `gordon_setter` (14)
- `great_dane` (15)
- `great_pyrenees` (14)
- `greater_swiss_mountain_dog` (17)

- `greenland_sled_dog` (11)
- `greyhound` (30)
- `havanese` (18)
- `irish_setter` (15)
- `irish_wolfhound` (21)
- `italian_greyhound` (13)
- `jack_russell_terrier` (22)
- `labrador_retriever` (65)
- `lagotto_romagnolo` (4)
- `leonberger` (51)
- `lhasa_apso` (12)
- `maltese` (20)
- `mastiff` (15)
- `miniature_pinscher` (16)
- `miniature_schnauzer` (30)
- `newfoundland` (17)
- `norfolk_terrier` (12)
- `norwegian_elkhound` (14)
- `norwich_terrier` (4)
- `nova_scotia_duck_tolling_retriever` (15)
- `old_english_sheepdog` (13)
- `papillon` (17)
- `pekingese` (13)
- `pembroke_welsh_corgi` (22)
- `pomeranian` (21)
- `poodle` (29)
- `portuguese_water_dog` (15)
- `pug` (23)
- `rhodesian_ridgeback` (12)
- `rottweiler` (28)
- `saint_bernard` (17)
- `saluki` (12)
- `samoyed` (14)
- `schipperke` (13)
- `scottish_terrier` (15)
- `shar_pei` (12)
- `shetland_sheepdog` (17)
- `shiba_inu` (17)
- `shih_tzu` (18)
- `siberian_husky` (21)
- `sloughi` (4)
- `soft_coated_wheaten_terrier` (15)
- `staffordshire_bull_terrier` (15)
- `standard_schnauzer` (4)
- `tibetan_mastiff` (12)
- `tibetan_spaniel` (12)
- `tibetan_terrier` (17)
- `toy_poodle` (28)
- `vizsla` (12)
- `weimaraner` (15)
- `west_highland_white_terrier` (21)
- `whippet` (14)
- `wire_fox_terrier` (16)
- `wirehaired_pointing_griffon` (13)

- yorkshire_terrier (71)

Regional Village Dogs

- village_dog_china (55)
- village_dog_nigeria (4)
- village_dog_vietnam (5)

Wolf Populations

- wolf_eurasian (25)
- wolf_north_american (19)