



Cat Ancestry Methods Overview

Kasia Bryc, PhD

Director of Research, Darwin's Ark

Senior Research Scientist II, Broad Institute of Harvard and MIT

Published: September 14, 2026

Summary

This methods paper provides a brief overview of how Darwin's Ark cat ancestry insights are produced. These ancestry insights showcase ancestry estimates across 42 populations, including populations corresponding to geographic locations, cat breeds, and other species of wild cats. An individual cat's genetic data is compared to the patterns of genetic variation in cats of known ancestry belonging to these 42 reference populations. The cat's ancestry estimates reflect the mixture of reference populations that best models the cat's individual genetic data.

Low-Pass Sequencing

Darwin's Ark uses low-pass sequencing data to generate genetic information on each participating cat. This low-pass sequence data is then filtered, aligned to the cat genome, and compared to a set of cats with high-coverage, high-quality genetic information to determine the most likely genotype, or genetic variants, carried at each site. The set of deeply sequenced cats is referred to as our "imputation panel", which is available by request for research use through the Darwin's Ark website: <https://darwinsark.org/collaborate>.

Cat Imputation Panel

The Darwin’s Ark cat imputation panel is composed of felines with high coverage Illumina sequence data publicly available in the Sequence Read Archive (SRA), a major bioinformatics database run by the National Institutes of Health (NIH) that houses raw sequencing data from genomic research. The imputation panel includes cat sequences contributed by the 99 Lives Cat Genome Project [Lyons, 2016, Buckley et al., 2020, 99 Lives Cat Genome Consortium et al., 2021], and the Mars Petcare Biobank, among others. After significant variant and sample QC, the final imputation panel is composed of 992 cats, mainly domestic cats, about 20% of which are pedigreed cats from 44 breeds, as well as a small number of animals from six wild cat species. The panel contains 99.2 million high-quality variants, the majority of which are SNPs (single nucleotide polymorphisms), but also includes 12 million small genomic insertions and deletions (indels). By comparing the genetic information for each Darwin’s Ark cat against this imputation reference panel, we are able to obtain high-quality genetic variant calls at 99.2 million sites for each participating cat. Read more about the construction of this imputation panel with our computational collaborator Fulcrum Genomics [here](#).

Ancestry Reference Panel

After analyzing the population structure of the first thousand Darwin’s Ark participating cats, together with the imputation panel cats, we were able to detect significant population substructure among domestic cats and other species of wild cat. The population structure among domestic cats corresponded largely to differences among cats born in different geographic regions. This substructure suggests the majority of population structure is driven by isolation-by-distance. In other words, cats living close together are more genetically similar, and cats living further apart are genetically more differentiated. Additional structure corresponded to genetic differences from cats from across cat breeds. We selected representative individuals with ancestry primarily from one of the geographic subpopulations, cat breeds, and wild cat species, to create an “ancestry reference panel”, or set of representative cats from each of 42 reference populations. Counts of each population’s representatives and their sources are described in Table 1.

Ancestry Estimation

To obtain ancestry estimates for each participating cat, we utilize their genetic information across 155,210 genome-wide variants. This genetic dataset has been thinned for linkage disequilibrium and filtered for missingness and minor allele frequency. On this pruned

genetic dataset, we compare an individual cat’s genetic information against the genetic information of the ancestry reference panel cats. We use the algorithm ADAMIXTURE [Saurina-i Ricos et al., 2026], which models each cat as a mixture of ancestry from each of the reference populations, finding the optimal mixture that best explains that individual’s genetic patterns. In the case where a cat does not have ancestry from any of the reference panel populations, the model either assigns the closest available population, or a mixture of proxy populations to attempt to recreate the missing ancestry.

Because this first ancestry reference panel has some populations represented by very few individuals, the genetic picture of those small populations is more fuzzy, and thus ancestry proportion estimates for those populations are less reliable. After inference, any estimate less than 2% is dropped and aggregated under “unknown,” except the European Wildcat. The European Wildcat, *Felis silvestris*, is both well-represented in our ancestry reference dataset and highly genetically diverged from the domestic cat, and therefore we show estimates down to 0.5%. In cases where a cat has many populations with less than 2% fractions, these fractions are aggregated and the final estimate of unknown ancestry can be as much as 10% or 20%. Ancestry estimates with large portions of “unknown” ancestry often occur when the cat’s true ancestry is not represented in the ancestry reference panel. In these cases, the algorithm attempts to reconstruct unrepresented populations through many small mixtures.

Finally, we compute precision and recall by 10-fold cross validation, where 10% of reference panel individuals are excluded from the reference panel and their ancestry estimates are inferred by the remaining 90% of individuals. The recall represents the amount of time that an ancestry is correctly assigned in reference panel individuals from that population, and precision reflects the fraction of ancestry assigned to the population that belongs to individuals from that reference group. Though these are best-case scenarios, and real-world admixed individual performance is likely much lower, these precision and recall estimates provide a relative estimate of performance among the various populations.

Table 1: Reference panel population counts, sources, together with precision and recall statistics for out-of-panel ancestry estimates for each population of reference individuals.

Population	Darwin’s Ark	Public Datasets	Total	Prec. (%)	Rec. (%)
<i>North American Breeds</i>					
Bengal & Toyger	9	6	15	99.86	99.99
Maine Coon	16	18	34	99.96	99.58

Continued on next page

Table 1 continued from previous page

Population	Darwin's Ark	Public Datasets	Total	Prec. (%)	Rec. (%)
Ragdoll	7	5	12	99.80	100.00
Savannah	1	1	2	99.34	99.98
Sphynx	0	12	12	100.00	99.99
<i>International Breeds</i>					
Abyssinian & Somali	4	3	7	99.07	100.00
Birman	7	7	14	99.90	100.00
British Shorthair & Chartreux	2	11	13	99.99	100.00
Burmese	0	11	11	99.81	100.00
Devon Rex	0	3	3	99.73	100.00
Egyptian Mau	2	1	3	98.77	99.99
Exotic, Himalayan, & Persian	2	12	14	99.82	100.00
Korat	0	3	3	98.46	100.00
Norwegian Forest Cat	2	3	5	98.95	100.00
Ocicat	3	1	4	98.81	100.00
Oriental and Siamese	1	9	10	99.52	100.00
Chinchilla & Persian	4	2	6	99.99	99.99
Russian Blue	1	3	4	99.72	100.00
Scottish Fold	0	4	4	99.08	100.00
Siberian	17	1	18	98.89	100.00
Toybob	0	2	2	99.80	100.00
<i>International Regions</i>					
East Asia	0	9	9	99.99	99.99
Central Eurasia & N. Africa	5	5	10	99.99	97.67
Northern Europe	1	13	14	97.59	100.00
<i>North American Regions</i>					
Northeast	42	3	45	99.27	99.47
Northeastern Core	20	11	31	100.00	98.86
Mid-Atlantic	11	16	27	100.00	99.91
Lower South Atlantic	12	17	29	99.69	98.73
Ohio River Valley	7	6	13	100.00	97.87
Great Lakes	19	3	22	99.80	100.00
Western Great Lakes	2	26	28	100.00	99.97
Northern Plains	20	7	27	99.78	99.90
Central US	16	8	24	100.00	99.78
Gulf Coast	17	3	20	98.28	99.83

Continued on next page

Table 1 continued from previous page

Population	Darwin's Ark	Public Datasets	Total	Prec. (%)	Rec. (%)
Southwest	32	13	45	100.00	99.35
Pacific Northwest	12	5	17	100.00	99.99
Hawaii	11	0	11	98.83	100.00
Wild Cats					
Sand Cat	0	2	2	99.96	100.00
Black-Footed Cat	0	3	3	99.98	100.00
European Wildcat	0	6	6	99.99	99.99
Pallas's Cat	0	3	3	99.98	100.00
Leopard Cat	0	9	9	99.99	100.00
Grand Total	305	286	591		

References

- 99 Lives Cat Genome Consortium, L. A. Lyons, R. M. Buckley, and R. J. Harvey. Mining the 99 Lives Cat Genome Sequencing Consortium database implicates genes and variants for the Ticked locus in domestic cats (*Felis catus*). *Animal Genetics*, 52(3):321–332, 2021. doi: 10.1111/age.13059. URL <https://doi.org/10.1111/age.13059>.
- Reuben M. Buckley, Brian W. Davis, Wesley A. Brashear, Fabiana H. G. Farias, Kei Kuroki, Tina Graves, LaDeana W. Hillier, Milinn Kremitzki, Gang Li, Rondo P. Middleton, Patrick Minx, Chad Tomlinson, Leslie A. Lyons, William J. Murphy, and Wesley C. Warren. A new domestic cat genome assembly based on long sequence reads empowers feline genomic medicine and identifies a novel gene for dwarfism. *PLOS Genetics*, 16(10):e1008926, 2020. doi: 10.1371/journal.pgen.1008926. URL <https://doi.org/10.1371/journal.pgen.1008926>.
- Leslie A. Lyons. Precision medicine and the 99 Lives Cat Genome Sequencing Initiative. *Journal of Animal Science*, 94(suppl_4):171–172, 2016. doi: 10.2527/jas2016.94supplement4171a. URL <https://doi.org/10.2527/jas2016.94supplement4171a>.
- Joan Saurina-i Ricos, Daniel Mas Montserrat, and Alexander G. Ioannidis. ADAMIXTURE: adaptive first-order optimization for biobank-scale genetic clustering. *Bioinformatics*, 42 (Supplement_1):btg236, 2026. doi: 10.1093/bioinformatics/btg236. URL <https://doi.org/10.1093/bioinformatics/btg236>. PMID: 42412790.