REPLY TO WANG ET AL.:

Sequencing datasets do not refute Central Asian domestication origin of dogs

Laura M. Shannon[a], Ryan H. Boyko[b], Marta Castelhano[c], Elizabeth Corey[c], Jessica J. Hayward[a], Corin McLean[d], Michelle E. White[a], Mounir R. Abi Said[e], Baddley A. Anita[f], Nono Ikombe Bondjengo[g], Jorge Calero[h], Ana Galov[i], Marius Hedimbi[j], Bulu Imam[k], Rajashree Khalap[l], Douglas Lally[m], Andrew Masta[n], Kyle C. Oliveira[a], Lucía Pérez[o], Julia Randall[p], Nguyen Minh Tam[q], Francisco J. Trujillo-Cornejo[o], Carlos Valeriano[h], Nathan B. Sutter[r], Rory J. Todhunter[c], Carlos D. Bustamante[s], and Adam R. Boyko[a,1]

We welcome the additional data and analyses of Wang et al. (1), but believe there are some misunderstandings regarding the methods and findings of Shannon et al. (2). First, although we merged Nepal and Mongolia when plotting linkage disequilibrium (LD) decay in figure 5B of ref. 2 for legibility, we did not assume Nepal and Mongolia represented a single, interbreeding population, and indeed computed separate LD scores for each population (figure 5A of ref. 2), matching Wang et al.'s (1) observation of slightly lower LD in Nepal than Mongolia. Although Nepal (along with India) is commonly considered part of South Asia, Nepal borders Central Asia. Dog populations in two Central Asian countries, Mongolia and Afghanistan, both have lower LD than India. Nepal does not border Southeast Asia. Because we cannot, given the resolution of current sampling data, say the center of diversity is precisely within Nepal, the LD pattern observed is most consistent with an origin of dogs in or near Central Asia.

Second, comparison of our LD curves with published resequencing data is problematic due to the low coverage of existing village dog genomes (3). Genotyping error at this coverage depth using the Samtools pipeline is ~1% in human datasets (4); in our laboratory, we see a similar level of genotyping discordance between the Illumina CanineHD array (2) and variants called from whole-genome sequencing using the Genome Analysis Toolkit (GATK) (5). As a consequence, LD curves computed from sequencing data are lower than LD curves computed from array data, even when the same dogs are considered at the same markers (Fig. 1). Indeed, we see even lower LD in sequenced Vietnamese or Indian village dogs (5) than in the Chinese village dogs reported by Wang et al. (1) (Fig. 2), but this is likely a reflection of the lower coverage depth of the non-Chinese populations (8–10× vs. 15×) and not representative of the true geographic patterning of genetic variation.

Last, Wang et al. (1) are correct that Y haplogroup 8 (HG8) was previously observed by Natanaelsson et al. (6) in a single Southeast Asian sample (a Thai Ridgeback). Genome-wide studies of Thai Ridgebacks are lacking, so it is unclear whether this HG8 haplogroup is likely present in indigenous Southeast Asian village dog populations or was brought to Southeast Asia by the common ancestor of Thai and African Ridgeback dogs. Even accepting the presence of HG8 in Southeast Asia, African village dogs still contain more Y haplogroups than Southeast Asian dogs because HG9 is present in Africa (and Central Asia and Siberia) but not Southeast Asia.

We appreciate the reanalysis and incorporation of new data by Wang et al. (1), and agree that consistent genetic data generated from additional samples from Central Asia (including western China) and neighboring regions are needed for better characterization of patterns of genetic diversity in Asian dogs. However, without stringent controls for the biases created by errors in low- and intermediate-coverage sequencing, comparisons of population genetic statistics like LD between array-based and sequence-based data can lead to spurious conclusions and should be interpreted with great caution.
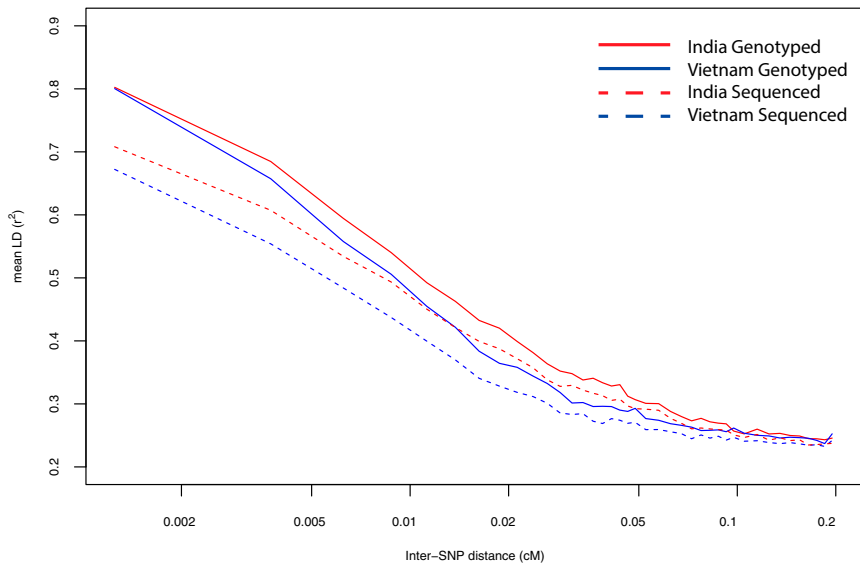
Fig. 1. Decay of LD curves computed from variant calls from whole-genome sequencing [Auton et al. (5), dashed lines] and from Illumina array genotypes [Shannon et al. (2), solid lines]. For each population, the same six dogs were used for sequencing and genotyping, and curves are computed based on the 156,000 autosomal markers passing quality control in both datasets. Thus, differences in LD decay estimates for the same population are due solely to genotyping discordances between the sequenced and genotyped data.
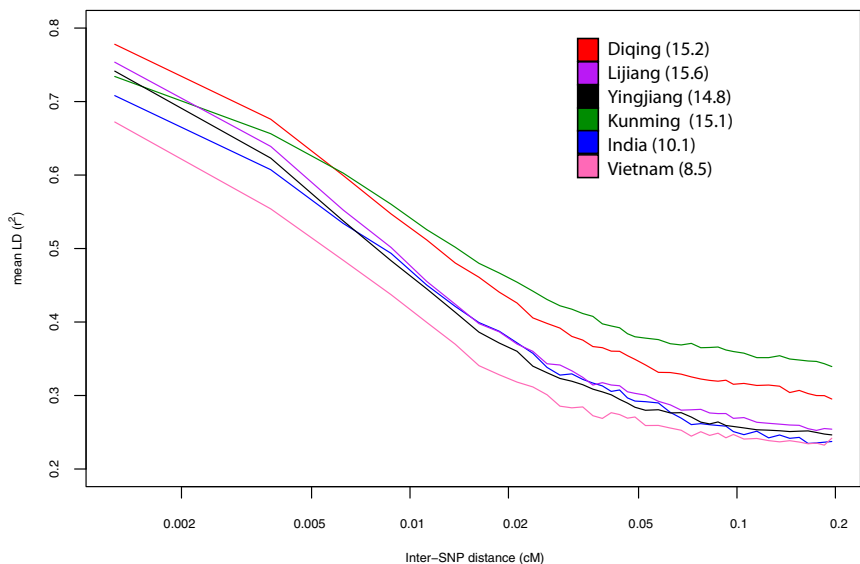

Fig. 2. Decay of LD in village dog populations based on six resequenced samples based on variant calls only at positions on the Illumina array. Mean depth of sequencing coverage is given in parentheses.

1 Wang G-D, Peng M-S, Yang H-C, Savolainen P, Zhang Y-P (2016) Questioning the evidence for a Central Asian domestication origin of dogs. *Proc Natl Acad Sci USA* 113:E2554–E2555.

2 Shannon LM, et al. (2015) Genetic structure in village dogs reveals a Central Asian domestication origin. *Proc Natl Acad Sci USA* 112(44):13639–13644.

3 Gou X, et al. (2014) Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res* 24(8): 1308–1315.

4 Cheng AY, Teo YY, Ong RT (2014) Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* 30(12):1707–1713.

5 Auton A, et al. (2013) Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet* 9(12):e1003984.

6 Natanaelsson C, et al. (2006) Dog Y chromosomal DNA sequence: identification, sequencing and SNP discovery. *BMC Genet* 7:45.