**Genomic ancestry of 85-lb New York canid**

Report by:

Dr. Bridgett vonHoldt
Associate Professor of Ecology & Evolutionary Biology
Princeton University
Princeton, NJ USA
vonHoldt@princeton.edu
http://canineancestry.princeton.edu

**Samples and DNA sequencing**

I received one sample (referred to as "Butera") from a New York location. I obtained genomic DNA using Qiagen DNEasy kits for blood and tissue. I prepared the genomic DNA for restriction-site associated and DNA capture sequencing (referred to as "RADseq"; Ali et al. 2015) by first digesting DNA with the *SbfI* restriction enzyme followed by ligation of a unique 8-bp barcoded biotinylated adapter. I then pooled equal nanograms of DNA from 48 samples, followed by random shearing to 400bp in a Covaris LE220 and enriched for the adapter ligated fragments using a Dynabeads M-280 streptavidin binding assay. I prepared the enriched pools the NEBnext Ultra II DNA Library Prep Kit for Illumina NovaSeq 6000 paired-end (2x150nt) sequencing at Princeton University's Lewis-Sigler Genomics Institute core facility. I used Agencourt AMPure XP magnetic beads for all steps of library cleaning or retaining 300-400bp fragments.

**Bioinformatic processing**

I retained sequence reads pairs that contained the unique barcode and remnant *SbfI* recognition site. Using *STACKS* v2.6 (Catchen et al. 2013; Rochette et al. 2019), I first rescued barcoded reads in the *process_radtags* module (a 2bp mismatch) and discarded reads with <10 quality score. I next removed PCR duplicates in the *clone_filter* module. I mapped sequence reads to the reference dog genome CanFam3.1 assembly (Lindblad-Toh et al. 2005) and the Y chromosome (KP081776.1; Li et al. 2013a) using *bwa-mem* (Li 2013b). I discarded mapped reads with MAPQ<20 and converted the SAM files to BAM format in *Samtools* v0.1.18 (Li et al. 2009).

**SNP discovery and genotyping**

I included canids representing each of the major demographic lineages and an enrichment for canids from the neighboring geography. I used the *gstacks* and *populations* modules in *STACKS* to discover and genotype SNP variants. I increased the minimum significance threshold in *gstacks* and used the marukilow model flags --vt-alpha and --gt-alpha with $p$=0.01. I conducted an initial filtering with *VCFtools* v0.1.17 (Danecek et al. 2011) to exclude singleton and private doubleton alleles, remove loci with more than 10% missing data across all samples, and remove individuals with more than 20% missing data. I filtered to exclude sites with a minor allele frequency (MAF<0.03) and allowed up to 80% genotyping rate per locus in *PLINK* v1.90b3i (Chang et al. 2015). Demographic estimates are most reliably obtained from statistically neutral and unlinked loci. Hence, I further filtered to exclude loci within 50-SNP windows that exceeded a genotype correlation of $r$=0.5 (*--indep-pairwise* 50 5 0.5; a proxy for linkage disequilibrium or LD) and significantly deviated from Hardy-Weinberg Equilibrium (HWE) ($p$<0.001).

I discovered 11,120,040 loci with a per-sample pre-filtering average coverage of 10.5x (s.d.=6.6) for 607 after removing 24 samples due to missing data. This dataset is composed of coyotes (n=230 representing US States of AZ, CA, ID, ME, MI, MN, NJ, NM, NV, NY, PA, VT, WI, and Canada provinces of New Brunswick, Ontario, and Saskatchewan), western gray wolves (n=57 representing CA, OR, WA, and WY), Great Lakes gray wolves (n=269 representing MI, MN, Ontario, and WI), eastern wolves (n=23 representing Ontario), red wolves (n=20 representing the captive breeding population and North Carolina), and domestic dogs (n=8 representing Ontario). After initial filtering, I obtained genotypes for 311,614 loci from which I further retained 43,139 loci after MAF and missing data filtering. After filtering to establish a statistically neutral and unlinked SNP set, I retained 27,756 SNP loci.

**Population genetic analysis**

I conducted a non-model cluster analysis using principal component analysis (PCA) of 27,756 SNP loci genotyped in 607 canids in *FlashPCA v2.1* (Abraham et al. 2017). I used the resulting cluster pattern to select specific canid lineages for global ancestry inference for the canid in question ("Butera"). First, I find the expected patterning of wild canid lineages populating each "tip" of a larger triangular shape: gray wolves (western), gray wolves (Great Lakes) and coyotes (**Fig. 1**). The split of the gray wolves into western and Great Lakes is due to their distinct demographic history. Gray wolves that inhabit the Great Lakes have a recent and possibly ongoing history of admixture with coyotes. Great Lakes gray wolf populations and genomes have on average ~10-20% coyote ancestry (vonHoldt et al. 2011). If any Great Lakes wolf genetics contributed to the canids in question, it would be important to assess if they also carried coyote genetics. I also find the expected placement of red and eastern wolves intermediate of the Great Lakes gray wolves and coyotes. This is due to their history of admixture and shared ancestry. The cluster of domestic dogs is spatially adjacent to the two gray wolf clusters, reflecting that dogs and gray wolves share a common ancestry but also the lack of coyote admixture found in the dog genome.

The "Butera" sample clustered squarely within the Great Lakes gray wolf cluster, suggesting I next test for coyote and gray wolf ancestry. The demographic history that eastern wolves are found in the greater Great Lakes region tells me it is also critical to capture that information thus include the eastern wolves in the ancestry inference.

**Global ancestry inference**

I implemented a two-layer hidden Markov model in the program *ELAI* to infer global (i.e. genome-wide) ancestry proportions with respect to five reference populations: gray wolves (western), gray wolves (Great Lakes), eastern wolves, dogs, and coyotes (Guan 2014). Due to the uncertain nature of historic admixture, I inferred ancestry at two time points (5 and 10 generations) in triplicate. I then averaged results over all independent analyses and only report autosomal ancestry proportions. The X chromosome was not included.

Results for each sample's ancestry percentages per reference canine group is listed in the table below. I find that the "Butera" sample carries a combined 97.8% gray wolf (*C. lupus*) ancestry with nearly all of that derived from gray wolves of the Great Lake and only 1.4% identified as Eastern wolf (*C. lycaon*).

| Sample | Coyote | Great Lakes gray wolf | Eastern wolf | Gray wolf | Dog |
|---|---|---|---|---|---|
| Butera | <1% | 96.2% | 1.4% | 1.6% | <1% |

**Literature cited**


Abraham G, Qiu Y, Inouye M (2017) FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* **33**(17), 2776-2778.

Ali OA, et al. (2015) RAD Capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics* **202**, 389-400.

Catchen J, et al. (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**(11), 3124-3140.

Chang CC, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**. doi:10.1186/s13742-015-0047-8

Danecek P, et al. (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Guan Y (2014) Detecting structure of haplotypes and local ancestry. *Genetics* **196**, 625-642.

Kierepka EM, Kilgo JC, Rhodes OE Jr (2017) Effect of compensatory immigration on the genetic structure of coyotes. *Journal of Wildlife Management* **81**(8), 1394-1407.

Li H, et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li H (2013b) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2.

Li G, et al. (2013a) Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Research* **23**, 1486-1495.

Lindblad-Toh K, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819.

Rochette NC, Givera-Colon AG, Catchen JM (2019) Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology* doi:10.1111/mec.15253.

vonHoldt BM, et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research* 21, 1294-1305

**Figure 1.** Principal component analysis of 27,756 SNP loci genotyped in 607 canids. The black star denotes the canid sample in question.