# Whole genome resequencing of the Iranian native dogs and wolves to unravel variome during dog domestication

**Zeinab Amiri Ghanatsaman**
Shahid Bahonar University of Kerman
**Guo-Dong Wang**
Kunming Institute of Zoology Chinese Academy of Sciences
**Masood Asadi Fozi**
Shahid Bahonar University of Kerman
**Min-Sheng Peng**
Kunming Institute of Zoology Chinese Academy of Sciences
**Ali Esmailizadeh** ( aliesmaili@uk.ac.ir )
Shahid Bahonar University of Kerman https://orcid.org/0000-0003-0986-6639
**Ya-Ping Zhang**
Kunming Institute of Zoology Chinese Academy of Sciences

**Research article**

**Keywords:** Single nucleotide variant, Copy number variant, Structural variant, Fertile crescent.

**Posted Date:** October 11th, 2019

**DOI:** https://doi.org/10.21203/rs.2.15926/v1



**Version of Record:** A version of this preprint was published on March 4th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6619-8.

# Abstract

Background  Advances in genome technology have simplified a new comprehension of the genetic and historical processes crucial to rapid phenotypic evolution under domestication. To get new insight into the genetic basis of the dog domestication process, we conducted whole-genome sequence analysis of three wolves and three dogs from Iran which covers the eastern part of the Fertile Crescent located in Southwest Asia where the independent domestication of most of the plants and animals has been documented and also high haplotype sharing between wolves and dog breeds has been reported.  Results Higher diversity was found within the wolf genome compared with the dog genome. A total of 10.45, 7.82, 3.11 and 2.24 million SNPs and small Indels were identified within wolf and dog genomes, respectively. A total of 10,571 copy number variation regions (CNVRs) were detected across the 6 individual genomes, covering 154.65 Mb, or 6.41%, of the reference genome. The genes related to olfactory and immune systems were enriched in the set of structural variants (SVs) identified in this work. Annotation of genomic variations showed that in general the proportion of genomic variations in the intron and intergenic regions in wolf genome is higher than that in dog genome while their proportion in the coding sequences and 3'-UTR in dog genome is higher than that in wolf genome. Generally, genes engaged in digestion and metabolism and neurological process had an important role in the process of dog domestication.  Conclusions By providing the first Iranian dog and wolf variome map, our findings contribute to understanding the genetic architecture of the dog domestication.


# Background

The dog (*Canis familiaris*) was likely the earliest domesticated animal and the only one humans' friend in the past [17,58]. Genetic studies and archaeology findings show that dogs have a common ancestor with the gray wolf (*Canis lupus*) [18, 56, 60]. In the Southwest Asia, major–scale farming extended within the so-named Fertile Crescent (FC) where the independent domestication of plants and animals had led to shifting from gathering and hunting to sedentary farming following expansion of the fist complex societies [19, 65]. Mostly, agricultural *developments* happened in the eastern horn of FC especially Elam (covering a region of southern Iraq and Iran), joining Mesopotamia and Iranian plateau [4]. Dogs are often drawn in art at *ancient* times in several parts of Southwest Asia [17, 44]. Therefore, one of the most theories about the geographical origin of the domestic dog has been that they originated in Southwest Asia, presumably in the FC [17]. In addition, Middle East has been proposed as the beginning of domestic dog for great haplotype sharing between Middle Eastern wolves and dog breeds [55] although this hypothesis has been questioned due to dog-wolf introgression in the Middle East [6, 7, 25] rather than an indication of Middle Eastern origins. The dog is a notable instance of variation under domestication, however the evolutionary processes underlying the genesis of this diversity are weakly realized.

In recent years, advance in high-capacity genome examining *techniques,* especially whole genome sequencing, SNP genotyping array and comparative genomic hybridization (CGH) arrays have authorized the recognition of genome-wide structural variants. The *array methods* have limited resolution and low sensitivity because their performance is strongly depending on the marker frequency and particularly

constructed non polymorphic markeres [5, 36, 46], thus they cannot detect small CNVs (< 10 kb) and cannot precisely identify boundaries of CNVs [64]. Compared to the other *methods*, next-generation sequencing methods provide a high-accuracy base-by-base vision of the genome and capture all variants by different size that might otherwise be missed, and all these are important and have significant effects on an extensive range of traits in domesticated animals. For examples: Increased transcription of *GRIK2* led to the increased fear response in domesticated animals compared with the wild counterparts including rabbit, guinea pig, dog and chicken [33], *MC1R* makes coat color variants in pig [23] and mutation in TSHR *influences* seasonal reproduction in chicken [48]. Copy number variations (CNVs) can also have major phenotypic changes in animals. For example, pea-comb in chickens is produced with the CNVs in the *SOX5* gene [61], The late feathering in chicken is caused with the CNVs *in the SPEF2 and PRLR* [22], The polledness in goats is produced by deletion variation [42], The hair ridge phenotype in ridgeback dogs causing with the CNV in *FGF* gene [28], highly duplicated *APOL3* gene engaged in lipid shifting has been reported in *breeds* of *beef* cattle [12] and *increasing AMY2B* and *AKR1B1* copy numbers make adaptability to a starch- high diet in dog [9, 59].

In this work for the *first time*, we sequenced the whole genomes of 6 canids from the same geographical range (three Iranian wolves and three Iranian dogs) with relatively high coverage (14.51x to 17.15x). One of the sequenced dogs, Qahderijani, is a mastiff ecotype dog originating in Qahderijan, Iran, that is located in FC belt (surrounding areas of FC) and the other two sequenced dogs were sampled from the Saluki breed, a dog breed that originated in the FC and is a hunting dog breed and is considered as the long marathon runner of the canine as its incredible endurance enables the dog to run for many miles.

In our analysis of the Iranian dog and wolf sequences, we applied assembly version canFam3.1 as a reference sequence [34]. SNPs and small Indels were detected in this research as differences between the recently *gained* genome sequences and reference sequence, and detected 12.45 million SNPs and 3.48 million small Indels. Valid algorithms were applied to analyze 6 genomes to get highly reliable CNVs and SVs. The potentially breed-specific CNVRs were defined and the functional relation of the SV and CNVR-covering genes was further evaluated with GO enrichment. Genome-wide analysis indicates more genetic diversity in dog genome than that in wolf genome. Disclosed annotation of the results from different types of genomic variations proposed that increasing the percentage of genomic variations in the coding and the regulatory regions of genes than that in intron and intergenic regions during domestication is the substantial contributor to the currently detected difference between dog and wolf. Also, comparison of effect genomic variations between dog and wolf genomes showed that generally genes engaged in neurological and digestion and metabolism *processes* had a considerable effect in the progress of dog domestication. The CNVs reported in this research are enriched for olfactory and immune system genes.

# Results

### Sequencing output

Illumina Paired-end sequencing was performed for 6 individuals (Additional file 1: Table S1 and Figure S1). After filtering, the range of total high-quality sequence data for 6 individuals was from 42.1 Gb (Sample ID: #GW1) to 51 Gb (#DogQI) and the coverage varied from 14.51 (#GW1) to 17.15 (#GW2) (Additional file 1: Table S2). For increasing reliability of CNV calling, we used uniform depth of coverage across the 6 individual genomes (Additional file 1: Table S2) as suggested formerly [1]. The mean insert size longer than the lengths of both reads with a Poisson-like distribution insert size and a small standard deviation across the 6 individual genomes (Additional file 1: Table S1 and Figure S1) have increased the amount of utilizable sequences in our dataset for detecting of genomic variations in this work [53]. To increase confidence of base calls and a*ccuracy of detecting genomic variations,* sequencing was done with relatively high mean depth for 6 individuals (Additional file 1: Table S2). Relatively high mean depth can increase the accuracy of CNV calling through read depth method [1], and using the paired-*end* DNA *sequencing* reads together with the relatively long read length will be useful to identify Indels [39, 54].

## SNP detection and annotation

SNPs were detected through *aligning* sequences to the reference genome. The number of SNPs was calculated for all individuals (Additional file 1: Table S3 and Figure S2). We a total of 12.45 million SNPs in six individuals, of which 10.45 million SNPs were identified within the 3 wolves and 7.82 million SNPs within the 3 dogs. We obtained the ratio of *transitions to transversions* (*Ti/ Tv*) for SNPs and the number of heterozygous and homozygous in SNPs across the 6 individual genomes (Additional file 1: Table S4). The number of heterozygous SNPs was higher than the number of homozygous SNPs in 6 individuals. The *Ti/ Tv ratio* in SNPs varied from 1.99 (#DogQI) to 2.07 (#GW3) (Additional file 1: Table S4). Annotation of results from SNPs showed that most of the SNPs are located in intergenic and intron regions (Additional file 1:  Tables S5). Of the total number of single-nucleotide polymorphisms, 53.57, 31.99, 0.81, 0.001, 4.83, 4.63, 0.44 and 0.12% were located within intergenic, introns, exon, transcript, upstream, downstream, three prime untranslated region (*3'-UTR)* and *five prime untranslated region* (*5'-UTR*) regions, respectively (Figure 1). Also, the total number of synonymous SNPs (silent SNPs) were more than the total number of non-synonymous SNPs (nonsense and missense SNPs) (Additional file 1: Table S6). Annotation of results from SNPs showed that the proportion of SNPs in intron and intergenic regions in wolf genome was higher than that in dog genome while the percentage of the SNPs in exon regions and 3'-UTR in the dog genome was higher than that in the wolf genome.

## Small Indels detection, annotation and gene ontology

Indels were detected using *aligning* sequences to the reference genome. The number of Indels was calculated for all individuals (Additional file 1: Table S3). A total of 3.48 million Indels were detected across the 6 individual genomes, 2.24 million and 3.11 million of which were for 3 dogs and 3 wolves, respectively. We calculated the number of heterozygous and homozygous Indels across the 6 individual genomes (Additional file 1: Table S4). The number of heterozygous Indels was higher than the number of homozygous Indels in 6 individuals. The total number of small insertions across the 6 individual genomes was 1.58 million, also the total number of small deletions across the 6 individual genomes was

1.9 million (Additional file 1: Table S7). We drew i*ndel length* histogram for 3 dogs (Additional file 1: Figure S3), 3 wolves (Additional file 1: Figure S4) and *across six individual genomes (*Additional file 1: *Figure S5)*. The results showed that *Indels* of *1* bp in *length* across the 6 individual genomes had the highest percentage and *in the same size* deletions had more percentage than the insertions. Annotation of results from small Indels showed that most of the Indels are located in intergenic and intron regions (Additional file 1: Tables S8). Of the total number of small Indels, 53.79, 34.778, 0.25, 0.002, 5.54, 4.95, 0.46, and 0.14% were located within the intergenic, introns, exon, transcript, upstream, downstream, 3'-UTR and 5'-UTR regions, respectively. The percentage of small Indels that are located in upstream, 5'-UTR, 3'-UTR, exon and transcript regions across 3 dog genomes was higher than that across 3 wolf genomes, but the percentage of Indels that are located in downstream, introns and intergenic regions across 3 wolf genomes was higher than that across 3 dog genomes. We obtained 21,104 genes from ensemble through annotation a total 3.48 million small Indels. After, we carried out gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis for these genes (Additional file 1: Table S9). Gene Ontology (GO) analysis categorized genes related to small Indels in the *three main* classes (molecular function, biological process and cellular component) (Additional file 1: Table S9). The KEGG pathway analysis showed that two pathways related to cancer and Melanoma (usually but not always, a cancer of the skin) were enriched among the small Indels in both dog and wolf.

## Structural variants detection, annotation and gene ontology

We obtained genomic structural variants including insertions, deletions, translocations (inter and *intra* chromosomal) and inversions for three dogs and three wolves (Additional file 1: table S10; Additional file 2: table S16 and Additional file3**:** S17). The total number of deletions, insertions, inversions, inter chromosomal translocations and intra chromosomal translocations, across the 6 individuals was 14,321, 566, 469, 798 and 637 respectively (Additional file 4**:** Table S18). The total number of all structural variants except insertions in the *wolf genome was higher than those in* the dog genome (Additional file 2 :Table S16 and Additional file 3**:** S17). To obtain potential functional roles related to the different types of structural variants, all genes that completely or partially have overlapped with them were retrieved from Ensemble. We obtained 470, 163 and 228, 6,466, 191 and 269 genes from annotation for the total number of Indels (insertion and deletion), inventions and complex structural variants (inter and *intra* chromosomal translocations), respectively in dog and wolf (Additional file 1: Table S11). Annotation of results from structural variants showed in general the percentage of intergenic and noncoding transcript variants in wolf genome is higher than that in dog genome while the proportion of coding sequences and 3'-UTR variants in dog genome is higehr than that in wolf genome (Additional file 1: Figures S6-S13). Gene ontology (GO) analysis categorized genes related to structural variations in three *classes*, covering molecular function, biological process and cellular component (Additional file 1: Table S12). The genes related to olfactory and immune systems were enriched in the set of SVs identified in this work (Additional file 1: Table S12). The most conspicuous cluster terms in dog and wolf were “cellular carbohydrate metabolic process” and “nervous system development.

## CNV detection

We obtained putative CNVs for 6 individuals using CNVnator program and the mean number of CNVs per individual was 4143.83 ranging from 2871 to 5437 (Additional file 1: Table S13). For all the autosomal CNVs categorized as gain, the mean copy number value of six individuals was 3.57 and the maximum copy number assessment was 174.472 on chromosome 7 (chr7) of wolf. The results showed that the number of gains in the three dog genomes was higher than those in the three wolf genomes (Additional file 1: Table S13). A total of 10571 CNVRs were obtained from overlapping of all CNVs across the 6 individuals (Additional file 5: Table S19), including 1-38 and X chromosomes, ranging in size from 1.05 kb to 3433.35 kb with an average of 14.63 kb and a median of 7.05 kb, covering 154.65 Mb, or 6.41%, of the assayed CanFam genome (Table 1). CNVRs were divided into three groups, including 6400 loss, 3916 gain and 255 both (gain and loss) events (Additional file 5: Table S19). Deletion:duplication ratio in the total CNVRs was 1.96. Among all CNVRs, 6,105 (57.75%) were found in a single individuals (singleton), 1,522 (14.39%) shared in two individuals, and 2,944 (27.84%) shared in at least three individuals (Figure 2B). A number of 6702 (63.4%) CNVR events were less than 10 Kb while 494 (4.7%) of the CNVRs were longer than 50 kb in size (Table 1 and Figure 2A). The highest and lowest numbers of CNVRs belonged to chromosomes 18 and 35, respectively (Figure 4 and Additional file 6: Table S20).

## CNV annotation and gene ontology analysis

*The annotation of results from* CNVs showed that the percentage of CNVs in coding sequences (14% vs. 6%) and 3'-UTR (6% vs 0) in the dog genome was greatly higher than that in the wolf genome, but the percentage of CNVs in the intergenic regions (22% vs. 14%) in wolf genome was greatly higher than that in the dog genome (Additional file1: Figures S14 and S15). To achieve potential functional roles related to the putative CNVs, all genes that completely or partially have overlapped with these CNVs were detected from Ensemble. A total of 8595 genes were retrieved, including 6703 of the CNVs. Results of gene ontology (Go) analysis showed that in general genes associated with olfactory and immune systems are enriched among the CNV gains in dog and wolf (Additional file 1: Table S14). All the terms related to olfactory system are over-represented ($P<0.01$) in the wolf compared with those in the dog (Additional file 1: Table S14). The term “Starch and sucrose metabolism” is enriched in the dog CNV gains. The terms “cardiac conduction”, “Cardiac muscle contraction”, “regulation of heart contraction”, “heart development”, “muscle filament sliding”, “regulation of smooth muscle cell proliferation”, “ATP binding”, “calcium ion binding” and “muscle cell development” are enriched among the CNV gains in Saluki dogs (Additional file 1: Table S14).

## Comparison with previous dog CNV studies

To compare the identified CNVRs in this work with those of the published studies, all previous CNVR coordinates from canFam2 were migrated to canFam3 using the UCSC leftover program. In our results, 4454 CNVRs (42.1%) were overlapped by four previous studies, and the remaining 6117 (57.865%) were considered as novel CNVRs (Additional file 1: Tables S15 and Additional file 7: S21).

## Visualization of Structural Genomic Variation

*For v*isualizing similarities and differences of positional relationships and genome structure between dog and wolf genomes, we drew *maps* of *circular genomes for dog and wolf (Figure 3).*

# Discussion

Analysis of high-quality next-generation sequencing data clearly showed the difference of the distribution and impact of the genomic variations between dog and wolf. The ratio of *transition to transversion* (ti/tv) *is an* indicator of false positive ratio for SNP calling [10, 27], the ratios calculated for all individuals (1.99 to 2.07) (Supplementary Table S4) *indicate* the precision of the identification of single-nucleotide mutations in our research. In addition, the results of this research similar to previous studies [50] showed that most of the SNPs belong to within introns or between genes and the number of synonymous SNPs was higher than non-synonymous SNPs. The majority of small Indels (95.89 % in dog and 95.64.% in wolf) were less than 10 bp in *length*, similar results were reported in study of Indels in chicken [63].

We detected 10571 CNVRs with a mean of 4143.83 CNVs per sample in the canine genome. Similar to those reported in dog and wolf [15, 38, 40, 41], human [20, 47] and mouse [26], loss events were more prevalent than gain events in our results (1.63 fold). This may mirror the greater relative hardness of identifying gains because of the smaller relative alteration in copy number (3:2 versus 2:1). Loss events included shorter genomic sequences than gains on median (4.499 kb vs. 11.699 kb), mean (7.387625 kb vs. 21.38724 kb) and total (47.280800 Mb vs. 83.752434 Mb) (Table 1). This could show that duplications are less likely to be cleaned by purifying selection [5]. A total of 4466 (42.25%) CNVRs are seen in at least two individuals and 6105 (57.75%) CNVRs present in only one individual. Percentage of singletons was obtained in this work is in agreement with that reported in previous studies related to identification of CNV in human [47], dog [40] and chicken [64]. We realized that the CNVRs were non-randomly distributed across the canid genome (Table S20). Chromosome 32, for example, has 2.03% of sequences displaying copy number variable, whereas chromosome 18 has 42.79% of sequences with copy number variation (Supplementary Table S20). In general, the chromosomes 9 (13.03%), 26 (14.97%) and 18 (42.78%) showed a high percentage of the CNVRs.

The terms “sensory perception of smell”, “detection of chemical stimulus” and “Olfactory transduction” are involved in sensory perception and were enriched among the CNV gains in dog and wolf and all of them were over-represented among the CNV gains in wolf ($P<0.01$). Both wolf and dog develop olfaction, audition and vision by 2 weeks, 4 weeks and 6 weeks of age on average, respectively [35]. Wolf pups start to investigate their environment at 2 weeks of age while they are blind and deaf, and must depend mainly on sense of smell, while dog pups start to investigate their environment at 4 weeks of age [35]. In a previous study, the fraction of olfactory receptor pseudogenes in dog and wolf was 17.78 and 12.08%, respectively, however, difference between these values in dog and wolf was not significant [67]. In another study, no difference in the olfactory capacity of the dog breeds that have been choosed for their smelling ability and the hand- breaded grey wolves was reported [43]. However, our results emphasize the importance of olfaction during dog domestication.

Many of GO terms belonged to CNV gains in this research are similar to those that were presented using aCGH method in dog [11]. Gene ontology go enrichment analysis showed that gene families involved in sense of smell and immune system commonly rapid growing for their importance in the organism answering to fast changes in the environment and fitness, also they have been frequently identified in CNV regions of multiple mammalian *genomes [2, 62, 69].* Go terms related to *heat function* such as “cardiac conduction” and “regulation of heart contraction” were only enriched in the CNV gains in Saluki dog. These results can be expected because Saluki is a hunting dog breed which is considered as the long marathon runner of the canine world and its incredible endurance enables the dog to run for many miles. It has been presented that endurance exercise training makes a number of of cardiac adaptations to marathon running [51].

A fundamental number of the CNVs from this work (42.13%) are compatible with those identified in previous studies in dogs and wolves. In addition, a substantial number of the Go terms that are enriched among the CNV in this study are concordant with the Go terms related to studies of copy number variations in dogs and wolves. This compatibility with the previous studies, in conjunction with the identification of the CNVs specific to the Saluki breed, lends more support to the CNVs identified in this work. The difference between the CNVs detected in the study herein and those described previously can be related to the particular breeds studied and also the difference between the methods used. Generally, the CNVs that are identified by read-depth analysis are on average much smaller than those detected by aCGH.

The total numbers of SNPs, Indels, deletions, inversions, inter and intera chromosomal translocations in the wolf genome were higher than those in the dog genome while the total number of duplications or insertions in the dog genome was higher than those in the wolf genome. It has been accepted that gene duplication can be a chief source of recentness in evolution [68].

Our results from the genome analysis of dog and wolf revealed reduction of *genomic diversity* during dog domestication. A population bottleneck occurred in the wolves thousand years ago after a population expansion occurred by human through artificial selection on specific traits leading to different breeds of dogs [3, 25]. The effective population size in wolves is higher than that in dogs so higher genome diversity in wolves is expected compared to dogs [3, 25]. Our results from two components of genetic variation sources including SVs and CNVs confirmed that the novel adaptations permitted the primal ancestors of recent dogs to live on a diet high starch compared to the carnivorous diet of wolves and formed a essential step in the primal domestication of dogs [8, 9, 25, 52, 60]. The terms “Negative regulation of neuron apoptotic process”, “positive regulation of dendrite development” and “nervous system development” were enriched among SVs in wolf and are indicative of reducing aggression in the first steps of animal domestication. “Nervous system development" is defined as a process that particular result is the development of nervous tissue over time from its production to its developed shape.

In previous studies, the terms “axon” and “Nervous system development” were enriched among the genes related to the regions under selection during dog domestication [9, 57].

Annotation of results from different types of genomic variations showed that in general the percentage of genomic variations in intron and intergenic regions in wolf genome is higher than that in dog genome while in *coding sequences* and 3'-UTR in dog genome is higher than that in wolf genome. It seems that domestication and its related processes such as relaxed selection have an important role in increasing the percentage of genomic variation in the coding and the regulatory *sequences* of genes in dog. The relaxation of selection likely increases the functional genetic diversity throughout the genome of the dog and this diversity includes both the genes and the elements involved in gene expression [13, 21]. However, it should be noted that mammalian genomes possess a complex structure with a diverseness of repetitive elements that complicates extensive genome-wide analyses [54]. To better acknowledge this result, there is still the need for using mate pair sequences or merging long-insert mate pair and short-insert paired-end sequences to analyze the dog and wolf genomes and elucidate difference of the distribution and impact of the genomic variations between dog and wolf *during dog domestication*.

# Conclusions

We resequenced the whole genomes of 6 canids from the Middle East for the first time and we compared the effect and distribution of the genomic variations between dog and wolf genomes. Whole genome resequencing of three dogs and three wolves detected 7.82 million and 10.45 million SNPs, respectively. Numerous putatively CNVs were identified through an analysis of read depth difference. Furthermore, we have identified SVs which could be useful for marker based population genetic investigation. Downstream analysis of the identified SVs and CNVs revealed the changes between dog and wolf genome during dog domestication. More work is needed to unravel the significance of the higher proportion of CNV gains in the Saluki dog.

# Methods

## Sampling and sequencing

The source of the animals used in this study were as follow: one wolf was sampled from Kerman zoo, South of Iran, two wolves were used from Eram Park Zoo, Tehran, Iran; two Saluki dogs were sampled from private farms in Kurdistan province, west of Iran and one Qahderijani dog was used from a private farm in Isfahan, Iran. We collected blood samples from three captive Iranian wolves (Additional file 1: S17) and three Iranian dogs including a Qahderijani (Additional file1: Figure S17) and two Saluki dogs (Additional file1: Figure S18) with the consent of the owners. Sampling locations are reported in Table 2. DNA was prepared with phenol/chloroform technique. Pair-end sequence data for all 6 individuals were generated using Hiseq 2500 Illumina company in China (www.berrygenomics.com).

## Quality control and mapping

The quality of reads was evaluated with FastQC program, outputs of quality control showed that all reads had high-quality and were without adaptor contamination. *Aligning data against* the genome assembly

canfam3. 1 was done with burrows wheeler aligner program (bwa) [31]. The SAMtools [32] was applied to change the Sequence Alignment MAP (SAM) files to the Binary Alignment MAP (BAM) files and sort and index them. All of the .bam files were cleaned from PCR duplicates with Picard program. The accuracy of mapping was evaluated using of two criteria including percentage of *aligning against* the reference genome and mean depth with SAMtools

### Short indel and SNP detection

*Genome Analysis Toolkit* (GATK) program [37] was applied to detect SNPs and Indels. All .bam files were preprocessed in two steps; i) local realignment around Indels was done using known Indels, ii) recalibrating base quality scores was done to increase quality score for each base. The purified data belonged to the same individual were jointly used to create genome variant call format (gVCF) files by GATK HaplotypeCaller, followed by merging the gVCF files belonged to all individuals employing the GATK GenotypeGVCFs. Finally, SNPs and Indels were separated from the resulted raw variant file and filtered using GATK Select Variants and GATK Variant Filtration, respectively.

### SVs detection

SVs including deletions, inversions, translocations (inter and intra chromosomal) and insertions were detected using BreakDancer-1.1 [14] software. SVs were filtered using BreakDancer with read coverage >=10, the score>=80 and size>=50 bp.

### SNP and Indel annotation

Functional consequence analysis of SNPs and short INDELs were studied using SnpEff 4.0e [16], also transition to transversion and homozygous to heterozygous ratios for single nucleotide variants and were calculated with SnpSift [49].

### CNV Calling

Putative CNVs on the 38 Canine autosomes and X chromosome were detected based on read depth method using CNVnator [1]. We run CNVnator with a bin size of 150 bp and GC correction (default) for our data. Filtering putative CNVs was done using different criteria including size > l kb, P-value < 0.01 and q0 (zero mapping quality) < 0.5. We removed all un-localized chromosome CNVs (chrUn). Putative CNVRs were obtained using Bedtools software [45] from overlapping of 1bp or greater CNVs on chromosomes 1-38 and X chromosome in 6 individuals as reported before [47]. All CNVRs were categorized into three classes, e.g., “Loss” (including deletion), “Gain” (including duplication) and “Both” (including both deletion and duplication). To compare the putative CNVRs from this study with the CNVRs reported in the previous studies, all coordinates related to CNVRs of the previous studies were converted from CanFam 2.0 to CanFam 3.1 using the lift over tools (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

### Gene contents and gene ontology analysis

Dog gene IDs that covered small Indels, SVs and CNVRs were retrieved from Ensemble annotation [24]. All dog gene IDs were changed to human gene IDs. Gene orthologous connection between dog and human was obtained from Ensembl. Gene ontology (GO) was done using DAVID program [29].

### Visualization of structural genomic variation

We drew the physical distribution of CNVRs on chromosomes 1-38 and X chromosomes using VCStools [30]. RCircos package [66] was used to draw *circular genetic maps for v*isualizing similarities and differences of positional relationships and genome structure between dog and wolf.

## List Of Abbreviations

BAM: Binary Alignment MAP; Bwa: burrows wheeler aligner program; chrUn: Un-localized chromosome; CGH: Comparative genomic hybridization; CNVRs: Copy number variation regions; CNVs: Copy number variations; FC: Fertile Crescent; GATK: Genome Analysis Toolkit; GO: Gene ontology; gVCF: Genome variant call format; GW: Gray wolf; Indels: Insertion and deletion; KEGG: Kyoto Encyclopedia of Genes and Genomes; QI: Qahderijani; SAM: Sequence Alignment MAP; SI: Saluki; SVs: Structural variants; 3'-UTR: Three prime untranslated region; 5'-UTR: Five prime untranslated region

## Declarations

### Ethics approval and consent to participate

This study had Institutional Animal Care and Use Committee (Kunming Institute of Zoology, approval ID: SYDW-2013021) approval. We collected peripheral blood samples from 3 Iranian dogs with the consent of owners and 3 gray wolves after obtaining authorization for research from the Department of Environmental Protection in Iran (No. 93/34089, dated 14 October 2014).

### Consent for publication

Not applicable.

### Availability of data and materials

**Data deposition**: Raw sequence reads data have been deposited in the Genome Sequence Archive (http://gsa.big.ac.cn/) under accession CRA0001324 for raw data of genomes.

### Competing interests

The authors express no competing interests.

### Funding

This research was funded by the National Natural Science Foundation of China (No. 91531303), the international cooperation program of bureau of international cooperation of Chinese Academy of Sciences (No.GJHZ1559), and the Animal Branch of the Germplasm Bank of Wild Species, Chinese Academy of Sciences (the Large Research Infrastructure Funding). A.E. was supported by the Chinese Academy of Sciences President's International Fellowship Initiative (No. 2016VBA050). MSP and GDW appreciate the assistances from the Youth Innovation Promotion Association, Chinese Academy of Sciences.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Authors' contributions

G-DW, AE and M-SP realized and *planed the study. ZAG and MAF provided samples.* G-DW prepared *the genomic DNAs of the six samples. ZAG and* G-DW *analyzed and interpreted the data. ZAG drafted the manuscript.* M-SP and AE revised the manuscript. *Y-PZ prepared* resequencing of data and *was the project leader.* All authors have read and approved the final version of the manuscript.

### Author details

[1]Department of Animal Science, Faculty of Agriculture, Shahid Bahonar University of Kerman, PB 76169-133, Kerman, Iran

[2]Yong Researchers Society, Shahid Bahonar University of Kerman, PB 76169-133, Kerman, Iran

[3]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences No. 32 Jiaochang Donglu, Kunming, Yunnan, 650223, China.

[4] State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan University, Kunming 650091, China.

[a¶]Co-first authors.

### Acknowledgements

This research was carried out as part of PhD thesis at Shahid Bahonar University of Kerman, Iran. We appreciate sampling assistance from the dog owners and staff from department of natural resources in assisted Tehran and Kerman, Kerman Zoo, Tehran Eram Zoo and Shiraz Zoo in Iran. Also we thank Dr. Hosein Rashidi and Dr. Iman Memarian for their assistance in sampling wolf in Kerman Zoo and Tehran Eram Zoo, respectively.

## References

1. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974-984. 2. Agam A, Yalcin B, Bhomra A, Cubin M, Webber C, Holmes C, et al. Elusive copy number variation in the mouse genome. PloS One. 2010;5:e12839. 3. Alam M, Han KI, Lee DH, Ha JH, Kim JJ. Estimation of effective population size in the Sapsaree: a Korean native dog (Canis familiaris). Asian Austral J Anim. 2012;25:1063-1072. 4. Alizadeh A. The rise of the highland Elamite state in southwestern Iran. Curr Anthropol. 2010;51:353–383. 5. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363-376. 6. Amiri Ghanatsaman Z, Adeola AC, Asadi Fozi M, Ma YP, Peng MS, Wang GD, et al. Mitochondrial DNA sequence variation in Iranian native dogs. Mitochondrial DNA A DNA Mapp Seq Anal. 2017;17:1-9. 7. Ardalan A, Kluetsch CF, Zhang AB, Erdogan M, Uhlén M, Houshmand M, et al. Comprehensive study of mtDNA among Southwest Asian dogs contradicts independent domestication of wolf, but implies dog–wolf hybridization. Ecol Evol. 2011;1:373-385. 8. Arendt M, Fall T, Lindblad-Toh K, Axelsson E. Amylase activity is associated with AMY2B copy numbers in dog: implications for dog domestication, diet and diabetes. Anim Genet. 2014;45:716-722. 9. Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013;495:360-364. 10. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 2011;12:R68. 11. Berglund J, Nevalainen EM, Molin AM, Perloski M, Andre C, Zody MC, et al. Novel origins of copy number variation in the dog genome. Genome Biol. 2012;13:R73. 12. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, et al. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 2012;22:778-790. 13. Björnerfeldt S, Webster MT, Vilà C. Relaxation of selective constraint on dog mitochondrial DNA following domestication. Genome Res. 2006;16:990-994. 14. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat methods. 2009;6: 677-681. 15. Chen WK, Swartz JD, Rush LJ, Alvarez CE. Mapping DNA structural variation in dogs. Genome Res. 2009;19:500-509. 16. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6:80-92. 17. Clutton-Brock J. Domesticated animals: from early times, Heinemann in assoc. with British Museum (natural history), London; 1981. 18. Clutton-Brock J. Originsofthedog:domesticationandearlyhistory.In: Serpell J, editor. The domestic dog: its evolution, behaviour, and interactions with people. New York: Cambridge University Press. p7–20;1995. 19. Colledge S, Conolly J, Shennan S, Bellwood P, Bouby L, Hansen J, et al. Archaeobotanical evidence for the spread of farming in the Eastern Mediterranean 1. Curr Anthropol. 2004; 45: S35–S58. 20. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464:704-712. 21. Cruz F, Vilà C, Webste MT. The legacy of domestication: accumulation of deleterious mutations in the dog genome. Mol Biol Evo. 2008; 25:2331-2336 22. Elferink MG, Vallée AA, Jungerius AP, Crooijmans RP, Groenen MA. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. BMC genomics. 2008;9:1-9. 23. Fang M, Larson G, Ribeiro HS. Contrasting mode of evolution at a coat color locus in wild and domestic pigs. PloS

Gene. 2009;5. e1000341. 24. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2011;40:D84-D90. 25. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet 2014;10: e1004016. 26. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 2007; 3:e3. 27. Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y, et al. The effect of strand bias in Illumina short-read sequencing data. BMC genomics. 2012;13:666. 28. Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, et al. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat Genet. 2007;39:1318-1320. 29. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44-57. 30. Kim H, Sung S, Cho S, Kim TH, Seo K, Kim H. VCS: Tool for Visualizing Copy Number Variation and Single Nucleotide Polymorphism. Asian Austral J Anim. 2014;27:1691-1694. 31. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754-1760. 32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078-2079. 33. Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. Domestication of the dog from the wolf was promoted by enhanced excitatory synaptic plasticity: a hypothesis. Genome Biol Evol. 2014;6:3115-3121. 34. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438:803–819. 35. Lord K. A comparison of the sensory development of wolves (Canis lupus lupus) and dogs (Canis lupus familiaris). Ethology. 2013;119:110-120. 36. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40:1166-1174. 37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a mapreduce frame work for analyzing next-generation dna sequencing data. Genome Res. 2010;20:1297-1303. 38. Molin AM, Berglund J, Webster MT, Lindblad-Toh K. Genome-wide copy number variant discovery in dogs using the CanineHD genotyping array. BMC genomics. 2014;15:1-10. 39. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. PLoS One. 2013;8:e77910. 40. Nicholas TJ, Baker C, Eichler EE, Akey JM. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. BMC Genomics. 2011;12:414. 41. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Res. 2009;19:491-499. 42. Pailhoux E, Vigier B, Chaffaux S, Servel N, Taourit S, Furet JP, et al. A 11.7-kb deletion triggers intersexuality and polledness in goats. Nat Genet. 2001;29:453-458. 43. Polgár Z, Kinnunen M, Újváry D, Miklósi Á, Gácsi M. A Test of Canine Olfactory Capacity: Comparing Various Dog Breeds and Wolves in a Natural Detection Task. PloS one. 2016;11: e0154087. 44. Przezdziecki XJ B, Paris G. Our levriers: the past, present and future of all sighthounds. Les Amis de Xavier Przezdziecki, La Colle-sur-Loup. France; 2001. 45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-842. 46. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs

identified and characterized by next generation sequencing technology. PloS one. 2009;4: e6524. 47. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444:444-454. 48. Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464:587-591. 49. Ruden DM, Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012;3:35 50. Stothard P, Choi JW, Basu U, Sumner-Thomson JM, Meng Y, Liao X, et al. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. BMC genomics. 2011;12:1-14. 51. Thompson PD. Cardiovascular Adaptations to Marathon Running. Sports Med. 2007; 37:444-447. 52. Tonoike A, Hori Y, Inoue-Murayama M, Konno A, Fujita K, Miyado M, et al. Copy number variations in the amylase gene (AMY2B) in Japanese native dog breeds. Anim Genet. 2015;46:580-583. 53. Turner FS. Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. Front Genet. 2014;5:5. 54. van Heesch S, Kloosterman WP, Lansu N, Ruzius FP, Levandowsky E, Lee CC, et al. Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC genomics. 2013;14:257. 55. VonHoldt BM et al. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature. 464:898–902. 56. Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, et al. Multiple and ancient origins of the domestic dog. Science. 1997;276:1687-1689. 57. Wang GD, Zhai W, Yang HC, Fan RX, Cao X, Zhong L, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat Commun. 2013;4:1860. 58. Wang GD, Zhai W, Yang HC, Wang L, Zhong L, Liu YH, et al. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Res. 2016;26:21–33. 59. Wang GD, Shao XJ, Bai B, Wang J, Wang X, Cao X, et al. Structural variation during dog domestication: insights from gray wolf and dhole genomes. Natl Sci Rev. 2018;6:110-122. 60. Wayne RK. Molecular evolution of the dog family. Trends Genet. 1993;9:218-224. 61. Wright D, Boije H, Meadows JR, Bed'Hom B, Gourichon D, Vieaud A, et al. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet. 2009;5: e1000512. 62. Xu L, Hou Y, Bickhart DM, Zhou Y, Song J, Sonstegard TS, et al. Population-genetic properties of differentiated copy number variations in cattle. Sci Rep-uk. 2016;6:23161 63. Yan Y, Yi G, Sun C, Qu L, Yang N. Genome-wide characterization of insertion and deletion variation in chicken using next generation sequencing. PloS one. 2014;8: e104652. 64. Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC genomics. 2014;15:962. 65. Zeder MA. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci U S A. 2008;105:11597-11604. 66. Zhang H, Meltzer P, Davis S. RCircos: an R package for Circos 2D track plots. BMC bioinformatics. 2013;14: 244. 67. Zhang HH, Wei QG, Zhang HX, Chen L. Comparison of the fraction of olfactory receptor pseudogenes in wolf (Canis lupus) with domestic dog (Canis familiaris). J For Res. 2011; 22:275-280. 68. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18: 292-298. 69. Zhang X, Wang K, Wang L, Yang Y, Ni Z, Xie X, et al. Genome-wide patterns of copy number variation in the Chinese yak genome. BMC genomics. 2016; 17:379.

# Tables

Table 1-Size distribution of CNVRs detected by CNVnator

| Summary statistic of CNVRs | Gain | Loss | Both (loss and gain) | Total |
|---|---|---|---|---|
| Number of CNVRs | 3916 | 6400 | 255 | 10571 |
| Total length(Mb) | 83.75 | 47.28 | 23.62 | 154.65 |
| Mean length(Kb) | 21.39 | 7.39 | 92.62 | 14.63 |
| Median length(Kb) | 11.70 | 4.49 | 38.99 | 7.05 |
| 1≥ Kb to <5 Kb | 555 (14.17%) | 60 (0.94%) | - | 3996 (37.80%) |
| 5≥ Kb to <10 Kb | 1119 (28.57%) | 3441(53.76%) | 14 (5.49%) | 2706 (25.59%) |
| 10≥ Kb to <20 Kb | 1160 (29.62%) | 1573 (24.57%) | 45 (17.64%) | 2252 (21.30%) |
| 20≥ Kb to <50 Kb | 750 (19.15%) | 1047 (16.35%) | 189 (74.11%) | 1123 (10.62%) |
| 50≥ Kb | 332 (8.47%) | 279 (4.35%) | 7 (2.74%) | 494 (4.67%) |

Table 2 - Sampling location and ecotypes

| The **latitude and longitude** of each location | Ecotype | Location | Sample ID | Sample |
|---|---|---|---|---|
| N, 46 59′ 32″ E 35 18′ 52″ | Saluki ((Tazi | Sanandaj, Iran | DogSI1 | Dog |
| N, 47 36′ 10″ E 35 52′ 22″ | Saluki ((Tazi | Bijar, Iran | DogSI2 | Dog |
| 32 38′ 0″ N, 51° 39′ 0″ E | Qahderijani | Esfahan, Iran | DogQI | Dog |
| N, 48° 31′ 0″ E 34 48′ 0″ | - | Hamadan, Iran | GW1 | Wolf |
| N, 51 25′ 23″ E 35 41′ 46″ | - | Tehran, Iran | GW2 | Wolf |
| N, 57 5′ 0″ E 30 17′ 0″ | - | Kerman, Iran | GW3 | Wolf |

# Additional Files

**Additional file 1:** Suplementary.doc. Included Tables S1-S15 and Figures S1- S18

**Additional file 2:** Table S16. Genomic structural variants including insertions, deletions, translocations (inter and *intra* chromosomal) and inversions for three dogs.

**Additional file 3:** Table S17. Genomic structural variants including insertions, deletions, translocations (inter and *intra* chromosomal) and inversions for three wolves.

**Additional file 4:** Table S18. The total number of deletions, insertions, inversions, inter chromosomal translocations and intra chromosomal translocations, across the 6 individuals.

**Additional file 5.** Table S19. The total number of CNVRs

**Additional file 6:** Table S20. S*tatistics* CNVs for Canine autosomes and X chromosome.

**Additional file 7:** Table S21. Comparison with previous dog CNV studies

# Figures

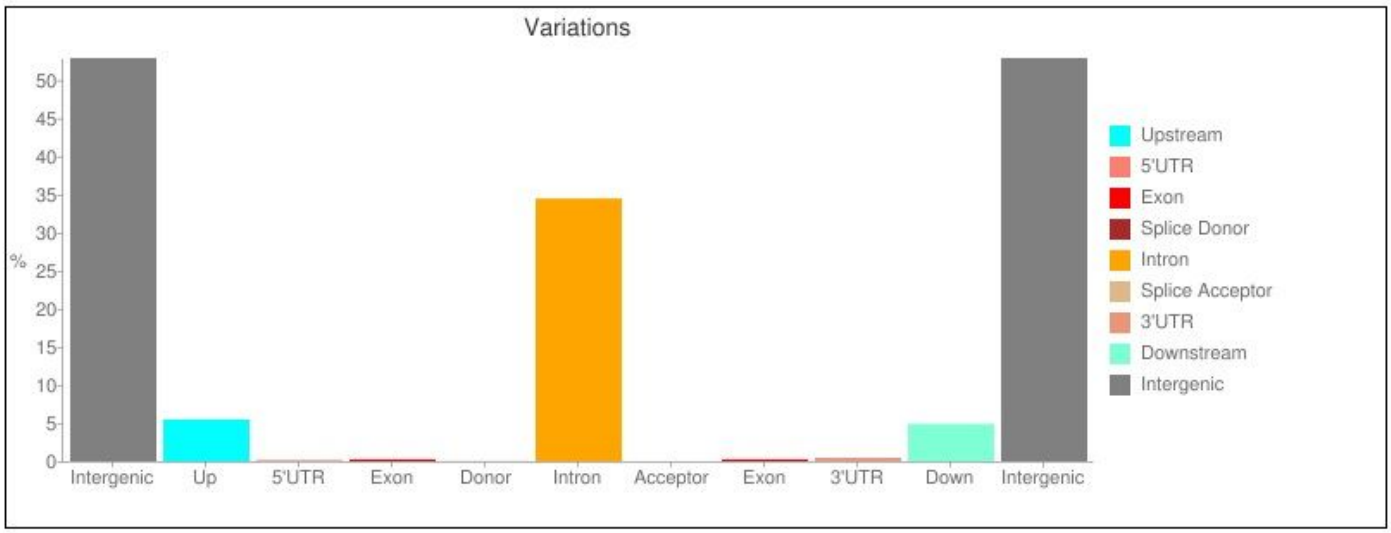
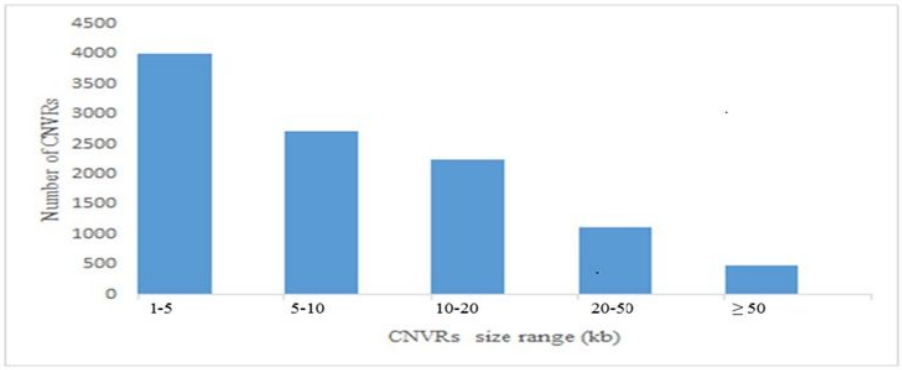
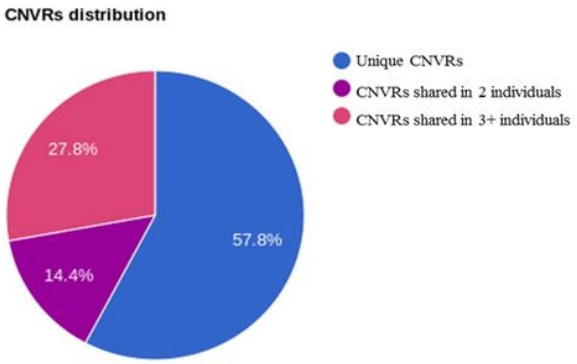



**Figure 1**

Number of SNP effects in different regions of genome.



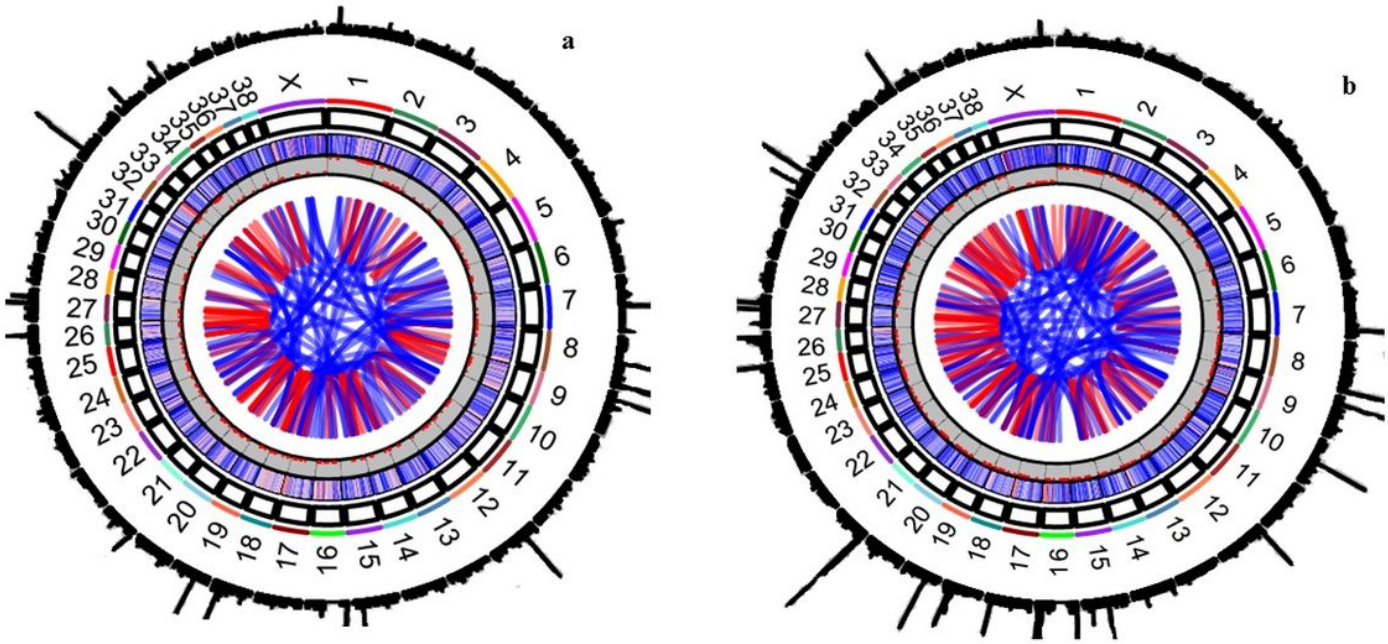

A




B

## Figure 2

The length and distribution of CNVRs. (a) a total of 6702 (63.39%) and 494 (4.67%) out of all CNVRs had sizes ranging from 1.049 to 10kb and longer than 50 kb in size, respectively. (b) 4466 (42.25%) CNVRs are shared in at least two individuals and 6105 (57.75%) CNVRs present in only one individual.




## Figure 3

Graphical visualization of predicted SVs for dog (a) and wolf (b). Starting from outside of the circle, the following features are shown: genomic locations of Indels, chromosome ideograms, heatmap plot of copy number variation with color according to the CNV value computed by CNVnator, genomic locations of inversions and genomic locations of intra (size > 1000 bp) and inter- chromosomal links (read coverage >=20, the score>=90).


## Figure 4

Distribution of CNVRs on chromosomes 1-38 and X chromosome.