Integrative genotyping and analysis of canine structural variation using long-read and short-read data

Peter Z. Schall[1] and Jeffrey M. Kidd[1]

[1]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA; Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

***Corresponding author:**

Jeffery M. Kidd, Ph.D.

jmkidd@umich.edu

734-764-6922

Short Title: Canine Structural Variants

Keywords: structural variants, long-read sequencing, mobile elements

Abstract

Structural variation makes an important contribution to canine evolution and phenotypic differences. Although recent advances in long-read sequencing have enabled the generation of multiple canine genome assemblies, most prior analyses of structural variation have relied on short-read sequencing. To offer a more complete assessment of structural variation in canines, we performed an integrative analysis of structural variants present in 12 canine samples with available long-read and short-read sequencing data along with genome assemblies. Use of long-reads permits the discovery of heterozygous variation that is absent in existing haploid assembly representations while offering a marked increase in the ability to identify insertion variants relative to short-read approaches. Examination of the size spectrum of structural variants shows that dimorphic LINE-1 and SINE variants account for over 40% of all deletions and insertions and identified 1,410 LINE-1s with intact open reading frames that show presence-absence dimorphism. Using a graph-based approach, we genotype newly discovered structural variants in an existing collection of 1,879 resequenced dogs and wolves, generating a variant catalog containing a 56.5% increase in the number of deletions and 705% increase in the number of insertions previously found in the analyzed samples. Examination of allele frequencies across admixture components present across breed clades identified 283 structural variants evolving with a signature of selection.

Significance statement

Existing studies of structural variation have focused on genomes sequenced with short-read technology. In this study, we systematically combine short-read and long-read data sets to identify and genotype structural variation across canines, resulting in an expanded catalog of structural variants. Analysis of allele frequencies identified structural variants in genic regions that may have been evolving under selection across breed clades.

Introduction

Structural variation, operationally defined as insertions, deletions, duplications, or inversions of genome sequence at least 50 base pairs (bps) in length, is a critical component of genome variation (Escaramis et al., 2015). Although fewer in number than single nucleotide polymorphisms (SNPs), structural variants (SVs) typically affect a larger percentage of the genome sequence (Pang et al., 2010, Genomes Project et al., 2015), and have been linked as a source of genetic diversity underpinning genome evolution, environmental adaptation, and differences in gene expression (Hujoel et al., 2022, Ho et al., 2020).

The past 15 years have seen marked improvement in algorithms to identify and genotype SVs, mostly focusing on widely prevalent short-read data from the Illumina platform (Mahmoud et al., 2019). These approaches rely upon split-read and discordant read-pair signatures obtained from aligning short-read data to an existing genome reference assembly (Ho et al., 2020, Kosugi et al., 2019). While short-read technologies, typically paired-end reads of ~150 bp in length, can readily identify SV breakpoints, fully resolving insertions longer than the read length remains a challenge. Identifying structural variation in genomic regions where short-reads cannot be unambiguously aligned (such as within duplications or regions with high repeat content) poses additional problems (Ho et al., 2020). With the recent advent of long-read based sequencing from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), sequenced reads can now reach lengths in the 100s of kb. This massive increase in read length has allowed for the closing of gaps in genome assemblies, enabling the identification of variants within these regions (Aganezov et al., 2022), and the full resolution of large SVs (Amarasinghe et al., 2020). Comparisons in humans and other species have shown that long-read approaches offer a more complete view of SVs, particularly of insertions longer than Illumina read lengths (Kosugi et al., 2019). Due to cost and library preparation requirements, short-read approaches continue to dominate large-scale analyses of genomic diversity. However, the development of graph-based analysis tools enables the genotyping of SVs identified from long-read data in samples with existing short-read information, offering a compelling analysis paradigm for a more comprehensive assessment of SVs (Yan et al., 2021, Chen et al., 2019).

Canines (dogs and wolves) have emerged as a powerful model for studies of genome evolution and disease. The differences in size, shape, coloration, and behavior among modern dogs makes canines an attractive system for studying the action of selection and for untangling the molecular basis of complex phenotypes (Boyko, 2011, Karlsson and Lindblad-Toh, 2008). Naturally occurring genetic variation in dogs is associated with traits, including morphological and behavior differences that have been the subject of selection (Shearin and Ostrander, 2010, Schlamp et al., 2016) as well as differences in longevity and susceptibility to multiple diseases, including cancer (Fleming et al., 2011). The canine genomics community has developed rich resources including the creation of multiple reference assemblies generated from long-read sequencing technologies (Dreger et al., 2016, Plassais et al., 2019, Freedman et al., 2016, Schall et al., 2023, Edwards et al., 2021, Sinding et al., 2021, Wang et al., 2021, Field et al., 2020, Field et al., 2022, Jagannathan et al., 2021, Player et al., 2021, Halo et al., 2021).

SVs are known to make important contributions to canine traits. Examples include deletions and duplications associated with body weight in Labrador Retrievers (Antkowiak and Szydlowski, 2023), a 133-kb duplication predisposing Ridgebacks to dermoid sinus and causing the distinctive hair ridge (Salmon Hillbertz et al., 2007), a small insertion causing globoid cell leukodystrophy Irish Setters (McGraw and Carmichael, 2006), a *FGF4* retrogene insertion causing the short-legged phenotypes found

in many breeds (Parker et al., 2009), and the well characterized copy-number variation of *AMY2B* leading to adaptation to a starch-rich diet (Axelsson et al., 2013).

Previous studies have found that mobile element insertions make a major contribution to genomic differences among dogs (Wang and Kirkness, 2005, Halo et al., 2021). Mobile elements (also called transposable elements) are DNA sequences which can move into new genomic locations. Retrotransposons, the most abundant type of mobile element in mammalian genomes, mobilize through a "copy-and-paste" mechanism that uses an RNA intermediate that is reverse transcribed and integrated into a new genomic location (Kazazian and Moran, 2017, Levin and Moran, 2011). Retrotransposons can be further subdivided into two classes, those that contain long terminal repeats (LTRs) and those that do not. Non-LTR retrotransposons include Long INterspersed Elements (LINEs), which are autonomous retrotransposons that encode the proteins required for their mobilization, and Short INterspersed Elements (SINEs), which are non-autonomous retrotransposons that require protein(s) encoded by LINEs for mobilization. LINE-1, the autonomous retrotransposon family active across mammals, is approximately 6 kb in size and encodes two proteins, ORF1p and ORF2p, required for its own mobilization (Beck et al., 2011).

Most copies of mobile elements in the genome were inserted many millions of years ago. These sequences have acquired multiple mutations that make them incapable of mobilization and are "fixed" in the species, meaning they are found in the same genomic location in all individuals. However, some elements that were inserted more recently in evolutionary time show presence-absence variation, known as dimorphism, among individuals. These variants manifest as insertions and deletions in a sample relative to a reference genome. Several previous studies have linked mobile element insertions to canine phenotypes, including cerebellar ataxia, retinal atrophy, and merle coloration (Hitti-Malin et al., 2020, Mauri et al., 2017, Langevin et al., 2018). A recent study identified a *de novo* LINE-1 mutation as causing X-linked muscular dystrophy in a Border Collie dog, confirming that at least one type of retrotransposon remains active in the canine genome (Van Poucke et al., 2024).

In this study, we perform a comprehensive analysis of SVs in 12 canine samples with available long-read and short-read sequencing data along with genome assemblies. We compare the ability of each modality to detect SVs and analyze the sequence content of identified variants, finding that dimorphic LINE-1 and SINE variants account for over 40% of all deletions and insertions. Finally, using a graph-based approach, we genotype newly discovered SVs in an existing collection of resequenced canine genomes analyzed by the Dog10K consortium and identify SVs with frequency patterns consistent with non-neutral evolution among clades of breed dogs, including several mobile element insertions. In total, this manuscript integrates long-read and short-read genome data to provide a comprehensive view of structural variation in canines.

Materials and Methods

Included samples with long-read sequencing and assembled genomes

We identified a total of 12 samples that had accessible long-read sequencing data and associated assembled genomes (Supplemental Table S1). The analyzed samples include 10 samples from 7 dog breeds: two Basenjis (Edwards et al., 2021), two Bernese Mountain Dogs (Schall et al., 2023), two German Shepherds (Wang et al., 2021, Field et al., 2020), a Cairn Terrier (Schall et al., 2023), a Boxer

(Jagannathan et al., 2021), a Great Dane (Halo et al., 2021), a Labrador Retriever (Player et al., 2021, Schall et al., 2023); a Dingo (Field et al., 2022); and a Greenland Wolf (Sinding et al., 2021). Illumina short-read sequencing was also available for each sample, except one Basenji (China). A detailed summary of the analyses undertaken in this manuscript can be found in Supplemental Figure S1.

Canine structural variant calling

Structural variants were identified from Oxford Nanopore Technologies (ONT) and PacBio long-reads using a method adapted from (Lemay et al., 2022). Reads were aligned to the UU_Cfam_GSD_1.0_ROSY genome reference with minimap2 (v2.26, RRID: SCR_018550) (Li, 2018) with standard settings, converting SAM output to BAM with SAMtools (v1.13, RRID:SCR_002105) (Li et al., 2009). Due to the incomplete nature of the assembled Y chromosome, our results are limited to assembled chromosomes (chr1-38 and X). Initial structural variants were called using sniffles2 (v2.4, RRID:SCR_017619) (Smolka et al., 2024) with a maximum deletion length of 10 Mb. To increase sensitivity, the coordinates of known tandem repeats were downloaded from the UCSC Genome Browser (RRID:SCR_005780) (Meyer et al., 2013) and included in the sniffles2 analysis. For each sample, a filtering step was applied removing: BND variants (breakend variants; where the SV breakpoints are not fully resolved), homozygous reference variants, insertions with unassembled or gap sequences ('N'), variants less than 50bp, and variants greater than 10Mb. Subsequently, the filtered VCF files were sorted, normalized, and filtered for duplicate entries with BCFtools (v1.12, RRID:SCR_005227) (Danecek et al., 2021). Refinement of SV breakpoints and sequences were conducted with the R script breakpoint_refinement.R (Lemay et al., 2022). In short, initial local assemblies were generated for each variant with wtdbg2 (v2.5, RRID:SCR_017225) (Ruan and Li, 2020) using reads mapped +/- 200bp from each SV followed by realignment of the same reads to each assembled sequence with minimap2, and the local assembly aligned to the region of the reference genome with AGE (v0.4, RRID:SCR_005253) (Abyzov and Gerstein, 2011). Since Nala and Sandy have both ONT and PacBio long-read sequencing data, an additional step of processing was applied to create a single long-read callset for each sample: the post-breakpoint refinement VCF files were merged with SURVIVOR (v1.0.7, RRID:SCR_022995) (Jeffares et al., 2017), with the following flags: 100 0 1 0 0 50. To facilitate genotyping SVs with short-read data, the final VCF files for each of the 12 samples were ported to SVmerge from the SVanalyzer package (v0.36) (https://github.com/nhansen/SVanalyzer) with the following flags: *-maxdist 15 -reldist 0.1 -relsizediff 0.1 -relshift 0.1 -seqspecific*. The SVmerge output VCF was processed with the R script merge_realigned_variants.R (Lemay et al., 2022), wherein SVs were selected from clusters, resulting in the final representative SV per loci.

Structural variants were also identified using Illumina short-read data. Read pre-processing and alignment was conducted using the dogmap pipeline (https://github.com/jmkidd/dogmap) (Meadows et al., 2023). In summary, the pipeline consists of mapping with bwa-mem2 (v2.1, RRID: SCR_022192) (Li, 2013), followed by marking and sorting duplicates with MarkDuplicates, base score recalibration with BaseRecalibrator and ApplyBQSR, from the GATK suite of tools (v4.2, RRID:SCR_001876) (DePristo et al., 2011). Structural variants were identified from the resulting alignment files using Manta (v1.6.0, RRID:SCR_022997) (Chen et al., 2016), using standard settings. The initial Manta SV calling of Sandy failed to complete, with further investigation noting an aberrant insert-size distribution. This necessitated an additional filter to remove those reads with an apparent fragment size less than 20; this filtering ameliorated the issue. Manta SV calls were merged with svimmer (v0.1) (https://github.com/DecodeGenetics/svimmer), with standard settings, and genotyped using

GraphTyper2 (v2.7.7) (Eggertsson et al., 2019). Additionally, to resolve large insertions that could not be rectified by the short-read sequencing data, structural variant genotypes were determined for the variants identified by long-reads (the output of SVmerge described above) from mapped Illumina data using Paragraph (v2.4a) (Chen et al., 2019). Within each sample, the output of the short-read manta SV calls and paragraph genotyping of long-read variants was merged with SURVIVOR, as previously described, filtering to include unique SVs only and retaining the genotype calls from Paragraph when SVs matched.

The third method for calling SVs consisted of comparing the published assemblies for each sample to the UU_Cfam_GSD_1.0_ROSY genome assembly. The genome alignment was conducted using minimap2 (v2.26) using the following settings: *-a -x asm5 --cs -r2k*. Resultant alignment output was sorted and indexed using SAMTools. Structural variants were called using the haploid caller from SVIM-asm (v1.0.3) (Heller and Vingron, 2021) with standard settings.

In summary, three SV call sets were generated for each sample compared to UU_Cfam_GSD_1.0_ROSY: long-read, short-read, and assembly. We additionally used short-read data to genotype SVs discovered by long-reads. To identify concordance across the three methodologies, within sample, the VCF files were merged with SURVIVOR. VCF files were then converted into table format to enabling porting to R (v4.2.0, RRID:SCR_001905) (Dessau and Pipper, 2008), for processing and plotting. Downstream analyses and annotations were further limited to insertions and deletions, due to the higher confidence in calling these events compared to the other SV types.

Annotation of genomic regions and structural variants

SVs were annotated using SnpEff (v5.2c, RRID:SCR_005191) (Cingolani et al., 2012), using the UU_Cfam_GSD_1.0 genomic data retrieved from Ensembl (Dyer et al., 2025), version 113, deriving genomic locations (exons, introns, intergenic, etc.) and applicable intersecting gene information. Only SVs that intersected with exons, or entire feature ablation, were considered to affect genes.

Repetitive elements within SVs were identified using blastn (v2.10, RRID:SCR_001653) (Ladunga, 2017), comparing canine and ancestral species repeat consensus sequences retrieved from RepBase (v29.05, RRID:SCR_021169) (Bao et al., 2015). The output was then annotated to include SV length, repeat sequence length, and reciprocal percent sequence overlap between repeat sequences and SVs. Aiming to identify SVs that correspond to dimorphic repeats, we identified SVs where 1) ≥95% of the SV corresponded to a given repeat as well as 2) SVs with a 95% reciprocal overlap with a given repeat, representing near-full length dimorphic repeats. Furthermore, insertions and deletions identified as putative full-length Long INterspersed Elements (LINEs) were analyzed using BLASTX (v2.10, RRID:SCR_001653) (Ladunga, 2017) to identify intact open-reading frames for ORF1p (QOV08756.1) and ORF2p (QOV08757.1), requiring >=99% amino acid identity.

Genotyping structural variants and selection analysis with the Dog10K dataset

The Dog10K dataset (Meadows et al., 2023) is a repository of short-read sequenced samples from nearly 2000 canines, covering well characterized breed dogs, mixed breeds, free breeding village dogs, and wolves. The output of the long-read SV pipeline, post SVmerge, was used for genotyping structural variants from the Dog10K dataset with Paragraph, with standard settings. The Dog10K dataset also included previously generated SV calls from Manta, which were retrieved and merged with the

Paragraph output, as described above. A total of 1,879 samples were used that previously passed all variant calling filters.

In the Dog10K manuscript, SNPs were used to identify signals of selection in the ancestral admixture components from 790 breed dogs that were categorized into nine broad clades (Belgian Herders, Mastiffs, Pointers, Scenthounds, Sighthounds, Spaniels, Spitz, UK Herders, and Waterdogs) using Ohana (Cheng et al., 2022). We replicated the initial generation of the allele frequency model based on the Dog10K SNP data, using the same breed-clades and number of ancestral components (K=5), and filtering parameters: minor allele frequency of 5% and no missing genotypes. We then applied the SNP-based model to the union of Manta and Paragraph SV genotypes. It should be noted that increased filtering metrics were required for the SV analysis in the Dog10K manuscript, resulting in a set of 781 dogs used for this present analysis. We applied thresholds to identify variants with allele frequency patterns inconsistent with neutral evolution as defined by Ohana. First, we applied a strict threshold of $1.67*10^{-7}$, corresponding to a nominal significance level of 0.05 after a Bonferroni correction accounting for the 299,116 deletion and insertions tested. Since the Bonferroni correction may be overly conservative as variant genotypes are not independent, we also applied a reduced threshold of 0.0001. A total of 283 variants passed this threshold, corresponding to the top 0.096% of all SVs evaluated.

Generation of Figures

Unless otherwise noted, all figures were produced in R (v4.2.0), using a combination of the following R packages: ggplot2 (v3.5.1, RRID:SCR_014601) (https://ggplot2.tidyverse.org), ggpubr (v0.6.0, RRID:SCR_021139) (https://github.com/kassambara/ggpubr), ggforce (v0.4.1, RRID:SCR_022575) (https://github.com/thomasp85/ggforce), ggrepel (v0.9.5, RRID:SCR_017393) (https://github.com/slowkow/ggrepel), ggnewscale (v0.4.9) (https://doi.org/10.5281/zenodo.2543762), ComplexHeatmap (v2.20.0, RRID:SCR_017270) (Gu et al., 2016), and paletteer (v1.6.0) (https://github.com/EmilHvitfeldt/paletteer).

Results

Canine structural variant detection

We identified structural variants from 12 canine samples using three modalities: genome assembly, long-read sequencing, and short-read sequencing (Figure 1, Supplemental Table S1). For increased readability, we refer to the different samples and assemblies by the name of the sequenced individual. All structural variants were identified compared to the UU_Cfam_GSD_1.0_ROSY (Mischka) genome, which is derived from a female German Shepherd Dog supplemented by partially assembled Y-chromosome sequences from a Labrador retriever. The Dog10K Consortium used the same assembly (Meadows et al., 2023). A detailed flowchart displaying the pipeline used for this manuscript can be found in Supplemental Figure S1. Regardless of input datatype, deletions and insertions were the most frequent class of detected structural variants, accounting for 52.7% and 46.8% of the total calls, respectively.

Structural variation identified by assembly comparison

The number of structural variants identified by assembly comparison, which due to the self-comparison resulted in zero Mischka calls, is dominated by deletions (mean=21,509) and insertions (mean=20,712), followed by inversions (mean=68), and duplications (mean=18) (Table 1, Figure 2 Panel A, Supplemental Figure S2, Supplemental Table S2). Querying the total number of impacted base pairs by sample found

an average of 21.9 Mb for deletions, 12.2 Mb for insertions, 4.6 Mb for duplications, and 17.4 Mb for inversions (Table 1, Figure 2 Panel B, Supplemental Figure S2, Supplemental Table S2). The length of SVs for both insertions and deletions displayed a similar distribution, with a mostly lefthanded dominance showing a prominent peak at ~200-250bp and a smaller peak at ~6,000bp (Figure 2 Panel C). The length distribution of both duplications and inversions displayed a lefthand bias (Supplemental Figure S2). SVs were filtered to include those confirmed by at least one other modality to avoid assembly errors, intersected with coding regions, and limited to genes with an annotated gene symbol. Deletions intersected with a median of 42 genes per sample (range=4-66 genes). Of the total 127 intersecting genes, 37.8% were variant in a single sample while only one, *ZCCHC10*, contained a deletion overlapping with an exon in all samples, excluding the self-comparison of Mischka. Total genes with insertions were less, numbering 70, with a median of 13 per sample (range=1-20 genes). Approximately 57.1% of the genes were variable in a single sample, and none were present in all samples (Supplemental Table S4).

Structural variation identified from long-reads

Similar to assembly-based calls, the greatest number of SVs identified from long-reads were deletions (mean=28,148) and insertions (mean=30,798), trailed by inversions (mean=205) and duplications (mean=26) (Table 1, Figure 2, Supplemental Figure S2, Supplemental Table S2). Average impacted base pairs by the different types of SVs totaled 12.4 Mb for deletions, 26.1 Mb for insertions, 13.8 Mb for inversions, and 2.1 Mb for duplications. The increase in the number of SVs from long-read-based calls compared to assembly comparison can largely be attributed to heterozygous SVs. On average, 51% of deletions (range=25-99%) and 65% of insertions (range=45-98%) were heterozygous (Supplemental Figure S3, Supplemental Table S3). As expected, the greatest proportion of heterozygous calls were found in Mischka. Zoey also displayed an increased number of heterozygous calls, possibly due to long-read data quality. The length distribution of deletions, insertions, and inversions were in line with assembly calls; however, the peak of the duplications was shifted rightward.

A median of 76 genes were impacted by deletions per sample (range=15-93 genes), with a total of 241 genes affected across the samples. Approximately 42% of the impacted genes were deleted in only a single sample. Only two genes were found to have exonic deletions in all samples: *ZCCHC10* and *EEF1A1.* A median of 32 genes per sample contained exonic insertions (range=3-157 genes), affecting a total of 364 genes. The majority (84.15%) of the genes were variable in a single sample and none present in all samples (Supplemental Table S4).

Structural variants identified using short-read sequencing

Short-read calling, which did not include China due the absence of Illumina data, was initially conducted using Manta. To better ascertain longer SVs, variants found by long-reads were genotyped using short-read from each data with the Paragraph software. The union of these discovered and genotyped SVs were defined as the short-read+Paragraph call set. The short-read deletions displayed a similar magnitude as previous methods (mean=20,121 per sample), with a reduced representation of insertions (mean=6,019), followed by duplications (mean=232), and inversions (mean=92) (Table 1, Figure 2, Supplemental Figure S2, Supplemental Table S2). The SV length distribution of deletions matched those resulting from assemblies and long-reads. Insertions displayed a similar distribution among shorter SVs while lacking longer, duplications were likewise similar to the assembly calls, while inversions displayed a much more flat and wider accounting of SV lengths. On average, 47% (range=32-93%) of deletions and 60% of insertions (range=4-97%) were heterozygous (Supplemental Figure S3, Supplemental Table S3). Sandy, the sample that had an aberrant insert-size distribution, was the sample with the fewest proportion of heterozygous insertion calls, and overall, the fewest insertions.

While a number of statistics show relative agreement in overall SV capture by short-reads, the union of short-read+Paragraph showed increased sensitivity, specifically for longer SVs. This was most apparent for insertions, with a median percent increase of 243% identified SVs, which had a mean length of 515bp (Figure 3, Supplemental Table S2). While less of a dramatic increase in SVs, deletion calls saw a median increase of 27%, and duplications and inversions remained relatively the same. The overall impacted base-pairs increased for deletions (mean=13.95 Mb, range=6.1-13.5 Mb) and insertions (mean=9.9 Mb, range=4.9-10.9 Mb). The distribution of heterozygous calls was in-line with other sequencing modalities with deletions averaging 51% and insertions 66%. A total of 128 genes had exonic overlapping deletions and 70 genes with insertions. The median number per sample of genes with exonic deletions was 36 (range=9-73 genes) and insertions numbered 11 (range=3-28 genes). Similar to the other modalities, a substantial proportion of deletion and insertion intersecting genes were singletons, 42.3% and 64.3%, respectively (Supplemental Table S4).

Comparison of SVs by modality

Within each sample, SV concordance was assessed as the fraction of shared deletions and insertions across sequencing modality, e.g., long-read vs. assembly, long-read vs. short-read, long-read vs. short-read+Paragraph (Figure 4, Supplemental Table S5). We found that on average 88% of deletions and 84% of insertions discovered by comparing genome assemblies were also found in the long-read data. Conversely, on average 64% of deletions and 53% of insertions discovered in the long-read data were also found in the genome assembly comparisons. Across the samples, a total of 336,121 long-read SVs were missing in the assembly calls, of which 81.5% were heterozygous. The relatively tight grouping of both deletions and insertions found in long-read vs assembly comparison was not found when comparisons involved calls from short-read or short-read+Paragraph, with short-read modalities consistently identifying a lower fraction of insertions. However, with the inclusion of Paragraph SV calling, there is an average increase in insertion agreement compared to assembly (4%) and long-read (19%) data (Figure 4).

Repeat annotation of SVs

All identified deletion and insertion sequences were extracted and annotated for matches to canine and ancestral repeat sequences from RepBase using blastn. The SV-repeat data was then filtered to identify those variants with at least 95% of length annotated as a given repeat. These SVs correspond to putatively variable repeats, as opposed to larger SVs that incidentally contain repetitive sequence. This filtered dataset contained a total of 18,082 deletions and 69,103 insertions, with a mean per sample of 5,492 deletions and 6,690 insertions (Figure 5). As expected, the repeat annotation was dominated by SINEC2A1 sequences, accounting for 72.5% of deletions and 67.4% of insertions that meet the criteria and 33.8% and 10.7% of the total deletions and insertions, respectively. The second largest category consisted of Long INterspersed Nuclear Elements-1 (LINE-1s), comprising 26.6% of deletions and 31.8% of insertions of the repeat annotated SVs, and 12.0% of total deletions and 5% of total insertions (Figure 5). Thus, variable SINEC and LINE-1 sequences account for 45.8% and 42.5% of all deletion and insertion structural variants.

The SV-repeat data was further filtered to identify variants putatively corresponding to full-length LINE-1s by requiring at least 95% reciprocal overlap with the element consensus sequence, which were primarily comprised of L1_CF sequences. A mean of 332 deleted and 709 inserted full-length LINE-1s were found across samples. Full-length LINE-1 sequences were further queried to identify those with intact open reading frames for the ORF1p and ORF2p proteins (>=99% sequence identity), splitting by SV type (Figure 6). A total of 1,410 variable LINEs with intact ORFs were identified with a sample median of

205 LINE-1s (Supplemental Table S6). The majority of the intact LINE-1s (82.1%) were present in a single sample, with 41.8% of the singletons originating from mCanLor, the Greenland Wolf. In fact, mCanLor had the largest number of intact LINE-1 sequences derived from SV events (n=707, deletions=233, insertions=474). Excluding Mischka, the samples with the fewest LINE-1 insertions (BD, OD, CA611, Yella, and China), were sequenced on the ONT platform only, possibly due to sequencing errors that disrupted apparent ORFs.

Signals of selection in structural variant allele frequencies

We leveraged the Dog10K dataset (Meadows et al., 2023), which consists of short-read sequencing of nearly 2,000 canines, comprised of characterized breed dogs, free breeding village dogs, and wolves, to assess frequency patterns of structural variation. We applied the Paragraph method to genotype the identified SVs in the Dog10K data, thus offering genotype information for variants discovered using both short-read and long-read data. Similar to the increase seen previously when applying the Paragraph method, we found a 56.5% increase in the number of deletions and 705% increase in the number of insertions relative to the Manta-only short-read calls (Figure 7).

We next used the comprehensive catalog of insertion and deletion variants genotyped in 1,867 canines (breed dogs=1575, village dogs=237, wolves=55) to search for allele frequency profiles indicating SVs that may be evolving under selection. Previously, the Dog10K consortium identified single nucleotide variants that may have been evolving under selection among major breed clades (Meadows et al., 2023). This analysis was performed using Ohana, a program which models allele frequencies among ancestry components to identify variants with frequency profiles that are not consistent with genome-wide profiles (Cheng et al., 2022).

Following the procedure used by the Dog10K consortium, we analyzed 781 breed dogs across nine clades (Belgian Herders, Mastiffs, Scenthounds, Sighthounds, Spaniels, Spits, Pointers, UK Herders, and Waterdogs). Using SNP genotype data, we replicated the ancestry proportions identified previously and found five ancestral components that largely corresponded to specific breed clades: 0 = Collie & Shetland Sheepdog (within the UK Herding clade), 1 = Spitz, 2 = Scenthounds, 3 = Mastiffs, and 4 = Pointers & Spaniels (Supplemental Figure S4 and Figure 8). We then fit the Ohana model to SV genotype data, from a total of 299,116 deletions and insertions, to identify SVs with allele frequencies that did not fit genome wide profiles identified using SNPs in each ancestral component. A total of 81 sites that exceeded a Bonferroni-corrected significance threshold ($1.67*10^{-7}$) were identified across the 5 ancestral components, although we note that this threshold is conservative since genotypes are not independent. The largest number of SVs were found in the ancestry component maximized in Collie & Shetland Sheepdog (n=64), followed by Mastiffs (n=10), Spitz (n=5), and Scenthounds (n=2) (Supplemental Table S7). For comparison, selection analysis was also repeated using SNP genotypes.

Specific breeds that were maximized within the selected clades included, Spitz: Japanese Spitz, Volpino Italiano, and the Pomeranian; Scenthounds: Billy, English Foxhound, Otterhound, Great Anglo-French White and Orange Hound, and Great Anglo-French Tricolour Hound; Mastiffs: Bull Terriers, English Bulldogs, Miniature Bull Terrier, and Boxer; Pointers & Spaniels: Sussex Spaniel, Ariège Pointer, Bourbonnais Pointing Dog, English Pointer, and Braque Français.

Several of the identified SVs overlap genes previously linked to breed specific traits or disorders. Of the 64 Collie & Shetland Sheepdog SVs, 34 were located within the introns of annotated genes. This included

389 a 7.8 kb intronic deletion within *NHEJ1*; this deletion has been previously associated with a disorder of
390 ocular development found in Collies and related breeds (Collie eye anomaly [*cea*]) (Parker et al., 2007)
391 (Figure 9, Supplemental Figure 5, Supplemental Table S8). We note that the identified deletion has a
392 more significant selection signal than other markers, including SNPs, in this region.

393 In the Spitz group, we found three SVs within introns of *HMGA2*, a gene which has been associated with
394 body weight (Rimbault et al., 2013), and ear type (Webster et al., 2015), although SNPs in the region also
395 show statistical evidence for selection (Supplemental Figure S6). SVs identified in the Mastiffs clade
396 intersected with the introns of three genes: *ABCB9*, *PITPNM2*, and *DENR*, that last of which was found to
397 have two intronic deletions and has been implicated in canine glioma susceptibility (Miller et al., 2019,
398 Truve et al., 2016) (Supplemental Figure S7). Each of the deleted sequences were identified as SINEC2A1
399 repeats with intact poly(A) tails and target site duplications. All four of the SINEC sequences are present
400 in the UU_Cfam_GSD_1.0 reference genome, derived from a German Shepherd. Consistent with the
401 recent origin of the SINE insertions, a high deletion frequency is found across canine populations, the
402 SINEs were absent in all available wolf samples, while found among Mastiffs. Interestingly, the two
403 polymorphic SINEs in *DENR* have approximately 96% sequence identity and are present in an opposite
404 orientation in adjacent introns, with one located near a coding exon, suggesting that they may affect
405 splicing (Supplemental Figure S8).

406 Relaxing the level of significance to p<=0.0001, increased the total number of identified SVs to 283. This
407 increased list included a Collie and Shetland Sheepdog intronic SV within *AP3B1*, where a homozygous
408 loss-of-function mutations have been associated with gray Collie syndrome (Benson et al., 2003), as well
409 as a large deletion covering multiple exons from *MX2. MX2* contributes to the innate immune defense
410 against viruses (Haller et al., 2018, Betancor, 2023) and regulates cell-cycle progression in melanoma
411 (Juraleviciute et al., 2020). Variation in *MX2* has also been linked to wool and skin traits in sheep
412 (Bolormaa et al., 2021) and increased expression of *MX2* has been found in canine atopic dermatitis
413 (Plager et al., 2012). Similarly, three intronic deletions were found below the relaxed threshold in the
414 Spitz group within *RCL1*, a gene implicated in snout ratio and curly tail phenotype (Vaysse et al., 2011).
415 Two additional Mastiff intronic SV deletions with signals of selection below the relaxed threshold were
416 found within *TRAF4* and *NSRP1*. Similar to the previous three genes identified with the Mastiffs clade,
417 each of the deleted sequences were identified as SINEC2A1 repeats with intact poly(A) tails and target
418 site duplications and not present within the mCanLor wolf reference genome (Supplemental Figure S9).
419 *TRAF4* plays an important role in bone development (Regnier et al., 2002) and promotes osteogenic
420 differentiation (Li et al., 2019). Overexpression of *TRAF4* has been implicated in the regulation of tumor
421 formation in osteosarcoma (Yao et al., 2014b, Yao et al., 2014a), a type of bone cancer that is common
422 in certain dog breeds, including Mastiffs (Edmunds et al., 2021). *NSRP1* is a splicing regulatory protein
423 (Kim et al., 2011). Homozygous loss-of-function alleles of *NSRP1* lead to a severe neurodevelopmental
424 disorder (Calame et al., 2021), while *NSRP1* function has been linked to adipogenesis and modulation of
425 weight in pigs (Liu et al., 2024, Yang et al., 2019). Mastiffs are a collection of large and giant dog breeds,
426 with a high incidence rate of cancer and short lifespan (Bell and Hesketh, 2021).

427 Discussion

428 Advances in long-read sequencing have ushered in a new era of genomics and have enabled the
429 generation of gap-free telomere-to-telomere assemblies (Marx, 2023). These resources offer a refined
430 view of complex genome variation, including structural variation involving duplicated and repetitive

431 sequences. Despite the power of long-read approaches, most existing catalogs of genetic variation are
432 derived from short-read data. The canine research community has generated a rich collection of
433 genomic resources including thousands of diverse samples sequenced using Illumina short-read
434 sequencing along with multiple samples with long-read data. In this study, we present a uniform
435 catalogue of SVs using a wealth of publicly available data covering short-read, long-read, and assembled
436 genomes from 12 canine samples. Further, we used recently developed graph-based algorithms to
437 genotype the discovered variants in a collection of 1,879 dog and wolf samples sequenced by the
438 Dog10K consortium.

439 As expected, the spectrum of SV found between genome assemblies is dominated by insertion and
440 deletions. However, the analyzed canine assemblies are haploid representations of diploid genomes and
441 offer an incomplete view of variation present in each sample. This is reinforced by comparison with SVs
442 identified through alignment of the raw long-read data used to generate each assembly, where the
443 increased number of identified SVs reflects heterozygous variants that are not represented in the
444 respective assemblies. For most samples, an excess of heterozygous insertions was found when
445 analyzing the long-reads, suggesting that existing assemblies may show a bias towards representing the
446 shorter, deletion allele at heterozygous sites. Differences in data quality among the samples likely also
447 contributes to these patterns.

448 The size distribution of identified insertion and deletions displayed peaks that correspond to the size of
449 dimorphic SINE and LINE-1 elements. Comparison with repeat element consensus sequences showed
450 that variable SINE and LINE-1 sequences from the most recent active families account for 45.8% and
451 42.5% of all deletion and insertion SVs. This includes 1,410 dimorphic LINE-1s that appear to have intact
452 open reading frames for both the ORF1p and ORF2p proteins required for retrotransposition. The
453 greatest contributor of intact LINE-1 sequences was found in the mCanLor Greenland Wolf genome, the
454 only analyzed sample sequenced using PacBio HiFi reads, highlighting the importance of sequencing
455 technology to the accurate base-pair level resolution of variation among repetitive sequences.

456 As expected, comparisons across sequencing technologies confirmed the dramatically reduced power of
457 short-read approaches to capture insertion variants. However, variants discovered with long-reads can
458 be genotyped using short-read data. To offer a more complete assessment of canine variation, we
459 therefore genotyped the expanded SV catalog in samples from the Dog10K collection, resulting in a
460 56.5% increase in the number of deletions and 705% increase in the number of insertions previously
461 reported in these samples (Meadows et al., 2023).

462 The Dog10K consortium previously analyzed single nucleotide variants to identify signatures of selection
463 in the ancestral components present across major clades of breed dogs. We repeated this analysis using
464 the expanded catalog of 299,116 deletions and insertions, revealing examples where a genotyped
465 insertion or deletion has a clearer signal of selection than flanking SNPs (Figure 9). Consistent with their
466 major contribution to canine SVs, we found that 90 of the 283 variants that exceed the reduced
467 selection threshold ($p<=0.0001$) corresponded to SINE or LINE deletions. Mobile element insertions can
468 disrupt gene function through multiple mechanisms, including disruption of protein-coding exons,
469 alterations to gene expression, and disrupted splicing processes (Hancks and Kazazian, 2016, Cordaux
470 and Batzer, 2009). Recent reports have highlighted the potential of pairs of SINEs located in inverted
471 orientation to disrupt splicing due to the formation of secondary structures in pre-mRNA, a mechanism
472 implicated as contributing to the tail loss in apes (Xia et al., 2024) and the evolution of skin pigmentation

variation in humans (Kamitaki et al., 2024). We identified a pair of dimorphic SINE sequences in introns of *DENR* that appear to have evolved under selection in Mastiffs and that have such an inverted orientation (Supplemental Figure S8). Additional studies will be required to determine whether improper splicing is found in dogs that carry the identified SINE insertions. Several of the genes containing SVs with signals of selection were also identified by an analysis of SNPs (Meadows et al., 2023), making it challenging to identify specific mutations that are functionally important. Several of the identified genes have been linked to diseases, making it unclear how to interpret statistical signals suggesting they were selected across breed clades. One possibility is that the identified variants are deleterious mutations that are tightly linked to variants that contribute to breed-defining phenotypes that have yet to be identified. Another possibility is that the variants have pleiotropic effects that both promote characteristics desired during breed formation while also contributing to the development of disease. More focused studies that link models of gene function with evolutionary dynamics across time periods will be required to identify the precise molecular changes underlying the formation of modern breeds.

Data availability

VCF files describing the identified structural variants have been deposited in the Zenodo data archive under accession DOI 10.5281/zenodo.14968873. A UCSC browser track hub depicting the detected SVs is available at https://github.com/KiddLab/dog-long-read-sv

Acknowledgements

The authors thank all members of the Kitano Lab for discussion related to this manuscript and members of the canine genomics community for creating and sharing valuable resources.

Author's Contributions

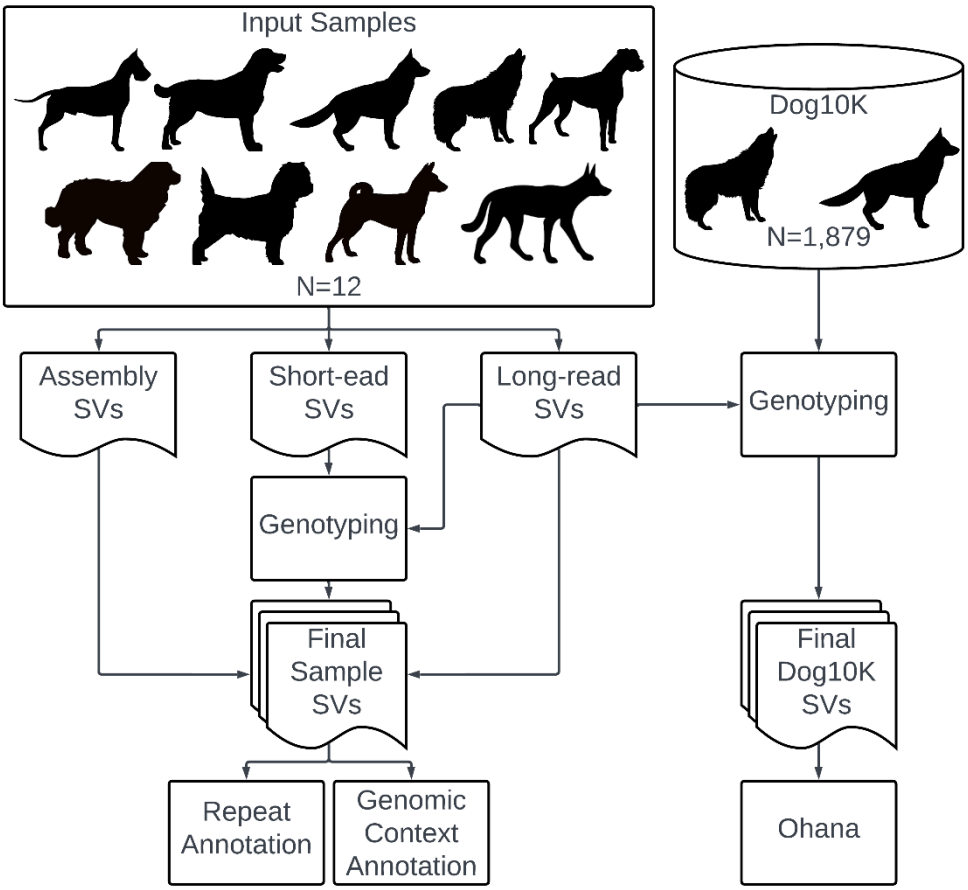
P.Z.S and J.M.K conceived and designed the study, P.Z.S. performed analyses and implemented computational tools, P.Z.S. and J.M.K wrote the manuscript.

Tables

		Number of SVs			Basepairs (Mb)		
	SV Type	Mean	Min	Max	Mean	Min	Max
Assembly	Deletions	21,509	14,775	32,112	21.9	12.8	33.4
	Insertions	20,712	12,686	28,262	12.2	8.6	17.7
	Inversions	68	19	210	17.4	0.0	160.0
	Duplications	18	7	34	4.6	0.0	44.8
Long-read	Deletions	28,148	16,021	39,021	12.4	5.3	16.9
	Insertions	30,798	25,992	45,520	26.1	12.9	102.6
	Inversions	205	86	289	13.8	0.5	147.9
	Duplications	26	7	45	2.1	0.2	53.9
Short-read	Deletions	20,121	12,634	24,369	12.2	12.8	33.4
	Insertions	6,019	732	9,361	12.2	8.6	17.7
	Inversions	92	4	114	17.4	0.0	160.0
	Duplications	232	110	319	4.6	0.0	44.8
Short-read + Paragraph	Deletions	25,525	15,542	36,853	14.0	6.1	20.7
	Insertions	23,730	9,646	40,930	9.9	4.9	14.4
	Inversions	92	4	114	111.0	9.1	167.9
	Duplications	234	110	323	1.3	0.5	2.0

Table 1 - Summary statistics of quantified SVs and impacted base pairs in 12 canine samples by modality

500 **Figures**



501

502 **Figure 1 – Overview of SV detection in canines**

503 The flowchart describes integrative analysis and detection of structural variants using genome assembly,
504 long-read, and short-read data from 12 samples. Additionally, genotyping of structural variants detected
505 using long-reads in the Dog10K repository is depicted. All variants were identified relative to the
506 UU_Cfam_GSD_1.0_ROSY genome assembly. Created in Lucidchart www.lucidchart.com.

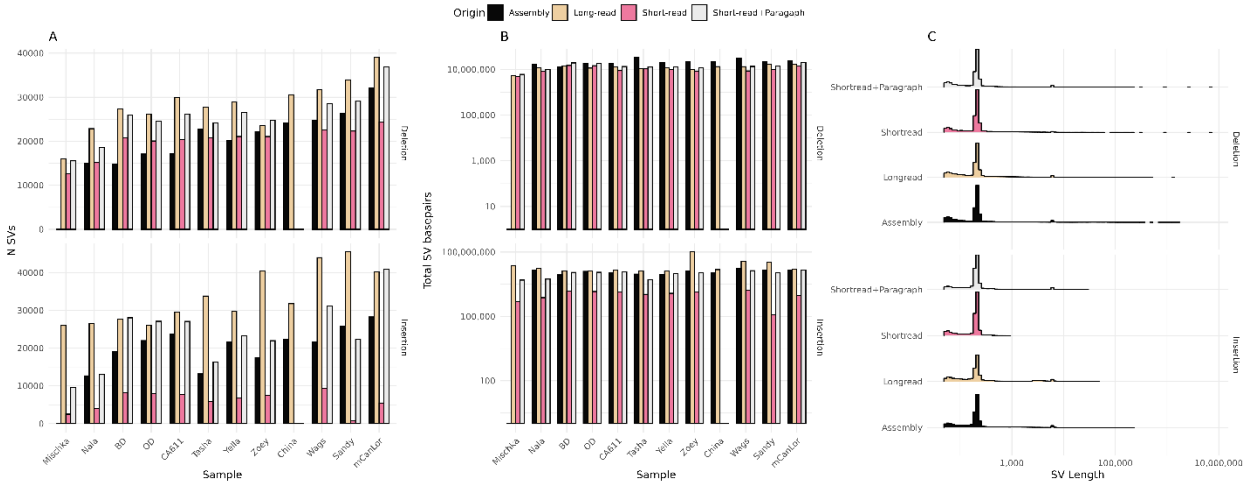


Figure 2 – Global quantification of deletions and insertions using different data types

Panel A plots the number of deletions and insertions found across the 12 samples and four different datatypes: assembly, long-read, short-read, and short-read+paragraph. Panel B displays the summation of the total impacted basepairs. Panel C plots the distribution of SV lengths. The fill color of each bar denotes datatype origin as noted in legend.

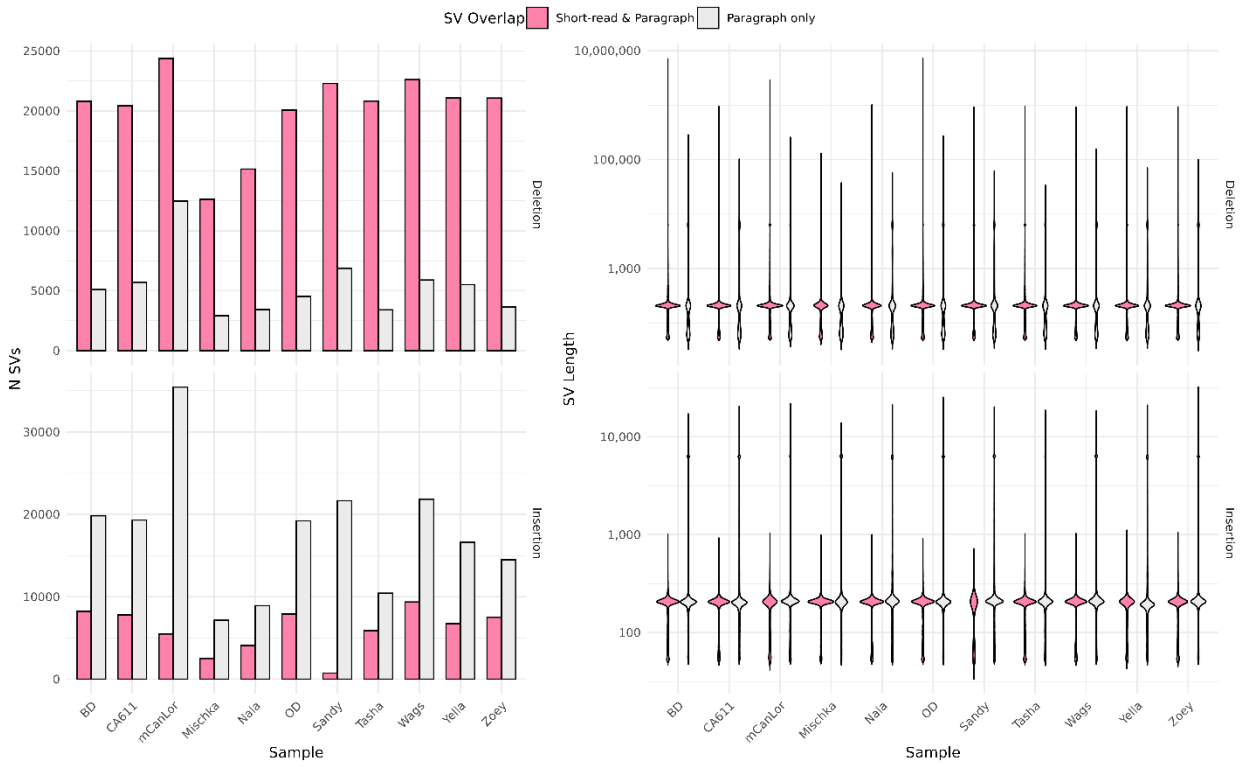


Figure 3 – Genotyping with paragraph enables increased detection of deletion and insertion variants using short-read data

The number and size of structural variants detected using short-read data only (Manta) and that can be genotyped using paragraph is shown. The first panel denotes the increased number of variants that can be detected when genotyping with paragraph. The pink bars display SVs found in both methods, and grey shows additional SVs detected from paragraph. The second panel shows the length distribution of detected SVs, illustrating the ability to genotype long insertions when using Paragraph.

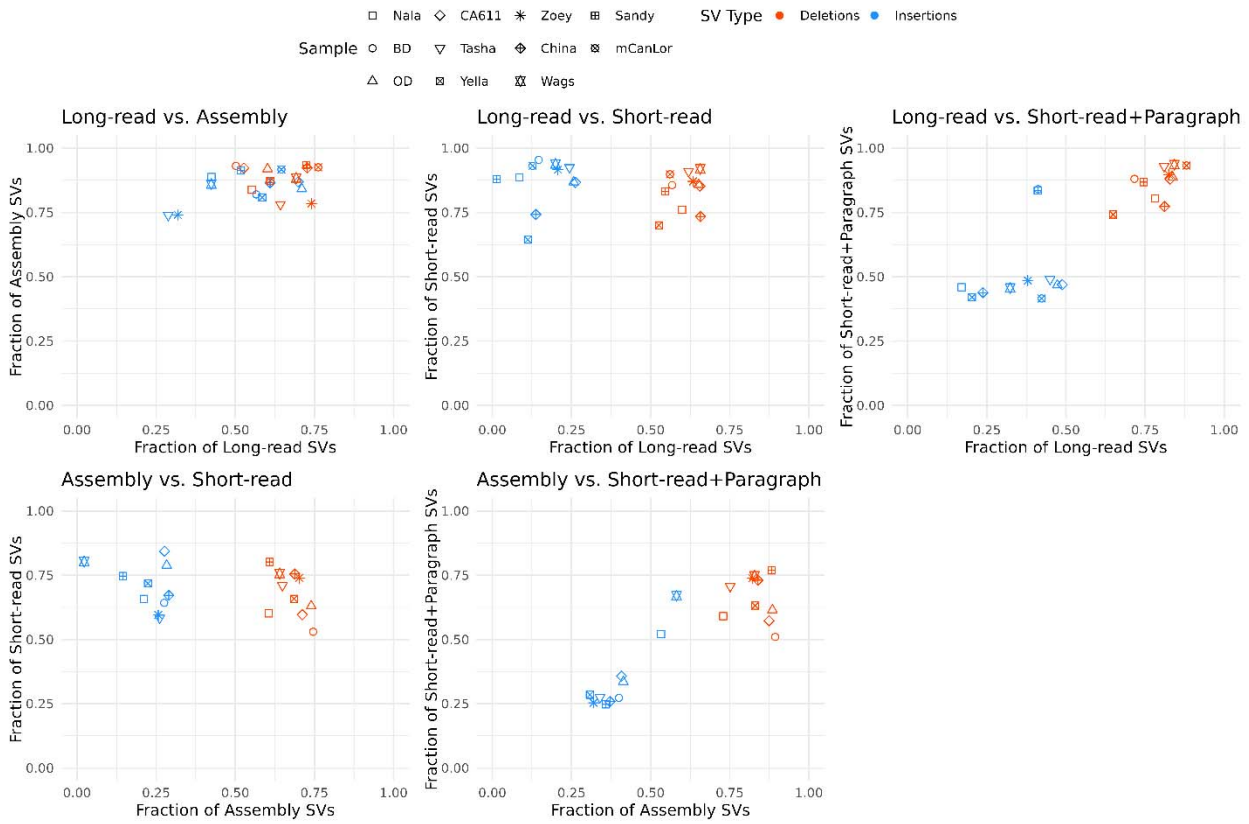


Figure 4 – Concordance of SV detection between datatypes

Scatter plots depict the fractional quantification of detected SVs when comparing different datatypes (e.g., long-read vs. assembly, long-read vs. short-read, etc.). Each panel depicts a specific comparison with the axis quantifying the fraction shared. The shape of each point denotes the sample and color denotes SV type: deletions in red and insertions in blue.

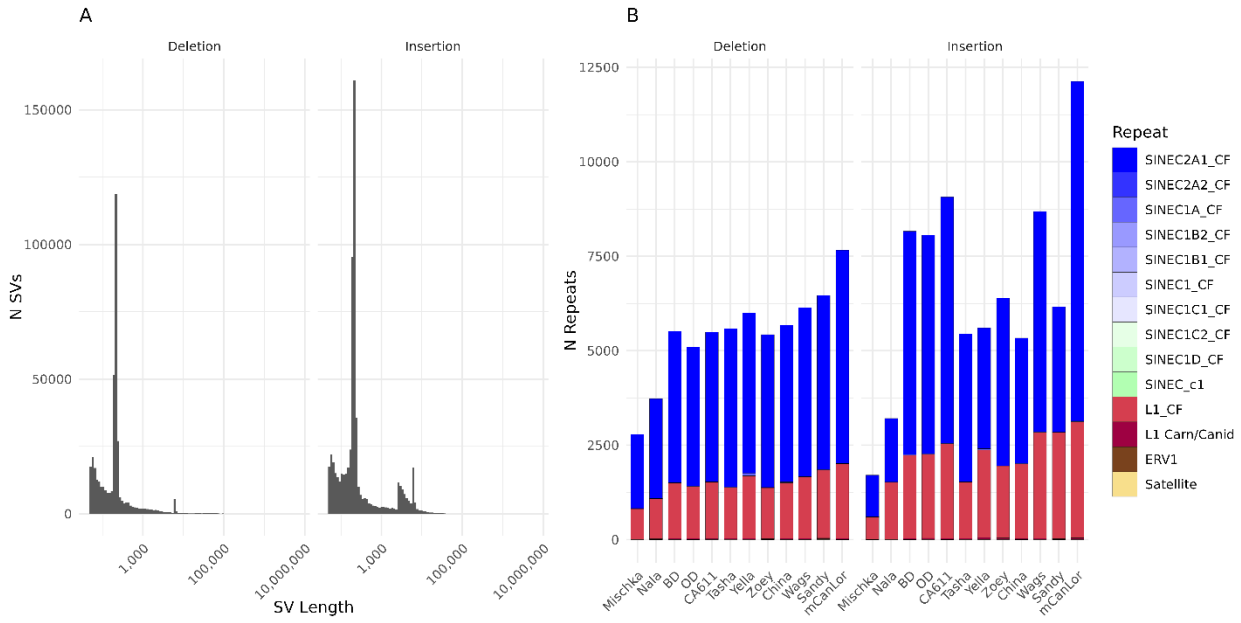
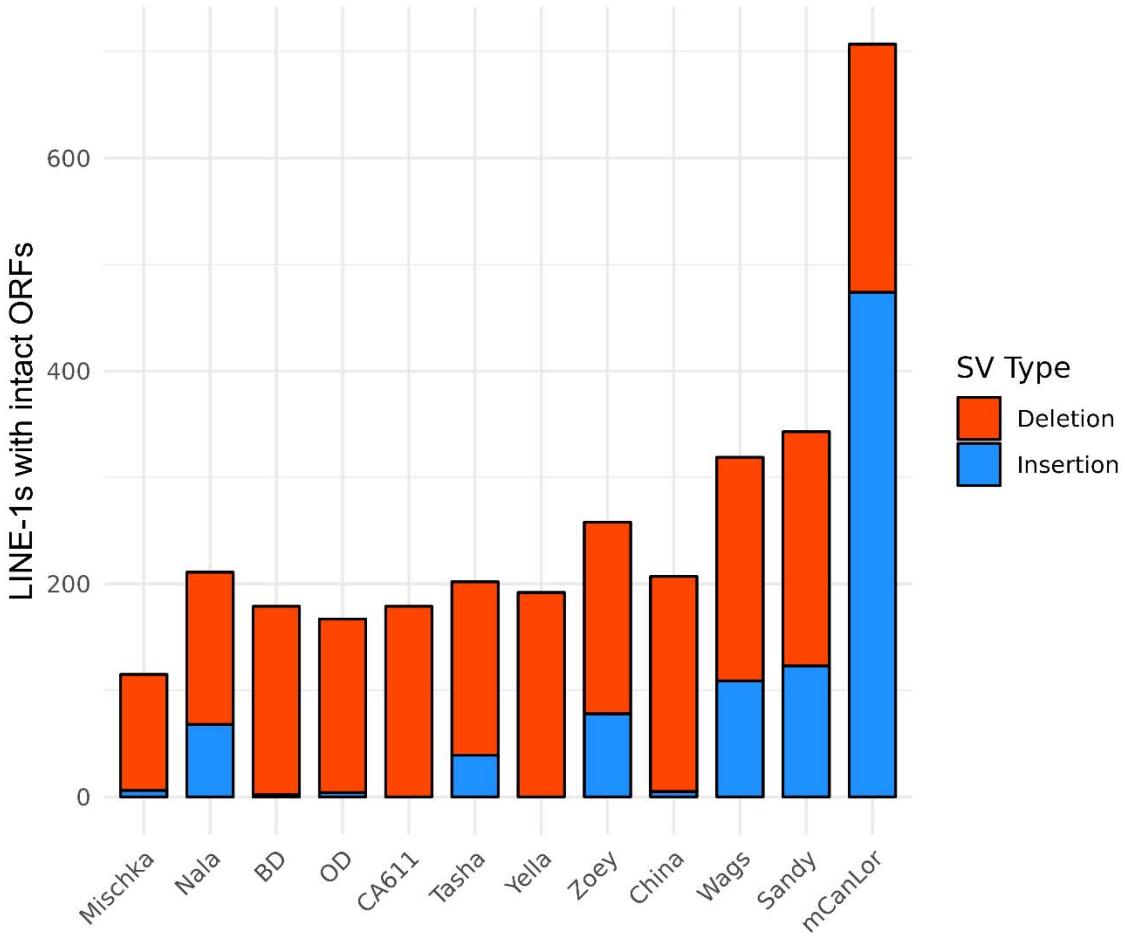


Figure 5 – Detection of SINEs and LINEs from SV data

All deletions and insertions, regardless of origin datatype, were scanned for the presence of repeat sequences via blastn using canine and ancestral SINE and LINE sequences from RepBase. To identify those SVs comprised of a singular repeat, SVs were only included with at least 95% of the length constituting a repeat sequence. The left panel denotes the distribution of SV lengths of all deletions and insertions. The right panel quantifies the number of detected SVs by repeat sequence name and/or family.



535

536 **Figure 6 – Census of dimorphic full-length LINE-1s with intact open reading frames**

537 SVs corresponding to full-length LINE-1 sequences were extracted and scanned for intact open reading
538 frames (>=99% sequence similarity) and quantified by either deletion or insertion for each sample. The
539 stacked barplot lists deleted LINE-1s in red, and inserted LINE-1s in blue.

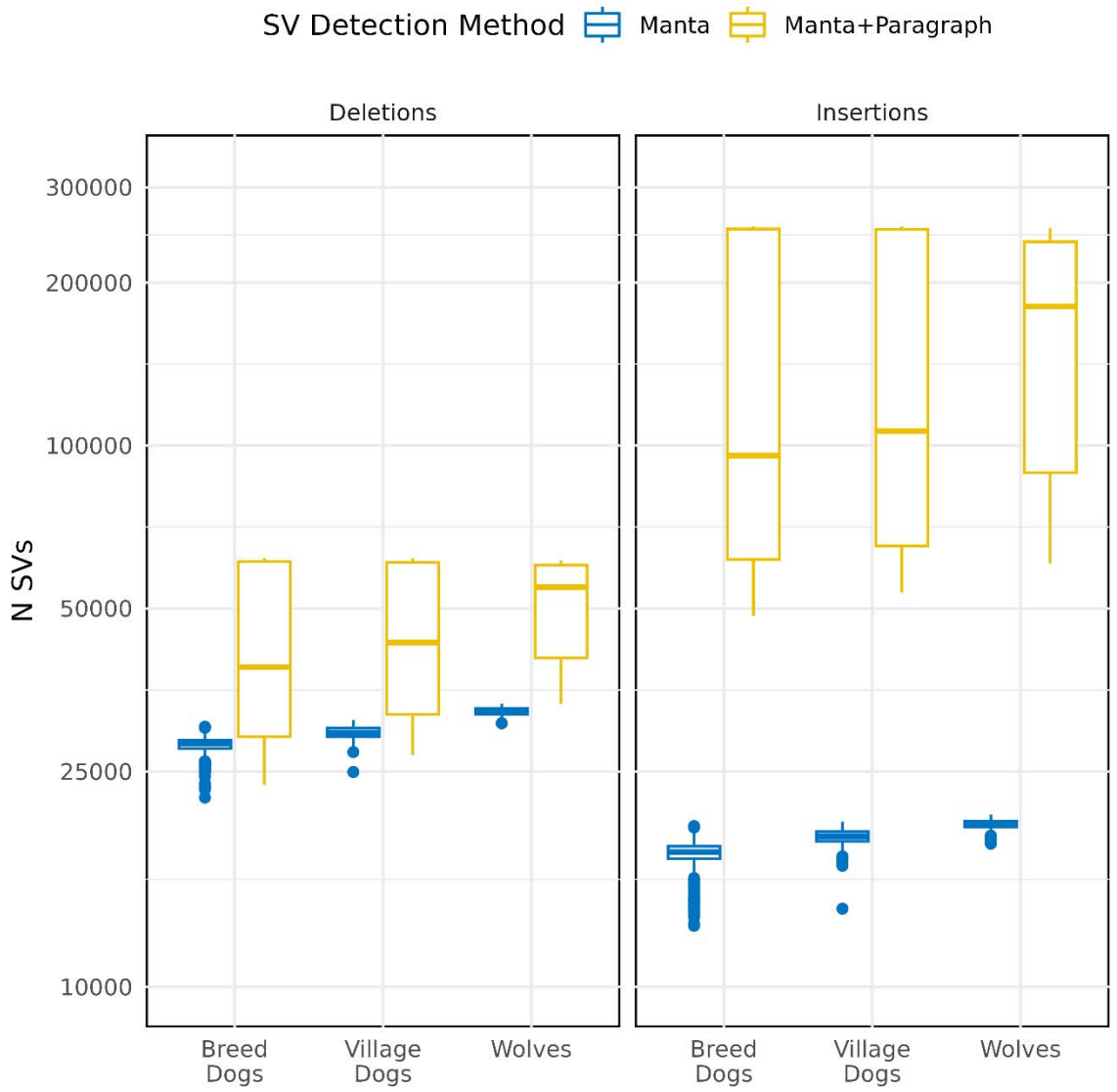
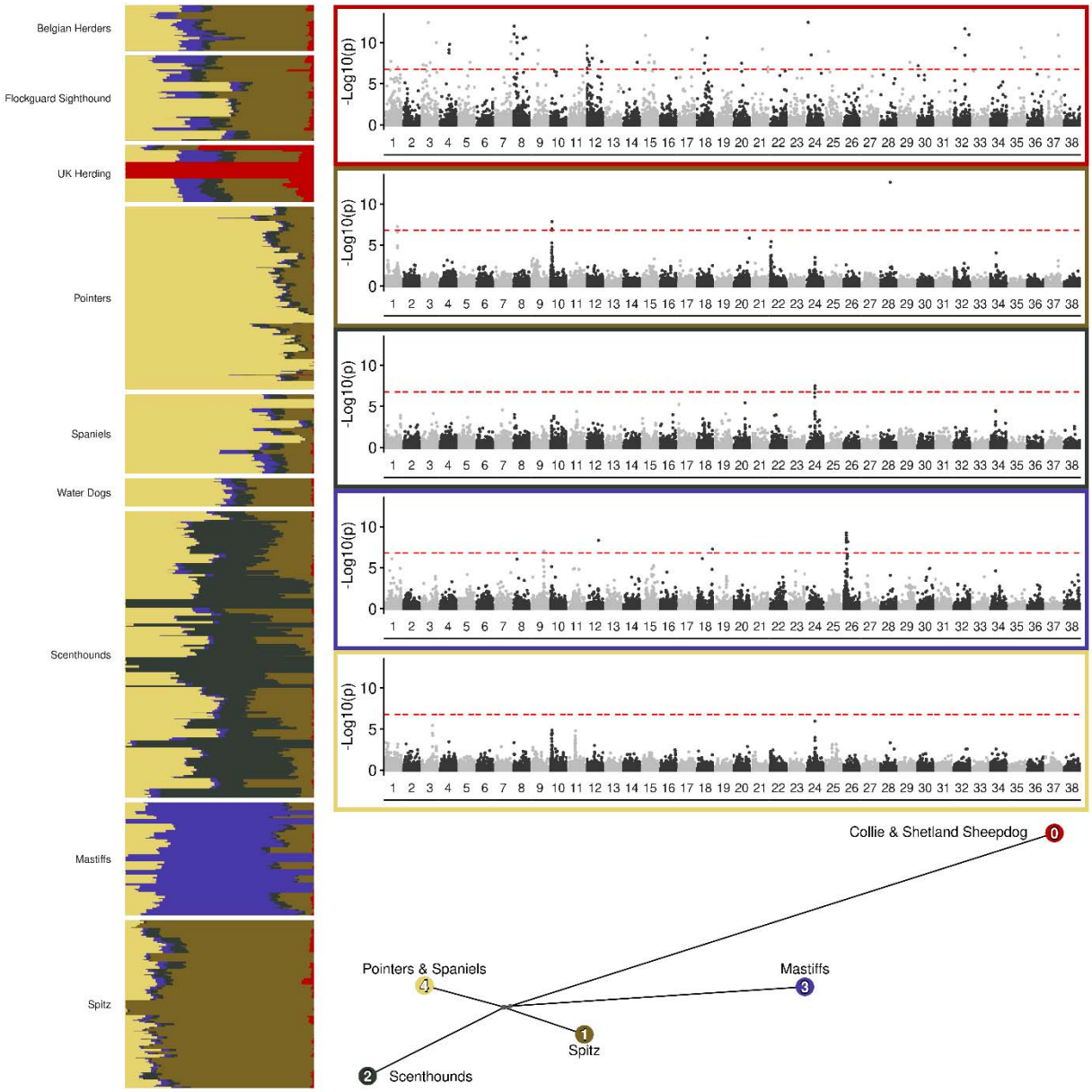


Figure 7 – Genotyping SVs identified using longreads in the Dog10K sample collection

Box plots depict the number of deletions and insertions genotyped across the Dog10K collection using the Manta and Manta+Paragraph approaches. The data are plotted by canine category (Breed Dogs, Village Dogs, and Wolves), as noted along the x-axis. The tabulation of the original Manta SVs is in blue, while Manta+Paragraph is in yellow.



546

547 **Figure 8 – Structural variants with signatures of selection detected by Ohana with five ancestral**
548 **components**

549 Left vertical panel plots the population structure for the nine dog-breed/clade group with *K*=5 ancestral
550 components based on the SNP data. Manhattan plots depict selection scans for SVs in each of the five
551 ancestral components, exterior border color corresponds to color of each ancestral component,
552 horizontal dashed red-line denotes Bonferroni threshold. The bottom network tree depicts the inferred
553 relationships among the ancestral components from the SNP data, labeled with the breeds in which
554 each component is maximized, filled color corresponding to both the Manhattan plot border and
555 population structure.

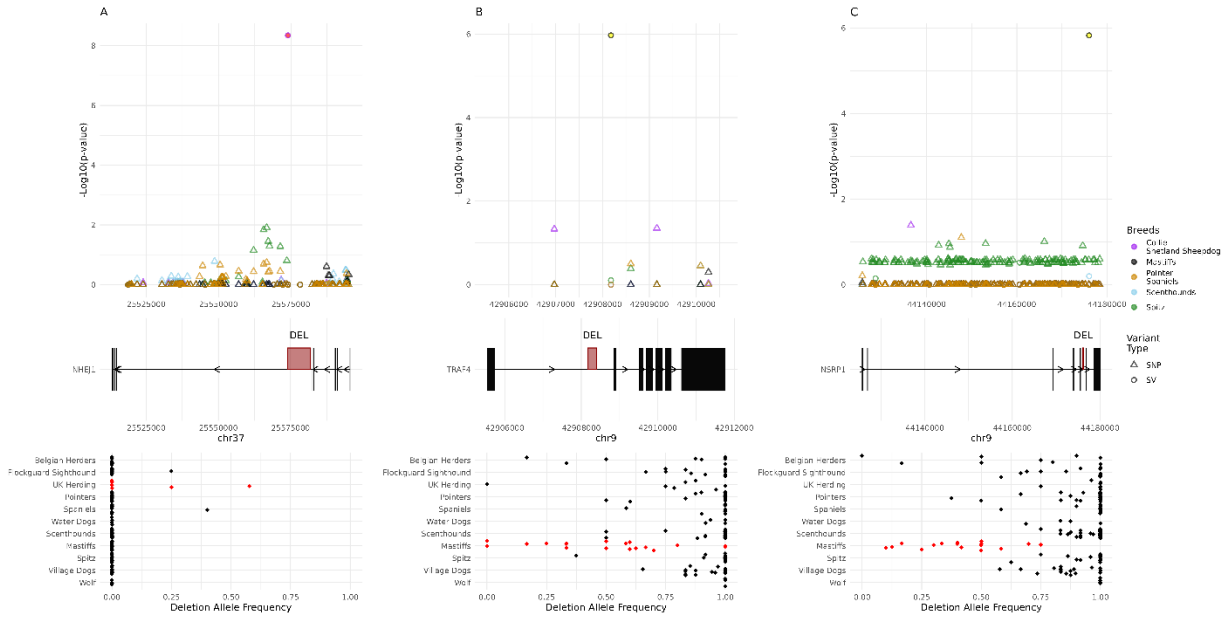


Figure 9 – Structural variants with signals of selection

The output of Ohana, combining SNP and SV data, for three examples in their specific genomic context across the included canine breed/clade categories is depicted. Each panel consists of three sub panels: the top displaying the genomic coordinates along the x-axis and the significance along the y-axis reported by Ohana, the middle displays the relevant gene model within the region, with thick boxes denoting the position of exons and the position of most significant SV given in red, the bottom displays the allele frequency by breed or geographic location for the most significant SV. For each top panel, the color denotes the canine grouping and shape delineating variant type. Within the bottom allele frequency panel, the red colored dots correspond to the clade identified via Ohana for the associated SV selection signal. Ohana output can be found in Supplemental Table S7, and allele frequencies in Supplemental Table S8.

References

ABYZOV, A. & GERSTEIN, M. 2011. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics,* 27**,** 595-603.

AGANEZOV, S., YAN, S. M., SOTO, D. C., KIRSCHE, M., ZARATE, S., AVDEYEV, P., TAYLOR, D. J., SHAFIN, K., SHUMATE, A., XIAO, C., WAGNER, J., MCDANIEL, J., OLSON, N. D., SAURIA, M. E. G., VOLLGER, M. R., RHIE, A., MEREDITH, M., MARTIN, S., LEE, J., KOREN, S., ROSENFELD, J. A., PATEN, B., LAYER, R., CHIN, C. S., SEDLAZECK, F. J., HANSEN, N. F., MILLER, D. E., PHILLIPPY, A. M., MIGA, K. H., MCCOY, R. C., DENNIS, M. Y., ZOOK, J. M. & SCHATZ, M. C. 2022. A complete reference genome improves analysis of human genetic variation. *Science,* 376**,** eabl3533.

AMARASINGHE, S. L., SU, S., DONG, X., ZAPPIA, L., RITCHIE, M. E. & GOUIL, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol,* 21**,** 30.

ANTKOWIAK, M. & SZYDLOWSKI, M. 2023. Uncovering structural variants associated with body weight and obesity risk in labrador retrievers: a genome-wide study. *Front Genet,* 14**,** 1235821.

AXELSSON, E., RATNAKUMAR, A., ARENDT, M. L., MAQBOOL, K., WEBSTER, M. T., PERLOSKI, M., LIBERG, O., ARNEMO, J. M., HEDHAMMAR, A. & LINDBLAD-TOH, K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature,* 495**,** 360-4.

BAO, W., KOJIMA, K. K. & KOHANY, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA,* 6**,** 11.

BECK, C. R., GARCIA-PEREZ, J. L., BADGE, R. M. & MORAN, J. V. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet,* 12**,** 187-215.

BELL, L. & HESKETH, S. 2021. Mastiff breed mortality: A study of owner experience, dog age and longevity. *Vet Anim Sci,* 13**,** 100194.

BENSON, K. F., LI, F. Q., PERSON, R. E., ALBANI, D., DUAN, Z., WECHSLER, J., MEADE-WHITE, K., WILLIAMS, K., ACLAND, G. M., NIEMEYER, G., LOTHROP, C. D. & HORWITZ, M. 2003. Mutations associated with neutropenia in dogs and humans disrupt intracellular transport of neutrophil elastase. *Nat Genet,* 35**,** 90-6.

BETANCOR, G. 2023. You Shall Not Pass: MX2 Proteins Are Versatile Viral Inhibitors. *Vaccines (Basel),* 11.

BOLORMAA, S., SWAN, A. A., STOTHARD, P., KHANSEFID, M., MOGHADDAR, N., DUIJVESTEIJN, N., VAN DER WERF, J. H. J., DAETWYLER, H. D. & MACLEOD, I. M. 2021. A conditional multi-trait sequence GWAS discovers pleiotropic candidate genes and variants for sheep wool, skin wrinkle and breech cover traits. *Genet Sel Evol,* 53**,** 58.

BOYKO, A. R. 2011. The domestic dog: man's best friend in the genomic era. *Genome Biol,* 12**,** 216.

CALAME, D. G., BAKHTIARI, S., LOGAN, R., COBAN-AKDEMIR, Z., DU, H., MITANI, T., FATIH, J. M., HUNTER, J. V., HERMAN, I., PEHLIVAN, D., JHANGIANI, S. N., PERSON, R., SCHNUR, R. E., JIN, S. C., BILGUVAR, K., POSEY, J. E., KOH, S., FIROUZABADI, S. G., ALEHABIB, E., TAFAKHORI, A., ESMKHANI, S., GIBBS, R. A., NOURELDEEN, M. M., ZAKI, M. S., MARAFI, D., DARVISH, H., KRUER, M. C. & LUPSKI, J. R. 2021. Biallelic loss-of-function variants in the splicing regulator NSRP1 cause a severe neurodevelopmental disorder with spastic cerebral palsy and epilepsy. *Genet Med,* 23**,** 2455-2460.

CHEN, S., KRUSCHE, P., DOLZHENKO, E., SHERMAN, R. M., PETROVSKI, R., SCHLESINGER, F., KIRSCHE, M., BENTLEY, D. R., SCHATZ, M. C., SEDLAZECK, F. J. & EBERLE, M. A. 2019. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol,* 20**,** 291.

CHEN, X., SCHULZ-TRIEGLAFF, O., SHAW, R., BARNES, B., SCHLESINGER, F., KALLBERG, M., COX, A. J., KRUGLYAK, S. & SAUNDERS, C. T. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics,* 32**,** 1220-2.

615 CHENG, J. Y., STERN, A. J., RACIMO, F. & NIELSEN, R. 2022. Detecting Selection in Multiple Populations by
616 Modeling Ancestral Admixture Components. *Mol Biol Evol,* 39.
617 CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN,
618 D. M. 2012. A program for annotating and predicting the effects of single nucleotide
619 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2;
620 iso-3. *Fly (Austin),* 6**,** 80-92.
621 CORDAUX, R. & BATZER, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nat*
622 *Rev Genet,* 10**,** 691-703.
623 DANECEK, P., BONFIELD, J. K., LIDDLE, J., MARSHALL, J., OHAN, V., POLLARD, M. O., WHITWHAM, A.,
624 KEANE, T., MCCARTHY, S. A., DAVIES, R. M. & LI, H. 2021. Twelve years of SAMtools and
625 BCFtools. *Gigascience,* 10.
626 DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A.,
627 DEL ANGEL, G., RIVAS, M. A., HANNA, M., MCKENNA, A., FENNELL, T. J., KERNYTSKY, A. M.,
628 SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A
629 framework for variation discovery and genotyping using next-generation DNA sequencing data.
630 *Nat Genet,* 43**,** 491-8.
631 DESSAU, R. B. & PIPPER, C. B. 2008. ["R"--project for statistical computing]. *Ugeskr Laeger,* 170**,** 328-30.
632 DREGER, D. L., RIMBAULT, M., DAVIS, B. W., BHATNAGAR, A., PARKER, H. G. & OSTRANDER, E. A. 2016.
633 Whole-genome sequence, SNP chips and pedigree structure: building demographic profiles in
634 domestic dog breeds to optimize genetic-trait mapping. *Dis Model Mech,* 9**,** 1445-1460.
635 DYER, S. C., AUSTINE-ORIMOLOYE, O., AZOV, A. G., BARBA, M., BARNES, I., BARRERA-ENRIQUEZ, V. P.,
636 BECKER, A., BENNETT, R., BERACOCHEA, M., BERRY, A., BHAI, J., BHURJI, S. K., BODDU, S.,
637 BRANCO LINS, P. R., BROOKS, L., RAMARAJU, S. B., CAMPBELL, L. I., MARTINEZ, M. C.,
638 CHARKHCHI, M., CORTES, L. A., DAVIDSON, C., DENNI, S., DODIYA, K., DONALDSON, S., EL
639 HOUDAIGUI, B., EL NABOULSI, T., FALOLA, O., FATIMA, R., GENEZ, T., MARTINEZ, J. G., GURBICH,
640 T., HARDY, M., HOLLIS, Z., HUNT, T., KAY, M., KAYKALA, V., LEMOS, D., LODHA, D., MATHLOUTHI,
641 N., MERINO, G. A., MERRITT, R., MIRABUENO, L. P., MUSHTAQ, A., HOSSAIN, S. N., PEREZ-SILVA,
642 J. G., PERRY, M., PILIZOTA, I., POPPLETON, D., PROSOVETSKAIA, I., RAJ, S., SALAM, A. I. A.,
643 SARAF, S., SARAIVA-AGOSTINHO, N., SINHA, S., SIPOS, B., SITNIK, V., STEED, E., SUNER, M. M.,
644 SURAPANENI, L., SUTINEN, K., TRICOMI, F. F., TSANG, I., URBINA-GOMEZ, D., VEIDENBERG, A.,
645 WALSH, T. A., WILLHOFT, N. L., ALLEN, J., ALVAREZ-JARRETA, J., CHAKIACHVILI, M., CHEEMA, J.,
646 DA ROCHA, J. B., DE SILVA, N. H., GIORGETTI, S., HAGGERTY, L., ILSLEY, G. R., KEATLEY, J.,
647 LOVELAND, J. E., MOORE, B., MUDGE, J. M., NAAMATI, G., TATE, J., TREVANION, S. J.,
648 WINTERBOTTOM, A., FLINT, B., FRANKISH, A., HUNT, S. E., FINN, R. D., FREEBERG, M. A.,
649 HARRISON, P. W., MARTIN, F. J. & YATES, A. D. 2025. Ensembl 2025. *Nucleic Acids Res,* 53**,** D948-
650 D957.
651 EDMUNDS, G. L., SMALLEY, M. J., BECK, S., ERRINGTON, R. J., GOULD, S., WINTER, H., BRODBELT, D. C. &
652 O'NEILL, D. G. 2021. Dog breeds and body conformations with predisposition to osteosarcoma in
653 the UK: a case-control study. *Canine Med Genet,* 8**,** 2.
654 EDWARDS, R. J., FIELD, M. A., FERGUSON, J. M., DUDCHENKO, O., KEILWAGEN, J., ROSEN, B. D.,
655 JOHNSON, G. S., RICE, E. S., HILLIER, D., HAMMOND, J. M., TOWARNICKI, S. G., OMER, A., KHAN,
656 R., SKVORTSOVA, K., BOGDANOVIC, O., ZAMMIT, R. A., AIDEN, E. L., WARREN, W. C. & BALLARD,
657 J. W. O. 2021. Chromosome-length genome assembly and structural variations of the primal
658 Basenji dog (Canis lupus familiaris) genome. *BMC Genomics,* 22**,** 188.
659 EGGERTSSON, H. P., KRISTMUNDSDOTTIR, S., BEYTER, D., JONSSON, H., SKULADOTTIR, A., HARDARSON,
660 M. T., GUDBJARTSSON, D. F., STEFANSSON, K., HALLDORSSON, B. V. & MELSTED, P. 2019.
661 GraphTyper2 enables population-scale genotyping of structural variation using pangenome
662 graphs. *Nat Commun,* 10**,** 5402.

663 ESCARAMIS, G., DOCAMPO, E. & RABIONET, R. 2015. A decade of structural variants: description, history
664 and methods to detect structural variation. *Brief Funct Genomics,* 14**,** 305-14.
665 FIELD, M. A., ROSEN, B. D., DUDCHENKO, O., CHAN, E. K. F., MINOCHE, A. E., EDWARDS, R. J., BARTON,
666 K., LYONS, R. J., TUIPULOTU, D. E., HAYES, V. M., A, D. O., COLARIC, Z., KEILWAGEN, J.,
667 SKVORTSOVA, K., BOGDANOVIC, O., SMITH, M. A., AIDEN, E. L., SMITH, T. P. L., ZAMMIT, R. A. &
668 BALLARD, J. W. O. 2020. Canfam_GSD: De novo chromosome-length genome assembly of the
669 German Shepherd Dog (Canis lupus familiaris) using a combination of long reads, optical
670 mapping, and Hi-C. *Gigascience,* 9.
671 FIELD, M. A., YADAV, S., DUDCHENKO, O., ESVARAN, M., ROSEN, B. D., SKVORTSOVA, K., EDWARDS, R. J.,
672 KEILWAGEN, J., COCHRAN, B. J., MANANDHAR, B., BUSTAMANTE, S., RASMUSSEN, J. A., MELVIN,
673 R. G., CHERNOFF, B., OMER, A., COLARIC, Z., CHAN, E. K. F., MINOCHE, A. E., SMITH, T. P. L.,
674 GILBERT, M. T. P., BOGDANOVIC, O., ZAMMIT, R. A., THOMAS, T., AIDEN, E. L. & BALLARD, J. W.
675 O. 2022. The Australian dingo is an early offshoot of modern breed dogs. *Sci Adv,* 8**,** eabm5944.
676 FLEMING, J. M., CREEVY, K. E. & PROMISLOW, D. E. 2011. Mortality in north american dogs from 1984 to
677 2004: an investigation into age-, size-, and breed-related causes of death. *J Vet Intern Med,* 25**,**
678 187-98.
679 FREEDMAN, A. H., SCHWEIZER, R. M., ORTEGA-DEL VECCHYO, D., HAN, E., DAVIS, B. W., GRONAU, I.,
680 SILVA, P. M., GALAVERNI, M., FAN, Z., MARX, P., LORENTE-GALDOS, B., RAMIREZ, O.,
681 HORMOZDIARI, F., ALKAN, C., VILA, C., SQUIRE, K., GEFFEN, E., KUSAK, J., BOYKO, A. R., PARKER,
682 H. G., LEE, C., TADIGOTLA, V., SIEPEL, A., BUSTAMANTE, C. D., HARKINS, T. T., NELSON, S. F.,
683 MARQUES-BONET, T., OSTRANDER, E. A., WAYNE, R. K. & NOVEMBRE, J. 2016. Demographically-
684 Based Evaluation of Genomic Regions under Selection in Domestic Dogs. *PLoS Genet,* 12**,**
685 e1005851.
686 GENOMES PROJECT, C., AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M.,
687 KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. & ABECASIS, G. R. 2015. A global
688 reference for human genetic variation. *Nature,* 526**,** 68-74.
689 GU, Z., EILS, R. & SCHLESNER, M. 2016. Complex heatmaps reveal patterns and correlations in
690 multidimensional genomic data. *Bioinformatics,* 32**,** 2847-9.
691 HALLER, O., ARNHEITER, H., PAVLOVIC, J. & STAEHELI, P. 2018. The Discovery of the Antiviral Resistance
692 Gene Mx: A Story of Great Ideas, Great Failures, and Some Success. *Annu Rev Virol,* 5**,** 33-51.
693 HALO, J. V., PENDLETON, A. L., SHEN, F., DOUCET, A. J., DERRIEN, T., HITTE, C., KIRBY, L. E., MYERS, B.,
694 SLIWERSKA, E., EMERY, S., MORAN, J. V., BOYKO, A. R. & KIDD, J. M. 2021. Long-read assembly
695 of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to
696 canine genomes. *Proc Natl Acad Sci U S A,* 118.
697 HANCKS, D. C. & KAZAZIAN, H. H., JR. 2016. Roles for retrotransposon insertions in human disease. *Mob*
698 *DNA,* 7**,** 9.
699 HELLER, D. & VINGRON, M. 2021. SVIM-asm: structural variant detection from haploid and diploid
700 genome assemblies. *Bioinformatics,* 36**,** 5519-5521.
701 HITTI-MALIN, R. J., BURMEISTER, L. M., RICKETTS, S. L., LEWIS, T. W., PETTITT, L., BOURSNELL, M.,
702 SCHOFIELD, E. C., SARGAN, D. & MELLERSH, C. S. 2020. A LINE-1 insertion situated in the
703 promoter of IMPG2 is associated with autosomal recessive progressive retinal atrophy in Lhasa
704 Apso dogs. *BMC Genet,* 21**,** 100.
705 HO, S. S., URBAN, A. E. & MILLS, R. E. 2020. Structural variation in the sequencing era. *Nat Rev Genet,* 21**,**
706 171-189.
707 HUJOEL, M. L. A., SHERMAN, M. A., BARTON, A. R., MUKAMEL, R. E., SANKARAN, V. G., TERAO, C. & LOH,
708 P. R. 2022. Influences of rare copy-number variation on human complex traits. *Cell,* 185**,** 4233-
709 4248 e27.

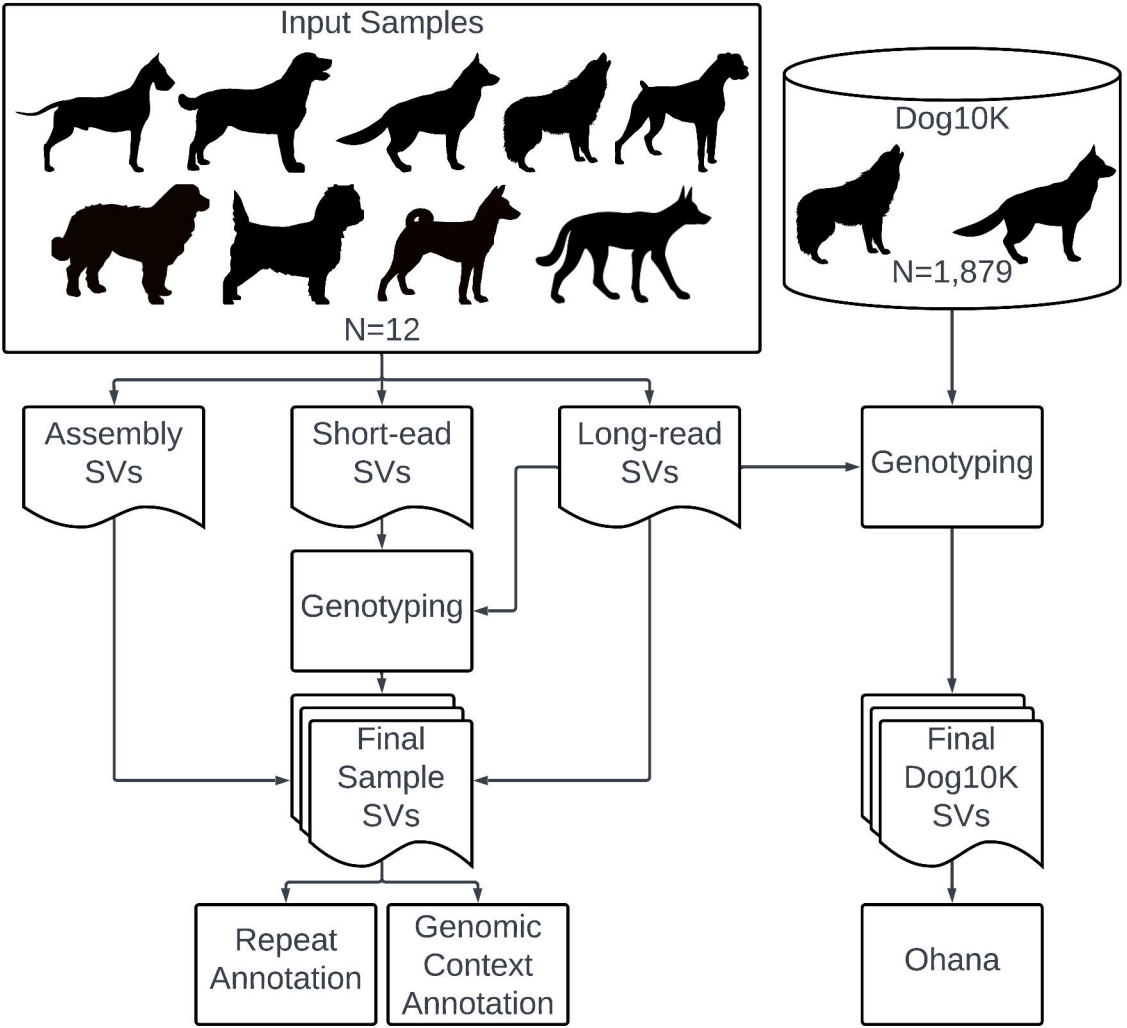
710 JAGANNATHAN, V., HITTE, C., KIDD, J. M., MASTERSON, P., MURPHY, T. D., EMERY, S., DAVIS, B.,
711 BUCKLEY, R. M., LIU, Y. H., ZHANG, X. Q., LEEB, T., ZHANG, Y. P., OSTRANDER, E. A. & WANG, G.
712 D. 2021. Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome.
713 *Genes (Basel),* 12.
714 JEFFARES, D. C., JOLLY, C., HOTI, M., SPEED, D., SHAW, L., RALLIS, C., BALLOUX, F., DESSIMOZ, C.,
715 BAHLER, J. & SEDLAZECK, F. J. 2017. Transient structural variations have strong effects on
716 quantitative traits and reproductive isolation in fission yeast. *Nat Commun,* 8**,** 14061.
717 JURALEVICIUTE, M., POZNIAK, J., NSENGIMANA, J., HARLAND, M., RANDERSON-MOOR, J., WERNHOFF,
718 P., BASSAROVA, A., OY, G. F., TROEN, G., FLORENES, V. A., BISHOP, D. T., HERLYN, M., NEWTON-
719 BISHOP, J. & SLIPICEVIC, A. 2020. MX 2 is a novel regulator of cell cycle in melanoma cells.
720 *Pigment Cell Melanoma Res,* 33**,** 446-457.
721 KAMITAKI, N., HUJOEL, M. L. A., MUKAMEL, R. E., GEBARA, E., MCCARROLL, S. A. & LOH, P. R. 2024. A
722 sequence of SVA retrotransposon insertions in ASIP shaped human pigmentation. *Nat Genet,* 56**,**
723 1583-1591.
724 KARLSSON, E. K. & LINDBLAD-TOH, K. 2008. Leader of the pack: gene mapping in dogs and other model
725 organisms. *Nat Rev Genet,* 9**,** 713-25.
726 KAZAZIAN, H. H., JR. & MORAN, J. V. 2017. Mobile DNA in Health and Disease. *N Engl J Med,* 377**,** 361-
727 370.
728 KIM, Y. D., LEE, J. Y., OH, K. M., ARAKI, M., ARAKI, K., YAMAMURA, K. & JUN, C. D. 2011. NSrp70 is a
729 novel nuclear speckle-related protein that modulates alternative pre-mRNA splicing in vivo.
730 *Nucleic Acids Res,* 39**,** 4300-14.
731 KOSUGI, S., MOMOZAWA, Y., LIU, X., TERAO, C., KUBO, M. & KAMATANI, Y. 2019. Comprehensive
732 evaluation of structural variation detection algorithms for whole genome sequencing. *Genome*
733 *Biol,* 20**,** 117.
734 LADUNGA, I. 2017. Finding Homologs in Amino Acid Sequences Using Network BLAST Searches. *Curr*
735 *Protoc Bioinformatics,* 59**,** 3 4 1-3 4 24.
736 LANGEVIN, M., SYNKOVA, H., JANCUSKOVA, T. & PEKOVA, S. 2018. Merle phenotypes in dogs - SILV SINE
737 insertions from Mc to Mh. *PLoS One,* 13**,** e0198536.
738 LEMAY, M. A., SIBBESEN, J. A., TORKAMANEH, D., HAMEL, J., LEVESQUE, R. C. & BELZILE, F. 2022.
739 Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean
740 structural variation biology. *BMC Biol,* 20**,** 53.
741 LEVIN, H. L. & MORAN, J. V. 2011. Dynamic interactions between transposable elements and their hosts.
742 *Nat Rev Genet,* 12**,** 615-27.
743 LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available:
744 https://ui.adsabs.harvard.edu/abs/2013arXiv1303.3997L [Accessed March 01, 2013].
745 LI, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 34**,** 3094-3100.
746 LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G.,
747 DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map
748 format and SAMtools. *Bioinformatics,* 25**,** 2078-9.
749 LI, J., WANG, P., XIE, Z., WANG, S., CEN, S., LI, M., LIU, W., TANG, S., YE, G., ZHENG, G., SU, H., MA, M.,
750 WU, X., WU, Y. & SHEN, H. 2019. TRAF4 positively regulates the osteogenic differentiation of
751 mesenchymal stem cells by acting as an E3 ubiquitin ligase to degrade Smurf2. *Cell Death Differ,*
752 26**,** 2652-2666.
753 LIU, L., WANG, W., LIU, W., LI, X., YI, G., ADETULA, A. A., HUANG, H. & TANG, Z. 2024. Comprehensive
754 Atlas of Alternative Splicing Reveals NSRP1 Promoting Adipogenesis through CCDC18. *Int J Mol*
755 *Sci,* 25.
756 MAHMOUD, M., GOBET, N., CRUZ-DAVALOS, D. I., MOUNIER, N., DESSIMOZ, C. & SEDLAZECK, F. J. 2019.
757 Structural variant calling: the long and the short of it. *Genome Biol,* 20**,** 246.

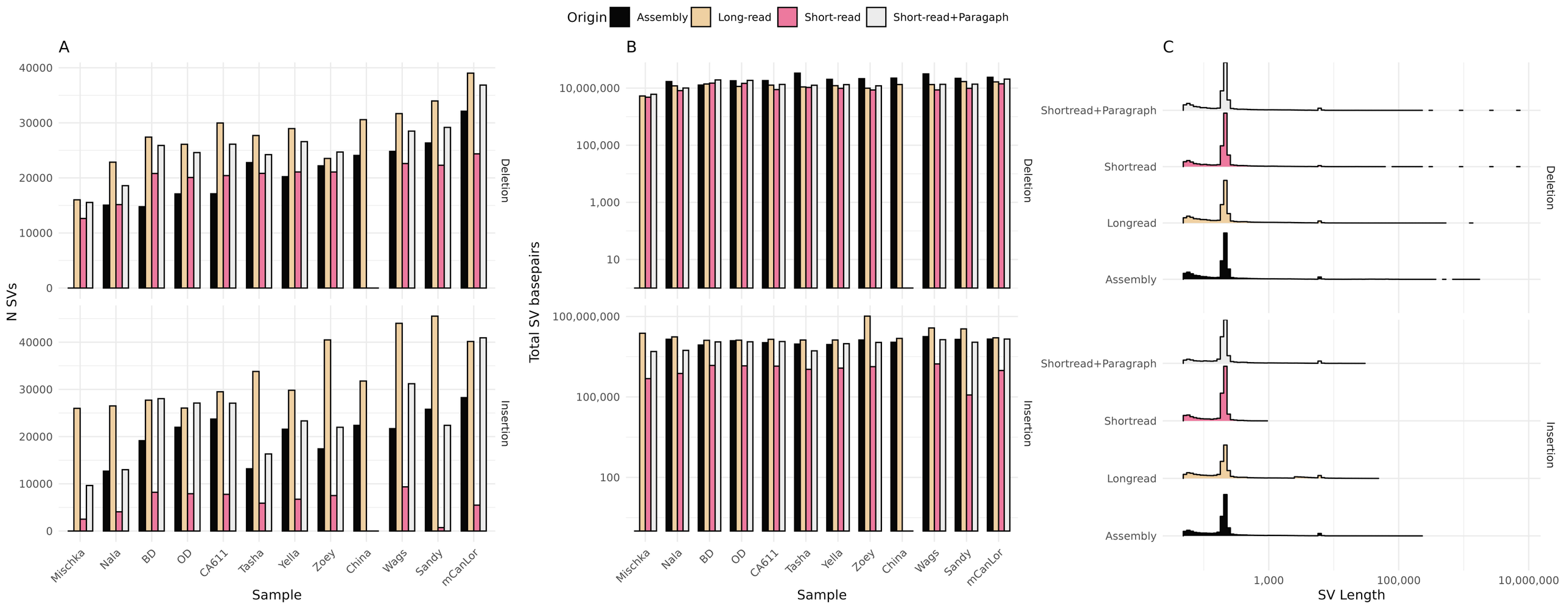
758 MARX, V. 2023. Method of the year: long-read sequencing. *Nat Methods,* 20**,** 6-11.
759 MAURI, N., KLEITER, M., DIETSCHI, E., LESCHNIK, M., HOGLER, S., WIEDMER, M., DIETRICH, J., HENKE, D.,
760 STEFFEN, F., SCHULLER, S., GURTNER, C., STOKAR-REGENSCHEIT, N., O'TOOLE, D., BILZER, T.,
761 HERDEN, C., OEVERMANN, A., JAGANNATHAN, V. & LEEB, T. 2017. A SINE Insertion in ATP1B2 in
762 Belgian Shepherd Dogs Affected by Spongy Degeneration with Cerebellar Ataxia (SDCA2). *G3*
763 *(Bethesda),* 7**,** 2729-2737.
764 MCGRAW, R. A. & CARMICHAEL, K. P. 2006. Molecular basis of globoid cell leukodystrophy in Irish
765 setters. *Vet J,* 171**,** 370-2.
766 MEADOWS, J. R. S., KIDD, J. M., WANG, G. D., PARKER, H. G., SCHALL, P. Z., BIANCHI, M., CHRISTMAS, M.
767 J., BOUGIOURI, K., BUCKLEY, R. M., HITTE, C., NGUYEN, A. K., WANG, C., JAGANNATHAN, V.,
768 NISKANEN, J. E., FRANTZ, L. A. F., ARUMILLI, M., HUNDI, S., LINDBLAD-TOH, K., GINJA, C.,
769 AGUSTINA, K. K., ANDRE, C., BOYKO, A. R., DAVIS, B. W., DROGEMULLER, M., FENG, X. Y.,
770 GKAGKAVOUZIS, K., ILIOPOULOS, G., HARRIS, A. C., HYTONEN, M. K., KALTHOFF, D. C., LIU, Y. H.,
771 LYMBERAKIS, P., POULAKAKIS, N., PIRES, A. E., RACIMO, F., RAMOS-ALMODOVAR, F.,
772 SAVOLAINEN, P., VENETSANI, S., TAMMEN, I., TRIANTAFYLLIDIS, A., VONHOLDT, B., WAYNE, R.
773 K., LARSON, G., NICHOLAS, F. W., LOHI, H., LEEB, T., ZHANG, Y. P. & OSTRANDER, E. A. 2023.
774 Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of
775 demography, genome function and architecture. *Genome Biol,* 24**,** 187.
776 MEYER, L. R., ZWEIG, A. S., HINRICHS, A. S., KAROLCHIK, D., KUHN, R. M., WONG, M., SLOAN, C. A.,
777 ROSENBLOOM, K. R., ROE, G., RHEAD, B., RANEY, B. J., POHL, A., MALLADI, V. S., LI, C. H., LEE, B.
778 T., LEARNED, K., KIRKUP, V., HSU, F., HEITNER, S., HARTE, R. A., HAEUSSLER, M., GURUVADOO,
779 L., GOLDMAN, M., GIARDINE, B. M., FUJITA, P. A., DRESZER, T. R., DIEKHANS, M., CLINE, M. S.,
780 CLAWSON, H., BARBER, G. P., HAUSSLER, D. & KENT, W. J. 2013. The UCSC Genome Browser
781 database: extensions and updates 2013. *Nucleic Acids Res,* 41**,** D64-9.
782 MILLER, A. D., MILLER, C. R. & ROSSMEISL, J. H. 2019. Canine Primary Intracranial Cancer: A
783 Clinicopathologic and Comparative Review of Glioma, Meningioma, and Choroid Plexus Tumors.
784 *Front Oncol,* 9**,** 1151.
785 PANG, A. W., MACDONALD, J. R., PINTO, D., WEI, J., RAFIQ, M. A., CONRAD, D. F., PARK, H., HURLES, M.
786 E., LEE, C., VENTER, J. C., KIRKNESS, E. F., LEVY, S., FEUK, L. & SCHERER, S. W. 2010. Towards a
787 comprehensive structural variation map of an individual human genome. *Genome Biol,* 11**,** R52.
788 PARKER, H. G., KUKEKOVA, A. V., AKEY, D. T., GOLDSTEIN, O., KIRKNESS, E. F., BAYSAC, K. C., MOSHER, D.
789 S., AGUIRRE, G. D., ACLAND, G. M. & OSTRANDER, E. A. 2007. Breed relationships facilitate fine-
790 mapping studies: a 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog
791 breeds. *Genome Res,* 17**,** 1562-71.
792 PARKER, H. G., VONHOLDT, B. M., QUIGNON, P., MARGULIES, E. H., SHAO, S., MOSHER, D. S., SPADY, T.
793 C., ELKAHLOUN, A., CARGILL, M., JONES, P. G., MASLEN, C. L., ACLAND, G. M., SUTTER, N. B.,
794 KUROKI, K., BUSTAMANTE, C. D., WAYNE, R. K. & OSTRANDER, E. A. 2009. An expressed fgf4
795 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science,* 325**,**
796 995-8.
797 PLAGER, D. A., TORRES, S. M., KOCH, S. N. & KITA, H. 2012. Gene transcription abnormalities in canine
798 atopic dermatitis and related human eosinophilic allergic diseases. *Vet Immunol Immunopathol,*
799 149**,** 136-42.
800 PLASSAIS, J., KIM, J., DAVIS, B. W., KARYADI, D. M., HOGAN, A. N., HARRIS, A. C., DECKER, B., PARKER, H.
801 G. & OSTRANDER, E. A. 2019. Whole genome sequencing of canids reveals genomic regions
802 under selection and variants influencing morphology. *Nat Commun,* 10**,** 1489.
803 PLAYER, R. A., FORSYTH, E. R., VERRATTI, K. J., MOHR, D. W., SCOTT, A. F. & BRADBURNE, C. E. 2021. A
804 novel canis lupus familiaris reference genome improves variant resolution for use in breed-
805 specific GWAS. *Life Sci Alliance,* 4.

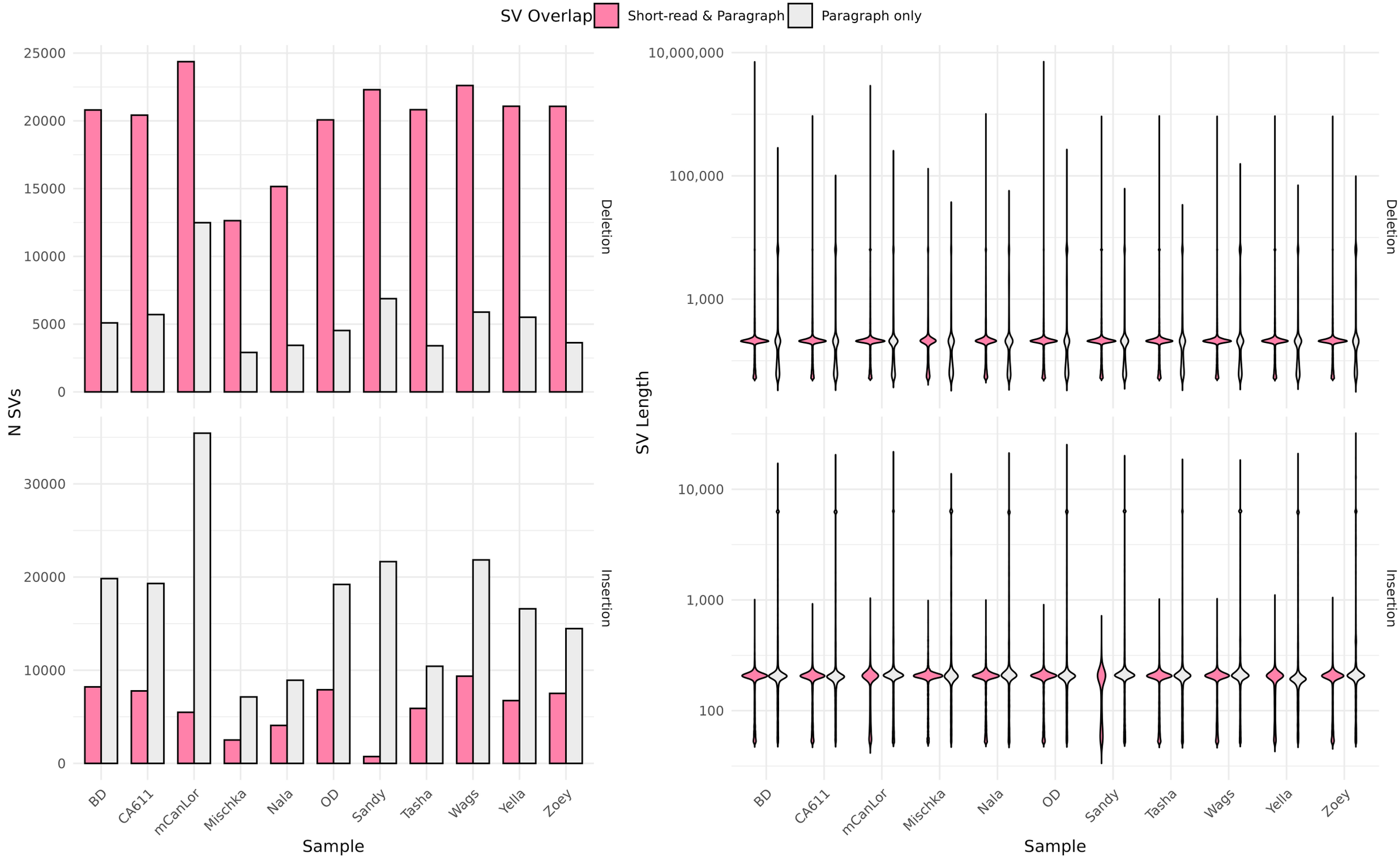
806 REGNIER, C. H., MASSON, R., KEDINGER, V., TEXTORIS, J., STOLL, I., CHENARD, M. P., DIERICH, A.,
807 TOMASETTO, C. & RIO, M. C. 2002. Impaired neural tube closure, axial skeleton malformations,
808 and tracheal ring disruption in TRAF4-deficient mice. *Proc Natl Acad Sci U S A,* 99**,** 5585-90.
809 RIMBAULT, M., BEALE, H. C., SCHOENEBECK, J. J., HOOPES, B. C., ALLEN, J. J., KILROY-GLYNN, P., WAYNE,
810 R. K., SUTTER, N. B. & OSTRANDER, E. A. 2013. Derived variants at six genes explain nearly half of
811 size reduction in dog breeds. *Genome Res,* 23**,** 1985-95.
812 RUAN, J. & LI, H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods,* 17**,** 155-158.
813 SALMON HILLBERTZ, N. H., ISAKSSON, M., KARLSSON, E. K., HELLMEN, E., PIELBERG, G. R., SAVOLAINEN,
814 P., WADE, C. M., VON EULER, H., GUSTAFSON, U., HEDHAMMAR, A., NILSSON, M., LINDBLAD-
815 TOH, K., ANDERSSON, L. & ANDERSSON, G. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1
816 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat Genet,* 39**,** 1318-
817 20.
818 SCHALL, P. Z., WINKLER, P. A., PETERSEN-JONES, S. M., YUZBASIYAN-GURKAN, V. & KIDD, J. M. 2023.
819 Genome-wide methylation patterns from canine nanopore assemblies. *G3 (Bethesda),* 13.
820 SCHLAMP, F., VAN DER MADE, J., STAMBLER, R., CHESEBROUGH, L., BOYKO, A. R. & MESSER, P. W. 2016.
821 Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Mol*
822 *Ecol,* 25**,** 342-56.
823 SHEARIN, A. L. & OSTRANDER, E. A. 2010. Canine morphology: hunting for genes and tracking mutations.
824 *PLoS Biol,* 8**,** e1000310.
825 SINDING, M. S., GOPALAKRISHNAN, S., RAUNDRUP, K., DALEN, L., THRELFALL, J., DARWIN TREE OF LIFE
826 BARCODING, C., WELLCOME SANGER INSTITUTE TREE OF LIFE, P., WELLCOME SANGER
827 INSTITUTE SCIENTIFIC OPERATIONS, D. N. A. P. C., TREE OF LIFE CORE INFORMATICS, C., DARWIN
828 TREE OF LIFE, C. & GILBERT, T. 2021. The genome sequence of the grey wolf, Canis lupus
829 Linnaeus 1758. *Wellcome Open Res,* 6**,** 310.
830 SMOLKA, M., PAULIN, L. F., GROCHOWSKI, C. M., HORNER, D. W., MAHMOUD, M., BEHERA, S., KALEF-
831 EZRA, E., GANDHI, M., HONG, K., PEHLIVAN, D., SCHOLZ, S. W., CARVALHO, C. M. B., PROUKAKIS,
832 C. & SEDLAZECK, F. J. 2024. Detection of mosaic and population-level structural variants with
833 Sniffles2. *Nat Biotechnol,* 42**,** 1571-1580.
834 TRUVE, K., DICKINSON, P., XIONG, A., YORK, D., JAYASHANKAR, K., PIELBERG, G., KOLTOOKIAN, M.,
835 MUREN, E., FUXELIUS, H. H., WEISHAUPT, H., SWARTLING, F. J., ANDERSSON, G., HEDHAMMAR,
836 A., BONGCAM-RUDLOFF, E., FORSBERG-NILSSON, K., BANNASCH, D. & LINDBLAD-TOH, K. 2016.
837 Utilizing the Dog Genome in the Search for Novel Candidate Genes Involved in Glioma
838 Development-Genome Wide Association Mapping followed by Targeted Massive Parallel
839 Sequencing Identifies a Strongly Associated Locus. *PLoS Genet,* 12**,** e1006000.
840 VAN POUCKE, M., LEDEGANCK, L., GUO, L. T., SHELTON, G. D., BHATTI, S. F. M., CORNELIS, I. & PEELMAN,
841 L. 2024. Exonisation of an intronic L1 element in the dystrophin gene associated with X-linked
842 muscular dystrophy in a Border Collie dog. *Anim Genet,* 55**,** 733-743.
843 VAYSSE, A., RATNAKUMAR, A., DERRIEN, T., AXELSSON, E., ROSENGREN PIELBERG, G., SIGURDSSON, S.,
844 FALL, T., SEPPALA, E. H., HANSEN, M. S., LAWLEY, C. T., KARLSSON, E. K., CONSORTIUM, L.,
845 BANNASCH, D., VILA, C., LOHI, H., GALIBERT, F., FREDHOLM, M., HAGGSTROM, J., HEDHAMMAR,
846 A., ANDRE, C., LINDBLAD-TOH, K., HITTE, C. & WEBSTER, M. T. 2011. Identification of genomic
847 regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS*
848 *Genet,* 7**,** e1002316.
849 WANG, C., WALLERMAN, O., ARENDT, M. L., SUNDSTROM, E., KARLSSON, A., NORDIN, J., MAKELAINEN,
850 S., PIELBERG, G. R., HANSON, J., OHLSSON, A., SAELLSTROM, S., RONNBERG, H., LJUNGVALL, I.,
851 HAGGSTROM, J., BERGSTROM, T. F., HEDHAMMAR, A., MEADOWS, J. R. S. & LINDBLAD-TOH, K.
852 2021. A novel canine reference genome resolves genomic architecture and uncovers transcript
853 complexity. *Commun Biol,* 4**,** 185.

854 WANG, W. & KIRKNESS, E. F. 2005. Short interspersed elements (SINEs) are a major source of canine
855 genomic diversity. *Genome Res,* 15**,** 1798-808.
856 WEBSTER, M. T., KAMGARI, N., PERLOSKI, M., HOEPPNER, M. P., AXELSSON, E., HEDHAMMAR, A.,
857 PIELBERG, G. & LINDBLAD-TOH, K. 2015. Linked genetic variants on chromosome 10 control ear
858 morphology and body mass among dog breeds. *BMC Genomics,* 16**,** 474.
859 XIA, B., ZHANG, W., ZHAO, G., ZHANG, X., BAI, J., BROSH, R., WUDZINSKA, A., HUANG, E., ASHE, H., ELLIS,
860 G., POUR, M., ZHAO, Y., COELHO, C., ZHU, Y., MILLER, A., DASEN, J. S., MAURANO, M. T., KIM, S.
861 Y., BOEKE, J. D. & YANAI, I. 2024. On the genetic basis of tail-loss evolution in humans and apes.
862 *Nature,* 626**,** 1042-1048.
863 YAN, S. M., SHERMAN, R. M., TAYLOR, D. J., NAIR, D. R., BORTVIN, A. N., SCHATZ, M. C. & MCCOY, R. C.
864 2021. Local adaptation and archaic introgression shape global diversity at human structural
865 variant loci. *Elife,* 10.
866 YANG, Q., WU, P., WANG, K., CHEN, D., ZHOU, J., MA, J., LI, M., XIAO, W., JIANG, A., JIANG, Y., BAI, L.,
867 ZHU, L., LI, X. & TANG, G. 2019. SNPs associated with body weight and backfat thickness in two
868 pig breeds identified by a genome-wide association study. *Genomics,* 111**,** 1583-1589.
869 YAO, W., WANG, X., CAI, Q., GAO, S., WANG, J. & ZHANG, P. 2014a. Knockdown of TRAF4 expression
870 suppresses osteosarcoma cell growth in vitro and in vivo. *Int J Mol Med,* 34**,** 1655-60.
871 YAO, W., WANG, X., CAI, Q., GAO, S., WANG, J. & ZHANG, P. 2014b. TRAF4 enhances osteosarcoma cell
872 proliferation and invasion by Akt signaling pathway. *Oncol Res,* 22**,** 21-28.

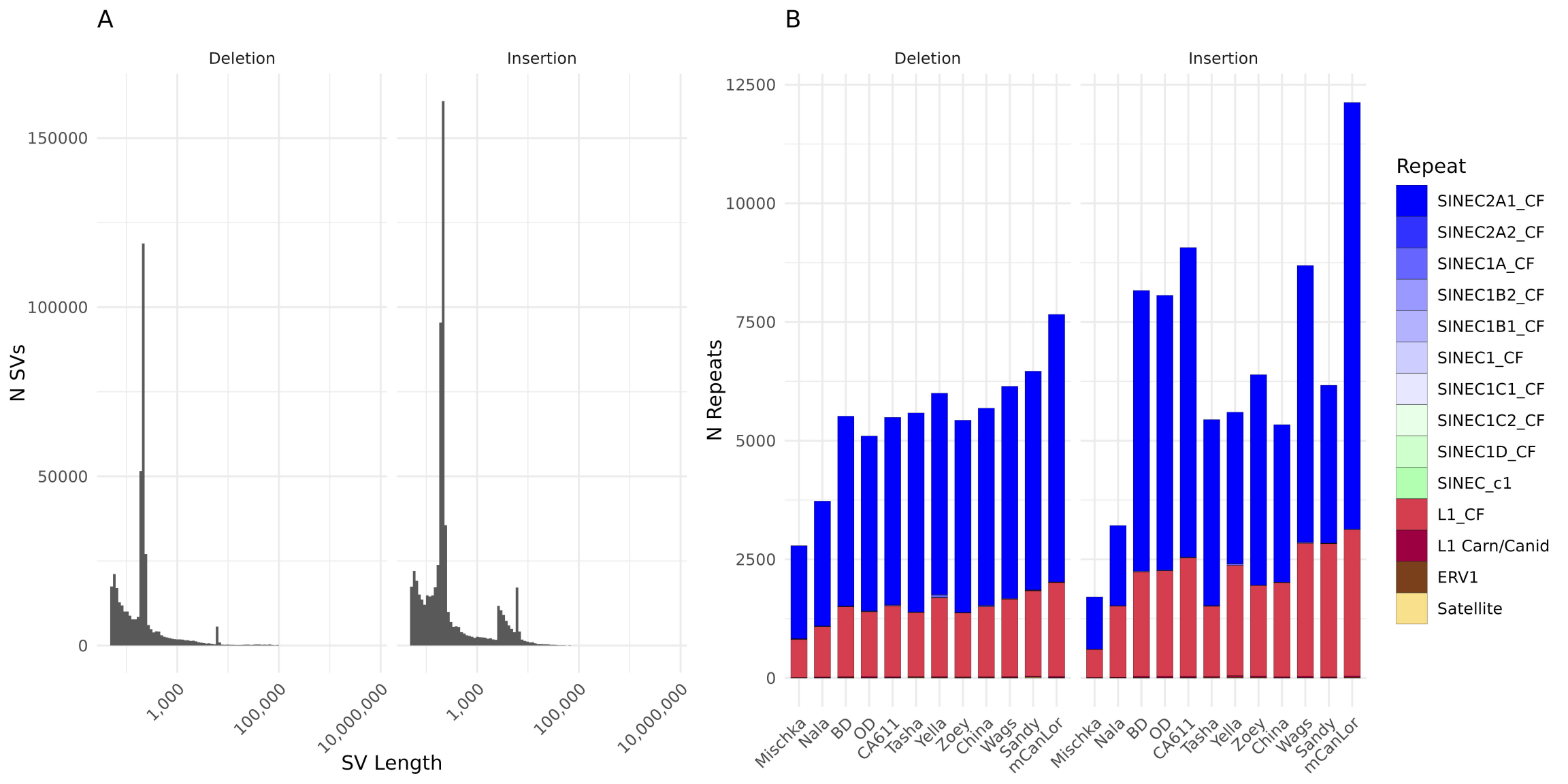
873

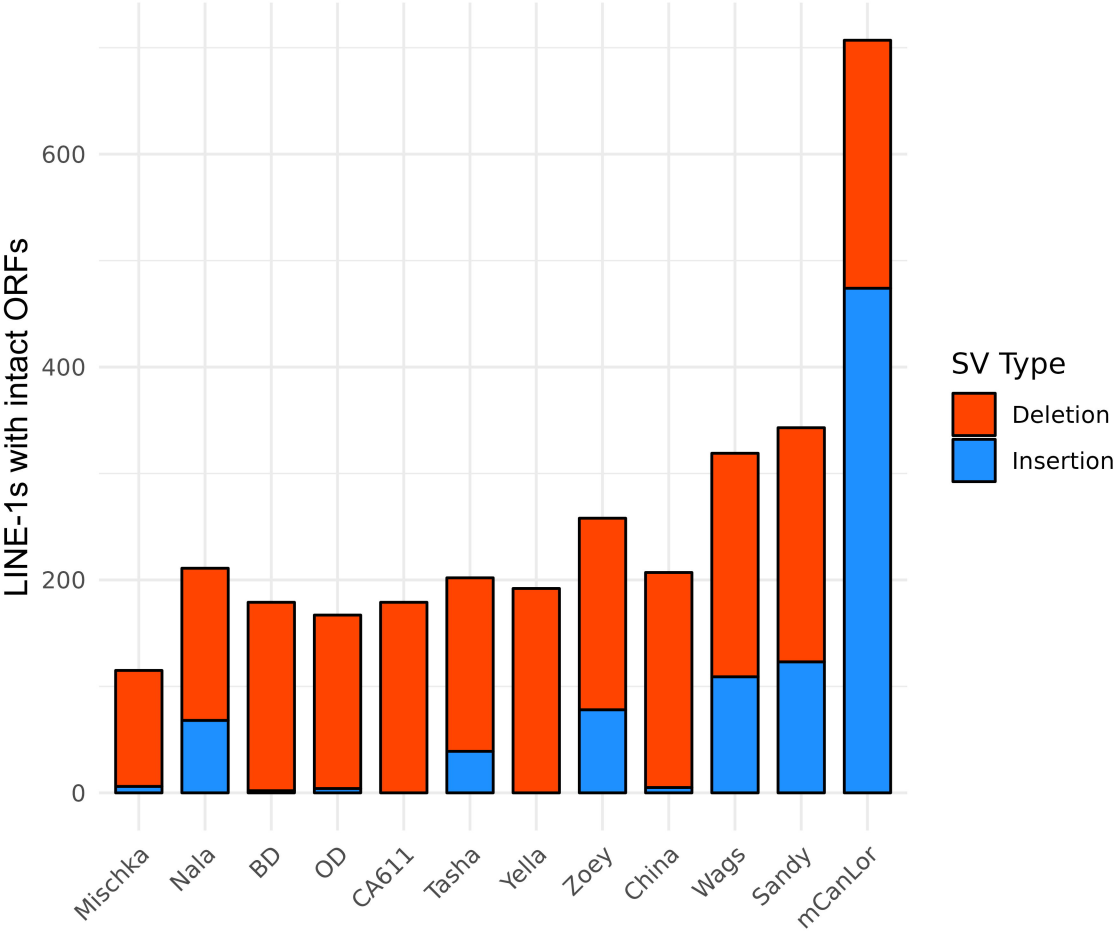

Input Samples
N=12
Dog10K
N=1,879
Assembly SVs
Short-ead SVs
Long-read SVs
Genotyping
Genotyping
Final Sample SVs
Final Dog10K SVs
Repeat Annotation
Genomic Context Annotation
Ohana

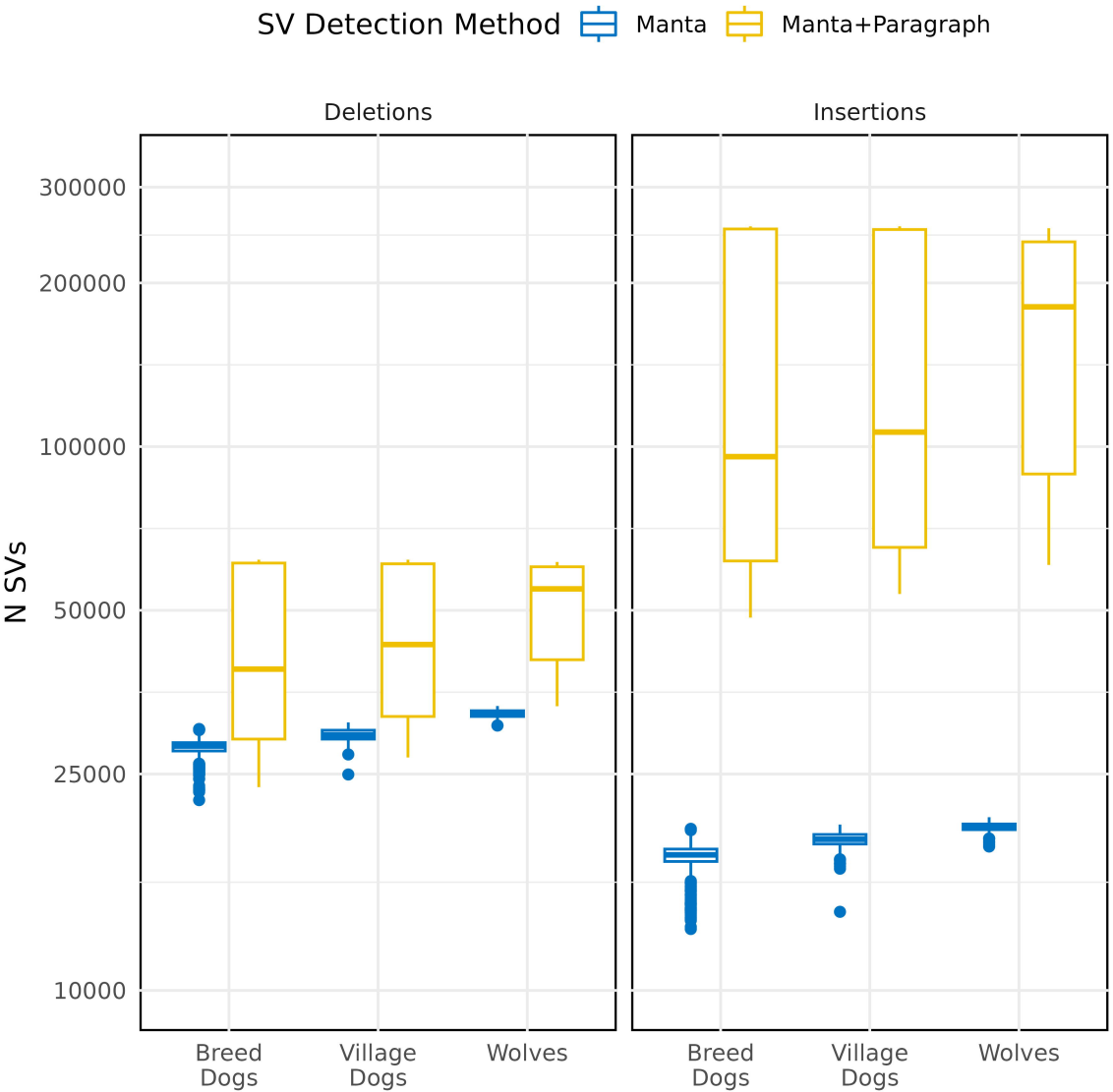

Origin
Assembly
Long-read
Short-read
Short-read+Paragaph
A
B
C
N SVs
Total SV basepairs
Deletion
Insertion
Mischka
Nala
BD
OD
CA611
Tasha
Yella
Zoey
China
Wags
Sandy
mCanLor
Sample
Shortread+Paragraph
Shortread
Longread
Assembly
SV Length
1,000
100,000
10,000,000

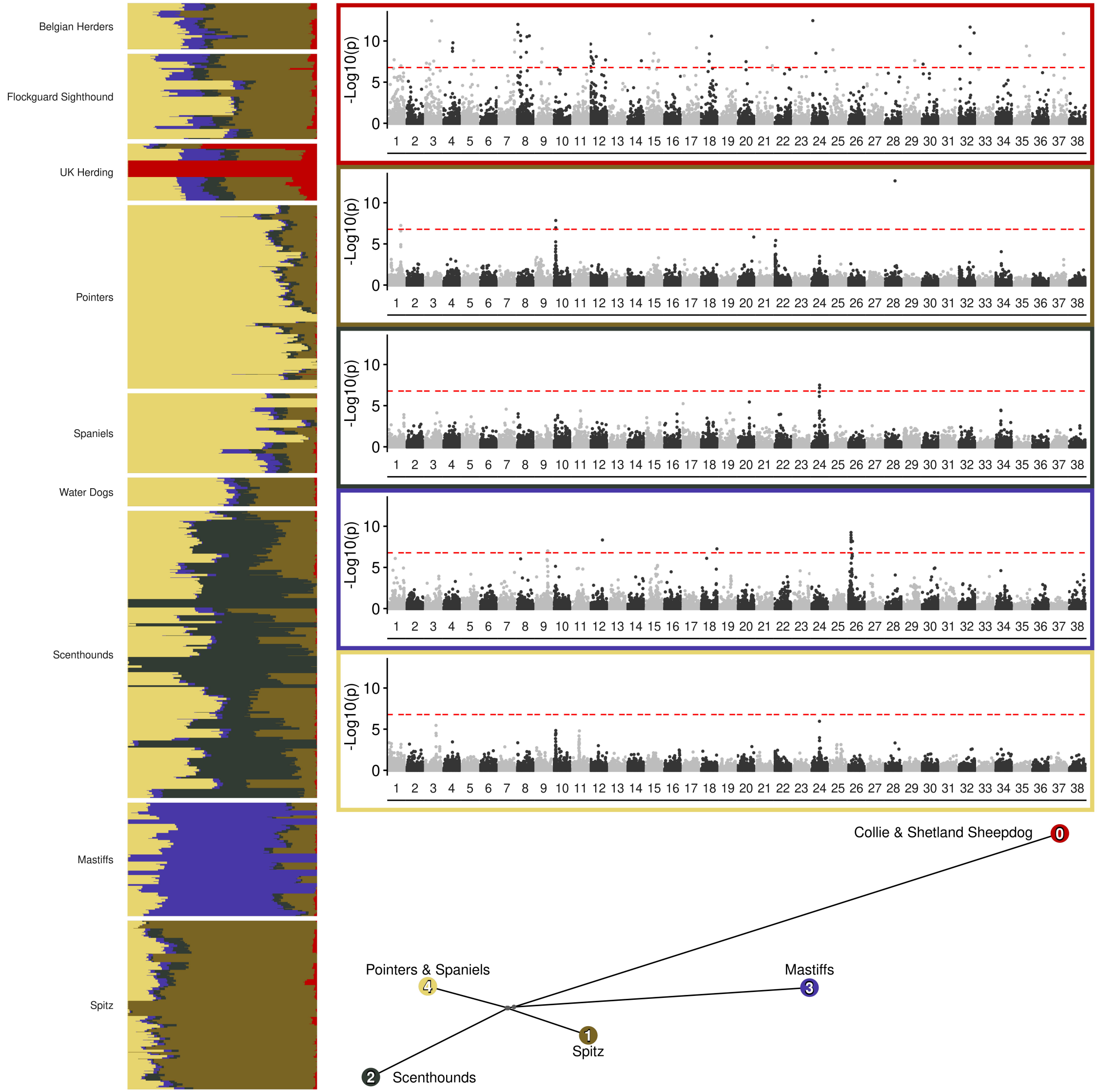

SV Overlap
Short-read & Paragraph
Paragraph only
N SVs
Deletion
Insertion
SV Length
Sample
BD
CA611
mCanLor
Mischka
Nala
OD
Sandy
Tasha
Wags
Yella
Zoey

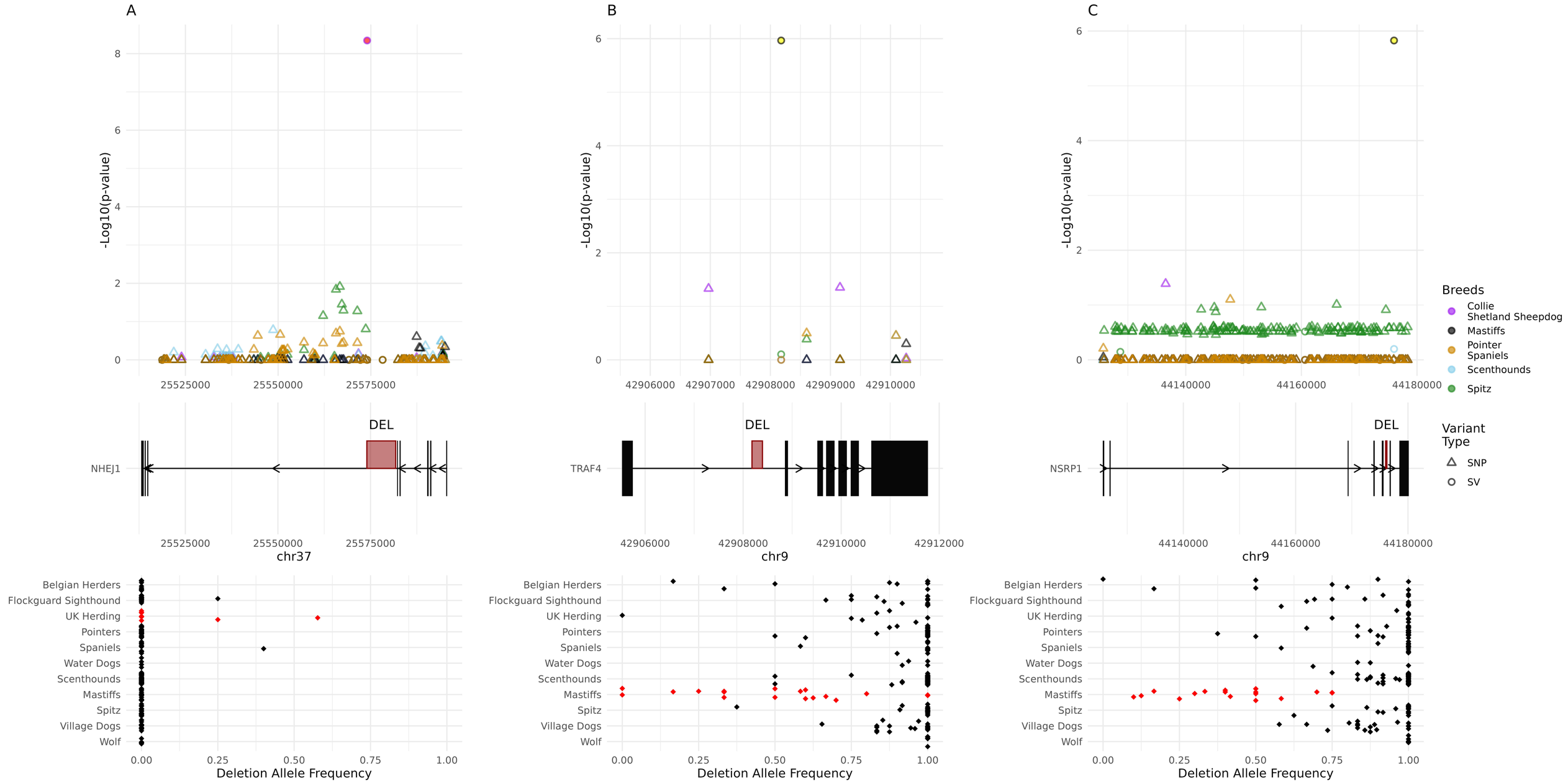
Sample
Nala
CA611
Zoey
Sandy
BD
Tasha
China
mCanLor
OD
Yella
Wags
SV Type
Deletions
Insertions
Long-read vs. Assembly
Fraction of Assembly SVs
Fraction of Long-read SVs
Long-read vs. Short-read
Fraction of Short-read SVs
Fraction of Long-read SVs
Long-read vs. Short-read+Paragraph
Fraction of Short-read+Paragraph SVs
Fraction of Long-read SVs
Assembly vs. Short-read
Fraction of Short-read SVs
Fraction of Assembly SVs
Assembly vs. Short-read+Paragraph
Fraction of Short-read+Paragraph SVs
Fraction of Assembly SVs
0.00
0.25
0.50
0.75
1.00


A
Deletion
Insertion
N SVs
150000
100000
50000
0
1,000
100,000
10,000,000
SV Length
B
Deletion
Insertion
N Repeats
12500
10000
7500
5000
2500
0
Mischka
Nala
BD
OD
CA611
Tasha
Yella
Zoey
China
Wags
Sandy
mCanLor
Repeat
SINEC2A1_CF
SINEC2A2_CF
SINEC1A_CF
SINEC1B2_CF
SINEC1B1_CF
SINEC1_CF
SINEC1C1_CF
SINEC1C2_CF
SINEC1D_CF
SINEC_c1
L1_CF
L1 Carn/Canid
ERV1
Satellite


LINE-1s with intact ORFs
0
200
400
600
Mischka
Nala
BD
OD
CA611
Tasha
Yella
Zoey
China
Wags
Sandy
mCanLor
SV Type
Deletion
Insertion


SV Detection Method
Manta
Manta+Paragraph
Deletions
Insertions
N SVs
300000
200000
100000
50000
25000
10000
Breed Dogs
Village Dogs
Wolves
Breed Dogs
Village Dogs
Wolves


Belgian Herders
Flockguard Sighthound
UK Herding
Pointers
Spaniels
Water Dogs
Scenthounds
Mastiffs
Spitz
-Log10(p)
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Collie & Shetland Sheepdog
0
Pointers & Spaniels
4
Mastiffs
3
1
Spitz
2
Scenthounds


A
B
C
-Log10(p-value)
25525000
25550000
25575000
42906000
42907000
42908000
42909000
42910000
44140000
44160000
44180000
Breeds
Collie
Shetland Sheepdog
Mastiffs
Pointer
Spaniels
Scenthounds
Spitz
DEL
NHEJ1
TRAF4
NSRP1
Variant Type
SNP
SV
chr37
chr9
Belgian Herders
Flockguard Sighthound
UK Herding
Pointers
Spaniels
Water Dogs
Scenthounds
Mastiffs
Spitz
Village Dogs
Wolf
0.00
0.25
0.50
0.75
1.00
Deletion Allele Frequency