Feline Breed Analysis

Basepaws launched a feline breed analysis product based on the comparison of a sample’s DNA to a genomic database of mixed-breed and purebred cats. Basepaws owns the largest feline database in the world (tens of thousands of cats) and is the only company that offers feline breed analysis. As our database grows and increases in diversity and complexity, we will continuously update and refine the breed insights that our report provides.

Contents:

Why Cat is Not Dog - Challenges in Cat Breed Identification

Domestication Versus Selective Breeding

Archeological findings indicate that cat domestication started around 10,000 years ago[1]. This process was most likely connected to the adoption of a more agricultural lifestyle by humans in the Fertile Crescent[2]. Cats were seen as the perfect remedy for controlling crop-destroying rodent pests[2,3]. The considerable benefit of cats in an agricultural setting led to them spreading around the world through common trade routes. Throughout this process, interbreeding between modern cats and feral or wild sub-species **did not** decrease, but rather increased the value cats brought to agricultural human societies[4,5,6]. Therefore, until modern times, there were very few concentrated efforts to selectively breed cats for certain traits[6].

Three historical factors make cat breed identification difficult. The first one is related to the fact that systematic cat breeding only appeared over the past 50 years[6]. In evolutionary terms, this is an extremely short period of time for robust genetically different sub-populations within any species to form.

The second factor has to do with cat breeding methodology. Selective cat breeding has been historically focused on aesthetic features, instead of body structure and functional and behavioural traits. Unlike complex behaviours and body structure, which are polygenic traits (i.e., defined by a cluster of genetic variants), aesthetic traits, such as coat colour and fur length, are usually monogenic traits (i.e., defined by a single gene mutation). This means that cat breeds often differ by a single gene variant (allele), while sharing the majority of alleles associated with other life history factors, such as geographic origin. Consequently, breed definition has been mostly driven by phenotypic presentation (traits) rather than genotype. A byproduct of this is cats with extremely diverse genotypes often being classified as the same breed simply because they have the same monogenic trait.

The final factor contributing to cat breed identification difficulties is related to some cat breeds having an extremely small founder population which had been outcrossed with other breeds in an attempt to avoid inbreeding and to regain certain desirable traits. One such example is the British Shorthair breed. Food shortages during World Wars I and II drastically decreased populations of British Shorthair cats and brought them close to extinction[7,8]. This compelled breeders to outcross the breed with Persian cats. Such instances of outcrossing further complicate the already complex genetics of cat breeds.

The short period of selective breeding, combined with the mostly aesthetics-driven breeding criteria and breed outcrossing, make cats a very unusual case of domesticated animal. This is particularly evident when we compare them to dogs whose domestication started ~14,000 years ago and which have undergone centuries of selective breeding focused on traits defined by complex gene interactions[9]. Dog are an example of extreme selective breeding, where the variation between dog breeds is much greater than the variation between different species in the phylogenetic family that dogs belong to (Canidae). The vast differences between cats' and dogs' evolutionary histories mean that breed analysis based on genotype will yield different conclusions for the two species.

Under-Representation of Cats in Genomic Research

There is a marked disparity in the amount of resources and research effort dedicated to feline genomics versus canine genomics. At the 2019 Conference on Canine and Feline Genetics and Genomics, the presentations focused on canine genomics outnumbered those focused on feline genomics almost five to one[10]. There is also a stark contrast in the genome sequencing goals set for the two fields. While researchers from the 99 Lives cat genome project celebrated when they sequenced the genomes of 200 domestic cats (double their initial goal)[11], the Dog10K Consortium is aiming to sequence the genomes of 10,000 dogs and wild canids[12]. The Dog10K project also aims to sequence all known dog breeds at high depth, which would allow different breeds to have their own high-quality genome assemblies. In contrast, until the 99 Lives project, there had been very little systematic effort to understand genome-wide differences between cat breeds.

Ancestry Versus Genomic Similarity

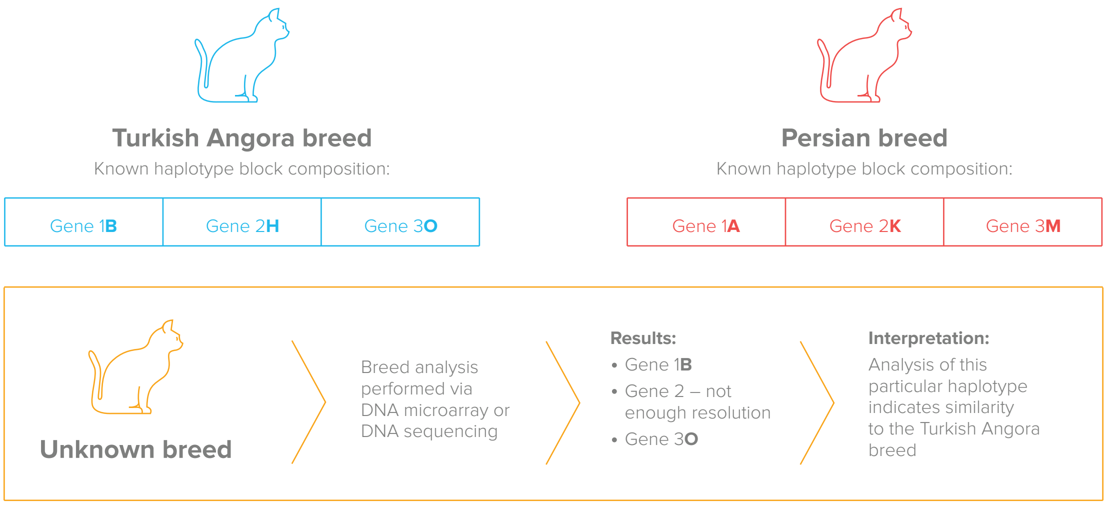
The evolution of cat breeds is inextricably connected to the species' ancestral and geographic history[6]. That being the case, cat breed analysis uses most of the same techniques used for human ethnic ancestry analysis. Commercial genetic tests for human ancestry rely on a comparison between the customer's DNA and a database containing DNA from people with known origins (reference panel)[13]. The goal is to see which parts of the customer's DNA are most similar to the DNA of populations represented in the reference panel. If a DNA segment is most similar to a DNA segment in the reference panel frequently seen in Italian people, this customer DNA segment is assigned to Italy as a place of origin. In the end, the customer gets a percentage breakdown of their DNA's similarity to different ethnicities represented in the reference panel. Since only modern day representatives of different races/ethnicities are used for the reference panel, the customer's ancestry results are simply a proxy for their actual historical ancestry. Our modern day genetic definitions of different races and ethnicities are not based on genetically pure populations and do not (and cannot) accurately account for the numerous human population inter-breeding events that occurred throughout the history of humankind. Ultimately, the customer's results can only tell them how similar their DNA is to different modern day races and ethnicities. The same problem exists when analyzing breeds of modern day cats (and dogs). For example, due to the almost complete historical extinction of British Shorthair cats and their outcrossing with Persian cats, it is impossible to say that a modern day cat is x% similar to the original British Shorthair breed. It is only possible to conclude that a modern day cat is x% similar to the modern British Shorthair breed represented in the genomic reference panel.

A subtle important nuance in both human ancestry analysis and animal breed analysis relates to the way DNA similarity to the reference panel is assessed. Analyses assess the sample of interest's DNA similarity to chunks of DNA (haplotype blocks), rather than the small individual units comprising the genome (nucleotides). Gene alleles are usually inherited together in discrete haplotype block units showing very low amount of 'genetic shuffling' across generations[14]. Because every species has their own haplotype inheritance pattern (also known as a multi-generational haplotype map or a linkage disequilibrium map), haplotype blocks can be used to assess a cat's similarity to a particular breed using a limited amount of data. As **Figure 1** shows, different breeds have a characteristic combination of alleles inherited together within each haplotype block.

A breed analysis has to take into account the sample's genetic similarity to all known feline haplotype blocks before judging the cat's overall genetic proximity to a particular breed. The better the quality of the feline haplotype map, the more accurate the breed analysis. In order to build a thorough high-resolution haplotype map, the reference panel should contain genome sequencing data from thousands of cats representing different breeds and geographic locations. If the reference panel has a small sample size or an obvious bias in population sampling, the allele frequency and allele co-segregation estimates on which the haplotype map is built will be inaccurate. Given the previously discussed limitations of the feline genomics effort, cats are disadvantaged when it comes to having sufficient publically available genomic data for the creation of a highly detailed haplotype map (and therefore breed identification).

Known Feline Haplotype Block: Genes 1, 2 and 3 are inherited together

Gene 1	Gene 2	Gene 3
Known **Gene 1** alleles: **A, B, C, L**	Known **Gene 2** alleles: **H, I, K, S**	Known **Gene 3** alleles: **M, O, Q, Z**



Turkish Angora breed

Known haplotype block composition:

Gene 1**B**	Gene 2**H**	Gene 3**O**

Persian breed

Known haplotype block composition:

Gene 1**A**	Gene 2**K**	Gene 3**M**

Unknown breed

Breed analysis performed via DNA microarray or DNA sequencing

Results:

- Gene 1**B**
- Gene 2 – not enough resolution
- Gene 3**O**

Interpretation:

Analysis of this particular haplotype indicates similarity to the Turkish Angora breed

Figure 1. Illustrative example of the use of haplotype blocks in cat breed analysis.

Low-Pass Whole Genome Sequencing and Basepaws' Approach to Cat Breed Analysis

Currently, all commercially performed pet breed analysis tests are based on a genotyping array where a set of known gene variants are assessed. While genotype arrays can have as many as 2 million gene variants[15], arrays used for direct-to-consumer pet genetic tests have fewer than a few hundred thousand variants represented. While economically favorable for high-volume commercial use, genotype microarrays can only be used for profiling known gene variants and detection of novel variants is not possible unless the hardware associated with the assay is re-designed.

An alternative to genotyping arrays is DNA sequencing where instead of focusing on pre-selected gene variants, the entire genome is queried with some average depth of coverage. Unlike microarrays, genome sequencing allows discovery of new gene variants. With dropping sequencing costs, low-pass DNA sequencing (typically defined as <1X coverage of the genome) presents an attractive substitute for microarrays. For comparison, 0.4X coverage translates to around one read covering each of ~30 million genetic variants of the human genome, while genotyping arrays provide information on orders of magnitude fewer genomic loci[15].

Low-pass sequencing is particularly useful when combined with imputation analysis, which allows to fill in the data gaps and impute missing data. Obtaining information on a few different variants in a haplotype block allows imputing the remaining known variants within the same block. A 0.4X genome coverage combined with imputation analysis was found to be 98.2% concordant with a genotype array-based analysis, while 1X coverage showed 99.2% concordance[15]. Therefore, low-pass sequencing in combination with imputation analysis can provide at least the same level of accuracy as a genotyping array.

Due to the fact that feline genomics is at very early stages compared to other organisms, such as humans and dogs, using DNA sequencing to study the cat genome is the only available scientifically justifiable option. It allows for discovery and the building of a robust multi-generational haplotype map. Once such a map is in place, low-pass genome sequencing and bioinformatic imputation can be used effectively and accurately for cat breed analysis. Basepaws has built (and is continuing to build) an extensive reference panel of purebred and mixed-breed cats' genome sequencing data. We have used this data to build a multi-generational haplotype map to serve as the foundation of our breed analysis.

Establishing a Reference Panel of Cats

In order to build and continuously improve our genomic reference panel, we are sequencing multiple cats from different breeds, including mixed-breed cats (i.e., polycats), from across the world. As of Q3 2019, we have sequenced 18 breeds with multiple representatives at high-depth genome coverage (average 15X). **Table 1** provides a list of the breeds included in our reference panel, together with their respective breed groups.

Breed group	Breed
Western	Norwegian Forest Cat
	American Shorthair
	Siberian
	Maine Coon
	Ragdoll
	Russian Blue
	Abyssinian
Eastern	Oriental Shorthair
	Birman
	Burmese
	Peterbald
Exotic	Savannah
	Bengal
	Egyptian Mau
Persian	British Shorthair
	Exotic Shorthair
	Himalayan
	Persian
Mixed-breed (polycat)	

Table 1. Breeds currently represented in Basepaws' genomic reference panel

Using all of our high-coverage genomic data, we filtered for loci with minor allele frequency (MAF) > 0.05. We then fed the data through a linkage disequilibrium (LD) pruning pipeline, keeping only loci with a correlation value $R^2 < 0.7$ within 1Mb of each other. The variants that passed this filter (~300,000) were used to perform a Principal Component Analysis (PCA) to observe breed clusters based on genetic similarity (forming Eastern, Western, Exotic, Persian, and Polycat breed group clusters). Highly correlated variants ($R^2 > 0.7$) were removed from the PCA to avoid biasing the analysis towards any one set of correlated variants.

Our reference panel is continuously maintained and enriched with new cat samples. When adding new samples to our panel, we first sequence them at low coverage as an initial screen to decide whether to sequence them at a higher depth. We use the low-pass sequencing data from these samples to map onto the PCA generated from our existing reference panel. We use Projection Procrustes Analysis[16] to do this and have observed that as few as 10,000 - 20,000 LD pruned loci are sufficient to recover population structure with less than 10% deviation (**Figure 2**).

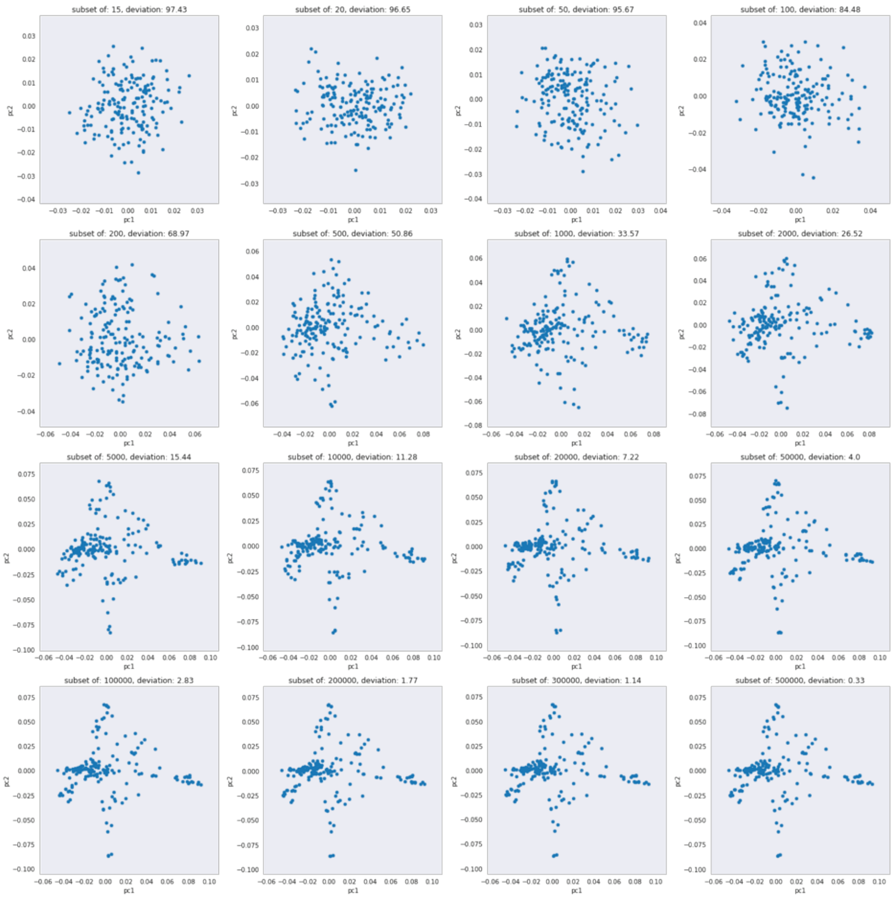


Figure 2. Projection Procrustes Analysis validation of low-pass sequencing data.

The Projection Procrustes Analysis allows us to see how well our new samples cluster with our high coverage reference samples. If we observe that, for example, a cat that is claimed to be a Turkish Angora does not cluster with other Turkish Angora cats in our reference panel, we do not include it in our reference panel. Conversely, if it clusters with other Turkish Angora cats and has a matching breed certificate, we sequence the sample at a higher depth and add it to our reference panel and PCA (**Figure 3**). In order to avoid biasing our reference panel towards our founder set of cats, as we accumulate more samples, we periodically re-analyse our reference data together with the samples that were initially excluded from the reference panel.

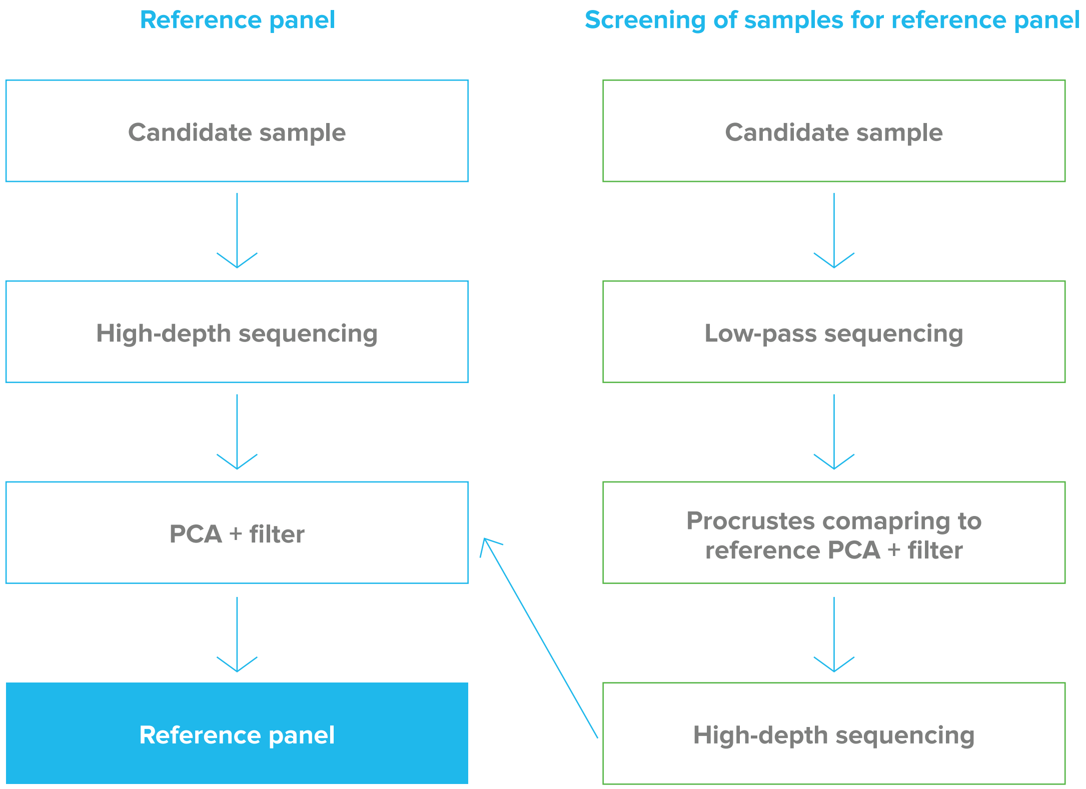


Figure 3. Workflow for building and updating Basepaws' breed reference panel.

Building a Multi-Generational Haplotype Map

We used the genomic data from our reference panel, together with our customer low-pass sequencing data (average coverage of 0.44X) and identified ~18 million variants. We further filtered these variants for MAF > 0.05 and a Hardy-Weinberg Error < 0.001. We then used the resulting ~11 million variants to generate a feline multi-generational haplotype map using the program LDMAP[17].

Since LDMAP uses LD units as a genetic linkage measure, we next performed a conversion of LD units to centimorgans (cM), the commonly accepted unit for measuring genetic linkage. In order to do this conversion, we used Purina's feline genetic linkage map[18] (in cM units) and overlaid it on top of our map. This showed a monotonic relationship between LD units and cM and allowed us to convert our map to cM using a monotonic cubic spline interpolator (**Figure 4**).

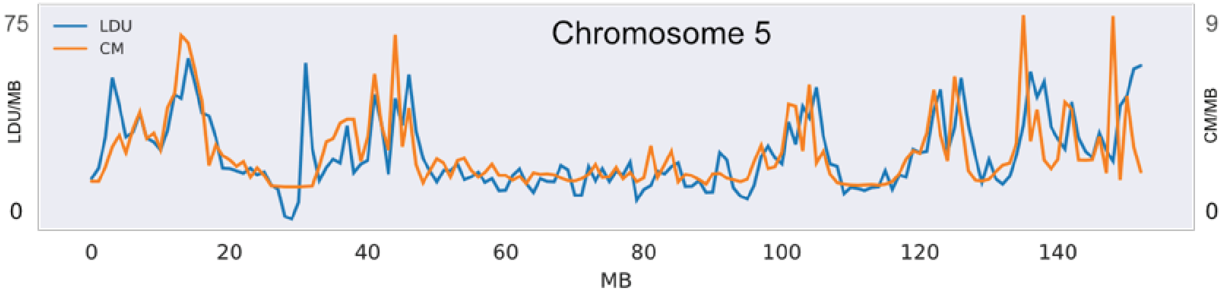


Figure 4. Purina's feline genetic linkage map for chromosome 5 in cM (orange) overlaid on top of Basepaws' multi-generational haplotype map for chromosome 5 in LD units (blue). The sharp LDU peak between 20 and 40MB is indicative of a centromeric region.

Analysis of Customer Samples

Every customer sample prepared for breed analysis undergoes DNA extraction and sequencing library preparation compatible with processing on an Illumina sequencing instrument. Samples are sequenced to an average of 0.44X genomic coverage. Every analysis begins with sample de-multiplexing and assigning sequencing reads to the correct sample. Following this, PCR duplicates are removed and the remaining sequencing reads are mapped to the latest version of the domestic cat's genome assembly (felix_catus_9.0). Next, variant calling using the program Genome Analysis Tool Kit HaplotypeCaller (GATK-HC)[19] is performed, followed by an imputation analysis of un-genotyped markers performed by the Beagle program[20].

There are two types of analysis included in our breed report (**Figure 5**):

- Overall sample DNA similarity to breed groups and individual breeds (given as a percentage)
- Chromosome painting analysis demonstrating particular genomic regions similar to particular breed groups and individual breeds

Both of these types of analysis are based on sample comparison to our reference panel. We use our multi-generational haplotype map to segment the genome into 10cM blocks. Each 10 cM block of sample DNA is compared to each 10 cM block of our reference panel using Ensemble's Random Forest machine learning classification algorithm[21]. The Random Forest algorithm is used in a stacked fashion - once, to identify the sample's similarity to a breed group, and a second time - to identify the sample's similarity to different breeds within a breed group.

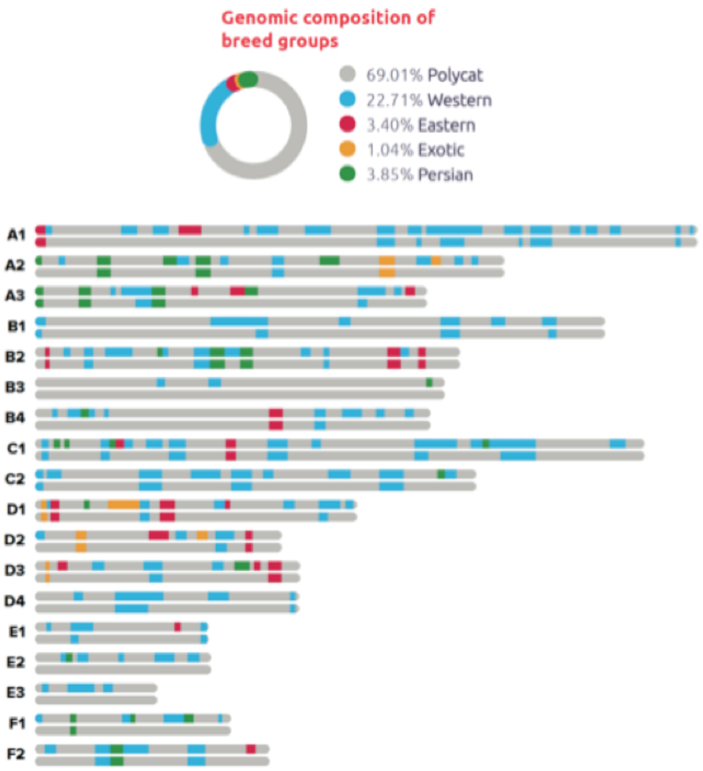


Figure 5. An example percentage breed group similarity and chromosome painting analyses included in the Basepaws' report.

About Basepaws

Basepaws is an animal health company specializing in genetics. In 2018, we launched the world's first at-home consumer DNA test for cats focused on delivering health and breed-related actionable insights. Our feline breed analysis product is the only one on the market. We are committed to providing a cutting edge service and are continuously updating our breed reference panel adding samples from under-represented breeds, as well as including new breeds that are currently not present in our panel. We want to accelerate discovery in feline health and believe that understanding breed is a key first step towards achieving this goal. Our mission is to improve the health and well-being of every pet.

References

1) Vigne, J.D., Guilaine, J., Debue, K., Haye, L. and Gérard, P., 2004. Early taming of the cat in Cyprus. Science, 304(5668), pp.259-259

2) Gupta, A.K., 2004. Origin of agriculture and domestication of plants and animals linked to early Holocene climate amelioration. CURRENT SCIENCE-BANGALORE-, 87, pp.54-59

3) Zohary, D. and Hopf, M., 2000. Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe and the Nile Valley (No. Ed. 3). Oxford University Press

4) Dobney, K. and Larson, G., 2006. Genetics and animal domestication: new windows on an elusive process. Journal of Zoology, 269(2), pp.261-271

5) Randi, E., Pierpaoli, M., Beaumont, M., Ragni, B. and Sforzi, A., 2001. Genetic identification of wild and domestic cats (Felis silvestris) and their hybrids using Bayesian clustering methods. Molecular Biology and Evolution, 18(9), pp.1679-1693

6) Lipinski, M.J., Froenicke, L., Baysac, K.C., Billings, N.C., Leutenegger, C.M., Levy, A.M., Longeri, M., Niini, T., Ozpinar, H., Slater, M.R. and Pedersen, N.C., 2008. The ascent of cat breeds: genetic evaluations of breeds and worldwide random-bred populations. Genomics, 91(1), pp.12-21

7) http://cfa.org/Breeds/BreedsAB/BritishShorthair/BritishShorthairArticle(2002).aspx

8) http://www.pethealthnetwork.com/cat-health/cat-breeds/british-shorthair

9) Adams, J., 2008. Genetics of Dog Breeding. Nature Education 1(1):144

10) https://www.catdog2019.unibe.ch/program/#pane693504

11) http://felinegenetics.missouri.edu/99lives

12) Wang, G.D., Larson, G., Kidd, J.M., vonHoldt, B.M., Ostrander, E.A. and Zhang, Y.P., 2019. Dog10K: The International Consortium of Canine Genome Sequencing. National Science Review

13) Ethnicity Estimate 2018 White Paper (Ancestry) - https://www.ancestrycdn.com/dna/static/images/ethnicity/help/WhitePaper_Final_091118dbs.pdf

14) https://www.broadinstitute.org/international-haplotype-map-project/haplotype-map

15) Wasik, K., Berisa, T., Pickrell, J.K., Li, J.H., Fraser, D.J., King, K. and Cox, C., 2019. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. bioRxiv, p.632141

16) Wang, C., Zhan, X., Liang, L., Abecasis, G.R. and Lin, X., 2015. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. The American Journal of Human Genetics, 96(6), pp.926-937.

17) Kuo, T.Y., Lau, W. and Collins, A.R., 2007. LDMAP. In Linkage Disequilibrium and Association Mapping (pp. 47-57). Humana Press

18) Li G, Hillier LW, Grahn RA, et al. A High-Resolution SNP Array-Based Linkage Map Anchors a New Domestic Cat Draft Genome Assembly and Provides Detailed Patterns of Recombination. G3 (Bethesda). 2016;6(6):1607–1616. Published 2016 Jun 1. doi:10.1534/g3.116.028746

19) McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20(9), pp.1297-1303

20) https://faculty.washington.edu/browning/beagle/beagle.html

21) https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Thank you for your attention