

1 Selective sweep analysis using village dogs highlights the
2 pivotal role of the neural crest in dog domestication

3

4 Amanda L. Pendleton^{1†}, Feichen Shen^{1†}, Angela M. Taravella¹, Sarah Emery¹, Krishna R.
5 Veeramah², Adam R. Boyko³, Jeffrey M. Kidd^{1,4}

6

7 ¹Department of Human Genetics, University of Michigan, Ann Arbor, MI, 48109 USA.

8 ²Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA.

9 ³Department of Biomedical Sciences, Cornell University, Ithaca, New York, 14853 USA.

10 ⁴Department of Computational Medicine and Bioinformatics, University of Michigan, Ann
11 Arbor, MI 48109 USA.

12

13 † These authors contributed equally to this work.

14 Abstract

15 Dogs (*Canis lupus familiaris*) were domesticated from gray wolves between 20-40kya in
16 Eurasia, yet details surrounding the process of domestication remain unclear. The vast array of
17 phenotypes exhibited by dogs mirror numerous other domesticated animal species, a
18 phenomenon known as the Domestication Syndrome. Here, we use signatures persisting in the
19 dog genome to identify genes and pathways altered by the intensive selective pressures of
20 domestication. We identified 37 candidate domestication regions containing 17.5Mb of genome
21 sequence and 172 genes through whole-genome SNP analysis of 43 globally distributed village
22 dogs and 10 wolves. Comparisons with three ancient dog genomes indicate that these regions
23 reflect signatures of domestication rather than breed formation. Analysis of genes within these
24 regions revealed a significant enrichment of gene functions linked to neural crest cell migration,
25 differentiation and development. Genome copy number analysis identified regions of localized
26 sequence and structural diversity, and discovered additional copy number variation at the
27 amylase-2b locus. Overall, these results indicate that primary selection pressures targeted genes
28 in the neural crest as well as components of the minor spliceosome, rather than genes involved in
29 starch metabolism. Smaller jaw sizes, hairlessness, floppy ears, tameness, and diminished
30 craniofacial development distinguish wolves from domesticated dogs, phenotypes of the
31 Domestication Syndrome that can result from decreased neural crest cells at these sites. We
32 propose that initial selection acted on key genes in the neural crest and minor splicing pathways
33 during early dog domestication, giving rise to the phenotypes of modern dogs.

34

35

36 Keywords: domestication, canine, neural crest, selective sweep

37 **Background**

38 Spanning thousands of years, the process of animal domestication by humans was complex, and
39 multi-staged, resulting in disparate appearances and behaviors of domesticates relative to their
40 wild ancestors [1–3]. In 1868, Darwin noted that numerous traits are shared among domesticated
41 animal species, an observation that has since been classified as the “Domestication Syndrome”
42 (DS) [4]. DS is a phenomenon where diverse phenotypes are shared among phylogenetically
43 distinct domesticated species but absent in their wild ancestors. Such traits include increased
44 tameness, shorter muzzles/snouts, smaller teeth, more frequent estrous cycles, floppy ears,
45 reduced brain size, depigmentation of skin or fur, and loss of hair.

46

47 Due to the selection in favor of specific desired traits by humans during the domestication
48 process, genetic signatures exist that distinguish domesticated and wild animals of the same
49 species, such as alterations in allele frequencies [5–10], signals of relaxed and/or positive
50 selection [11–13], and linkage disequilibrium patterns [14,15]. Though numerous genome
51 selection scans have been performed within a variety of domesticated animal taxa [5–10], no
52 single “domestication gene” either shared across domesticates or unique to a single species has
53 been identified [16,17]. However, this is not unexpected given the diverse behavioral and
54 complex physical traits that fall under DS. Rather, numerous genes with pleiotropic effects likely
55 contribute to DS traits through mechanisms which act early in organismal development [16,17].
56 One potential explanation was presented in Wilkins et al. 2014, that highlighted the parallel
57 between DS phenotypes (e.g. craniofacial abnormalities, ear malformations, etc.) and those
58 exhibited in neurocristopathies, disorders caused by aberrant development of tissues that derive
59 from the embryonic neural crest. This hypothesis states that deficits during the development,

60 proliferation, and migration of neural crest cells (NCCs) can explain the phenotypic patterns
61 shared across domesticated animals [16].

62

63 To further explore the genetic underpinnings of DS phenotypes, we have searched for genomic
64 signatures of early domestication in the domestic dog (*Canis lupus familiaris*), a species that
65 represents the first known domesticated animal and that also starkly contrasts with its ancient
66 ancestor the gray wolf (*Canis lupus*) in numerous DS traits. Putative signals of domestication
67 have been identified through comparisons of dog and wolf genomes, with breed dogs either fully
68 or partially representing dog genetic diversity in these studies [5,8,18,19]. However, most
69 modern breed dogs arose ~300 years ago [20] and thus contain only a small portion of the
70 genetic diversity found among the vast majority of extant dogs. Instead, semi-feral village dogs
71 are the most abundant and genetically diverse modern dog group, and have undergone limited
72 targeted selection by humans since initial domestication [21,22]. These two dog groups represent
73 products of the two severe bottlenecks in the evolution of the domestic dog, the first resulting
74 from the initial domestication of gray wolves, and the second from modern breed formation
75 [23,24]. To distinguish signatures of early domestication from breed formation we have limited
76 our analyses to genomic comparisons between village dogs and wolves. Village dogs are more
77 genetically similar to ancient dogs [25,26], and have been used to discriminate between breed
78 and domestication sweeps. We recently showed [26] that modern village dogs do not exhibit
79 selective sweep haplotypes at 18 of 30 swept loci previously identified using breed dogs [5].
80 Furthermore, swept haplotypes were not found at many previously identified loci in two ancient
81 Neolithic German dog genomes, suggesting that rather than candidate domestication regions

82 (CDRs), Axelsson et al. 2013 largely identified regions associated with breed formation or other
83 post-domestication selection [26].

84

85 In conjunction with traditional SNP-based selection scans based on extreme differences in allele
86 frequency, we implement copy number (CN) scans to ascertain regions of the dog genome that
87 have undergone intensive selection during domestication. To accomplish this, we survey
88 differences in variants from a diverse panel of globally-distributed village dogs compared to gray
89 wolves, and subsequently assess whether the swept alleles are present in the genomes of three
90 ancient European dogs that are at least ~5,000 years old. With this approach, we identify CDRs
91 that exhibit unusual divergence between village dogs and wolves at both the single nucleotide
92 and copy-number level. Gene annotations and enrichment results indicate that numerous CDRs
93 harbor genes integral in the initiation, differentiation, and migration of neural crest cells (NCCs)
94 including the Wnt, FGF, and TGF- β pathways that are integral to the neural crest (NC).
95 Additionally, we have isolated selective sweeps that contain three of the seven subunits of the
96 minor (U11/U12) spliceosome, highlighting a possible role of alterations in minor splicing in
97 establishing the domestic dog phenotype. More specifically, patterns of allele divergence suggest
98 that ancient selective pressures may have first targeted the 65KDa minor spliceosomal subunit
99 gene, *RNPC3*, which is positioned adjacent to the more recently duplicated [26–28] starch
100 metabolizing gene, amylase-2b (*AMY2B*). Altogether, we argue that the dog domestication
101 process altered the activity of genes associated with the embryological NC pathway and minor
102 splicing, contributing to the array of DS phenotypes found in the modern dog including reduced
103 aggression, craniofacial alterations, floppy ears, and decreased tooth sizes.

104

105 **Results**

106 **Sample selection and sequence variant identification**

107 We used ADMIXTURE and identity-by-state (IBS) analysis to identify a collection of 43 village
108 dog and 10 gray wolf whole genomes that excludes closely related samples and that shows less
109 than 5% admixed ancestry (Additional File 1: Note 1). Principal component analysis illustrates
110 the genetic separation between village dogs and wolves and largely reflects the geographic
111 distribution of the wolf and village dog populations (Figure 1B and 1C). We created two SNP
112 call sets to identify regions with unusually large allele frequency differences between village
113 dogs and wolves. First, we identified a total of 7,315,882 SNPs (including 53,164 on the non-
114 PAR region of the X chromosome) that are variable among the 53 analyzed samples. This total
115 SNP call set represents the densest set of identified variation and underlies the bulk of the
116 subsequent analysis. Since differences in sequencing coverage and uneven sample sizes may lead
117 to biased variant discovery, we additionally created a second SNP call set of 2,761,165 SNPs
118 (including 17,851 on the non-PAR region of the X chromosome) limited to sites ascertained as
119 being variable among three New World wolf samples [26]. Subsequent analyses utilizing this
120 NWW-ascertained set were based on only the 5 Eurasian wolves and 43 village dogs.

121

122 **F_{ST} scans identify regions of differentiation between village dogs and wolves**

123 We performed several scans to localize genomic intervals with unusual levels of allele frequency
124 differentiation between the village dog and wolf populations. Following previous studies [5,8],
125 we first calculated average F_{ST} values in disjoint 200kb windows along the genome and Z-
126 transformed the resulting mean F_{ST} estimates. Due to differences in the effective population size
127 and corresponding expected levels of genetic drift, Z-transformations were performed separately

128 for the non-PAR region of the X chromosome. Setting Z-score cutoffs of 5 for the autosomes and
129 3 for the X, we identified 24 regions of extreme genetic differentiation, spanning a total of 7.8
130 Mb of sequence. Since this approach may fail to detect differentiated regions near window
131 boundaries, we additionally performed an F_{ST} scan with greater resolution, utilizing 200 kb
132 windows that slid along the genome in 50 kb increments. As expected, the sliding window
133 approach identified a greater number of candidate regions: 30 regions encompassing 12.3 Mb of
134 sequence. A 350kb region on chromosome 6 (chr6: 46,800,001-47,350,000) that contains the
135 *pancreatic amylase 2B (AMY2B)* and *RNA Binding Region Containing 3 (RNPC3)* genes had the
136 highest observed ZF_{ST} score using the total SNP call set ($ZF_{ST} = 9.97$). We completed analogous
137 scans using the NWW SNP set, identifying 26 outlier regions (6.6 Mb) using non-overlapping
138 windows and 34 regions (12.1 Mb) using sliding windows. A region on chromosome 16 that
139 contains numerous genes including *Taste Receptor 2 Member 38 (TAS2R38)*, *Maltase-*
140 *Glucoamylase 1 and 2 (MGAM and MGAM2)*, and a serine protease (*PRSS37*) is the most
141 significant region identified using the NWW call set ($ZF_{ST} = 7.53$). Interestingly, the *AMY2B*
142 region only achieved the thirteenth highest ZF_{ST} score within the NWW SNP set.

143

144 **Identifying 37 candidate domestication regions**

145 The union of the outlier regions from the total and NWW SNP sets yields 37 regions showing
146 unusual levels of allele frequency differences among village dogs and wolves, 19 of which were
147 outliers from analysis of both SNP sets (Table 1; Additional File 2: Table S1). Of these 37
148 candidate domestication regions (CDRs) only 17 intersect with previously reported dog CDRs
149 (Figure 2A). Within our sample and SNP set, we assessed whether the dog or wolf haplotype is
150 present at the 36 and 18 domestication sweeps reported in [5] and [8], respectively, in 46

151 additional canine samples, including three ancient dogs ranging in age from 5,000-7000 years
152 old (see Methods; [25,26]). Likely due to the absence of sampled village dogs in their study,
153 some Axelsson CDRs [5] appear to contain selective sweeps associated with breed formation, as
154 evidenced by the lack of ancient and village dogs with the breed (reference) haplotype (example
155 in Figure 2B). Although all autosomal sweeps identified by [8] intersected with CDRs from our
156 study, six X chromosome Cagan and Blass windows did not meet the thresholds of significance
157 from our SNP sets (example in Figure 2C). The authors of this study performed F_{ST} scans and Z
158 transformations for windows on autosomes and X chromosome together, which may falsely
159 inflate F_{ST} values on the X. We additionally compared the SNPs identified in [29] that exhibit
160 diversifying selection signals shared between modern village dogs (or free-breeding dogs) and
161 breed dogs with the locations of our CDRs. None of these sites were located within any of our
162 CDR boundaries, suggesting that our approach may have identified sites resulting from early
163 selection pressures of domestication rather than from subsequent selection events.

164

165 The genotypes of ancient dogs at these CDRs can aid in determining the age of these putative
166 selective sweeps, thus discriminating domestication loci from breed sweeps. Visual genotype
167 matrices (Additional File 3: Dataset 1) and non-reference (or “wild”) allele proportions
168 (Additional File 4: Dataset 2) identified thirteen selective sweeps where three ancient dogs were
169 outliers, or did not share modern village dog haplotypes (see Methods). More specifically, HXH
170 (~7,000 year old German dog) and NGD (~5,000 year old Irish dog) were outliers at five CDRs
171 each, while a known wolf-admixed sample, CTC (~5,000 year old German dog [26]), was wild-
172 like at nine windows. Interestingly, all three dogs are homozygous for the wild haplotype at only
173 one putatively swept region (CDR15), which contains only a single gene, *Transmembrane*

174 *Protein 131 (TMEM131)*. A human homolog of the gene, *TMEM131L*, is a regulator of the
175 differentiation and proliferation of thymocytes and is an antagonist of the *Wnt* signalling
176 pathway[30].

177
178 To pinpoint functional sequence variants that may be driving the patterns at each CDR we
179 calculated per-SNP F_{ST} values and annotated SNP effects on coding sequence using Variant
180 Effect Predictor (VEP; [31]) and SNPEff [32]. Although we did not observe any sites with
181 significant ZF_{ST} scores within the CDRs that confer nonsense or frameshift mutations, we
182 identified four missense variants with high F_{ST} values including a V/I change in *MGAM*
183 (chr16:7156695; $ZF_{ST} = 6.18$), a T/A mutation in *CCNB3* (chrX:42982379; $ZF_{ST} = 3.46$), a K/E
184 mutation in ENSCAFG00000016002 (chrX:43018605; $ZF_{ST} = 3.46$), and a H/R mutation in
185 *SNX19* (chr5:4046840; $ZF_{ST} = 6.72$). Given these limited results, we further searched for
186 nonsense mutations at sites with $F_{ST} > 0.35$ (similar to [17]). No nonsense variants were identified
187 from this SNP set, but three SNPs with elevated F_{ST} values were annotated as conferring a stop
188 gain. These include a site in ENSCAFG00000024996 (chr25:10257734; $F_{ST} = 0.684$; $ZF_{ST} = 4.65$),
189 *SPICE1* (chr33:17770461; $F_{ST} = 0.482$; $ZF_{ST} = 3.10$), and *TRIM38* (chr35:23931336; $F_{ST} = 0.406$;
190 $ZF_{ST} = 2.52$).

191

192 **Pathway enrichment analysis highlights morphological development and the neural crest**

193 We searched for gene functions that are overrepresented in the 172 genes found in the 37 CDRs
194 using the BLAST2GO [33] and topGO ([34]) pipelines. A subset of the 41 categories with $P <$
195 0.05 using the Parent-Child (Fisher's) enrichment test in topGO are listed in Table 2. Analysis
196 using alternative enrichment tests (e.g. Classic Fisher's) yields similar results (see Additional
197 File 1: Note 3 and Additional File 5: Table S2). Though many of these categories detail
198 enrichments of pathways or functions at the subcellular level (e.g. secretion and establishment of
199 polarity), additional top enriched pathways include those linked to skeletal and craniofacial
200 development, brain and nervous system function, and metabolism.

201
202 Bone development and ossification is a strongly overrepresented process (GO:0001649; $p =$
203 0.004), with six of the 207 genes in the dog genome ascribed to this GO category located in a
204 CDR. Additional multi-gene enriched categories involved in skeletal development include
205 osteoblast proliferation (GO:0033687; $p = 0.001$), palate development (GO:0060021; $p = 0.018$),
206 as well as the differentiation (GO:002062; $p = 0.03$) and development regulation (GO:0061181;
207 $p = 0.032$) of chondrocytes (cells required for cartilaginous tissues such as the external ear) were
208 also enriched. Altogether, genes associated with these five GO categories are located within nine
209 unique CDRs and include key genes involved in embryonic cranial development including *Axin*
210 *2* [35], *Protein Kinase C Alpha (PRKCA)*; [36], and both members (*WNT9B* and *WNT3*; [37])
211 and regulators (*WFKINNI*; [38]) of neural crest associated signaling pathways.

212
213 Determination of polarity and both the anteroposterior and dorsoventral axes is critical for proper
214 neural tube and crest development, as well as supplying positional identity to neural crest cells

215 along these axes [39,40]. We observe an overrepresentation of CDR genes that establish cellular
216 polarity and symmetry (Table 2), belonging to GO categories including determination of
217 dorsal/ventral axis specification (GO:0009950; $p = 0.037$), establishment of planar polarity
218 involved in nephron morphogenesis (GO:0072046; $p = 0.011$), and determination of left/right
219 pancreatic asymmetry (GO:0035469; $p = 0.045$). A gene with wide-reaching associations is
220 *Smoothened* (*SMO*; CDR19), which has roles in determination of the ventral midline
221 (GO:0007371; $p = 0.016$) and specification of anterior/posterior patterns (GO:0009952; Classic
222 Fisher's $p = 0.021$), a multi-gene GO category with four other genes (Table 2). *SMO* also
223 regulates the hedgehog transcription factor pathway (GO:0007228; $p = 0.025$), and the transition
224 of epithelial cells to mesenchymal cells in metanephric renal vesicle formation (GO:0072285;
225 Classic Fisher's $p = 0.008$), a complex process closely linked to apical/basolateral polarity [41].
226 Finally, *SMO* is also a negative regulator of hair follicle development (GO:0051799; $p = 0.035$)
227 and osteoblast differentiation.

228
229 We also observe an overrepresentation of genes involved in brain and nervous system
230 development within the CDRs. *Fibroblast growth factor 13* (*FGF13R*; CD37) is linked to the
231 establishment of polarity in neuroblasts (GO:0045200; $p = 0.032$), while *SMO* (CDR19), *TLN1*
232 (CDR16), and *PRKCA* (CDR13) are linked to the activation of astrocytes (GO:0048143; Classic
233 Fisher's $P = 0.024$), which, as glial cells, are derivatives of the neural crest [42]. *WNT3* (CDR11)
234 has been linked to axonogenesis, as it positively regulates collateral sprouting in absence of
235 injury (GO:0048697; Classic Fisher's $P = 0.016$). Finally, three CDR genes are involved in the
236 development of the substantia nigra in the midbrain (GO:0021762; Classic Fisher's $P = 0.017$),
237 which include *proteolipid protein 1* (*PLP1*; CDR34), *ATP5F1* (an ATP synthase; CDR7), and

238 *myelin basic protein (MBP; CDR1)*, a protein that is the major constituent of the myelin sheath
239 of Schwann cells and oligodendrocytes.

240

241 Congruent with previous work (Axelsson et al. 2013), we found a significant enrichment of
242 genes relating to metabolism, though no overrepresentation of categories related to digestion was
243 observed. Two genes with roles in one-carbon metabolism (GO:0006730; $p = 0.034$),
244 *Adenosylhomocysteinase Like 2 (AHCYL2)* and *Carbonic Anhydrase 9 (CA9)* are located in
245 CDRs 19 and 16, respectively. Also, *PLP (CDR34)* is involved in long-chain fatty acid
246 biosynthesis (GO:0042759; Fisher's Classic $p = 0.04$). The region harboring *MGAM (CDR20)*,
247 which encodes an enzyme responsible for the second step of starch metabolism, is a previously
248 identified sweep region [5,8]. *MGAM* belongs to many enriched categories including starch
249 catabolic process (GO:0005983; $p = 0.042$), polysaccharide catabolic process (GO:000272; $p =$
250 0.028), and maltose metabolism (GO:0000023; Classic Fisher's $p = 0.008$), and is situated within
251 a region of elevated per SNP F_{st} that distally extends over two additional genes in this locus, the
252 bitter taste receptor *TAS2R38* and *CLEC5A (C-type lectin domain family 5 member A)* (Figure
253 3A). Closer analysis of the CDR20 region highlights the challenge in identifying the underlying
254 targets of selection from variation data. Although ZF_{st} scores sharply decrease just downstream
255 of *MGAM*, near its directly adjacent paralog, *MGAM2*, a pattern of highly differentiated SNPs
256 extends upstream across multiple genes including *CLEC5A*, which belongs to multiple highly
257 enriched ontologies representing broadly acting processes that may contribute to the
258 domestication syndrome including osteoblast development (GO:0002076), negative regulation of
259 myeloid cell apoptotic process (GO:0033033), and positive regulation of cytokine secretion
260 (GO:0050715). At another starch digestion gene locus, a strong signal of SNP differentiation is

261 observed at the *AMY2B* region on chromosome 6, overlapping a previously reported expansion
262 of the *AMY2B* gene found in modern dogs [5,43]. As in CDR20, however, interpretation of
263 underlying causal genes at this region is made more complicated by the observation that ancient
264 dog samples contain the selected haplotype without the amylase expansion (Additional File 6:
265 Figure S1; [26]), and that the region of high F_{st} extends >4.5 Mb distally, thereby encompassing
266 the adjacent *RNPC3* (Figure 3B), the 65 KDa subunit of the U11/U12 minor spliceosome, as is
267 visible in the genotype matrix for the region (Additional File 6: Figure S1A).

268

269 **Selection scans for CN variation between dogs and wolves**

270 Copy-number variants have also been associated with population-specific selection and
271 domestication in a number of species [5,44,45]. Since regions showing extensive copy-number
272 variation may not be uniquely localized in the genome reference and may have a deficit of SNPs
273 passing our coverage thresholds, we directly estimated genome copy number along the reference
274 assembly and searched for regions of unusual copy number differences using two approaches:
275 fastCN, a method based on depth of sequencing reads which tolerates mismatches and combines
276 copy-number estimates across related paralogs, and QuicK-mer, a mapping-free approach that
277 resolves paralog-specific copy-number estimates.

278

279 Both fastCN and QuicK-mer estimation pipelines were applied to the 53 village dog and wolf
280 samples analyzed in the F_{st} analysis. We assessed the quality of our CN estimations by
281 calculating the signal-to-noise ratio (SNR) in genomic intervals that are not CNV, finding a
282 higher SNR in wolf samples that have higher sequencing depth. Importantly, many village dogs
283 with considerably lower depth (~4-10x) still achieve similar SNRs to these higher coverage

284 wolves (Additional File 6: Figure S2). Next, we validated the accuracy of our CN estimates
285 through comparison with probe intensity data from a previous Comparative Genome
286 Hybridization Array (aCGH) study that analyzed 23 wolf-like canids [46], seven of which have
287 genome sequence data analyzed in our F_{ST} sample set. After accounting for the noisy sequence
288 data from the sample used as the reference in the aCGH experiments (Additional File 1: Note
289 6.3), we find strong correlation between *in silico* aCGH and actual probe data, with R^2 values
290 ranging from 0.43-0.84 and 0.34-0.70 for fastCN and QuicK-mer (Additional File 6: Figure S3;
291 Additional File 7: Dataset 3), and mean R^2 values at 0.71 and 0.55, respectively. The higher
292 correlation with fastCN copy-number likely reflects this method's tolerance of mismatches,
293 which is analogous to aCGH probe hybridization. The accuracy of our CN estimates is further
294 supported by targeted validation at the *AMY2B* locus described below.

295
296 V_{ST} scans, a modified F_{ST} approach that calculates deviations of copy-number between populations
297 [44], were completed using CN estimates within 380,731 and 642,203 windows from fastCN and
298 QuicK-mer, respectively, for the 53 canine set (Additional File 1: Note 7). Following Z
299 transformation, 597 fastCN windows obtained significant ZV_{ST} values ($ZV_{ST} > 5$ on autosomes
300 and chrX-PAR, $ZV_{ST} > 3$ on chrX-NonPAR) while 250 QuicK-mer windows were significant
301 outliers (Additional File 1: Note 7.3). The outlier regions from both pipelines were merged based
302 on overlapping coordinates using methods analogous to F_{ST} CDRs to generate 202 outlier regions
303 (Additional File 8: Table S3). Due to low genome coverage of some dog samples, confident
304 detection of small CNVs using read depth is difficult. Therefore, the 202 outlier regions were
305 further filtered to require at least two adjacent CN windows from either fastCN or QuicK-mer, or
306 combined, resulting in 67 filtered VCDRs (Additional File 9: Table S4) containing 39 genes.

307 Once this filtration was applied, VCDRs were either supported by both pipelines (N=35) or
308 fastCN only (N=32), but no region was identified by QuicK-mer alone. Of these filtered VCDRs,
309 five intersected with a F_{st} CDR including VCDR20 which is contained within F_{st} CDR (CDR8), a
310 previously published sweep locus [5,8], which encompasses the well-documented copy number
311 variable *AMY2B* locus. VCDRs 27, 28 and 31 intersect with two F_{st} CDRs 10 and 11 on
312 chromosome 9, while VCDR 126 and CDR 32 co-localize on chromosome X. Additionally,
313 previously identified SNPs under putative diversifying selection shared between breed dogs and
314 village dogs [29] were not located in any VCDR. Finally, likely due to the sparsity of genes in
315 the VCDRs, no prominent gene enrichment patterns were observed (Additional File 10: Table
316 S5).

317

318 **Characterization of a large-scale structural variant located at the *AMY2B* locus**

319 Outliers from both the F_{st} (chr6: 46800001-47350000) and V_{st} (chr6: 46945638-46957719)
320 selection scans encompass the *AMY2B* gene, which at increased CN confers greater starch
321 metabolism efficiency due to higher pancreatic amylase enzyme levels [5,43]. As stated
322 previously, examination of per SNP F_{st} results suggest that *RNPC3* may be an additional (or
323 alternative) target of selection at this locus (Figure 3B). Furthermore, read-depth CN analysis has
324 illustrated that the *AMY2B* tandem expansion was absent in three analyzed ancient dog samples
325 [26]. Instead, large-scale segmental duplication at the locus accounts for some detected *AMY2B*
326 CN increases [26].

327

328 Using methods implemented previously [26], we recapitulated the elevated CN estimates across
329 the locus for some samples (including the ancient Newgrange dog; Additional File 1: Figure S4).

330 Fine-scale CN estimates reveal the presence of two duplications with unique breakpoints (Figure
331 4C). Read-depth based CN patterns show a proximal extension of ~55kb (Figure 4A) and a distal
332 extension of ~20kb (Figure 4B) that differentiate the two large duplications (designated the 1.9
333 Mb and 2.0 Mb duplications based on approximate lengths).

334

335 We utilized droplet digital PCR (ddPCR) to survey CN across this region in 90 dogs to determine
336 the validity of read-depth based copy number estimations and ascertain whether the large-scale
337 duplications can account for observed differences in *AMY2B* copy-number. Primers were
338 designed to target a region unique to the proximal end of the 2.0 Mb duplication, within the 1.9
339 Mb duplication and shared with the 2.0 Mb duplication, and targeting the *AMY2B* gene (see
340 Figure 5). The ddPCR results (Additional File 11: Table S6 and Additional File 12: Table S7)
341 were found to be strongly correlated with the *AMY2B* (Additional File 6: Figure S5A) and large-
342 scale duplication (Additional File 6: Figure S5B) CN estimates for 42 of the 90 ddPCR-sampled
343 dogs for which we had sequence data. The ddPCR results confirm the standing variation of
344 *AMY2B* copy-number across dogs, and distinguish the two large-scale duplications that
345 encompass this location (Additional File 6: Figure S6). Per ddPCR, total *AMY2B* copy number
346 ranged from 2-18 copies per dog (average $2n_{AMY2B} = 11$), relative to a diploid control region on
347 chromosome 18 (see Methods). Furthermore, the *AMY2B* CN expansion appears to be
348 independent of the large-scale duplications, as ddPCR results show that some dogs without the
349 large duplications still maintain very high *AMY2B* CN. Both ancient dog samples from Germany
350 (5 and 7kya) did not indicate CN increases for either the duplication or the tandem *AMY2B*
351 expansions [26]. However, based on read-depth patterns at the duplication breakpoints, the

352 Newgrange Irish dog (5kya; [25]) harbored the 2.0 Mb duplication while $2n_{AMY2B} = 3$, indicating no
353 tandem *AMY2B* expansion, but rather CN increases due to the presence of the large duplication.

354

355 **A region of extreme CNV on chromosome 9 co-localizes with F_{st} selective sweeps**

356 Co-localization analysis indicated a clustering of VCDR and CDR windows within the first
357 25Mb of chromosome 9. Upon closer inspection of the copy-number and F_{st} data at this region,
358 we observed anomalous patterns not found elsewhere in the genome for our datasets. Average
359 CN values from fastCN and QuicK-mer both indicate significantly higher CN in wolves within
360 19 VCDR windows here (Additional File 9: Table S4). Notably, boundaries of the VCDRs are
361 directly adjacent to regions undergoing significant allele frequency differentiation, as highlighted
362 in per site ZF_{st} peaks in Figure 5A. Such a pattern of extended divergence is reminiscent of
363 inverted haplotypes which have been characterized in several species [47–49]. To further
364 characterize this locus, we identified candidate inversions in the dogs and wolves separately
365 using *inveR*sion [50] which relies on SNP genotypes to locally phase alleles and determine
366 haplotype blocks for inversion breakpoint estimations (see Methods). Although no inversions
367 were detected in the wolves on chromosome 9, five potential inversions were identified in village
368 dogs clustered within this region of interest. Interestingly, predicted breakpoints of two
369 inversions are situated at the transition point between the elevated F_{st} region (chr9: ~9.0-16.7
370 Mb) and major copy-number peaks. Correlations between copy number states of VCDRs (per
371 fastCN and QuicK-mer) and SNP genotypes of the 53 samples on chromosome 9 indicates two
372 loci 8 Mb apart in the reference genome share elevated R^2 value, patterns consistent with genome
373 rearrangements at this region (Figure 5B; Additional File 13: Dataset 4 and Additional File 14:

374 Dataset 5). More specifically, the copy number states of VCDR 31 and VCDR 48 (Additional
375 File 6: Figure S7) share similar correlation even though the location are separated by 8 Mb.

376 **Discussion**

377 Genetic and archaeological data indicate that the dog was first domesticated from Eurasian gray
378 wolves by hunter-gatherers between 20-40 kya [25,26,51,52]. Though the exact events
379 associated with domestication will likely never be unearthed, evidence suggests that the process
380 was complex, and may have spanned thousands of years [3,25]. Through genome analysis of
381 unrelated modern village dog and wolf samples with low dog-wolf admixture, we have identified
382 37 CDRs and 67 VCDRs that are strongly deviated between dogs and wolves in allele frequency
383 and estimated CN. Our sampling strategy, which utilized only village dogs, and comparisons
384 with ancient dog genome data indicate that these selection events likely reflect early
385 domestication selection pressures.

386

387 **F_{st} interpretation**

388 Most regions detected from sequence selection scans (CDRs) highlighted in this study are
389 unique, and no earlier works have identified genomic loci exhibiting copy-number selection
390 signals (VCDRs) in dogs with a comparable method of analysis. Though 43% of our CDRs
391 corresponded with at least one previously determined swept region [5,8], our overall gene
392 enrichment patterns are not consistent with these earlier studies, and we emphasize this is most
393 likely due to differences in the studied samples. Previous selection scans have either solely relied
394 on comparisons of breed dogs to gray wolves [5,19], or a mixture of village and breed dogs
395 [8,18]. By avoiding spurious signals resulting from breed formation, we argue that our sweeps

396 identified in modern village dogs are more likely signals of ancient selection events that arose
397 thousands of years ago, and we emphasize the strength of sampling village dogs with high
398 genetic diversity from broad geographic distributions in order to avoid capturing local selection
399 events present in populations from limited geographic locations.

400

401 A further strength in our methodology was the analysis of sites under two distinct ascertainment
402 schemes. More specifically, variants from the NWW SNP set were either present in the ancestral
403 population of New and Old World wolves or are private to modern New World wolves, which
404 permitted SNP analysis at sites with an unbiased ascertainment with regard to Eurasian dog and
405 wolf populations [26]. For this reason, detection of significant windows from analysis of both
406 total and NWW SNP sets is likely robust to any potential ascertainment bias.

407

408 However, we acknowledge limitations in the experimental design of this study, and attribute
409 some uncertainty in our inference as a result of the current dog reference genome, as well as
410 sample size and availability. The current dog assembly (CanFam3.1) remains substantially
411 incomplete (>23,000 gaps), which can impair accurate copy-number estimation and lead to the
412 misidentification of selection signals in duplicated regions. A higher quality reference genome
413 may overcome such limitations in selection scans, allowing direct examination of variation
414 among gene paralogs.

415

416 Altogether, our methods defined 37 distinct genomic regions (CDRs) containing 172 genes that
417 display significant sequence deviation in dogs relative to wolves (Table 1), of which 54 genes
418 belong to significantly enriched gene ontology categories including developmental, metabolism,

419 and reproductive pathways (Table 2). It is important to note that the genomic position of most
420 enriched CDR genes are not co-localized, which implies that the observed enrichment is not an
421 artifact of a selective sweep containing a cluster of paralogous genes. Rather, we observe
422 enrichment of genes belonging to the same pathway (e.g. skeletal development, polarity
423 determination, or brain development) that localize to distinct regions of differentiated allele
424 frequencies. Such distributions suggest that parallel selective pressures occurred at numerous
425 genomic positions for genes involved in a shared biological function, reflecting selective sweeps
426 that likely arose from differential pressures that have acted on village dogs and wolves, possibly
427 occurring as early as initial domestication. Additionally, similar to selection scans in chickens
428 [6], pigs [9], and rabbits [17], the scarcity of significantly deviating protein sequence altering
429 SNPs in the CDRs indicates that gene loss did not have a significant role in the domestication of
430 dogs, possibly implicating alterations in gene regulation instead during domestication, as has also
431 been hypothesized for artificial selection during breed formation [53].

432

433 **V_{st} interpretation**

434 In conjunction with surveys of SNP deviation, V_{st} selection scans were used to identify 202
435 windows with significant copy-number deviation (VCDRs) between dogs and wolves, and the
436 resulting discrepancy in the number of windows discovered by fastCN and QuicK-mer is
437 intriguing. Although QuicK-mer interrogates a higher percentage of the genome in terms of the
438 number of base pairs, this pipeline discovers fewer VCDRs. A possible source of this disparity is
439 the incomplete nature of the current canine assembly. During QuicK-mer's 30-mer generation
440 process, all chromosomes are considered, including the 3,329 unplaced contigs, some of which
441 may be redundant. Thus, any regions that appear duplicated due to assembly error will lose 30-

442 mer coverage. Secondly, QuicK-mer is sensitive to single base pair changes due to the usage of
443 unique k-mers (Additional File 1: Note 5). Without the resource of a wolf genome assembly,
444 generation of unique 30-mers based on the genome of an inbred dog decreases the likelihood for
445 discovering regions with divergence in wolves. However, QuicK-mer complements the fastCN
446 pipeline by only considering unique regions. For example, QuicK-mer clearly distinguishes two
447 regions of major structural variation on chromosome 9 that show related copy number states,
448 while fastCN suggests more complicated substructures (Additional File 6: Figure S8).

449

450 **The role of the neural crest in dog domestication**

451 The collective traits displayed by dogs and other domestic animals that constitute the
452 domestication syndrome (DS) are diverse, can be manifested in vastly different anatomical
453 zones, and appear seemingly disconnected. However, regardless of the sampling or
454 methodologies implemented, no survey of selection sweeps in the dog genome have isolated a
455 singular genomic region as the sole contributor of the complex domestic dog phenotypes that can
456 explain the DS [5,8,18,19]. Instead, the results of these previous studies mirror those presented
457 here, with numerous swept regions distributed about the genome, possibly arising from selection
458 that occurred independently at multiple loci, which gave rise to each of the observed DS traits.
459 Alternatively, selection could have acted on considerably fewer genes that are members of an
460 early-acting developmental pathway.

461

462 For these reasons, the pivotal role of the neural crest cells (NCCs) in animal domestication has
463 gained support from researchers over recent years [16,54]. In 2014, Wilkins et al. (2014)
464 established that the vast array of phenotypes displayed in the animal DS mirror those exhibited in

465 mild human neurocristopathies, whose pathology stems from aberrant differentiation, division,
466 survival, and altered migration of NCCs [55]. NCCs are multipotent, transient, embryonic stem
467 cells that are initially located at the crest (or dorsal border) of the neural tube. Following
468 induction and transition from epithelial to mesenchymal cell types, NCCs will migrate along
469 defined pathways to various sites in the developing embryo. Tissues that derive from NCCs
470 include most of the skull, the adrenal medulla, sympathetic ganglia, odontoblasts (tooth
471 precursors), pigmentation-associated melanocytes, bone, cartilage, and many others [16,54].

472

473 Support for the role of the NCCs in the domestication process has been strengthened by results of
474 silver fox breeding experiments. Tameness or reduced fear toward humans was likely the earliest
475 trait selected for by humans during domestication [3,56,57]. Mirroring this, when researchers
476 only selected for tameness in the fox breeding population, numerous physiological and
477 morphological characteristics appeared within twenty generations, including phenotypes
478 associated with DS such as floppy ears, fur pigmentation changes, altered craniofacial
479 proportions, and unseasonal timing for mating [1,58]. As the progenitor for the adrenal medulla
480 that produces hormones associated with the “fight-or-flight” response in the sympathetic nervous
481 system, hypofunction of NCCs can lead to changes in the tameness of animals [16]. Altogether,
482 the link between tameness and the NC indicates that changes in neural crest development could
483 have arisen first, either through direct selection by humans for desired behaviors or via the “self-
484 domestication” [59,60] of wolves that were more docile around humans.

485

486 The initiation and regulation of neural crest development is a multi-stage process requiring the
487 actions of many early-expressed genes including the Fibroblast Growth Factor (Fgf), Bone

488 Morphogenic Protein (Bmp), Wingless (Wnt), and *Zic* gene families [61]. Members of both the
489 *Fgf* (*Fgf13* in CDR37) and *Wnt* (*Wnt9B* and *Wnt3* in CDR11) gene families that aid in the
490 origination of the neural crest are found in our CDRs, along with *ZIC3* (CDR37) which promotes
491 the earliest stage of NC development [62]. Assignment of identity and determination of
492 migration routes for NCCs relies on positional information provided by external signaling cues
493 [39,40]. In addition to the previously described genes (*Wnt9B*, *Wnt3*, *Fgf13*, *ZIC3*), *AXIN2*
494 (CDR13) and *SMO* (CDR19) are essential for the determination of symmetry, polarity, and axis
495 specification (Table 2). Since these genes are located in putative sweeps, our results suggest that
496 early selection may have acted on genes essential to the initiation and regulation of the NC, thus
497 altering the proper development of NCC-derived tissues linked to the DS.

498
499 Ancient dog remains indicate that body size, snout lengths, and cranial proportions of dogs
500 considerably decreased compared to the wolf ancestral state following early domestication [63].
501 Further, ancient remains indicate jaw size reduction also occurred, as evidenced by tooth
502 crowding [63]. These alterations are consistent with the DS, and implicate aberrant NCC
503 migration since decreases in the number of NCCs in facial primordia is directly correlated with
504 reductions in mid-face and jaw sizes [16,64]. Overall, we observed an enrichment of CDR genes
505 associated with bone and palate development (Table 2). At critical stages of midfacial
506 development, the Wnt pathway is activated during lip fusion and facial outgrowth, and
507 expression of both *WNT3* (CDR11) and *WNT9B* (CDR11) have been detected in the developing
508 facial ectoderm, and linked to cleft palates [37]. *WNT9B* is also critical for Fgf signaling in the
509 developing nasal and maxillary processes [65]. Craniofacial malformations have also been
510 attributed to insufficient levels of *NOL11* (CDR12; [66]), while decreases in *AXIN2* (CDR13), a

511 negative regulator of Wnt signaling [67], causes craniosynostosis or premature fusion of the skull
512 [35]. We hypothesize that the altered cranial morphology of modern village dogs is linked to
513 altered activities of NCC regulators, resulting in a decline of NCCs migrating NCCs to the
514 developing skull.

515

516 Drooping or “floppy” ears are a hallmark feature of domesticates. Compared to the pricked ear
517 phenotype of wolves, village and breed dogs predominantly have floppy ears [68]. In humans,
518 insufficient cartilage in the pinna, or outer ear, results in a drooping ear [69] phenotype linked to
519 numerous NC-associated neurocristopathies (e.g. Treacher Collins, Mowat-Wilson, etc.).
520 Cartilage is a NCC-derived tissue [16] that solely consists of chondrocytes, and our enrichment
521 results highlighted three genes involved in the differentiation and regulation of chondrocyte
522 development (Table 2). This enrichment, coupled with established connections between the NC
523 and cartilage formation, provides support to the hypothesis that the floppiness of domesticated
524 dog ears arose due to reductions in the number of NCCs targeted to the developing ear and its
525 cartilage, thereby lessening support of the external ear [16].

526

527 Finally, two additional DS phenotypes, depigmentation and reduction in hair (or hairlessness),
528 are linked to pathways that regulate NC development. Excluding retinal, all pigmented cells are
529 derived from the neural crest [70], and aberrant pigmentation has been linked to
530 neurocristopathies such as von Recklinghausen neurofibromatosis, neurocutaneous melanosis,
531 Waardenburg syndrome, and albinism [42]. Though no gene directly implicated in pigmentation
532 was identified in our sweeps, some CDR genes that either belong to or interact with key NC
533 regulation families such as Wnt (*WNT3* (CDR11), *WNT9B* (CDR11)), *Bmp* (*WFIKKN1*

534 (CDR6)), or *Fgf* (FGF13 (CDR37)) could alter the development of or delimit the migration of
535 NCCs, causing depigmented phenotypes. A further NC regulation network, the Hedgehog (hh)
536 pathway, requires and is positively regulated by *SMO* [71,72]. Ablation of *SMO* prevents
537 effective hh signalling, disrupts hair follicle development, and, along with the *Wnt* pathway,
538 likely is a determining factor in epidermal cell fate [71,72]. Again, perturbation of crucial NC
539 signaling pathways results in DS phenotypes with systemic expression.

540

541 **The uncertain role of selection at starch metabolism loci during domestication**

542 Clear distinctions in the efficiencies of starch digestion have been illustrated in dogs compared to
543 their wild wolf ancestors, either from copy-number expansion or genic mutations in starch
544 metabolizing genes [5,27,43]. Earlier work highlighted the importance of efficient starch
545 digestion in early dog domestication, exhibited by selective sweeps at starch metabolizing genes
546 [5]. Since then, recent population genetic studies have pushed back the estimated time of dog
547 domestication to a time that predates human establishment of agriculture [23,26,52,73], the point
548 in which human diets would have also become more starch-rich. Given these new estimates,
549 selection for enhanced starch processing must have occurred since domestication [26,74].
550 Congruent with this, our selection scans suggest that two core starch metabolism genes, *MGAM*
551 and *AMY2B*, may not have been the initial targets of selection. Though we see enrichment of
552 starch metabolism categories in our F_{ST} analyses, this enrichment is largely driven by *MGAM*,
553 which is assigned to three of the five categories associated with carbon metabolism. Per site F_{ST}
554 patterns at the *MGAM* locus indicate that primary selection may have been on adjacent genes
555 *CLEC5A* and *TAS2R38* (Figure 3A). Similar patterns were observed at the *AMY2B* locus (Figure
556 3B), where F_{ST} peaks are shifted distally toward the 65KDa minor spliceosomal subunit, *RNPC3*,

557 a gene that may have a role in early development and growth, as mutations in humans are linked
558 to dwarfism [75]. As first noted by Cagan and Blass (2016), *RNPC3* may have been the possible
559 first driver of selection at this locus since selection scans from their study did not identify
560 significant windows containing the *AMY2B* gene. A recent study has detected *AMY2B* copy-
561 number greater than two in some ancient Romanian dogs (6,000-7,000 years old) using qPCR
562 [28], but the presence of encompassing larger duplications (such as that found in the Newgrange
563 dog) was not tested. However, the *AMY2B* tandem expansion was not detected in HXH, a
564 German sample of the same age as the Romanian samples from Ollivier et al. 2016, suggesting
565 that the tandem duplication may have been a locally-evolved mutation that was either
566 subsequently introduced further into Europe or arose independently at a subsequent time.

567

568 **A role for the minor spliceosome in altered cell differentiation and development**

569 Intriguingly, *RNPC3* is not the only minor spliceosome subunit represented in our CDRs. Of the
570 seven experimentally validated subunits identified in the human minor (U11/U12) spliceosome
571 [76], the 20 (*ZMAT1*), 25 (*SNRNP25*), and 65 (*RNPC3*) KDa subunits are located within CDRs
572 33, 7, and 8, respectively. Complementary to the CDR enrichment results, mutations in other
573 minor spliceosomal subunits and components have been linked to developmental disorders and
574 neurocristopathies, disorders arising from abnormalities in NC development. First, mutations in
575 *Sf3b1*, a U12 spliceosome component [77], alters the expression of critical NC regulators such as
576 *Sox* and *Snail* family members in zebrafish [78]. *U4atac*, another snRNA associated with the
577 U12 spliceosome, has been linked to the Taybi-Lindner Syndrome (TALS) of which phenotypes
578 include craniofacial, brain, and skeletal abnormalities [79]. Finally, mutations in a further minor
579 spliceosome constituent, the human *Zrsr2* gene, disrupts proper cell differentiation pathways,

580 and is linked to myelodysplastic syndrome [80]. This gene is also highly conserved as mutants
581 for the *Zrsr2* plant homolog, *Rough Endosperm 3*, display improperly spliced U11/U12 introns
582 and whose phenotype includes disrupted stem cell differentiation in maize, highlighting the
583 ancient role of the minor spliceosome in cell differentiation pathways [81]. Altogether, these
584 findings that link minor splicing to aberrations in cell differentiation, migration, and neural crest
585 development, substantiate hypotheses that primary selection during domestication first occurred
586 for genes involved in critical, far-reaching pathways such as splicing. It is possible that key
587 genes in the NC pathway are spliced by the minor spliceosome.

588 **Conclusions**

589 Selection scans of the dog genome have yielded 37 regions of sequence deviation between
590 village dogs and their wild ancestors, the gray wolves. We argue that timing inconsistencies with
591 human diet shifts as well as aberrant F_{ST} patterns at loci harboring starch metabolizing genes,
592 does not support the theory that early selection pressures were targeted to genes related to diet.
593 Instead, we believe that the primary targets of selection are nearby genes that are associated with
594 the complex, far-reaching neural crest development pathway. We have discovered enrichment of
595 genes in CDRs that are linked to the early establishment of the neural crest, as well as critical for
596 the migration and differentiation of resulting neural crest cells, that may explain many of the
597 traits attributed to the Domestication Syndrome including the reduction in skull and jaw size,
598 decrease in body size, loss of hair, and floppy ears. Additionally, three regions with unusual
599 levels of allele frequency differences harbor subunits of the minor spliceosome, components of
600 the larger minor splicing pathway associated with impaired cell differentiation across diverse
601 taxa. Together, our results suggest that the NC and minor splicing mechanisms, two early acting

602 pathways with extensive phenotypic effects, had a pivotal role in the early domestication process
603 of dogs.

604 **Methods**

605 **Sample Processing and Population Structure Analysis**

606 Genomes of canids (Additional File 15: Table S8) were processed using the pipeline outlined in
607 [26]. Using GATK [82], a resulting dataset of single nucleotide polymorphisms (SNPs) was
608 produced [26]. Thirty-seven breed dogs, 45 village dogs, and 12 wolves were selected from the
609 samples described in Botigue et al., and ADMIXTURE [83] was utilized to estimate the levels of
610 wolf-dog admixture within this subset. The data was thinned with PLINK v1.07 (--indep-
611 pairwise 50 10 0.1; Additional File 1: Note 1.2; [84]) to leave 1,030,234 SNPs for admixture
612 analysis. Based on five replicates, three clusters (K=3) was determined to best explain the data
613 (average cross validation error = 0.3073). Two dogs and one wolf were eliminated from the
614 sample set because they exhibited over 5% admixture (Additional File 1: Note 1.3).

615

616 Following elimination of admixed samples, we called SNPs in 43 village dogs and 11 gray
617 wolves using GATK which resulted in 7,657,272 sites (Additional File 1: Note 1.4). Using these
618 SNPs, we calculated relatedness and subsequently removed samples that exhibited over 30%
619 relatedness following identity by state (IBS) analysis with PLINK v1.90 (--min 0.05;
620 Additional File 1: Note 1.5; [84]). Only one sample, a mexican wolf, was removed from the
621 sample set as it was highly related to another mexican wolf in the dataset. Principle component
622 analyses were completed on the remaining 53 samples (43 dogs and 10 wolves) using
623 smartpca, a component of Eigensoft package version 3.0 [85] after randomly thinning the

624 total SNP set to 500,000 sites using PLINK v.1.90 [86]. Once PCA confirmed clear genetic
625 distinctions between the dogs and wolves, this final sample set was used for subsequent F_{ST}
626 selection scans. To ensure that SNP choice did not bias the detection of selective sweeps, we
627 generated a second SNP set under a particular ascertainment scheme, which only sampled sites
628 variable in New World wolves (NWWs; a Great Lakes wolf, Mexican wolves (N=2), and
629 Yellowstone wolves (N=2)), generating a SNP set containing 2,761,165 sites.

630

631 F_{ST} Selection Scans

632 To ensure that our inferences were robust to SNP choice, F_{ST} pipelines were completed on the
633 two differentially called SNP sets described above. Allele frequencies were identified for dogs
634 and wolves separately using VCFtools (`--min-alleles 2 --max-alleles 2 --`
635 `recode`; [87]) for both the total and NWW SNP sets. To minimize biases associated with the
636 SNP filtration of the NWW SNP set, the allele frequencies of the NWWs were not included in
637 the F_{ST} calculations for this SNP set.

638

639 Dog and wolf allele counts were used to calculate the fixation index (F_{ST}) using the estimator
640 developed in [88]. For all F_{ST} selection scans, the autosomes and the pseudoautosomal region (X-
641 PAR) of the X chromosome (chrX: 1-6650000) were analyzed separately from the non-
642 pseudoautosomal region (X-NonPAR) of the X (chrX: 6650001-123869142). The Hudson F_{ST}
643 value was calculated in 200kb sliding windows that either did not overlap or were in tiled with
644 50kb spacing across the genome. Each window was required to contain at least 10 SNPs.
645 Additionally, a per site F_{ST} was calculated for each SNP that did not have missing data in any
646 sample. For all three F_{ST} approaches, Z-scores were obtained for each window or site. Z-scores

647 greater than or equal to 5 standard deviations was deemed significant for autosomal and X-PAR
648 loci, and 3 for the X-NonPAR. For additional methods and results, see Additional File 1: Note 2.

649

650 **Genotyping additional canines at CDRs**

651 Forty-six additional canines (e.g. breeds, jackals, coyotes, etc.) were genotyped at CDRs
652 identified in this study, Axelsson et al. 2013, and Cagan and Blass 2016, using autosomal SNPs
653 previously called in [26] (Additional File 15: Table S8). These samples include ancient German
654 dogs, HXH and CTC, that are approximately 7 and 5ky old, respectively [89]. An ancient dog
655 from Newgrange, Ireland (~5ky old; [25]) was also used. SNPs within CDRs of interest were
656 extracted from the [26] dataset using the PLINK make-bed tool with no missing data filter. We
657 note that only CDR SNPs examined in our F_{ST} scans were utilized to create matrices and non-
658 reference allele proportions, and these analyses were only completed using the SNP set (total or
659 NWW) that achieved significance from the F_{ST} analysis (Table 1). If both sets were significant for
660 a CDR, then the total SNP set was used since this set has higher SNP counts.

661

662 Per sample, each SNP was classified as 0/0, 0/1, or 1/1 at all CDRs (1 representing the non-
663 reference genome allele), and this genotype data was stored in Eigenstrat genotype files,
664 which were generated per window using `convertf` (EIGENSOFT package; [90]). Custom
665 scripts then converted the Eigenstrat genotype files into matrix formats for visualization
666 using `matrix2png` [91]. The full set of genotype matrices for the autosomes can be found in
667 Additional File 3: Dataset 1.

668

669 We classified a sample as dog-like or wolf-like based on the proportion of non-reference or
670 “wild” alleles (0/1 or 1/1) within a CDR. The threshold for determining outlier samples was
671 calculated based on the average proportion observed in the 43 village dogs used in the F_{ST}
672 analysis, of which the outlier threshold was determined to be greater than one standard deviation
673 above the village dog mean. Additional File 15: Table S8 provides full results for non-reference
674 allele proportions (Spreadsheets 1 and 2) and outlier identifiers (Spreadsheet 3) per Pendleton
675 CDR, and all plots are within Additional File 4: Dataset 2.

676

677 **Gene Enrichment and Variant Annotation**

678 Coordinates and annotations of dog gene models were obtained from the Ensembl release 81
679 (<ftp://ftp.ensembl.org/pub/release-81/> and http://useast.ensembl.org/Canis_familiaris/Info/Index,
680 respectively), and a non-redundant annotation set was determined. The sequence of each
681 Ensembl protein was BLASTed against the NCBI non-redundant database (`blastp -outfmt`
682 `5 -evaluate 1e-3 -word_size 3 -show_gis -max_hsps_per_subject 20 -`
683 `num_threads 5 -max_target_seqs 20`) and all blastp outputs were processed through
684 BLAST2GO [92] with the following parameters: minimum annotation cut-off of 55, GO weight
685 equal to 5, BLASTp cut-off equal to $1e-6$, HSP-hit cut-off of 0, and a hit filter equal to 55.
686 Positions of all predicted F_{ST} and V_{ST} domestication loci (CDRs) were intersected with the
687 coordinates of the annotated Ensembl canine gene set to isolate genes within the putatively swept
688 regions. Gene enrichment analyses were performed on these gene sets using topGO [34]. The
689 predicted effects of SNP variants were obtained by the processing of the total variant VCF file of
690 all canine samples by SNPEff [32] and Variant Effect Predictor (VEP; [31]). For full methods,
691 please see Additional File 1: Note 3.

692

693 **Copy Number Estimation Using QuickK-mer and fastCN**

694 We implemented two CN estimation pipelines to assess copy-number for the 43 village dogs and
695 10 gray wolves using the depth of sequencing reads. The first, fastCN, is a modified version of
696 existing pipelines that considers multi-mapping reads to calculate CN within 3kb windows.
697 (Additional File 1: Note 4). Related pipelines based on this mapping approach have been
698 successfully used to study CNV in diverse species [93–96]. By considering multi-mapping reads,
699 copy-number profiles will be shared among related gene paralogs, making it difficult to identify
700 specific sequences that are potentially variable. This limitation is addressed by the second
701 pipeline we employed, QuickK-mer, a map-free approach based on k-mer counting which can
702 accurately assess CN in a paralog-sensitive manner (Additional File 1: Note 5). Both pipelines
703 analyze sequencing read-depth within pre-defined windows, apply GC-correction and other
704 normalizations, and are able to convert read depth to a copy-number estimate for each window.
705 The signal-to-noise ratio (SNR), defined as the mean depth in autosomal control windows
706 divided by the standard deviation, was calculated for the 53 dogs and wolves that were used for
707 F_{ST} analysis. The CN states called by both the QuickK-mer and fastCN pipelines are validated
708 through comparison with aCGH data from [46]. Regions with copy number variation between
709 samples in the aCGH or WGS data were selected for correlation analysis. (Additional File 1:
710 Note 6.3).

711

712 **V_{ST} Selection Scans**

713 V_{ST} values [44] were calculated for genomic windows that have evidence of copy number
714 variation using both the QuickK-mer and fastCN pipelines. A higher V_{ST} value indicates a

715 divergent copy number state between the wolf and village dog populations. We identified outlier
716 regions as windows exhibiting at least a 1.5 copy number range across all samples, and ZF_{ST}
717 scores greater than 5 on the autosomes and pseudoautosomal region on the X, or greater than 3 in
718 the X non-pseudoautosomal region. Prior to analysis, estimated copy numbers for male samples
719 on the non-PAR region of the X were doubled. Outlier regions with more than one window were
720 then classified as variant candidate domestication regions (VCDRs) (Additional File 9: Table
721 S4). A similar analysis was performed for the unplaced chromosomal contigs in the CanFam3
722 assembly (Additional File 17: Table S10). See Additional File 1: Note 7 for additional methods
723 and details.

724

725 **Amylase Structural Variant Analysis**

726 We estimated copy-number using short-read sequencing data from each canine listed in
727 Additional File 15: Table S8 with the methodology described in [26]. CN estimations for the
728 *AMY2B* gene using fastCN were based on a single window located at chrUn_AAEX03020568:
729 4873-8379. See Additional File 1: Note 8.1.1 for further methods and results.

730

731 Droplet digital PCR (ddPCR) primers were designed targeting overlapping 1.9Mb and 2.0Mb
732 duplications, the *AMY2B* gene, and a CN control region (chr18: 27,529,623-27,535,395) found to
733 have a CN of 2 in all sampled canines by QuickK-mer and fastCN. CN for each target was
734 determined from ddPCR results from a single replication for 30 village dogs, 3 New Guinea
735 singing dogs, and 5 breed dogs (Additional File 11: Table S6) and averaged from two replicates
736 for 48 breed dogs (Additional File 12: Table S7). For more details on primer design, methods
737 and results for the characterization of the *AMY2B* locus, see Additional File 1: Note 8.1.2.

738

739 **Inversion analysis**

740 The VCF file for the 53 samples analyzed in the F_{ST} pipeline was separated into village dog and
741 wolf files, and further split by chromosome. We ran the InveRision [97] program that utilizes the
742 inversion model from [97,98] to locally phase genotype data and link haplotype blocks to
743 positions of putative inversion breakpoints. These blocks are then used by InveRision to
744 distinguish samples containing each inversion. The total results for all inversions can be found in
745 Additional File 18: Table S11.

746 **Acknowledgements**

747 We thank Shiya Song for advice and assistance in the processing of canid variation data and
748 Laura Botigue for discussion of results utilizing ancient DNA.

749 **Funding**

750 This work was supported by R01GM103961 (AB and JMK) and T32HG00040 (AP). DNA
751 samples and associated phenotypic data were provided by the Cornell Veterinary Biobank, a
752 resource built with the support of NIH grant R24GM082910 and the Cornell University College
753 of Veterinary Medicine.

754

755 **Availability of data and materials**

756 The datasets supporting the conclusions of this article are available in the article and its
757 additional files, as well as a custom UCSC track hub
758 ([https://raw.githubusercontent.com/KiddLab/Pendleton_2017_Selection_Scan/master/Selection_t](https://raw.githubusercontent.com/KiddLab/Pendleton_2017_Selection_Scan/master/Selection_track_hub.txt)
759 [rack_hub.txt](https://raw.githubusercontent.com/KiddLab/Pendleton_2017_Selection_Scan/master/Selection_track_hub.txt)). Software (fastCN and QuicK-mer) implemented in this article are available for

760 download in a GitHub repository (<https://github.com/KiddLab/>). Pre-computed 30-mers from the
761 dog, human, mouse, and chimpanzee genomes can be downloaded from
762 <http://kiddlabshare.umms.med.umich.edu/public-data/QuicK-mer/Ref/> for QuicK-mer
763 processing. Genome sequence data for three New Guinea singing dogs was published under
764 project ID SRP034749 in the Short Read Archive.

765 **Author contributions**

766 JMK, ALP, and FS designed the study. JMK oversaw the study. Selection scans were performed
767 by ALP, AT, and FS. AT and ALP assessed population structure. CNVs were estimated by FS
768 and JMK. Functional annotations and enrichment analyses were performed by ALP. FS
769 processed aCGH data. KV processed ancient dog samples. Samples and genome sequences were
770 provided by KV, AB and JMK. SE performed the DNA extractions, library generation, and
771 ddPCR analyses. ALP, FS, and JMK wrote the paper with input from the other authors.

772 **Competing interests**

773 ARB is a cofounder and officer of Embark Veterinary, Inc., a canine genetics testing company.

774 **Materials & Correspondence**

775 Correspondence and material requests should be addressed to Jeffrey M. Kidd
776 (jmkidd@med.umich.edu).

777

778 **Abbreviations**

779 aCGH: array comparative genomic hybridization; CDR: candidate domestication region; chrUn:
780 chromosome unknown; CN: copy number; CNV: copy number variation; ddPCR: droplet digital
781 polymerase chain reaction; GO: gene ontology; NC: neural crest; NCC: neural crest cell; qPCR:
782 quantitative polymerase chain reaction; SNP: single-nucleotide polymorphism; SNR: signal to
783 noise ratio; VCDR: V_{ST} candidate domestication region.

784

785

786

787

788

789

790

791

792

793

794

795

796 **Additional Files**

797 **Additional File 1:** Supplementary methods and results. (PDF 73kb)

798 **Additional File 2: Table S1.** Coordinates and annotations of F_{ST} candidate domestication
799 regions (CDRs). (XLSX 8.3Mb)

800 **Additional File 3: Dataset 1.** Genotype matrices of autosomal F_{ST} candidate domestication
801 regions (CDRs) in 98 canines. (PDF Xkb)

802 **Additional File 4: Dataset 2.** Non-reference allele proportions of autosomal F_{ST} candidate
803 domestication regions (CDRs) in 98 canines. (PDF 2.2Mb)

804 **Additional File 5: Table S2.** Gene enrichment results for F_{ST} candidate domestication regions.
805 (XLSX 60kb)

806 **Additional File 6:** Supplementary Figures S1-S7. (PDF 1.8Mb)

807 **Additional File 7: Dataset 3.** fastCN and QuicK-mer copy number validations with Ramirez et
808 al. 2014 aCGH probe intensities. (PDF Kb)

809 **Additional File 8: Table S3.** Total V_{ST} outlier regions. (XLSX 111Kb)

810 **Additional File 9: Table S4.** Coordinates of V_{ST} candidate domestication regions (VCDRs).
811 (XLSX 85Kb)

812 **Additional File 10: Table S5.** Gene enrichment results for V_{ST} candidate domestication regions
813 (VCDRs). (XLSX 64Kb)

814 **Additional File 11: Table S6.** ddPCR results from 30 village, 3 New Guinea singing, and 5
815 breed dog samples of AMY2B segmental duplications. (XLSX 60Kb)

816 **Additional File 12: Table S7.** ddPCR results from 48 breed dog samples of AMY2B segmental
817 duplications. (XLSX 55Kb)

818 **Additional File 13: Dataset 4.** fastCN chr9 SNP correlation plots. (PDF 5.6Mb)

819 **Additional File 14: Dataset 5.** QuicK-mer chr9 SNP correlation plots. (PDF 2.8Mb)

820 **Additional File 15: Table S8.** Description and accession numbers for canine genomes processed
821 in this study. (XLSX 75Kb)

822 **Additional File 16: Table S9.** Non-reference allele proportion tables for F_{ST} CDRs. (XLSX
823 121Kb)

824 **Additional File 17: Table S10.** Coordinates of V_{ST} candidate domestication regions on
825 chromosome unknown. (XLSX 48Kb)

826 **Additional File 18: Table S11.** Inversions called in 43 village dogs and 10 gray wolves. (XLSX
827 64Kb)

828 **Additional File 19: Dataset 6.** Supplemental QuicK-mer validation figures. (PDF 1Mb)

829 **Additional File 20: Dataset 7.** Copy number plots of outlier V_{ST} regions. (PDF 6Mb)

830

831

832 **Works Cited**

- 833 1. Trut L. Early Canid Domestication: The Farm-Fox Experiment. *Am. Sci.* 1999;87:160.
- 834 2. Germonpré M, Sablin MV, Lázničková-Galetová M, Després V, Stevens RE, Stiller M, et al.
835 Palaeolithic dogs and Pleistocene wolves revisited: a reply to Morey (2014). *J. Archaeol. Sci.*
836 2015;54:210–6.
- 837 3. Larson G, Burger J. A population genetics view of animal domestication. *Trends Genet.* 2013;29:197–
838 205.
- 839 4. Harlan JR. *Crops & Man* 7 Jack R. Harlan. 1975.
- 840 5. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic
841 signature of dog domestication reveals adaptation to a starch-rich diet. *Nature.* 2013;495:360–4.
- 842 6. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome
843 resequencing reveals loci under selection during chicken domestication. *Nature.* 2010;464:587–91.
- 844 7. Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, et al. Genomic analyses identify distinct patterns of
845 selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* 2013;45:1431–8.
- 846 8. Cagan A, Blass T. Identification of genomic variants putatively targeted by selection during dog
847 domestication. *BMC Evol. Biol.* 2016;16:10.
- 848 9. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong
849 signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U. S. A.* 2012;109:19529–36.
- 850 10. Qiu Q, Wang L, Wang K, Yang Y, Ma T, Wang Z, et al. Yak whole-genome resequencing reveals
851 domestication signatures and prehistoric population expansions. *Nat. Commun.* 2015;6:10283.
- 852 11. Fang M, Larson G, Ribeiro HS, Li N, Andersson L. Contrasting mode of evolution at a coat color

- 853 locus in wild and domestic pigs. *PLoS Genet.* 2009;5:e1000341.
- 854 12. Wang Z, Yonezawa T, Liu B, Ma T, Shen X, Su J, et al. Domestication relaxed selective constraints
855 on the yak mitochondrial genome. *Mol. Biol. Evol.* 2011;28:1553–6.
- 856 13. Cheng T, Fu B, Wu Y, Long R, Liu C, Xia Q. Transcriptome sequencing and positive selected genes
857 analysis of *Bombyx mandarina*. *PLoS One.* 2015;10:e0122837.
- 858 14. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, et al. Linkage disequilibrium
859 and demographic history of wild and domestic canids. *Genetics.* 2009;181:1493–505.
- 860 15. Amaral AJ, Megens H-J, Crooijmans RPMA, Heuven HCM, Groenen MAM. Linkage disequilibrium
861 decay and haplotype block structure in the pig. *Genetics.* 2008;179:569–79.
- 862 16. Wilkins AS, Wrangham RW, Fitch WT. The “domestication syndrome” in mammals: a unified
863 explanation based on neural crest cell behavior and genetics. *Genetics.* 2014;197:795–808.
- 864 17. Carneiro M, Rubin C-J, Di Palma F, Albert FW, Alföldi J, Barrio AM, et al. Rabbit genome analysis
865 reveals a polygenic basis for phenotypic change during domestication. *Science.* 2014;345:1074–9.
- 866 18. Wang G-D, Zhai W, Yang H-C, Fan R-X, Cao X, Zhong L, et al. The genomics of selection in dogs
867 and the parallel evolution between dogs and humans. *Nat. Commun.* 2013;4:1860.
- 868 19. vonHoldt BM, Pollinger JP, Lohmueller KE, Eunjung H, Parker HG, Pascale Q, et al. Genome-wide
869 SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature.* 2010;464:898–
870 902.
- 871 20. Boyko AR. The domestic dog: man’s best friend in the genomic era. *Genome Biol.* 2011;12:216.
- 872 21. Shannon LM, Boyko RH, Castelhana M, Corey E, Hayward JJ, McLean C, et al. Genetic structure in
873 village dogs reveals a Central Asian domestication origin. *Proc. Natl. Acad. Sci. U. S. A.*

- 874 2015;112:13639–44.
- 875 22. Pilot M, Malewski T, Moura AE, Grzybowski T, Oleński K, Ruś A, et al. On the origin of mongrels:
876 evolutionary history of free-breeding dogs in Eurasia. *Proc. Biol. Sci.* 2015;282:20152189.
- 877 23. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome
878 sequencing highlights the dynamic early history of dogs. *PLoS Genet.* 2014;10:e1004016.
- 879 24. Marsden CD, Ortega-Del Vecchyo D, O’Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks
880 and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc.*
881 *Natl. Acad. Sci. U. S. A.* 2016;113:152–7.
- 882 25. Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, et al. Genomic and
883 archaeological evidence suggest a dual origin of domestic dogs. *Science.* 2016;352:1228–31.
- 884 26. Botigue L, Laura B, Shiya S, Amelie S, Shyamalika G, Amanda P, et al. Ancient European dog
885 genomes reveal continuity since the early Neolithic [Internet]. 2017. Available from:
886 <https://doi.org/10.1101/068189>
- 887 27. Arendt M, Cairns KM, Ballard JWO, Savolainen P, Axelsson E. Diet adaptation in dog reflects spread
888 of prehistoric agriculture. *Heredity* . 2016;117:301–6.
- 889 28. Ollivier M, Tresset A, Bastian F, Lagoutte L, Axelsson E, Arendt M-L, et al. Amy2B copy number
890 variation reveals starch diet adaptations in ancient European dogs. *R Soc Open Sci.* 2016;3:160449.
- 891 29. Pilot M, Malewski T, Moura AE, Grzybowski T, Oleński K, Kamiński S, et al. Diversifying Selection
892 Between Pure-Breed and Free-Breeding Dogs Inferred from Genome-Wide SNP Analysis. *G3* .
893 2016;6:2285–98.
- 894 30. Maharzi N, Parietti V, Nelson E, Denti S, Robledo-Sarmiento M, Setterblad N, et al. Identification of
895 TMEM131L as a novel regulator of thymocyte proliferation in humans. *J. Immunol.* 2013;190:6187–97.

- 896 31. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect
897 Predictor. *Genome Biol.* 2016;17:122.
- 898 32. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and
899 predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*
900 *melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6:80–92.
- 901 33. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput
902 functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 2008;36:3420–35.
- 903 34. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. 2016.
- 904 35. Yu H-MI, Jerchow B, Sheu T-J, Liu B, Costantini F, Puzas JE, et al. The role of Axin2 in calvarial
905 morphogenesis and craniosynostosis. *Development.* 2005;132:1995–2005.
- 906 36. Kowalczyk MS, Hughes JR, Babbs C, Sanchez-Pulido L, Szumska D, Sharpe JA, et al. Npr13 is
907 required for normal development of the cardiovascular system. *Mamm. Genome.* 2012;23:404–15.
- 908 37. Lan Y, Ryan RC, Zhang Z, Bullard SA, Bush JO, Maltby KM, et al. Expression of Wnt9b and
909 activation of canonical Wnt signaling during midfacial morphogenesis in mice. *Dev. Dyn.*
910 2006;235:1448–54.
- 911 38. Szláma G, Kondás K, Trexler M, Patthy L. WFIKKN1 and WFIKKN2 bind growth factors TGF β 1,
912 BMP2 and BMP4 but do not inhibit their signalling activity. *FEBS J.* 2010;277:5040–50.
- 913 39. Bronner-Fraser M. Neural crest cell formation and migration in the developing embryo. *FASEB J.*
914 1994;8:699–706.
- 915 40. Santagati F, Rijli FM. Cranial neural crest and the building of the vertebrate head. *Nat. Rev. Neurosci.*
916 2003;4:806–18.

- 917 41. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.*
918 2009;119:1420–8.
- 919 42. Hall BK. *The Neural Crest in Development and Evolution.* 1999.
- 920 43. Arendt M, Fall T, Lindblad-Toh K, Axelsson E. Amylase activity is associated with AMY2B copy
921 numbers in dog: implications for dog domestication, diet and diabetes. *Anim. Genet.* 2014;45:716–22.
- 922 44. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy
923 number in the human genome. *Nature.* 2006;444:444–54.
- 924 45. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated
925 accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.*
926 2015;33:408–14.
- 927 46. Ramirez O, Olalde I, Berglund J, Lorente-Galdos B, Hernandez-Rodriguez J, Quilez J, et al. Analysis
928 of structural diversity in wolf-like canids reveals post-domestication variants. *BMC Genomics.*
929 2014;15:465.
- 930 47. Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. *Genetics.*
931 2006;173:419–34.
- 932 48. Yeaman S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl.*
933 *Acad. Sci. U. S. A.* 2013;110:E1743–51.
- 934 49. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of
935 adaptive evolution in threespine sticklebacks. *Nature.* 2012;484:55–61.
- 936 50. Cáceres A, Sindi SS, Raphael BJ, Cáceres M, González JR. Identification of polymorphic inversions
937 from genotypes. *BMC Bioinformatics.* 2012;13:28.

- 938 51. Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, et al. Worldwide patterns of genomic
939 variation and admixture in gray wolves. *Genome Res.* 2016;26:163–73.
- 940 52. Skoglund P, Ersmark E, Palkopoulou E, Dalén L. Ancient wolf genome reveals an early divergence of
941 domestic dog ancestors and admixture into high-latitude breeds. *Curr. Biol.* 2015;25:1515–9.
- 942 53. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of
943 artificial selection in the dog genome. *Proc. Natl. Acad. Sci. U. S. A.* 2010;107:1160–5.
- 944 54. Sánchez-Villagra MR, Geiger M, Schneider RA. The taming of the neural crest: a developmental
945 perspective on the origins of morphological covariation in domesticated mammals. *R Soc Open Sci.*
946 2016;3:160107.
- 947 55. Etchevers HC, Amiel J, Lyonnet S. Molecular bases of human neurocristopathies. *Adv. Exp. Med.*
948 *Biol.* 2006;589:213–34.
- 949 56. Agnvall B, Jöngren M, Strandberg E, Jensen P. Heritability and genetic correlations of fear-related
950 behaviour in Red Junglefowl—possible implications for early domestication. *PLoS One.* 2012;7:e35162.
- 951 57. Lindberg J, Björnerfeldt S, Saetre P, Svartberg K, Seehuus B, Bakken M, et al. Selection for tameness
952 has changed brain gene expression in silver foxes. *Curr. Biol.* 2005;15:R915–6.
- 953 58. Trut LN, Plyusnina IZ, Oskina IN. An Experiment on Fox Domestication and Debatable Issues of
954 Evolution of the Dog. *Russ. J. Genet.* 2004;40:644–55.
- 955 59. Hare B, Wobber V, Wrangham R. The self-domestication hypothesis: evolution of bonobo
956 psychology is due to selection against aggression. *Anim. Behav.* 2012;83:573–85.
- 957 60. Morey DF, Jeger R. Paleolithic dogs: Why sustained domestication then? *Journal of Archaeological*
958 *Science: Reports.* 2015;3:420–8.

- 959 61. Bronner ME, LeDouarin NM. Development and evolution of the neural crest: An overview. *Dev.*
960 *Biol.* 2012;366:2–9.
- 961 62. Nakata K, Nagai T, Aruga J, Mikoshiba K. *Xenopus* Zic3, a primary regulator both in neural and
962 neural crest development. *Proc. Natl. Acad. Sci. U. S. A.* 1997;94:11980–5.
- 963 63. Etchevers HC, Couly G, Vincent C, Le Douarin NM. Anterior cephalic neural crest is required for
964 forebrain viability. *Development.* 1999;126:3533–43.
- 965 64. Jin Y-R, Han XH, Taketo MM, Yoon JK. Wnt9b-dependent FGF signaling is crucial for outgrowth of
966 the nasal and maxillary processes during upper jaw and lip development. *Development.* 2012;139:1821–
967 30.
- 968 65. Griffin JN, Sondalle SB, Del Viso F, Baserga SJ, Khokha MK. The ribosome biogenesis factor Noll1
969 is required for optimal rDNA transcription and craniofacial development in *Xenopus*. *PLoS Genet.*
970 2015;11:e1005018.
- 971 66. Jho E-H, Zhang T, Domon C, Joo C-K, Freund J-N, Costantini F. Wnt/ β -Catenin/Tcf Signaling
972 Induces the Transcription of Axin2, a Negative Regulator of the Signaling Pathway. *Mol. Cell. Biol.*
973 *American Society for Microbiology*; 2002;22:1172–83.
- 974 67. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A simple genetic
975 architecture underlies morphological variation in dogs. *PLoS Biol.* 2010;8:e1000451.
- 976 68. Rapini RP, Warner NB. Relapsing polychondritis. *Clin. Dermatol.* 2006;24:482–5.
- 977 69. Amiel J, Benko S, Gordon CT, Lyonnet S. Disruption of long-distance highly conserved noncoding
978 elements in neurocristopathies. *Ann. N. Y. Acad. Sci.* 2010;1214:34–46.
- 979 70. Gritli-Linde A, Hallberg K, Harfe BD, Reyahi A, Kannius-Janson M, Nilsson J, et al. Abnormal hair
980 development and apparent follicular transformation to mammary gland in the absence of hedgehog

- 981 signaling. *Dev. Cell.* 2007;12:99–112.
- 982 71. St-Jacques B, Dassule HR, Karavanova I, Botchkarev VA, Li J, Danielian PS, et al. Sonic hedgehog
983 signaling is essential for hair development. *Curr. Biol.* 1998;8:1058–68.
- 984 72. Wang G-D, Zhai W, Yang H-C, Wang L, Zhong L, Liu Y-H, et al. Out of southern East Asia: the
985 natural history of domestic dogs across the world. *Cell Res.* 2016;26:21–33.
- 986 73. Freedman AH, Lohmueller KE, Wayne RK. Evolutionary History, Selective Sweeps, and Deleterious
987 Variation in the Dog. *Annu. Rev. Ecol. Evol. Syst.* 2016;47:73–96.
- 988 74. Argente J, Flores R, Gutiérrez-Arumí A, Verma B, Martos-Moreno GÁ, Cuscó I, et al. Defective
989 minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. *EMBO*
990 *Mol. Med.* 2014;6:299–306.
- 991 75. Will CL, Schneider C, Hossbach M, Urlaub H, Rauhut R, Elbashir S, Tuschl T, Lührmann R. The
992 human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome.
993 *RNA.* 2004;10:929–41.
- 994 76. Will CL, Schneider C, Reed R, Lührmann R. Identification of both shared and distinct proteins in the
995 major and minor spliceosomes. *Science.* 1999;284:2003–5.
- 996 77. An M, Henion PD. The zebrafish sf3b1b460 mutant reveals differential requirements for the sf3b1
997 pre-mRNA processing gene during neural crest development. *Int. J. Dev. Biol.* 2012;56:223–37.
- 998 78. Edery P, Marcaillou C, Sahbatou M, Labalme A, Chastang J, Touraine R, et al. Association of TALS
999 developmental disorder with defect in minor splicing component U4atac snRNA. *Science.* 2011;332:240–
1000 3.
- 1001 79. Madan V, Kanojia D, Li J, Okamoto R, Sato-Otsubo A, Kohlmann A, et al. Aberrant splicing of U12-
1002 type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat. Commun.* 2015;6:6042.

- 1003 80. Gault CM, Martin F, Mei W, Bai F, Black JB, Barbazuk WB, et al. Aberrant splicing in maize rough
1004 endosperm3 reveals a conserved role for U12 splicing in eukaryotic multicellular development. Proc.
1005 Natl. Acad. Sci. U. S. A. 2017;114:E2195–204.
- 1006 81. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
1007 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome
1008 Res. 2010;20:1297–303.
- 1009 82. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated
1010 individuals. Genome Res. 2009;19:1655–64.
- 1011 83. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for
1012 whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81:559–75.
- 1013 84. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.
- 1014 85. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for
1015 whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81:559–75.
- 1016 86. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format
1017 and VCFtools. Bioinformatics. 2011;27:2156–8.
- 1018 87. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data.
1019 Genetics. 1992;132:583–9.
- 1020 88. Botigue L, Laura B, Shiya S, Amelie S, Shyamalika G, Amanda P, et al. Ancient European dog
1021 genomes reveal continuity since the early Neolithic [Internet]. 2017. Available from:
1022 <http://dx.doi.org/10.1101/068189>
- 1023 89. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components
1024 analysis corrects for stratification in genome-wide association studies. Nat. Genet. nature.com;

- 1025 2006;38:904–9.
- 1026 90. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. *Bioinformatics*. Oxford Univ
1027 Press; 2003;19:295–6.
- 1028 91. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput
1029 functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 2008;36:3420–35.
- 1030 92. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy
1031 number and segmental duplication maps using next-generation sequencing. *Nat. Genet*. 2009;41:1061–7.
- 1032 93. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human
1033 copy number variation and multicopy genes. *Science*. 2010;330:641–6.
- 1034 94. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity,
1035 population stratification, and selection of human copy-number variation. *Science*. 2015;349:aab3761.
- 1036 95. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, et al. Evolution and
1037 diversity of copy number variation in the great ape lineage. *Genome Res*. 2013;23:1373–82.
- 1038 96. Cáceres A, Sindi SS, Raphael BJ, Cáceres M, González JR. Identification of polymorphic inversions
1039 from genotypes. *BMC Bioinformatics*. 2012;13:28.
- 1040 97. Sindi SS, Raphael BJ. Identification and frequency estimation of inversion polymorphisms from
1041 haplotype data. *J. Comput. Biol*. 2010;17:517–31.

1042

1043

1044

1045

1046

1047

1048 **Table 1**

1049 Final Candidate Domestication Regions (CDRs) identified from the union of significant
 1050 windows from the overlapping, and non-overlapping approaches of F_{ST} analysis of the Total and
 1051 New World Wolf (NWW) SNP sets. The coordinates, unique identifier, length (bp), maximum
 1052 ZF_{ST} score from both SNP sets, and whether or not a previously published CDR from Axelsson
 1053 et al. 2013 (AX) or Cagan and Blass 2016 (CB) intersects with our CDR.

| Coord. | CDR ID | CDR Length (bp) | Total Max ZF_{ST} | NWW Max. ZF_{ST} | Previously Pub. CDR |
|----------------------------|--------|-----------------|---------------------|--------------------|---------------------|
| chr1: 2300001 - 3250000 | CDR1 | 950000 | 6.86 | 6.35 | AX_1, CB_1, CB_2 |
| chr1: 79800001 - 80150000 | CDR2 | 350000 | 6.29 | 5.74 | AX_4 |
| chr3: 18800001 - 19050000 | CDR3 | 250000 | 5.51 | 2.52 | AX_7 |
| chr4: 40800001 - 41000000 | CDR4 | 200000 | 5.24 | 3.72 | AX_10 |
| chr5: 3850001 - 4250000 | CDR5 | 400000 | 6.00 | 4.32 | |
| chr6: 39900001 - 40100000 | CDR6 | 200000 | 5.39 | | |
| chr6: 40350001 - 40950000 | CDR7 | 600000 | 5.41 | 6.89 | |
| chr6: 46800001 - 47350000 | CDR8 | 550000 | 9.97 | 5.64 | AX_12, CB_3 |
| chr7: 5000001 - 5300000 | CDR9 | 300000 | 5.67 | 5.75 | |
| chr9: 8950001 - 9300000 | CDR10 | 350000 | 5.34 | 5.27 | |
| chr9: 9950001 - 10850000 | CDR11 | 900000 | 4.94 | 5.25 | |
| chr9: 12600001 - 12850000 | CDR12 | 250000 | 5.01 | 5.14 | |
| chr9: 13150001 - 14650000 | CDR13 | 1500000 | 6.28 | 6.91 | |
| chr10: 3950001 - 4150000 | CDR14 | 200000 | 4.64 | 5.05 | AX_17 |
| chr10: 44600001 - 44850000 | CDR15 | 250000 | 4.25 | 5.41 | |
| chr11: 52150001 - 52400000 | CDR16 | 250000 | 4.64 | 5.33 | |
| chr11: 52700001 - 52950000 | CDR17 | 250000 | 5.09 | 5.54 | |
| chr13: 3650001 - 3950000 | CDR18 | 300000 | 5.34 | 5.22 | |
| chr14: 7200001 - 7550000 | CDR19 | 350000 | 6.03 | 6.32 | AX_20 |
| chr16: 7050001 - 7400000 | CDR20 | 350000 | 8.28 | 7.53 | AX_23, CB_4 |
| chr16: 8800001 - 9000000 | CDR21 | 200000 | 3.86 | 5.01 | |
| chr16: 39750001 - 40050000 | CDR22 | 300000 | 6.24 | 5.77 | |
| chr18: 350001 - 1100000 | CDR23 | 750000 | 8.75 | 6.48 | AX_25, CB_5 |
| chr18: 1450001 - 1800000 | CDR24 | 350000 | 6.64 | 5.94 | |

| | | | | | |
|-----------------------------|-------|---------|-----------------|-----------------|---------------------------------|
| chr18: 2550001 - 3750000 | CDR25 | 1200000 | 8.22 | 7.42 | AX_26 |
| chr21: 1 - 200000 | CDR26 | 200000 | 3.56 | 5.06 | |
| chr25: 1050001 - 1350000 | CDR27 | 300000 | 4.93 | 5.61 | AX_30 |
| chr26: 12800001 - 13050000 | CDR28 | 250000 | 5.14 | 2.11 | |
| chr31: 50001 - 400000 | CDR29 | 350000 | 5.47 | 5.86 | |
| chr33: 30000001 - 30200000 | CDR30 | 200000 | 3.79 | 5.02 | |
| chrX: 42850001 - 43100000 | CDR31 | 250000 | 3.11 | | AX_36 |
| chrX: 65900001 - 66200000 | CDR32 | 300000 | 3.20 | 3.30 | |
| chrX: 75700001 - 76350000 | CDR33 | 650000 | 4.02 | 3.86 | |
| chrX: 770500011 - 77550000 | CDR34 | 500000 | 1.953 | 3.556 | CB_7, CB_8 |
| chrX: 78250001 - 78650000 | CDR35 | 400000 | 2.6955 79317 | 3.6766 67204 | AX_38, CB_9 |
| chrX: 105500001 - 105800000 | CDR36 | 300000 | 3.7070 80102 | 2.9315 17165 | CB_13 |
| chrX: 107800001 - 110050000 | CDR37 | 2250000 | 5.1353 83888 | 1.2882 81325 | AX_39 - AX_41, CB_15 - CB_18 |

1054
1055
1056
1057
1058
1059

1060 **Table 2**

1061 A subset of enriched GO categories with P-values less than or equal to 0.05 from Parent-Child
 1062 topGO analysis of CDR genes versus the total genome. The GO ID and term annotation, gene
 1063 count in this category in the genome versus in CDRs (in parentheses), the P-value, CDR gene
 1064 IDs, and the CDRs that the genes are located in are provided for each significant category.
 1065
 1066

| Biological Role | GO Category ID | GO Term Annotation | Total (CDR) Genes | P | Genes | CDRs |
|--|----------------|--|-------------------|-------|--|--|
| Skeletal Development | GO:0001649 | osteoblast differentiation | 207 (6) | 0.004 | SMO, FIGNL1, CLEC5A, AXIN2, MRC2, WNT3 | CDR19, CDR24, CDR20, CDR13, CDR11, CDR11 |
| | GO:0033687 | osteoblast proliferation | 25 (2) | 0.011 | FIGNL1, ENSCAFG00000013735 | CDR24, CDR10 |
| | GO:0060021 | palate development | 80 (3) | 0.018 | WNT9B, WFIKKN1, NPRL3 | CDR11, CDR6, CDR7 |
| | GO:0002062 | chondrocyte differentiation | 87 (3) | 0.03 | SNX19, AXIN2, PRKCA | CDR5, CDR13, CDR13 |
| | GO:0061181 | regulation of chondrocyte development | 3 (1) | 0.032 | AXIN2 | CDR13 |
| | GO:0044339 | canonical Wnt signaling pathway involved in osteoblast differentiation | 2 (1) | 0.022 | WNT3 | CDR11 |
| Secretory System | GO:0042147 | retrograde transport, endosome to Golgi | 66 (3) | 0.005 | RGP1, ENSCAFG00000013419, RAB9B | CDR16, CDR11, CDR34 |
| | GO:0048211 | Golgi vesicle docking | 1 (1) | 0.008 | NSF | CDR11 |
| | GO:2000707 | positive regulation of dense core granule biogenesis | 2 (1) | 0.013 | PRKCA | CDR13 |
| | GO:0006891 | intra-Golgi vesicle-mediated transport | 36 (2) | 0.026 | ENSCAFG00000013419, NSF | CDR11, CDR11 |
| Polarity and Axis Determination | GO:0035545 | determination of left/right asymmetry in nervous system | 1 (1) | 0.008 | ZIC3 | CDR37 |
| | GO:0072046 | establishment of planar polarity involved in nephron morphogenesis | 1 (1) | 0.011 | WNT9B | CDR11 |
| | GO:0007371 | ventral midline determination | 1 (1) | 0.016 | SMO | CDR19 |
| | GO:0045200 | establishment of neuroblast polarity | 3 (1) | 0.032 | FGF13 | CDR37 |
| | GO:0009950 | dorsal/ventral axis specification | 18 (2) | 0.037 | AXIN2, WNT3 | CDR13, CDR11 |
| | GO:0035469 | determination of pancreatic left/right asymmetry | 5 (1) | 0.045 | ZIC3 | CDR37 |

| | | | | | | |
|--------------------------|------------|----------------------------------|--------|-------|-------------|--------------|
| Starch Metabolism | GO:0000272 | polysaccharide catabolic process | 25 (2) | 0.028 | MGAM | CDR20 |
| | GO:0006730 | one-carbon metabolic process | 38 (2) | 0.034 | AHCYL2, CA9 | CDR19, CDR16 |
| | GO:0005983 | starch catabolic process | 1 (1) | 0.042 | MGAM | CDR20 |

| | | | | | | |
|--------------|------------|---|-------|-------|-----|-------|
| Other | GO:0007228 | positive regulation of hh target transcription factor activity | 3 (1) | 0.025 | SMO | CDR19 |
| | GO:0048143 | astrocyte activation | 3 (1) | 0.033 | SMO | CDR19 |
| | GO:0051799 | negative regulation of hair follicle development | 4 (1) | 0.035 | SMO | CDR19 |
| | GO:0021938 | smoothened signaling pathway involved in regulation of cerebellar granule cell precursor cell proliferation | 4 (1) | 0.048 | SMO | CDR19 |

1067
1068
1069
1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085 **Figure Titles and Legends**

1086 **Figure 1 - Origin and diversity of sampled village dogs and wolves.**

1087 (A) The approximate geographic origin of the village dog (circles) and gray wolf (triangles)
1088 genome samples included in our analysis. Results of a Principal Component Analysis of the
1089 filtered village dog (N=43) and gray wolf set (N=11) are shown. Results are projected on (B)
1090 PC1 and PC2 and (C) PC3 and PC4. Colors in in all figures correspond to sample origins and are
1091 explained in the PCA legends.

1092

1093 **Figure 2 - Comparison with previously published candidate domestication regions.**

1094 (A) Venn diagram of intersecting Pendleton, Axelsson et al. 2013 (AX), and Cagan and Blass
1095 2016 (CB) CDRs. (B) Genotype matrix for the 130 SNPs within chr7: 24,632,211-25,033,464 in
1096 AX_14 for 99 canine samples. Sites homozygous for the reference (0/0; blue) and alternate
1097 alleles (1/1; orange) are indicated along with heterozygous sites (0/1; white). (C) ZF_{ST} scores
1098 from the sliding window (bars) and per site (dots) F_{ST} analyses of SNPs belonging to the total
1099 SNP set for CB_14 on chromosome X. Significant per site SNPs ($ZF_{ST} > 3$) are red. The span of
1100 CB_14 relative to these SNPs is indicated. Gene models are situated along the bottom.

1101

1102 **Figure 3 - Selective sweeps at the *MGAM* and *AMY2B* loci.**

1103 ZF_{ST} scores from the sliding window (bars) and per site (dots) analyses of SNPs belonging to the
1104 total SNP set for (A) CDR20 and (B) CDR8. Sliding, overlapping windows meeting significant
1105 ZF_{ST} threshold ($ZF_{ST} > 5$) are red. The span of each CDR relative to these SNPs indicated (white
1106 box). Gene models are situated along the bottom with blue arrows indicate transcriptional
1107 direction.

1108

1109 **Figure 4 - Copy number (CN) estimates distinguish large-scale duplications at *AMY2B***

1110 **locus.**

1111 Read-depth based CN estimations for the (A) full chromosome 6 region of interest, as well as the

1112 (B) proximal and (C) distal margins of large-scale duplications surrounding the *AMY2B* locus.

1113 Estimates for three dogs with differential duplication genotypes are displayed in this figure

1114 including those with either the 1.9Mb (QA27; black) or 2.0Mb (2972; blue) duplication,

1115 compared to a dog without a large-scale duplication (LB79; gray). Relative positions of the

1116 1.9Mb (green bar) and 2.0Mb (purple bar) duplications are indicated along with their respective

1117 primers (darker boxes within duplications). Extensions of the 2.0Mb duplication relative to the

1118 1.9Mb duplication are highlighted in yellow.

1119

1120 **Figure 5 - Region of complex structural variation on chromosome 9.**

1121 (A) Relative to Ensembl gene models and reference assembly gaps (top two tracks), the co-

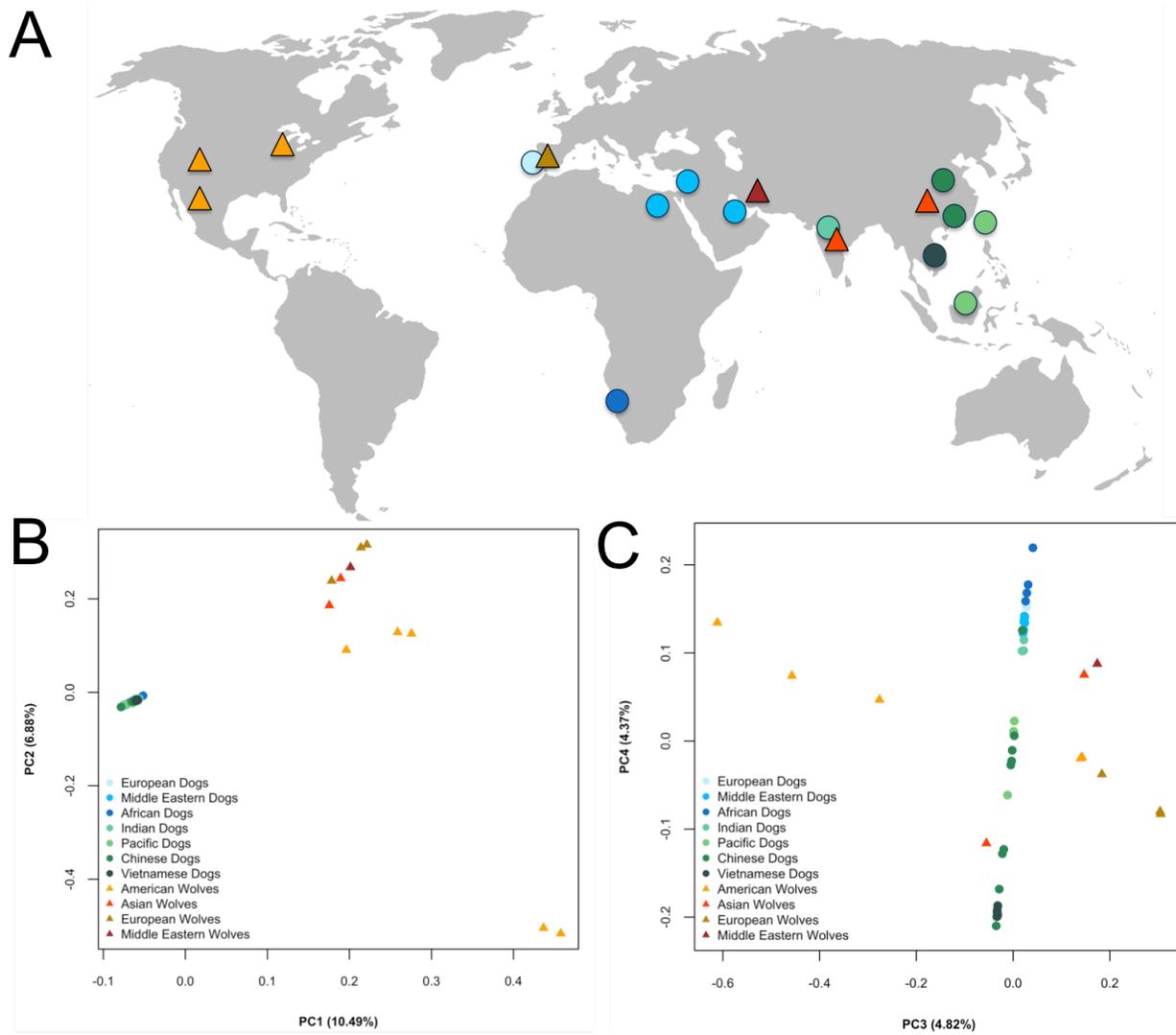
1122 localization of VCDRs (dark blue) with regions of copy number expansion can be observed.

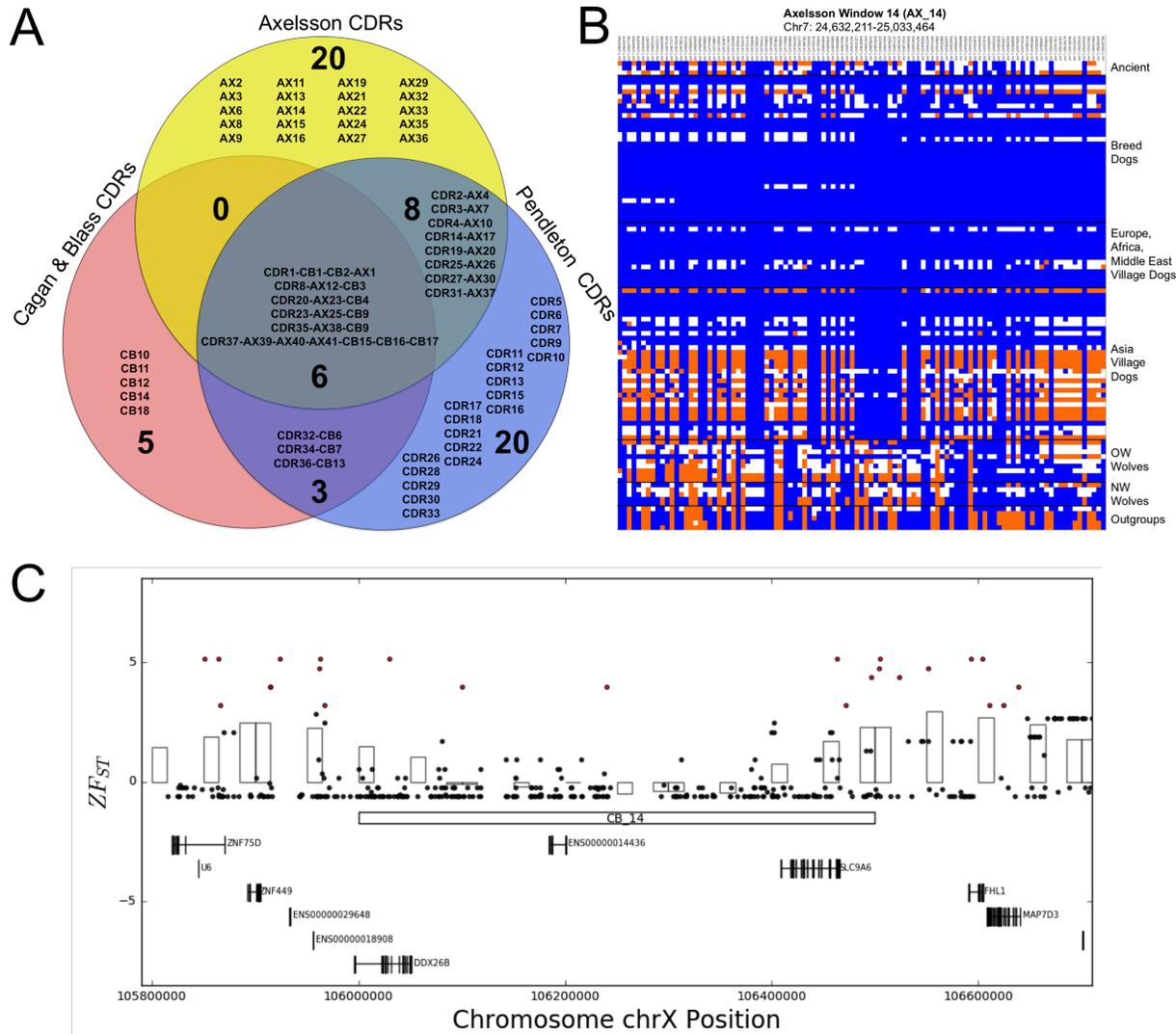
1123 Tracks 3 and 4 display the average copy number states of wolf (orange) and village dog (blue)

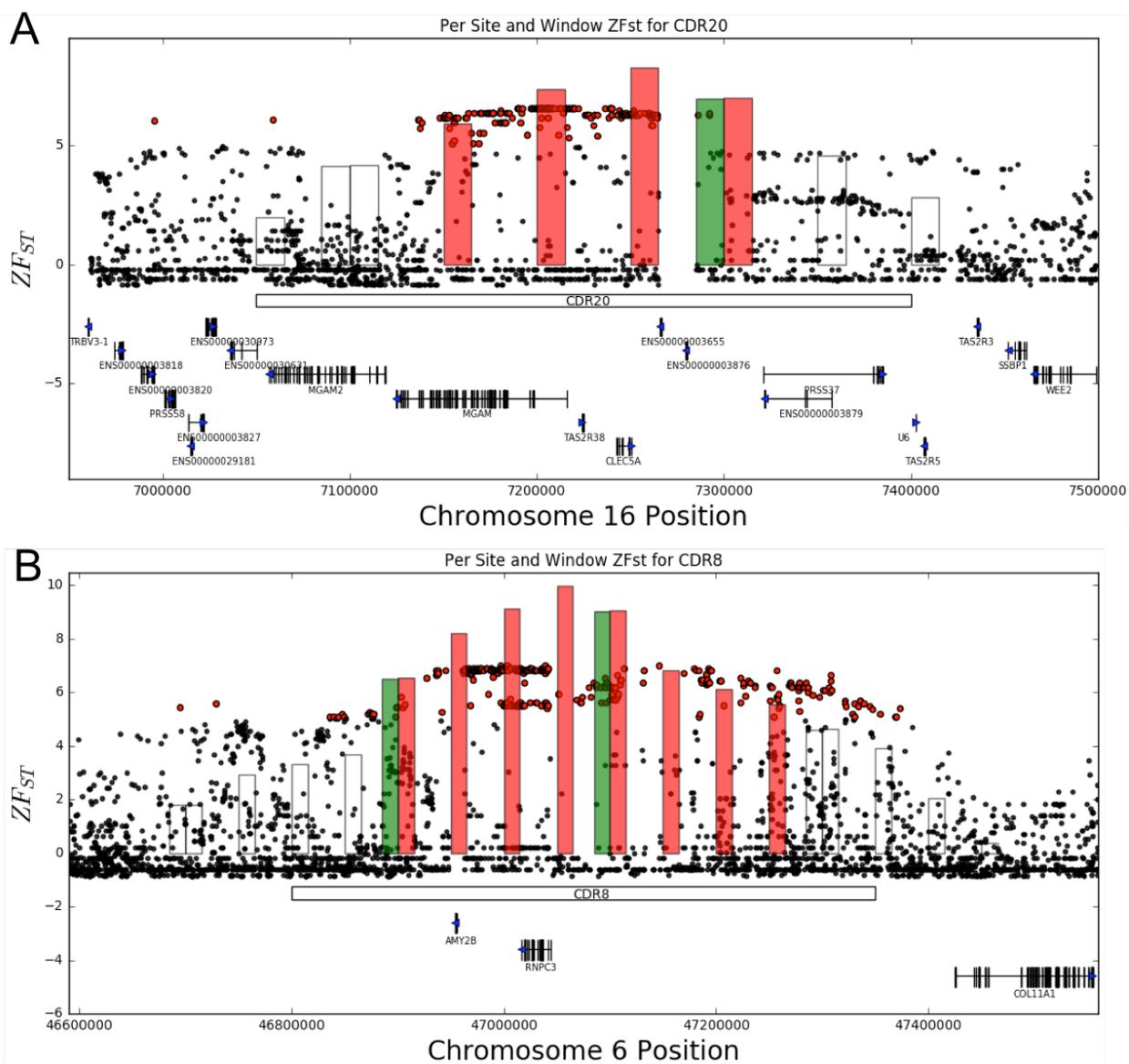
1124 populations as determined by fastCN and QuicK-mer, with regions of consistent CN between the

1125 populations as green. Putative inversions (purple bars), F_{ST} CDRs (red bars), and per site ZF_{ST}

1126 values from the total SNP set (red = $ZF_{ST} > 5$), are also provided in tracks 5-7.







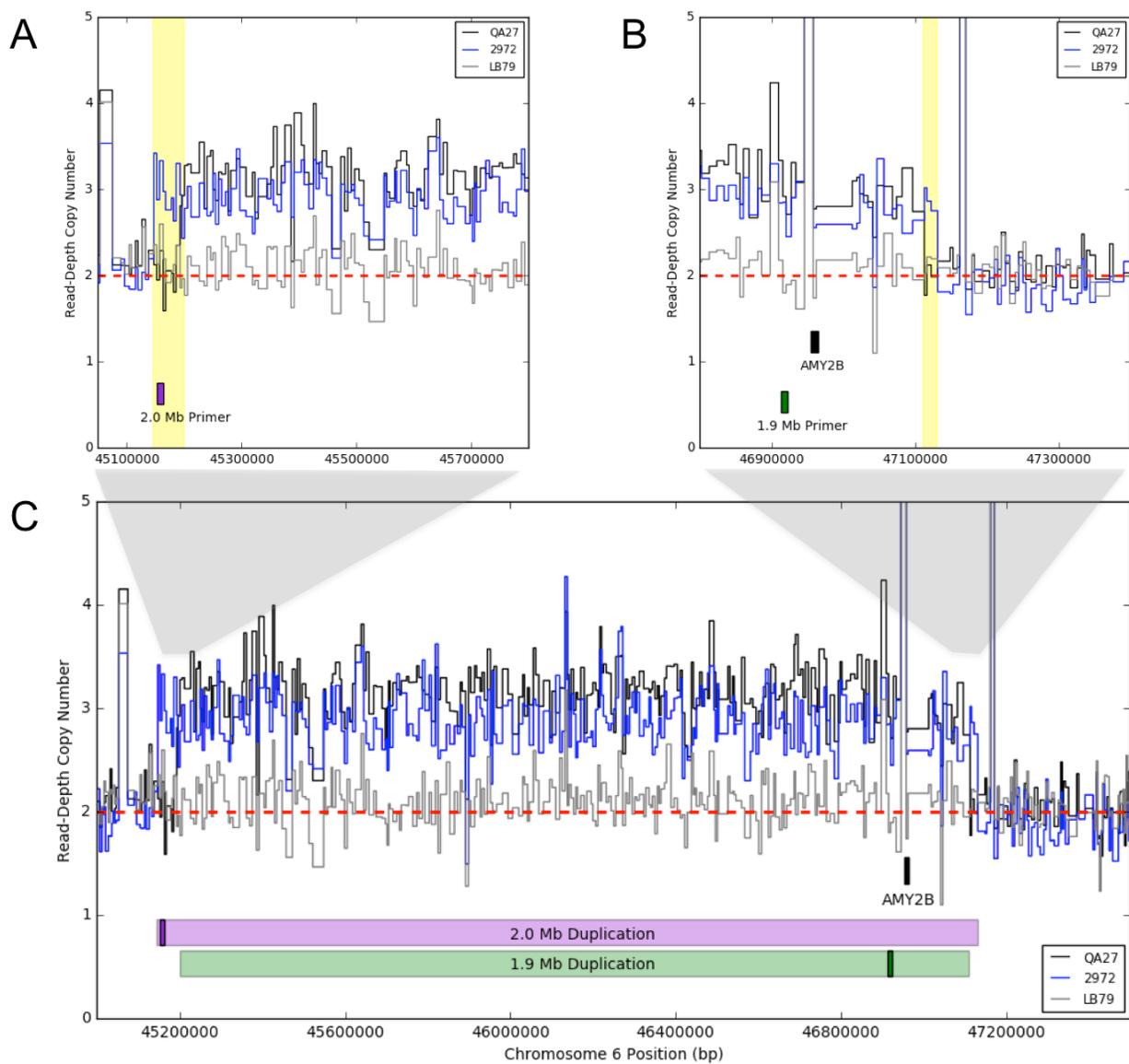


Figure 4

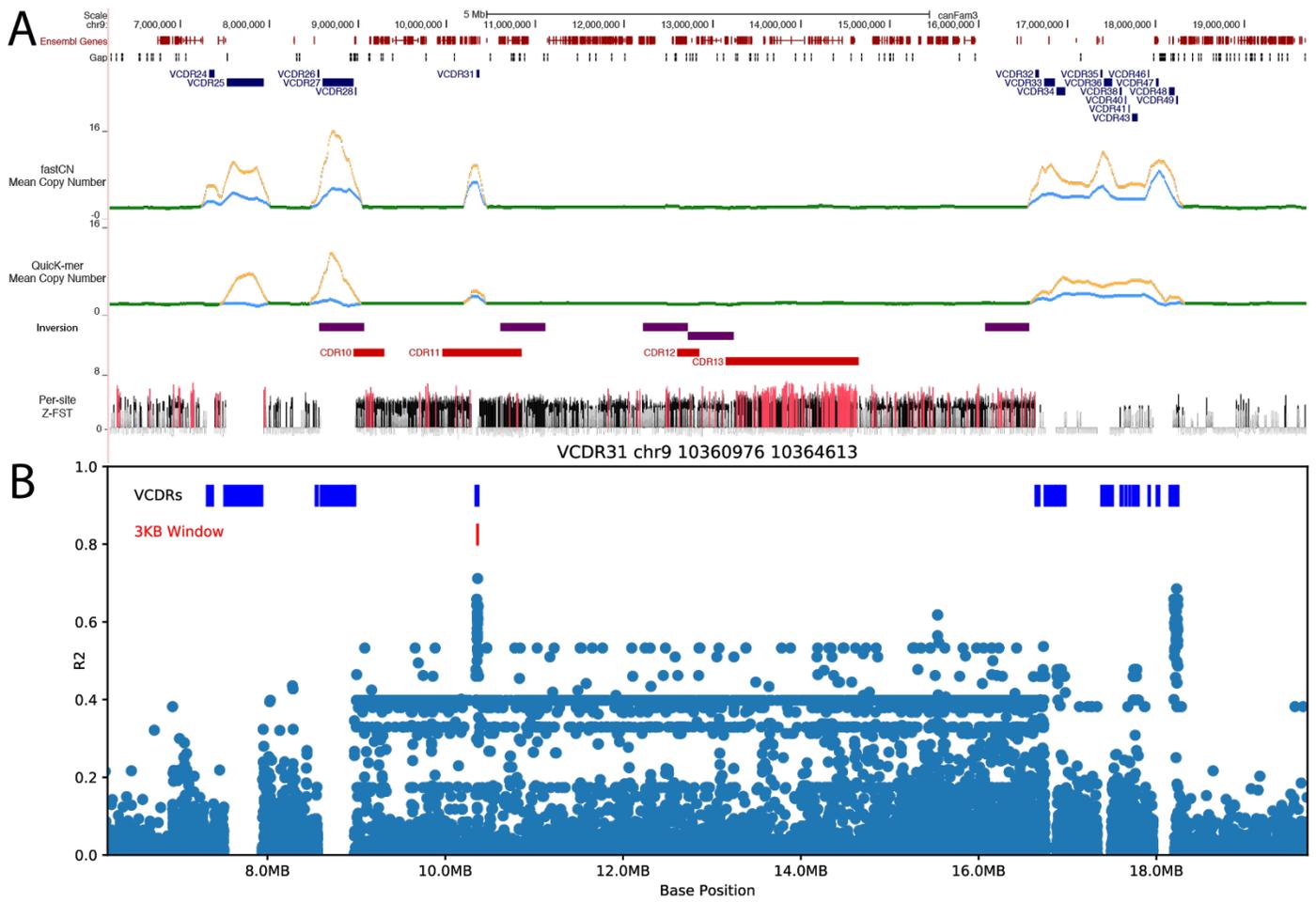


Figure 5