# A new domestic cat genome assembly based on long sequence reads empowers feline genomic medicine and identifies a novel gene for dwarfism.

Reuben M. Buckley[1], Brian W. Davis[2], Wesley A. Brashear[2], Fabiana H. G. Farias[3], Kei Kuroki[4], Tina Graves[3], LaDeana W. Hillier[3], Milinn Kremitzki[3], Gang Li[2], Rondo Middleton[5], Patrick Minx[3], Chad Tomlinson[3], Leslie A. Lyons[1], William J. Murphy[2], and Wesley C. Warren[5,6]

[1]Department of Veterinary Medicine and Surgery, College of Veterinary Medicine, University of Missouri, Columbia, Missouri 65211, USA.

[2]Department of Veterinary Integrative Biosciences, Interdisciplinary Program in Genetics, College of Veterinary Medicine, Texas A&M University, College Station, Texas 77843, USA

[3]McDonnell Genome Institute, Washington University, School of Medicine, St Louis, Missouri 63108, USA;

[4]Veterinary Medical Diagnostic Laboratory, College of Veterinary Medicine, University of Missouri, Columbia, Missouri, 65211, USA

[5]Nestlé Purina Research, Saint Louis, Missouri 63164, USA

[6]Division of Animal Sciences, School of Medicine, University of Missouri, Columbia, Missouri 65211, USA

Corresponding author: warrenwc@missouri.edu

Running title: **Feline genomic medicine resources**

Keywords: animal model, chondrodysplasia, dwarfism, *Felis catus*, long-read, Precision Medicine

**Abstract**

The domestic cat (*Felis catus*) numbers over 77 million in the USA alone, occupies households as a companion animal, and, like humans, suffers from cancer and common and rare diseases. However, genome-wide sequence variant information is limited for this species. To empower trait analyses, a new cat genome reference assembly was developed from PacBio long sequence reads that significantly improves sequence representation and assembly contiguity. The whole genome sequences of 54 domestic cats were aligned to the reference to identify single nucleotide variants (SNVs) and structural variants (SVs). Across all cats, 18 SNVs predicted to have deleterious impacts and in a singleton state were identified as high priority candidates for causative mutations. One candidate was a stop gain in the tumor suppressor *FBXW7*. The SNV is found in cats segregating for feline mediastinal lymphoma and is a candidate for inherited cancer susceptibility. SV analysis revealed a complex deletion coupled with a nearby potential duplication event that was shared privately across three unrelated dwarfism cats and is found within a known dwarfism associated region on cat chromosome B1. This SV interrupted *UDP-glucose 6-dehydrogenase (UGDH)*, a gene involved in the biosynthesis of glycosaminoglycans. Importantly, *UGDH* has not yet been associated with human dwarfism and should be screened in undiagnosed patients. Overall, the new high-quality cat genome reference and the compilation of sequence variation demonstrate the importance of these resources when searching for disease causative alleles in the domestic cat and for identification of feline biomedical models.

## Introduction

In the veterinary clinic, the practice of genomic medicine is impending (Mauler et al. 2017). With actionable genetic information in hand, companion animal therapeutic interventions are feasible, including treatment of animal patients prior to, or to prevent the appearance of, more severe symptoms and allow therapeutic administration of drugs with higher efficacy and fewer side effects. Genomic information can also alert veterinarians to imminent disease risks for diagnostic consideration. Each of these applications could significantly enhance veterinary medicine, however, none are currently in practice. As in human medicine, formidable challenges exist for the implementation of genomic-based medicine, including accurate annotation of the genome and databases of genetic variation from well-phenotyped individuals that include both single nucleotide variant (SNV) and structural variant (SV) discovery and annotation (Fang et al. 2017; Shendure et al. 2019; Wise et al. 2019). Targeted individual companion animal genome information is becoming more readily available, cost effective, and tentatively linked to the actionable phenotypes via direct-to-consumer DNA testing. Thus correct interpretation of DNA variants is of the upmost importance for communicating findings to clinicians practicing companion animal genomic medicine (Moses et al. 2018).

Companion animals suffer from many of the same diseases as humans, with over 600 different phenotypes identified as comparative models for human physiology, biology, development, and disease (Online Mendelian Inheritance in Animals (OMIA) ; Nicholas 2003). In domestic cats, at least 70 genes are shown to harbor single and multiple DNA variants that are associated with disease (Lyons 2015) with many more discoveries expected as health care improves.. The genetic and clinical manifestations of most of these known variants are described in the Online Mendelian Inheritance in Animals (Nicholas 2003). Examples include common human diseases such as cardiomyopathy (Kittleson et al. 2015), retinal degenerations (Menotti-Raymond et al. 2007), and polycystic kidney disease (Lyons et al. 2004). Veterinarians, geneticists and other

researchers are actively banking DNA from companion animals and attempting to implement genomic medicine (Mauler et al. 2017). Once coupled with the quickly advancing sequencing technology and exploitable results, genomic medicine in companion animals promises to expand the comparative knowledge of mechanisms of action across species. Despite the continuing successful discovery of feline disease variants using both candidate gene and whole genome sequencing (WGS) approaches (Mauler et al. 2017; Spycher et al. 2018; Wang et al. 2018; Hug et al. 2019; Jaffey et al. 2019), the understanding of normal and disease sequence variation in the domestic cat and interrogation of gene structure and function is limited by an incomplete genome assembly.

A fundamental hurdle hampering the interpretation of feline disease variant data is the availability of a high-quality, gapless reference genome. The previous domestic cat reference, Felis_catus_8.0, contains over 300,000 gaps, compromising its utility for identifying all types of sequence variation (Li et al. 2016), in particular SVs. In conjunction with various mapping technologies, such as optical resolved physical maps, recent advances in the use of long-read sequencing and assembly technology has produced a more complete genome representation (i.e., fewer gaps) for many species (Gordon et al. 2016; Bickhart et al. 2017; Ananthasayanam 2019; Low et al. 2019).

Another hurdle for performing feline genomic medicine is the availability of WGS data from various breeds of cat with sufficient sequencing depth to uncover rare alleles and complex structural variants. Knowledge of variant frequency and uniqueness among domestic cats is very limited and is crucial in the identification of causal alleles. As a result of the paucity of sequence variant data across breeds, the 99 Lives Cat Genome Sequencing Initiative was founded as a centralized resource with genome sequences produced of similar quality and techniques. The resource supports researchers with variant discovery for evolutionary studies and identifying the genetic origin of inherited diseases and can assist in the development of

high-density DNA arrays for complex disease studies in domestic cats (Aberdein et al. 2017; Mauler et al. 2017; Oh et al. 2017; Genova et al. 2018; Ontiveros et al. 2018).

Here we present a new version of the domestic cat genome reference (Cinnamon, an Abyssinian breed), generated from deep sequence coverage of long-reads and scaffolding from an optical map (BioNano) and a high-density genetic linkage map (Li et al. 2016). Published cat genomes from the 99 Lives Cat Genome Consortium (Aberdein et al. 2017; Mauler et al. 2017) were aligned to the Felis_catus_9.0 reference to discover a plethora of unknown SNVs and SVs (multi-base insertions and deletions), including a newly identified structural variant (SV) for feline disproportionate dwarfism. Our case study of dwarfism demonstrates when disease phenotypes are coupled with revised gene annotation and sequence variation ascertained from diverse breeds, the new cat genome assembly is a powerful resource for trait discovery. This enables the future practice of feline genomic medicine and improved ascertainment of biomedical models relevant to human health.

## Results

**Genome assembly.** A female Abyssinian cat (Cinnamon) was sequenced to high-depth (72-fold coverage) using real-time (SMRT; PacBio) sequence data and all sequence reads were used to generate a *de novo* assembly. Two PacBio instruments were used to produce average read insert lengths of 12 kb (RSII) and 9 kb (Sequel). The ungapped assembly size was 2.48 Gb and is comparable in size to other assembled carnivores (**Table 1**). There were 4,909 total contigs compared to 367,672 contigs in Felis_catus_8.0 showing a significant reduction in sequence gaps. The assembly contiguity metric of N50 contig and scaffold lengths were 42 and 84 Mb, respectively (**Table 1**). The N50 contig length of other PacBio sequenced carnivore assemblies are less contiguous, ranging from 3.13 Mb to 20.91 Mb (**Table 1**). Across carnivores, RepeatMasker showed consistent measures of total interspersed repeat content, (with 43% in Felis_catus_9.0; **Supplemental Table S1**) (Smit et al. 2013). Due to repetitive and other

genome architecture features, 1.8% (46 Mb) of all assembled sequences remained unassigned to any chromosome position. These sequences had an N50 scaffold length of 12,618 bp, demonstrating the assembly challenge of some repeat types in diploid genome assemblies, even of an inbred individual.

1 **Table 1.** Representative assembly metrics for various chromosome level assembled carnivore genomes[1].

| Assembly | Species | Breed | Isolate | Release date (MM/DD/YY) | Sequencing technology | Genome coverage | Total ungapped length (Gb) | Scaffold N50 (Mb) | Contig N50 (Mb) | Unplaced length (Mb) |
|---|---|---|---|---|---|---|---|---|---|---|
| Felis_catus_9.0 | *Felis catus* (domestic cat) | Abyssinian | Cinnamon | 11/20/17 | PacBio; 454 Titanium; Illumina; Sanger dideoxy sequencing | 72x | 2.48 | 83.97 | 41.92 | 46.02 |
| Felis_catus_8.0 | *Felis catus* (domestic cat) | Abyssinian | Cinnamon | 11/07/14 | Sanger; 454 Titanium; Illumina | 2x Sanger; 14x 454, 20x Illumina | 2.60 | 18.07 | 0.05 | 73.71 |
| mLynCan4_v1.p | *Lynx canadensis* (Canada lynx) | NA | LIC74 | 07/26/19 | PacBio Sequel I; 10X genome; Bionano Genomics; Arima Genomics Hi-C | 72x | 2.41 | 146.11 | 7.50 | 6.18 |
| PanLeo1.0 | *Panthera leo* (lion) | NA | Brooke | 10/07/19 | Illumina; Oxford Nanopore; 10X Genomics | 46x | 2.39 | 136.05 | 0.29 | 242.29 |
| ASM864105v1 | *Canis lupus familiaris* (dog) | German Shepherd | Nala | 09/25/19 | PacBio Sequel; Oxford Nanopore PromethION; Illumina (10X Chromium) | 30x | 2.40 | 64.35 | 20.91 | 22.83 |
| ASM488618v2 | *Canis lupus familiaris* (dog) | Basenji | MU ID 185726 | 08/16/19 | Sequel | 45x | 2.41 | 61.09 | 3.13 | 120.80 |
| UMICH_Zoey_3.1 | *Canis lupus familiaris* (dog) | Great Dane | Zoey | 05/30/19 | PacBio RSII | 50x | 2.34 | 64.20 | 4.72 | 16.82 |
| CanFam3.1 | *Canis lupus familiaris* (dog) | boxer | Tasha | 11/02/11 | Sanger | 7x plus >90Mb finished sequence | 2.39 | 45.88 | 0.27 | 75.10 |

2 [1]All species-specific assembly metrics derived from the NCBI assembly archive.

**Sequence accuracy and quality assessment**. Illumina whole-genome sequence data from Cinnamon was used to identify reference sequence errors as homozygous SNVs. These numbered 60,449 in total, indicating a high level of sequence accuracy across assembled contigs (>99.9%). Sequence order and orientation was also highly accurate (>98%), as only 1.2% of BAC-end sequence alignments derived from Cinnamon were identified as discordant. Felis_catus_9.0 sequence order and orientation was also supported by high levels of agreement between individual chromosome sequence alignment and ordered sequence markers from the published cat genetic linkage map (Li et al. 2016) (**Supplemental Data S1**). The raw sequence data, assembled contigs, and sequence order coordinates (AGP) were accessioned and are fully available by searching GCF_000181335.3 in GenBank.

**Gene annotation**. The number of annotated protein-coding genes was very similar between the NCBI and Ensembl pipelines at 19,748 and 19,409, respectively. Approximately 376 protein-coding genes (NCBI) were identified as novel with no matching annotations in Felis_catus_8.0 **(Supplemental Data S2)**. Conversely, 178 genes from Felis_catus_8.0 did not map to Felis_catus_9.0, of which the cause is unknown (**Supplemental Table S1**). A large portion of genes changed substantially (8.4%) in Felis_catus_9.0 during NCBI annotation (**Supplemental Data S2**). Aligned sequence of known same-species RefSeq transcripts (n = 420) to Felis_catus_9.0 is higher (99.5%) than Felis_catus_8.0 (97.8%) and the mean coverage of these same translated RefSeq proteins is also improved (90.1% versus 88.3% in Felis_catus_8.0). One important consequence of the less fragmented gene annotation is a 2% increase in aggregate sequence alignments of feline RNA-seq datasets to Felis_catus_9.0. These improvements are largely attributed to fewer assembly gaps. The various reported metrics of gene annotation quality conclusively show the protein-coding genes of the domestic cat are of high quality for all trait discovery studies.

**Genetic variation in cats.** To improve variant knowledge of the domestic cat, variants from a diverse set of 74 resequenced cats from the 99 Lives project were analyzed in depth. The average sequence coverage was 38.5x with a mean of 98% reads mapped per cat (**Supplemental Data S3**). Approximately 46,600,527 variants were discovered, 39,043,080 were SNVs with 93% as biallelic displaying a Ts/Tv ratio of 2.44, suggesting a relatively high level of specificity for variant detection (**Supplemental Table S2**). In addition, 97% of SNV positions from the feline 63K genotyping array mapped to Felis_catus_9.0 were detected as SNVs in the WGS call set (**Supplemental Table S3**; **Supplemental Data S4**) (Gandolfi et al. 2018). Using the variant data to estimate cat relatedness, 13 highly related cats (□ > 0.15), two cats with poor read quality, four bengals or bengal crosses, and Cinnamon, the reference, were removed from the sequence dataset to obtain a final set of 54 cats for all subsequent analyses (**Supplemental Data S3**). The average number of discovered SNVs per cat was 9.6 million (**Fig 1A**). Differences in SNV numbers varied according to whether cats were from a specific breed or were random bred (P-value < 0.005, Wilcoxon rank sum test), the two cats with the lowest number of SNVs (~8 million) were both Abyssinians, the same breed as Cinnamon, while random bred cats from either the Middle East or Madagascar each carried the highest number (> 10.5 million) (**Supplemental Data S5**). Individual singleton frequency and estimated inbreeding coefficients (*F* statistic) showed a similar trend with random bred cats generally having significantly more singletons and higher levels of heterozygosity than breed cats (P-value < 0.005, Wilcoxon rank sum test) (**Fig 1B, Fig 1C**). Breed cats with higher levels of variation and heterozygosity were either from newly established breeds or were outcrossed individuals. For example, Napoleon cats, which all had an *F* statistic at least one standard deviation below the mean (**Fig 1C; Supplemental Data S5**), are frequently outcrossed as their defining trait dwarfism is likely homozygous lethal *in utero* (Lyons et al. 2019).

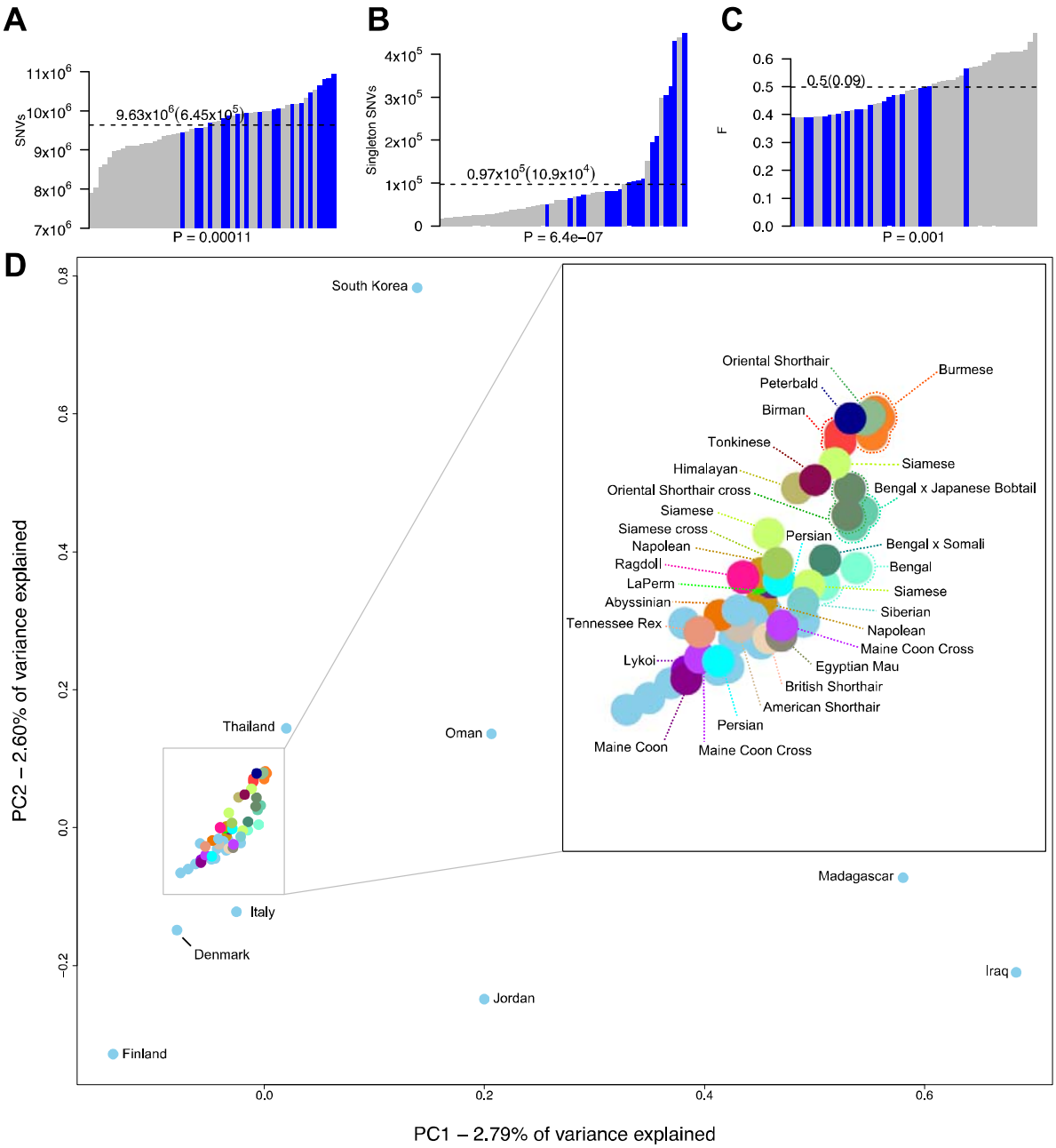



**Figure 1**. **Genetic variation in whole genome sequenced cats. (A)** Number of SNVs found in each genome, **(B)** singletons per genome, and **(C)** the per sample inbreeding coefficient *F*. Blue bars indicate each individual random bred cat and grey bars indicate each individual breed cat. P-values underneath each x axes were calculated using Wilcoxon rank-sum test and were used to compare breed cats to random bred cats. Dotted lines indicate the mean for each statistic, which is printed above along with standard deviation in braces. **(D)** Population structure of all unrelated cats (including Bengal breeds) estimated using principal components analysis. Random bred cats are colored light blue. Those that were sampled globally for diversity regions are named according to their sampling location. All other cats are named according to their breed or breed cross.

PCA analysis showed the expected distribution of genetic relatedness among cats when considering their geographical location and genetic origins (**Fig. 1D**). In general, most random bred cats displayed a scattered distribution consistent with previous studies on cat population diversity and origins (Lipinski et al. 2008; Gandolfi et al. 2018). Although tightly clustered, breed cats could also be distinguished according to their populations of origin. The Asian-derived breeds, Siamese, Burmese, Birman, and Oriental shorthairs were at one end of the spectrum, clustering closely with random bred cats from Thailand. Conversely, cats derived from western populations, such as Maine Coons and Persians, were at the opposite end of the spectrum grouping with random bred cats from Northern Europe, such as Denmark and Finland.

**Implications of feline genetic variation on human disease genes.** To characterize feline genetic variation in disease contexts, SNVs were annotated as having high, moderate, or low impacts. High impact SNVs were those that caused stop gain, stop loss, and start loss, moderate impact variants were missense mutations, and low impact variants were usually synonymous mutations (**Supplemental Table S4**). In addition to SNP annotation and classification, genes themselves were grouped according to their genetic constraint across human populations. Gene groups included all genes, orthologs, and intolerant orthologs. The "all genes" gene group contained all ensembl annotated genes in Felis_catus_9.0, the "orthologs" gene group contained all genes with human orthologs identified by reciprocal best hits, and the intolerant orthologs gene group contained orthologs that had a probability of loss of function intolerance (pLI) greater than 0.9 in humans (Methods).

Across each impact category, as expected, low impact SNVs were the most abundant, while high impact SNVs were the least abundant (**Fig. 2A**). The most frequently occurring high impact SNV was a stop gain, while stop lost SNVs were the least frequently occurring (**Supplemental Table S4)**. Importantly, the per kb density of SNVs of the same impact varied according to the respective gene groups. For example, between all genes and intolerant orthologs, the density of

low impact SNVs ranged from 4.14 to 3.25 variants per kb of coding sequence, while the density of high impact variants ranged from 0.038 to 0.006 variants per kb of coding sequence. The severely decreased variant density within intolerant orthologs suggests human and cat orthologs are similarly constrained. Consistently, nonsynonymous (moderate and high impact) SNVs were significantly depleted across both all orthologs and intolerant orthologs, where low impact SNVs were significantly more enriched in orthologs and intolerant orthologs compared to expected SNV densities (**Fig. 2A**). The distribution of variants and their impact across cats are consistent with selection against protein altering mutations in genes under selective constraint in human, supporting their further study as putative genetic models of human diseases.

To characterize the landscape of genetic variation across cats, SNV allele frequencies in cats were examined across each impact class and gene group. For all impact classes combined, 18.6% of SNVs in all genes and 19.1% of SNVs in orthologs had allele frequencies below 0.01. The majority of SNVs in these gene groups belonged to the more common allele frequency categories of '0.01 < AF < 0.1', which contained 35.6% of all gene SNVs and 35.9% of all ortholog SNVs, and '0.1 < AF < 1.0', which contained 45.8% of all gene SNVs and 45.0% of all ortholog SNVs. This distribution of SNVs across allele frequency categories was also similar for individual impact classes (**Fig. 2B**). Alternatively, for intolerant orthologs, nonsynonymous SNVs were more skewed toward lower allele frequencies than synonymous SNVs. For example, in intolerant orthologs, 37.4% of high impact and 27.7% of moderate impact SNVs were found at allele frequencies < 0.01, compared to 18.2% of low impact SNVs (**Fig. 2B**). To test whether differences in allele frequency distributions were significant, the fraction of SNVs in a singleton state in each gene group were compared to a random distribution generated by 10,000 permutations. Amongst orthologs and intolerant orthologs, the singleton fraction of nonsynonymous SNVs was significantly higher than expected and the singleton fraction for synonymous SNVs was significantly lower than expected (**Fig. 2A**).

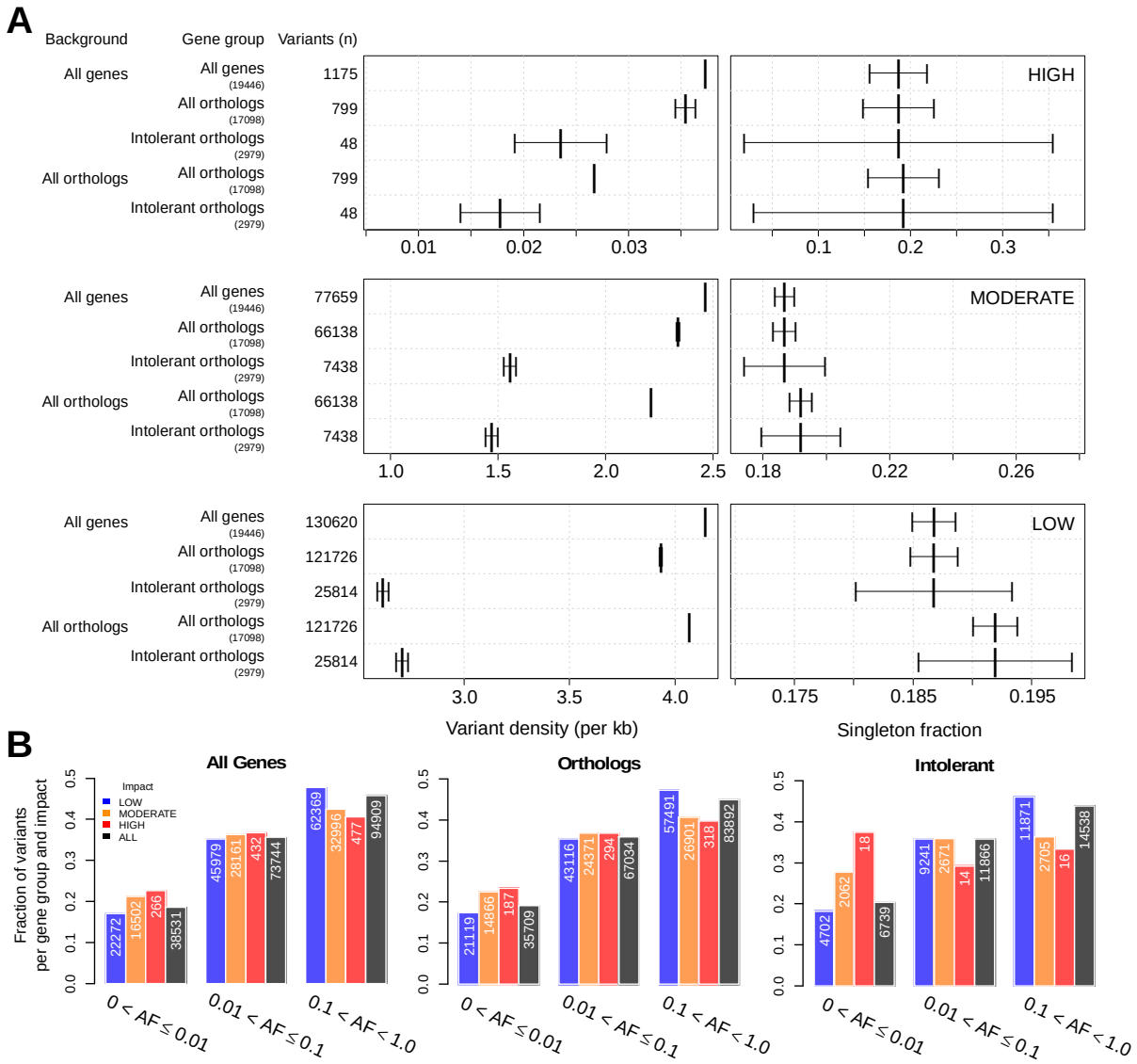



**Figure 2**. Deleterious SNVs in cats are uncommon and depleted from human constrained genes. **(A)** The observed and expected values for per kb variant density and the fraction of variants that were singletons. Large vertical bars represent the expected value for each statistic while smaller bars represent the 99% confidence intervals for the expected values. Confidence intervals were calculated using 10,000 permutations (methods). Filled in circles represent the observed values for each statistic. To the left of the plots is the number of variants belonging to each gene group and impact class. In parentheses is the number of genes that belong to each gene group. **(B)** The fraction of variants across different allele frequency (AF) intervals. Numbers in white within each bar show the raw count of variants for each AF interval, impact class, and gene group.

Overall, 18 high impact singleton SNVs were identified in intolerant orthologs and could be potential candidate disease causing variants. Since some cats within 99 lives had recorded disease statuses, these SNVs were assessed for their potential role in cat diseases (**Supplemental Table S5**). Of the 18 SNVs, four were supported by both Ensembl and NCBI annotations and were in cats segregating for particular diseases (**Table 2**). Of note is a stop gain in the tumor suppressor *F-box and WD repeat domain containing 7* (*FBXW7*) (Yeh et al. 2018), which was only found in a parent and child segregating for feline mediastinal lymphoma. Other high impact SNVs include stop gains found in *Family With Sequence Similarity 13 Member B* (*FAM13B*) in a random bred with ectodermal dysplasia, cytoplasmic FMR1 interacting protein 2 (*CYFIP2*) in an Egyptian Mau with urate stones, and *SH3 And PX Domains 2A* (*SH3PXD2A*) in a random bred cat with feline infectious peritonitis. Most candidates are not likely disease causing, as each cat carried a mean of 10.85 high impact SNVs in intolerant orthologs (**Supplemental Fig. S1**). However, while most of the high impact SNVs had allele frequencies > 10%, the mean number of high impact singleton SNVs in intolerant regions was 0.33 per cat (**Supplemental Fig. S2**). These results suggest gene intolerance to mutations may provide as a useful metric for reducing the number of candidate variants for certain diseases.

**Table 2. High impact singletons in intolerant orthologs for unrelated cats with disease traits.**

| SNV location[a] | Ref/Alt | Consequence | Gene symbol | pLI | Individual ID | Disease | Status |
|---|---|---|---|---|---|---|---|
| chrA1:116653102 | G/A | stop_gained | *FAM13B* | 0.92 | felCat.Fcat19194.Pudge | Ectodermal dysplasia[b] | Affected |
| chrA1:192824838 | G/A | stop_gained | *CYFIP2* | 1.00 | felCat.Fcat20406.Gannon | Stones | Affected |
| chrB1:77305060 | C/T | stop_gained | *FBXW7* | 1.00 | felCat.Fcat5012.Colorado[c] | Lymphoma | Carrier |
| chrD2:63980875 | G/A | stop_gained | *SH3PXD2A* | 1.00 | felCat.CR1397.Isabella | Infectious peritonitis[d] | Affected |

[a] SNV locations were only reported if they were supported by NCBI annotations.

[b] A second unrelated cat, felCat.Fcat19197.Kooki, was also affected.

[c] Affected offspring removed earlier from analysis inherited the same SNV.

[d] A second unrelated cat, felCat.CR1219.Tamborine, was also affected.

**Structural variant discovery**. The merging of the two independent SV call sets was performed across all individuals for variants occurring within 50 bp of the independent variant call position, with agreement on variant type and strand, and variant size within 500 bp. Per cat, an average of 44,990 SVs were identified, with variants encompassing 134.3 Mb across all individuals. Deletions averaged 905 bp, duplications 7,497 bp, insertions 30 bp, and inversions 10,993 bp. The breed and breed crosses (n = 36) compared to random bred cats (n = 18) showed comparable SV diversity (t-test p = 0.6) (**Supplemental Fig. S3**). In total, 208,135 SVs were discovered, of which 123,731 (60%) were deletions (**Fig. 3A**). SV population frequencies were similar across SV types, except for inversions. For deletions, duplications and insertions, 38% to 48% of each SV type was found at population frequencies of 0.02 – 0.10 and 0.10 – 0.50. Meanwhile, > 90% of inversions are found at a population frequency of 0.02 – 0.10 (**Fig. 3A)**. The majority of SVs identified are common across cats, suggesting their impacts are mostly tolerated.

SV density across autosomes was relatively constant with chromosome E1 carrying the largest SV burden at 96.95 SVs per Mb (**Supplemental Fig. S4)**. Approximately 6,096 SVs (3%, 10.1

Mb) were observed in >90% of the cat genomes (**Supplemental Data S6**), indicating the cat used for the Felis_catus_9.0 assembly, Cinnamon, an Abyssinian, likely carries a minor allele at these positions. SV annotation showed that SV counts per region were consistent with the fraction of the genome occupied by each region type. For example, 58.15% of SVs were intergenic, 40.22% of SVs were intronic, and 1.06% of SVs were exonic, potentially impacting 217 different protein coding genes (**Fig. 3B**) (**Supplemental Data S7**). Conversely, the proportion of some SV types found in certain gene regions varied from their genome-wide averages. For example, in regions 5 kb upstream and downstream of genes, duplications were increased approximately two-fold. For exonic regions, 74% of SVs were deletions, an increase form the genome wide level of 59.45%. For 5` UTRs, the majority of SVs were inversions, which only represent 17.02% of total SVs. These results suggest an interaction between the impact of SV types and the potential function of the gene regions in which they are found.

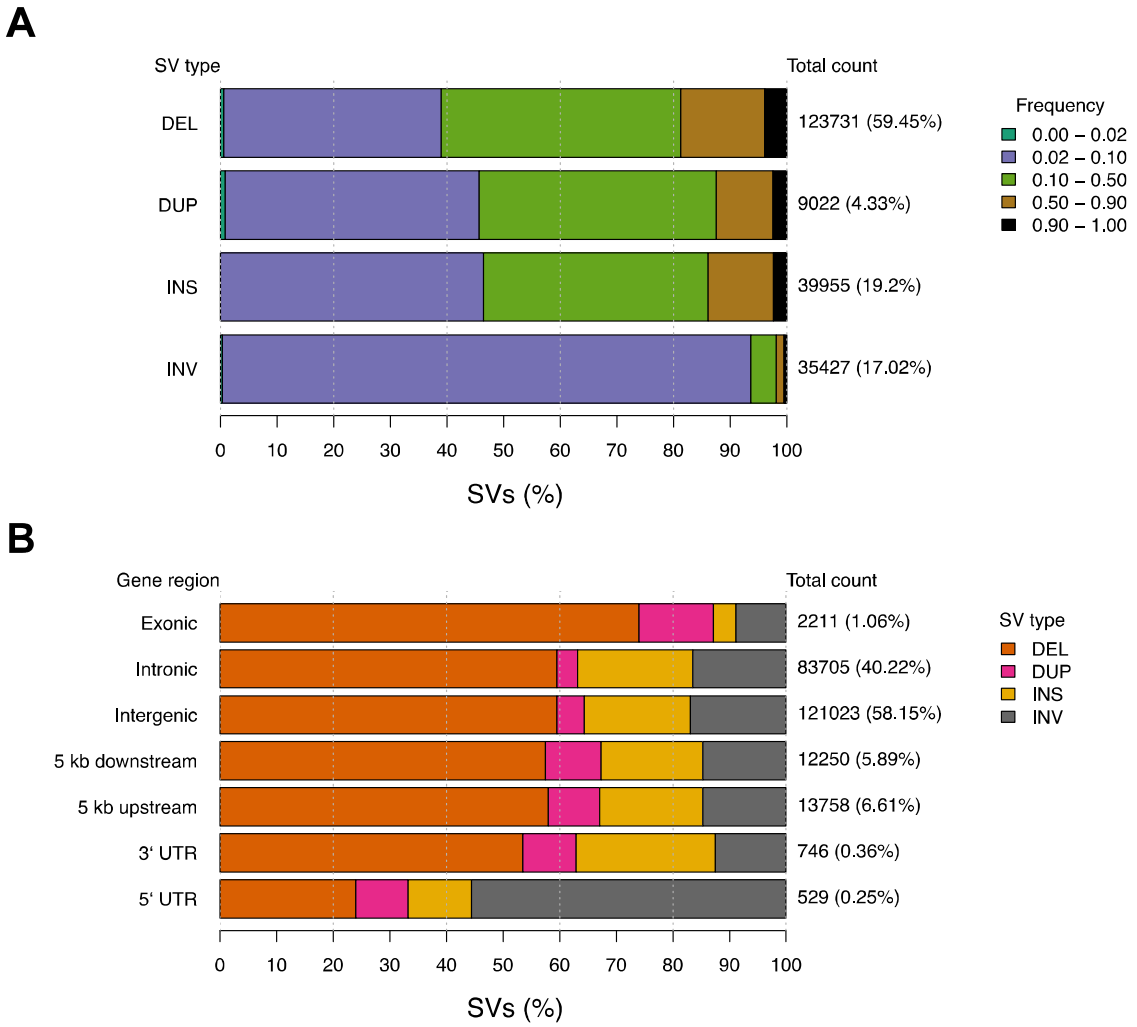



**Figure 3**. The structural variant landscape of cats. **(A)** Population frequency of each SV type. Colored bars represent the proportion of a given SV type found within various population frequency ranges. Total count values to the right of each bar represent the total number of SVs of each type, while percentage values in braces represent the proportion of all SVs that belong to each type. Frequency ranges shown in the legend are ordered along each bar from left to right. **(B)** The proportion of SVs found in each genomic region. Colored bars represent the proportion of SVs belonging to a particular type found across different genomic regions. Similar to above, the total count values to the right of each bar represent the total number of SVs found in each genomic region. The percentage values in braces represent the genome-wide proportion of all SVs found in each gene region. These values sum greater than 100 percent as a single SV can span multiple types of genomic regions. Structural variants are noted as: deletion (DEL); duplication (DUP); insertion (INS); inversion (INV).

**Genetics of feline dwarfism.** Dwarfism in cats is the defining feature of the munchkin breed and is characterized by shortened limbs and normal sized torso (**Fig. 4A**) (TICA 2015). Similar to analyses with cat assemblies Felis_catus-6.2 and Felis_catus_8.0, previous investigations for SNVs in the new Felis_catus_9.0 assembly did not identify any high priority candidate variants for disproportionate dwarfism (Lyons et al. 2019). However, SV analysis within the critical region previously identified by linkage and GWAS on chromosome B1:170,786,914-175,975,857 (Lyons et al. 2019) revealed a 3.3 kb deletion at position chrB1:174,882,897-174,886,198, overlapping the final exon of *UDP-glucose 6-dehydrogenase (UGDH)* (**Fig. 4B**). Upon manual inspection of this SV, a 49 bp segment from exon 8 appeared to be duplicated and inserted 3.5 kb downstream, replacing the deleted sequence. This potentially duplicated segment was flanked by a 37 bp sequence at the 5` end and a 20 bp sequence at the 3' end, both of unknown origin (**Fig. 4C**). Discordant reads consistent with the SV were private to all three unrelated WGS dwarf samples (**Supplemental Fig. S5**). The breakpoints surrounding the deletion were validated in WGS affected cats with Sanger sequencing. PCR-based genotyping of the 3.3 kb deletion breakpoints was conducted in an additional 108 cats including, 40 normal and 68 affected dwarf cats (**Supplemental Fig. S6**). Expected amplicon sizes and phenotypes were concordant across all cats, except for a "munchkin; non-standard (normal legs); Selkirk mix", which appeared to carry the mutant allele, suggesting an alternate causal gene or sampling error.

As *UGDH* has known roles in proteoglycan synthesis in chondrocytes (Clarkin et al. 2011; Wen et al. 2014), growth plates in two healthy and seven dwarf neonatal kittens were histologically analyzed for structural irregularities and proteoglycan concentrations (Methods). The articular cartilage or bone in the dwarf specimens had no significant histopathologic changes and epiphyseal plate was present in all specimens. In the two normal kittens, chondrocytes in the epiphyseal plate exhibited a regular columnar arrangement with organization into a zone of

reserve cells, a zone of proliferation, a zone of hypertrophy and a zone of provisional calcification (**Fig. 4D; Supplemental Fig. S7**). Moreover, the articular-epiphyseal cartilage complex and the epiphyseal plate in the distal radius from these two normal kittens also contained abundant proteoglycan as determined by toluidine blue stain. In contrast, in dwarf kittens, the epiphyseal plate had a disorganized columnar arrangement that was frequently coupled with proteoglycan depletion (**Fig. 4D; Supplemental Fig. S7**).

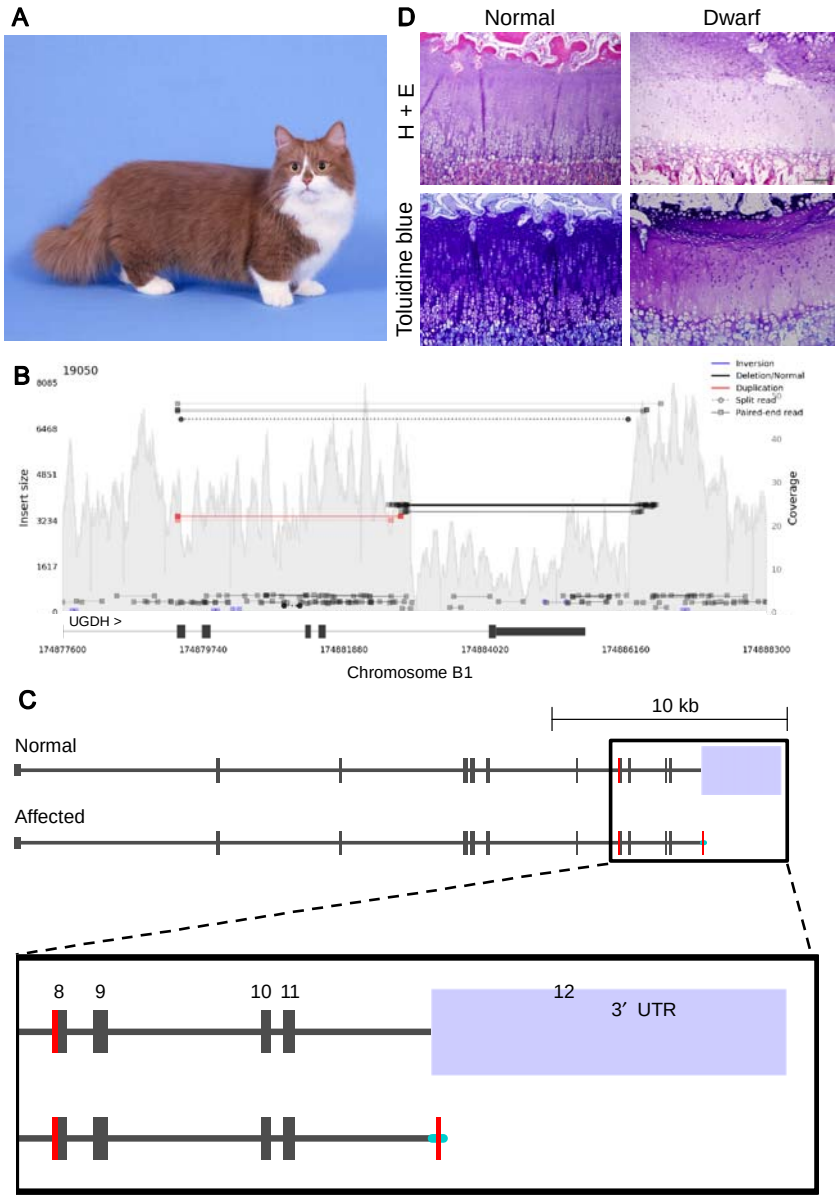



**Figure 4.** A complex structural variant is associated with feline dwarfism. **(A)** Image of dwarf cat from the munchkin breed. Notice that limbs are short while torso remains normal size. **(B)** Samplot output of discordant reads within an affected cat. Decreased coverage over the final exon of *UGDH* represents a deletion. Discordant reads spanning beyond the deleted region show newly inserted sequence sharing homology with *UGDH* exon 8. **(C)** Schematic illustrating the candidate structural variant for dwarfism. Blue rectangle overlaps deleted region, shaded portion of exon 8 shows potential duplication in the affected, and cyan lines indicate inserted sequence of unknown origin. **(D)** Hematoxylin and eosin and toluidine blue histologic samples of the distal radius epiphyseal cartilage plate from a neonatal kitten and an age-matched dwarf kitten. The normal kitten showed regular columnar arrangement of chondrocytes with abundant proteoglycan as determined by strong metachromasia with toluidine blue stain. The dwarf kitten showed disorganized columnar arrangement of chondrocytes with proteoglycan depletion as determined by weak metachromasia with toluidine blue stain.

## Discussion

Studies of the domestic cat show the amazing features of this obligate carnivore, including their incomplete domestication, wide range of coat color variation, and use as biomedical models (Lyons et al. 2004; Montague et al. 2014; Yu et al. 2019a; Yu et al. 2019b). Using a combination of long-reads, combined with long-range scaffolding methods, Felis_catus_9.0 was generated as a new cat genome resource, with an N50 contig length (42 Mb) surpassing all other carnivore genome assemblies. This sequence contiguity represents a 1000-fold increase of ungapped sequence (contigs) length over Felis_catus_8.0, as well as a 40% reduction in the amount of unplaced sequence. All measures of sequence base and order accuracy, such as the low number of discordant BAC end sequence alignments, suggest this reference will strongly support future resequencing studies in cats. Equally important are the observed improvements in gene annotation including the annotation of new genes, overall more complete gene models and improvements to transcript mappability. Only 178 genes were missing in Felis_catus_9.0 compared to Felis_catus_8.0, which will require further investigation. However, 376 predicted genes are novel to Felis_catus_9.0, i.e. not found in Felis_catus_8.0.

Using this Felis_catus_9.0 assembly, a vast new repertoire of SNVs and SVs were discovered for the domestic cat. The total number of variants discovered across our diverse collection of domestic cats was substantially higher than previous studies in other mammals, such as, cow (Daetwyler et al. 2014), dog (Bai et al. 2015), rat (Atanur et al. 2013; Hermsen et al. 2015; Teng et al. 2017), sheep (Yang et al. 2016; Chen et al. 2018), pig (Choi et al. 2015), and horse (Jagannathan et al. 2019). Even rhesus macaques, with twice as many variants as human, do not approach the same levels of cat SNV variation (Xue et al. 2016; Bimber et al. 2017). The cats had 36.6 million biallelic SNVs with individuals carrying ~9.6 million SNVs each. Conversely, humans roughly carry 4 to 5 million SNVs per individual (The 1000 Genomes Project Consortium et al. 2015). One possible explanation for this discrepancy may be the

unique process cats underwent for domestication. Rather than undergoing strong selective breeding leading to a severe population bottleneck, cats were instead “self” domesticated, never losing some ancestral traits such as hunting behavior (Vigne et al. 2004; Driscoll et al. 2007; Van Neer et al. 2014; Ottoni et al. 2017). The practice of strong selective breeding in cats only began in recent history and is based almost exclusively on aesthetic traits. Consistent with previous analyses, our evaluation of breeds and random bred cats showed tight clustering between cats from different breeds that suggest cat breeds were likely initiated from local random bred cat populations (Lipinski et al. 2008; Kurushima et al. 2013). In regards to per sample numbers of variants, most breed cats had fewer SNVs than random bred cats. Likewise, random bred cats also had higher numbers of singletons and a lower inbreeding coefficient than breed cats, suggesting breed cats may share a common genetic signature distinct from random bred populations.

For genetic discovery applications, animals with lower levels of genetic diversity are often more desirable as they reduce experimental variability linked to variation in genetic architecture (Festing 2010). However, with decreasing costs of genome or exome sequencing leading to exponential growth in variant discovery, animals with higher genetic diversity, such as cats, could enhance discovery of tolerable loss of function or pathogenic missense variants. Similar to cats, rhesus macaques from research colonies, exhibited per sample SNV rates more than two-fold greater than human (Xue et al. 2016). Importantly, macaque genetic diversity has been useful for characterizing non-deleterious missense variation in humans (Sundaram et al. 2018).

A major premise of our feline genomics analysis is that increased discovery of genetic variants in cats will help improve benign/pathogenic variant classification and reveal similarities between human and feline genetic disease phenotypes, aiding disease interpretation in both species. To gain insights into the burden of segregating variants with potentially deleterious impacts on protein function, SNV impacts were classified across 54 unrelated domestic cats. Overall,

nonsynonymous SNVs (high and moderate impact) were significantly depleted and more likely to be singleton in genes identified in human as intolerant to mutations than in genes under more relaxed constraint (**Fig. 2A**), showing both species harbor similar landscapes of genetic constraint and indicating the utility of Felis_catus_9.0 in modeling human genetic disease. An important limitation on the analysis was the available sample size. As there were only 54 cats, true rare variants, defined as variants with allele frequency of at least 1% in the population (Frazer et al. 2009), could not be distinguished from common variants that appear in only one of 54 randomly sampled cats by chance. Instead, singleton SNVs were focused on as candidates for rare variants. Altogether, 18.6% of all feline SNVs could be considered candidates for rare variants (**Fig. 2B)**. Alternatively, larger analyses in human reveal a much higher fraction of low frequency variants (Gudbjartsson et al. 2015; Lek et al. 2016). For example, 36% of variants in an Icelandic population of over 2000 individuals had a minor allele frequency below 0.1% (Gudbjartsson et al. 2015). Similarly sized analyses in other mammals reported ~20% of variants had allele frequencies < 1% (Xue et al. 2016; Chen et al. 2018; Jagannathan et al. 2019). As the number of cats sequenced increases, the resolution to detect rare variants, as well as the total fraction of low frequency variants, will continue to grow linearly as each individual will contribute a small number of previously undiscovered variants (The 1000 Genomes Project Consortium et al. 2010).

Focusing specifically on potentially rare variants with high impacts in likely constrained genes identified a potential cause for early onset feline mediastinal lymphoma, a stop gain in tumor suppressor gene *FBXW7* (Yeh et al. 2018). Feline mediastinal lymphoma is distinct from other feline lymphomas in its early onset and prevalence in Siamese cats and Oriental Shorthairs (Gabor et al. 1998), suggesting a genetic cause for lymphoma susceptibility specific to Siamese related cat breeds (Louwerens et al. 2005; Fabrizio et al. 2014). The stop gain was initially observed in a heterozygous state in a single cat identified as a carrier in the unrelated set of

cats. Subsequent analysis of the full set of cats revealed the SNV had been inherited in an affected offspring. Despite discordance between the presence of this mutant allele and the affection status of the cats it was found in, the variant still fits the profile as a susceptibility allele for mediastinal lymphoma. For example, homozygous knockout of *Fbxw7* is embryonic lethal in mice (Tetzlaff et al. 2004; Tsunematsu et al. 2004), while heterozygous knockout mice develop normally (Tsunematsu et al. 2004). However, irradiation experiments of *Fbxw7*$^{+/-}$ mice and *Fbxw7*$^{+/-}$*p53*$^{+/-}$ crosses identify *Fbwx7* as a haploinsufficient tumor suppressor gene that requires mutations in other cancer related genes for tumorgenesis (Mao et al. 2004), a finding supported in subsequent mouse studies (Maser et al. 2007; Perez-Losada et al. 2012). Similarly, in humans, germline variants in *FBXW7* are strongly associated with predisposition to early onset cancers, such as Wilms tumors and Hodgkin's lymphoma (Roversi et al. 2015; Mahamdallie et al. 2019). Screening of Siamese cats and other related breeds will validate the *FBWX7* stop gain as a causative mutation for lymphoma susceptibility and may eventually aid in the development of a feline cancer model.

An important concern with using human constraint metrics for identifying causative variants in cats is the potential for false positives. For example, out of 18 SNVs initially identified as potential feline disease candidates, many belonged to cats with no recorded disease status (**Supplemental Table S5**). However, since a large fraction of healthy humans also carried SNVs matching similar disease causing criteria as used in cats (Lek et al. 2016), evolutionary distance between humans and cats is unlikely to be a significant contributing factor to false positive candidate disease variant identification. Instead, the frequency of high impact variants in constrained genes is likely due to the limited resolution provided by analyses at the gene level. Recent strategies in humans have confronted this problem by focusing on constraint at the level of gene region. These analyses identified many genes with low pLI values that contained highly constrained gene regions that were also enriched for disease causing variants (Havrilla et al.).

Alternatively, another strategy for further refining constrained regions could involve combining genomic variation from multiple species. Many variants observed in cats, along with their impact on genes, are likely unique to cats and could potentially be applied to variant prioritization workflows in humans.

Presented here is the first comprehensive genome-wide SV analysis for Felis_catus_9.0. Previous genome-wide SV analyses in the cat were performed in Felis_catus-6.2 and focused solely on copy number variation (CNV) (Genova et al. 2018). Approximately 39,955 insertions, 123,731 deletions, 35,427 inversions, and 9,022 duplications were identified, far exceeding previous CNV calculations of 521 losses and 68 gains. This large discrepancy is likely due to a number of reasons including filtering stringency, sensitivity, and reference contiguity, where increased numbers of assembly gaps in previous reference genome assemblies hindered detection of larger SV events. An important difference regarding filtering stringency and sensitivity is that the average CNV length was 37.4 kb as compared to 2.7 kb for SVs in Felis_catus_9.0. Since CNV detection depends on read-depth (Abyzov et al. 2011; Klambauer et al. 2012), only larger CNVs can be accurately identified, as coverage at small window sizes can be highly variable. The use of both split-read and read-pair information (Rausch et al. 2012; Layer et al. 2014), allowed the identification of SV events at much finer resolution than read-depth based tools (Kosugi et al. 2019).

Improved sensitivity for smaller SV events, while helpful for finding new disease causing mutations, may also lead to increased false positive SV detection. In three human trios sequenced with Illumina technology, LUMPY and Delly were both used to identify 12,067 and 5,307 SVs respectively, contributing to a unified call set, along with several other tools, of 10,884 SVs per human on average (Chaisson et al. 2019). In cats, the average number of SVs per individual was 4 times higher than in humans, with 44,990 SVs per cat, suggesting the total number SVs in cats is likely inflated. However, the majority of the SVs were at population

frequencies below 0.5, ruling out poor reference assembly as a contributing factor. Instead, the difference in SV count between cats and humans is likely due to sample specific factors. For example, the majority of samples were sequenced using two separate libraries of 350 bp and 550 bp insert sizes (**Supplemental Data S2**). Despite the potentially high number of false positive SVs, increased SV sensitivity was useful for trait discovery. The deletion associated with dwarfism was only found in the Delly2 call set. If SV filtering were more stringent, such as requiring SVs to be called by both callers, the feline dwarfism SV may not have otherwise been detected. Ultimately, these results highlight the importance of high sensitivity for initial SV discovery and the use of highly specific molecular techniques for downstream validation of candidate causative SVs.

The improved contiguity of Felis_catus_9.0 was particularly beneficial for identifying a causative SV for feline dwarfism. In humans, approximately 70% of cases are caused by spontaneous mutations in *fibroblast growth factor 3* resulting in achondroplasia or a milder form of the condition known as hypochondroplasia (Horton et al. 2007; Foldynova-Trantirkova et al. 2012). The domestic cat is one of the few species with an autosomal dominant mode of inheritance for dwarfism that does not have other syndromic features. It can therefore provide as a strong model for hypochondroplasia. Previous GWAS and linkage analyses suggested a critical region of association for feline disproportionate dwarfism on cat chromosome B1 that spanned 5.2 Mb (Lyons et al. 2019). Within the critical region, SV analysis identified a 3.3 kb deletion that had removed the final exon of *UGDH*, which was replaced by a 106 bp insertion with partial homology to *UGDH* exon 8, suggesting a potential duplication event. Importantly, *UGDH* likely plays a role in proteoglycan synthesis in articular chondrocytes, as osteoarthritic human and rat cartilage samples have revealed reduced UGDH protein expression was associated with disease state (Wen et al. 2014). Similarly, in dwarf cat samples, histology of the distal radius showed irregularity of the chondrocyte organization and proteoglycan depletion. Collectively,

results suggest a disease model of reduced proteoglycan synthesis in dwarf cat chondrocytes caused by loss of function of *UGDH* resulting in abnormal growth in the long bones of dwarf cats. In ClinVar, only one missense variant in *UGDH* has been documented NM_003359.4(**UGDH**):c.950G>A (p.Arg317Gln) and is associated with epileptic encephalopathy. Seventeen other variants involving *UGDH* span multiple genes in the region. The SV in the dwarf cats suggest *UGDH* is important in early long bone development. As a novel gene association with dwarfism, *UGDH* should be screened for variants in undiagnosed human dwarf patients.

High-quality genomes are a prerequisite for unhindered computational experimentation. Felis_catus_9.0 is currently the most contiguous genome of a companion animal, with high accuracy and improved gene annotation that serves as a reference point for the discovery of genetic variation associated with many traits. This new genomic resource will provide a foundation for the future practice of genomic medicine in cats and for comparative analyses with other species.

## Methods

**Whole genome sequencing**. The same genome reference inbred domestic cat, Cinnamon, the Abyssinian, was used for the long-read sequencing (Pontius et al. 2007) (Montague et al. 2014). High molecular weight DNA was isolated using a MagAttract HMW-DNA Kit (Qiagen, Germantown, MD) from cultured fibroblast cells according to the manufacturer's protocol. Single molecule real-time (SMRT) sequencing was completed on the RSII and Sequel instruments (Pacific Biosciences, Menlo Park, CA).

**Genome assembly**. All sequences (~72x total sequence coverage) were assembled with the fuzzy Bruijn graph algorithm, WTDBG (https://github.com/ruanjue/wtdbg), followed by collective raw read alignment using MINIMAP to the error-prone primary contigs (Li 2016). As a result, contig coverage and graph topology were used to detect base errors that deviated from the majority haplotype branches that are due to long-read error (i.e. chimeric reads) or erroneous graph trajectories (i.e. repeats) as opposed to allelic structural variation, in which case, the sequence of one of the alleles is incorporated into the final consensus bases. As a final step to improve the consensus base quality of the assembly, from the same source DNA (Cinammon), short read sequences (150 bp) were generated from 400 bp fragment size TruSeq libraries to ~60X coverage on the Illumina HiSeqX instrument, which was then used to correct homozygous insertion, deletion and single base differences using PILON (Walker et al. 2014).

**Assembly scaffolding***.* To generate the first iteration of scaffolds from assembled contigs, the BioNano Irys technology was used to define order and orientation, as well as, detect chimeric contigs for automated breaks (Lam et al. 2012). HMW-DNA in agar plugs was prepared from the same cultured fibroblast cell line using the BioNano recommended protocol for soft tissues, where using the IrysPrep Reagent Kit, a series of enzymatic reactions lysed cells, degraded protein and RNA, and added fluorescent labels to nicked sites. The nicked DNA fragments were labeled with ALEXA Fluor 546 dye and the DNA molecules were counter-stained with YOYO-1

dye. After which, the labeled DNA fragments were electrophoretically elongated and sized on a single IrysChip, with subsequent imaging and data processing to determine the size of each DNA fragment. Finally, a *de novo* assembly was performed by using all labeled fragments >150 kb to construct a whole-genome optical map with defined overlap patterns. Individual maps were clustered, scored for pairwise similarity, and Euclidian distance matrices were built. Manual refinements were then performed as previously described (Lam et al. 2012).

**Assembly QC.** The scaffolded assembly was aligned to the latest cat linkage map (Li et al. 2016) to detect incorrect linkage between and within scaffolds, as well as discontinuous translocation events that suggest contig chimerism. Following a genome-wide review of interchromosomal scaffold discrepancies with the linkage map, the sequence breakpoints were manually determined and the incorrect sequence linkages were separated. Also, to assess the assembly of expanded heterozygous loci 'insertions', the same reference DNA sequences (Illumina short read inserts 300 bp) were aligned to the chromosomes to detect homozygous deletions in the read alignments using Manta (Chen et al. 2016), a structural variant detection algorithm. In addition, the contigs sequences were aligned to the Felis_catus_8.0 using BLAT (Kent 2002) at 99% identity and scored alignment insertion length at 0.5 to 50 kb length to further refine putatively falsely assembled heterozygous loci that when intersected with repeat tracks suggested either error in the assembly or inability to correctly delineate the repetitive copy.

**Chromosome builds**. Upon correction and completion of the scaffold assembly, the genetic linkage map (Li et al. 2016) was used to first order and orient all possible scaffolds by using the Chromonomer tool similarly to the previously reported default assembly parameter settings (Small et al. 2016). A final manual breakage of any remaining incorrect scaffold structure was made considering various alignment discordance metrics, including to the prior reference Felis_catus_8.0 that defined unexpected interchromosomal translocations and lastly paired end

size discordance using alignments of BAC end sequences from the Cinnamon DNA source (http://ampliconexpress.com/bac-libraries/ite).

**Gene Annotation**. The Felis_catus_9.0 assembly was annotated using previously described NCBI (Pruitt et al. 2012; Thibaud-Nissen et al. 2013) and Ensembl (Zerbino et al. 2018) pipelines, that included masking of repeats prior to *ab initio* gene predictions and evidence-supported gene model building using RNA sequencing data (Visser et al. 2019). RNA sequencing data of varied tissue types (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Felis_catus/104) was used to further improve gene model accuracy by alignment to nascent gene models that are necessary to delineate boundaries of untranslated regions as well as to identify genes not found through interspecific similarity evidence from other species.

**Characterizing feline sequence Variation.** Seventy-four cat WGSs from the 99 Lives Cat Genome Project were downloaded from the NCBI short read archive (SRA) affiliated with NCBI biosample and bioproject numbers (**Supplemental Data S2**). All sequences were produced with Illumina technology, on either an Illumina HiSeq 2500 or XTen instrument using PCR-free libraries with insert lengths ranging from 350 bp to 550 bp, producing 100 – 150 bp paired-end reads. WGS data was processed using the Genome analysis toolkit (GATK) version 3.8 (McKenna et al. 2010; Van der Auwera et al. 2013). BWA-MEM from Burrows-Wheeler Aligner version 0.7.17 was used to map reads to Felis_catus_9.0 (GCF_000181335.3) (Li 2013). Piccard tools version 2.1.1 (http://broadinstitute.github.io/picard/) was used to mark duplicate reads, and samtools version 1.7 (Li et al. 2009) was used to sort, merge and index reads. Tools used from GATK 3.8 consisted of IndelRealigner and RealignerTargetCreator for indel realignment, BaseRecalibrator for base quality score recalibration (BQSR) (DePristo et al. 2011), and HaplotypeCaller and GenotypeGVCFs for genotyping (Poplin et al. 2018). The variant database used for BQSR was built by first genotyping non-recalibrated BAMs and applying a

strict set of filters to isolate high confidence variants. To determine the final variant call set, post BQSR, a less-strict GATK recommended set of filters, was used. All filtering options are outlined in supplementary material (**Supplemental Table S6**). The set of unrelated cats was determined using vcftools' relatedness2 function on SNP genotypes to estimate the kinship coefficient, □, for each pair of cats (Danecek et al. 2011). First, related cats were identified as sharing potential sibling and parent-child relationships if □ > 0.15. Next, cats with the highest number of relatives were removed in an iterative fashion until no relationships with □ > 0.15 remained. To detect population structure among the sequenced cats, a principal component analyses was conducted using SNPRelate version 1.16.0 in the R Statistical Software package. The SNV set generated after the appropriate quality control measures were used was further filtered for non-biallelic SNVs and sites displaying linkage disequilibrium (r2 threshold = 0.2) as implemented in the SNPRelate package.

**Measuring coding variant impacts on human disease genes.** VCF summary statistics and allele counts were computed using various functions from vcftools (Danecek et al. 2011) and vcflib (https://github.com/vcflib/vcflib). Variant impact classes of high, moderate, and low were determined using Ensembl's variant effect predictor (VEP) with annotations from Ensembl release 94 (McLaren et al. 2016; Zerbino et al. 2018). Genes were placed into the following groups, all genes, orthologs, and intolerant orthologs. For the ortholog gene group, cat and human orthologs were identified using reciprocal best hits blast, where Ensembl 94 protein fasta sequences were used as queries. Intolerant orthologs were identified as human cat orthologs whose probability for loss of function (LoF) intolerance (pLI) in human genes was greater than 0.9, where pLI for human genes was obtained from gnomAD (Lek et al. 2016; Karczewski et al. 2019). The observed per kb variant density, $Y^D$, for each gene group and impact was calculated as, $Y_{GI}^D = X_{GI}/C_G \times 1000bp$, where $C$ is the total length of the coding sequence and $X$ is the number of variants within $C$. The subscript $G$ refers to the subset of either $C$ or $X$ that are found

within a particular gene group. The subscript $I$ refers to the subset of variants, $X$, from a particular impact class. For example, when $G$ represents intolerant orthologs and $I$ represents high impact variants, $X_{GI}$ would be all high impact variants within the coding sequence of intolerant orthologs. The expected per kb variant density, $E^D$, for each gene group and impact was calculated as, $E^D_{GIB} = \frac{X_{IB}\frac{C_{GB}}{C_B}}{C_{GB}} \times 1000bp$, where the subscript $B$ refers to background genes, which is either all genes or orthologs. The 99% confidence intervals surrounding the expected per kb variant densities were calculated from a random distribution generated by 10,000 permutations, where variant impacts were shuffled randomly across variant sites. The observed singleton fraction, $Y^F$, for each gene group and impact was calculated as, $Y^F_{GI} = X_{GIP}/X_{GI}$, where the subscript $P$ refers to variants identified as singletons. The expected singleton fraction was calculated as, $E^F_B = X_{PB}/X_B$, where the 99% confidence intervals surrounding the expected singleton fractions were also calculated from a random distribution generated by 10,000 permutations. Here, for each permutation, variant allele frequencies were shuffled so variants were randomly assigned "singleton" status.

**Structural variant identification and analysis.** To discover SVs in the size range of <100 kb, aligned reads from all cats to Felis_catus_9.0 were used as input for the LUMPY (Layer et al. 2014) and Delly2 (Rausch et al. 2012) SV callers. For LUMPY, the empirical insert size was determined using samtools and pairend_distro.py for each BAM. Discordant and split-reads extracted from paired-end data using SpeedSeq (Chiang et al. 2015) were used as input along with each aligned BAM, minimum mapping threshold of 20, and empirical mean and standard deviation of insert size. Samples were called independently. SVTyper (Chiang et al. 2015) was used to genotype each SV before merging all resulting VCFs using BCFtools (Li et al. 2009). For Delly2 (Rausch et al. 2012), all SVs were called for individual BAMs independently, then merged into a single BCF. For each sample, variants were then re-called using the merged results from all samples. The individual re-called VCFs were then merged into a single file using

BCFtools. Given the poor resolution of LUMPY for small insertions, only the Delly2 calls were considered and variants were required to be found in more than 2 individuals. A convergence of calls was determined by reciprocal overlap of 50% of the defined breakpoint in each caller as our final set. SVs were annotated using SnpEff (Cingolani et al. 2012), which was used to count gene region intersects. SVs were considered exonic if they were annotated as exon_region, frameshift, start_lost, or stop_gained.

**Disease variant discovery for dwarfism**. To discover causal variants associated with dwarfism, three unrelated affected cats with disproportionate dwarfism from the 99 Lives genome dataset were examined for SVs. Identified SVs were considered causal candidates if they were, 1) concordant with affection status and an autosomal dominant inheritance pattern, 2) and located within the ~5.2 Mb dwarfism critical region located on cat chromosome B1:170,786,914 – 175,975,857 (Lyons et al. 2019). After initial identification, candidate variants were prioritized according to their predicted impact on protein coding genes. High priority candidate SVs were further characterized manually in affected individuals using the integrated genomics viewer (IGV) (Robinson et al. 2011). STIX (structural variant index) was used to validate candidate SVs by searching BAM files for discordant read-pairs that overlapped candidate SV regions (https://github.com/ryanlayer/stix). After manual characterization, SV breakpoints were validated with PCR amplification and Sanger sequencing. For further genotyping of candidate SVs, all sample collection and cat studies were conducted in accordance with an approved University of California, Davis Institutional Animal Care and Use protocols 11977, 15117, and 16691 and University of Missouri protocols 7808 and 8292. DNA samples from dwarf and normal cats were genotyped for the candidate SV identified in the three sequenced dwarfism cats (Lyons et al. 2019). PCR primers were designed using the known SV sequence breakpoints (**Supplemental Fig. S6A**). For validation, PCR amplification products were sanger sequenced and compared against Felis_catus_9.0. For screening, all samples

from previous linkage and GWAS studies (Lyons et al. 2019) were genotyped using the primers, UGDH_mid_F, UGDH_del_R, and UGDH_dn_R (**Supplemental Table S7**) in a single reaction. PCR products were separated by gel electrophoresis (80V, 90 minutes) in 1.25% (w/v) agarose in 1X TAE. A 622 bp amplicon was expected from the normal allele and a 481 bp amplicon from the affected allele (**Supplemental Fig. S6**).

**Histological characterization of dwarf cat growth plates**. Cat owners voluntarily submitted cadavers of stillborn dwarf and musculoskeletally normal kittens (7 and 2 respectively, dying from natural causes) via overnight shipment on ice. Distal radius including physis (epiphyseal plate) were collected and fixed in 10% neutral buffered formalin. These tissues were decalcified in 10% EDTA solution. After complete decalcification the tissues were dehydrated with gradually increasing concentrations of ethanol and embedded in paraffin. Frontal sections of distal radial tissues were cut to 6 µm and mounted onto microscope slides. The samples were then dewaxed, rehydrated, and stained with hematoxylin and eosin to evaluate the tissue structure and cell morphology. The sections were also stained with toluidine blue to determine the distribution and quantity of proteoglycans. These samples were subjectively assessed for chondrocyte and tissue morphology and growth plate architecture by a pathologist (KK) who was blinded to the sample information.

## Data access

The whole genome sequence data generated in this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA16726. Illumina WGS data used in this study can be found under the NCBI BioProject accession PRJNA308208.

## Acknowledgments

Funding for this project has been provided in part by Nestlé Purina, Wisdom Health unit of Mars Veterinary and Zoetis. We appreciate the donation of funding and samples from cat breeders, especially Terri Harris. We appreciate Thomas R. Juba for his technical assistance of the variant validation and genotyping. We thank Susan Brown at Kansas State University for the generation of the BioNano map.

*Author contributions:* Conceptualization: L.A.L., W.J.M., and W.C.W. Data Curation: R.M.B., B.W.D., F.H.G.F, Formal analysis: R.M.B., B.W.D., F.H.G.F., W.A.B, K.K., Funding Acquisition: L.A.L., R.M., W.J.M., and W.C.W. Supervision: L.A.L. Validation: L.A.L. Writing (original draft): R.M.B., L.A.L., and W.C.W. Writing (review and editing): all authors.

**Disclosure declaration**

The authors disclose there are no conflicts of interest.

## References


Aberdein D, Munday JS, Gandolfi B, Dittmer KE, Malik R, Garrick DJ, Lyons LA, 99 Lives Consortium. 2017. A FAS-ligand variant associated with autoimmune lymphoproliferative syndrome in cats. *Mamm Genome* **28**: 47-55.

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974-984.

Ananthasayanam S, Kothandaraman, H., Nayee, N., Saha, S., Baghel, D.S., Gopalakrishnan, K., Peddamma, S., Singh, R.B., Schatz, M. 2019. First near complete haplotype phased genome assembly of River buffalo (Bubalus bubalis). *bioRxiv*.

Atanur SS, Diaz AG, Maratou K, Sarkis A, Rotival M, Game L, Tschannen MR, Kaisaki PJ, Otto GW, Ma MC et al. 2013. Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell* **154**: 691-703.

Bai B, Zhao WM, Tang BX, Wang YQ, Wang L, Zhang Z, Yang HC, Liu YH, Zhu JW, Irwin DM et al. 2015. DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res* **43**: D777-783.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**: 643-650.

Bimber BN, Ramakrishnan R, Cervera-Juanes R, Madhira R, Peterson SM, Norgren Jr RB, Ferguson B. 2017. Whole genome sequencing predicts novel human disease models in rhesus macaques. *Genomics* **109**: 214-220.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220-1222.

Chen ZH, Zhang M, Lv FH, Ren X, Li WR, Liu MJ, Nam K, Bruford MW, Li MH. 2018. Contrasting Patterns of Genomic Diversity Reveal Accelerated Genetic Drift but Reduced Directional Selection on X-Chromosome in Wild and Domestic Sheep Species. *Genome Biol Evol* **10**: 1282-1297.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**: 966-968.

Choi JW, Chung WH, Lee KT, Cho ES, Lee SW, Choi BH, Lee SH, Lim W, Lim D, Lee YG et al. 2015. Whole-genome resequencing analyses of five pig breeds, including Korean wild and native, and three European origin breeds. *DNA Res* **22**: 259-267.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. 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 (Austin)* **6**: 80-92.

Clarkin CE, Allen S, Kuiper NJ, Wheeler BT, Wheeler-Jones CP, Pitsillides AA. 2011. Regulation of UDP-glucose dehydrogenase is sufficient to modulate hyaluronan

production and release, control sulfated GAG synthesis, and promote chondrogenesis. *J Cell Physiol* **226**: 749-761.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**: 858-865.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**: 491.

Driscoll CA, Menotti-Raymond M, Roca AL, Hupe K, Johnson WE, Geffen E, Harley EH, Delibes M, Pontier D, Kitchener AC et al. 2007. The Near Eastern origin of cat domestication. *Science* **317**: 519-523.

Fabrizio F, Calam AE, Dobson JM, Middleton SA, Murphy S, Taylor SS, Schwartz A, Stell AJ. 2014. Feline mediastinal lymphoma: a retrospective study of signalment, retroviral status, response to chemotherapy and prognostic indicators. *J Feline Med Surg* **16**: 637-644.

Fang H, Wu Y, Yang H, Yoon M, Jiménez-Barrón LT, Mittelman D, Robison R, Wang K, Lyon GJ. 2017. Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine. *BMC medical genomics* **10**: 10.

Festing MF. 2010. Inbred strains should replace outbred stocks in toxicology, safety testing, and drug development. *Toxicologic pathology* **38**: 681-690.

Foldynova-Trantirkova S, Wilcox WR, Krejci P. 2012. Sixteen years and counting: the current understanding of fibroblast growth factor receptor 3 (FGFR3) signaling in skeletal dysplasias. *Hum Mutat* **33**: 29-41.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**: 241.

Gabor L, Malik R, Canfield P. 1998. Clinical and anatomical features of lymphosarcoma in 118 cats. *Australian Veterinary Journal* **76**: 725-732.

Gandolfi B, Alhaddad H, Abdi M, Bach LH, Creighton EK, Davis BW, Decker JE, Dodman NH, Ginns EI, Grahn JC et al. 2018. Applications and efficiencies of the first cat 63K DNA array. *Sci Rep* **8**: 7024.

Genova F, Longeri M, Lyons LA, Bagnato A, 99 Lives Consortium, Strillacci MG. 2018. First genome-wide CNV mapping in FELIS CATUS using next generation sequencing data. *BMC Genomics* **19**: 895.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**: 435-444.

Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome.

Hermsen R, de Ligt J, Spee W, Blokzijl F, Schafer S, Adami E, Boymans S, Flink S, van Boxtel R, van der Weide RH et al. 2015. Genomic landscape of rat strain and substrain variation. *BMC Genomics* **16**: 357.

Horton WA, Hall JG, Hecht JT. 2007. Achondroplasia. *Lancet* **370**: 162-172.

Hug P, Kern P, Jagannathan V, Leeb T. 2019. A TAC3 Missense Variant in a Domestic Shorthair Cat with Testicular Hypoplasia and Persistent Primary Dentition. *Genes* **10**: 806.

Jaffey JA, Reading NS, Giger U, Abdulmalik O, Buckley RM, Johnstone S, Lyons LA, 99 Lives Cat Genome Consortium. 2019. Clinical, metabolic, and genetic characterization of hereditary methemoglobinemia caused by cytochrome b5 reductase deficiency in cats. *Journal of veterinary internal medicine*.

Jagannathan V, Gerber V, Rieder S, Tetens J, Thaller G, Drogemuller C, Leeb T. 2019. Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Anim Genet* **50**: 74-77.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*: 531210.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kittleson MD, Meurs KM, Harris SP. 2015. The genetic basis of hypertrophic cardiomyopathy in cats and humans. *J Vet Cardiol* **17 Suppl 1**: S53-73.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* **40**: e69.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**: 117.

Kurushima JD, Lipinski MJ, Gandolfi B, Froenicke L, Grahn JC, Grahn RA, Lyons LA. 2013. Variation of cats under domestication: genetic assignment of domestic cats to breeds and worldwide random-bred populations. *Anim Genet* **44**: 311-324.

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M. 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology* **30**: 771.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Li G, Hillier LW, Grahn RA, Zimin AV, David VA, Menotti-Raymond M, Middleton R, Hannah S, Hendrickson S, Makunin A et al. 2016. A High-Resolution SNP Array-Based Linkage Map Anchors a New Domestic Cat Draft Genome Assembly and Provides Detailed Patterns of Recombination. *G3 (Bethesda)* **6**: 1607-1616.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103-2110.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lipinski MJ, Froenicke L, Baysac KC, Billings NC, Leutenegger CM, Levy AM, Longeri M, Niini T, Ozpinar H, Slater MR et al. 2008. The ascent of cat breeds: genetic evaluations of breeds and worldwide random-bred populations. *Genomics* **91**: 12-21.

Louwerens M, London CA, Pedersen NC, Lyons LA. 2005. Feline lymphoma in the post-feline leukemia virus era. *J Vet Intern Med* **19**: 329-335.

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat Commun* **10**: 260.

Lyons LA. 2015. DNA mutations of the cat: the good, the bad and the ugly. *J Feline Med Surg* **17**: 203-219.

Lyons LA, Biller DS, Erdman CA, Lipinski MJ, Young AE, Roe BA, Qin B, Grahn RA. 2004. Feline polycystic kidney disease mutation identified in PKD1. *J Am Soc Nephrol* **15**: 2548-2555.

Lyons LA, Fox DB, Chesney KL, Britt LG, Buckley RM, Coates JR, Gandolfi B, Grahn RA, Hamilton MJ, Middleton JR. 2019. Localization of a feline autosomal dominant dwarfism locus: a novel model of chondrodysplasia. *bioRxiv*: 687210.

Mahamdallie S, Yost S, Poyastro-Pearson E, Holt E, Zachariou A, Seal S, Elliott A, Clarke M, Warren-Perry M, Hanks S et al. 2019. Identification of new Wilms tumour predisposition genes: an exome sequencing study. *Lancet Child Adolesc Health* **3**: 322-331.

Mao JH, Perez-Losada J, Wu D, Delrosario R, Tsunematsu R, Nakayama KI, Brown K, Bryson S, Balmain A. 2004. Fbxw7/Cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. *Nature* **432**: 775-779.

Maser RS, Choudhury B, Campbell PJ, Feng B, Wong KK, Protopopov A, O'Neil J, Gutierrez A, Ivanova E, Perna I et al. 2007. Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447**: 966-971.

Mauler DA, Gandolfi B, Reinero CR, O'Brien DP, Spooner JL, Lyons LA, and 99 Lives Consortium. 2017. Precision Medicine in Cats: Novel Niemann-Pick Type C1 Diagnosed by Whole-Genome Sequencing. *J Vet Intern Med* **31**: 539-544.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122.

Menotti-Raymond M, David VA, Schäffer AA, Stephens R, Wells D, Kumar-Singh R, O'Brien SJ, Narfström K. 2007. Mutation in CEP290 discovered for cat model of human retinal degeneration. *Journal of Heredity* **98**: 211-220.

Montague MJ, Li G, Gandolfi B, Khan R, Aken BL, Searle SM, Minx P, Hillier LW, Koboldt DC, Davis BW et al. 2014. Comparative analysis of the domestic cat genome reveals

genetic signatures underlying feline biology and domestication. *Proc Natl Acad Sci U S A* **111**: 17230-17235.

Moses L, Niemi S, Karlsson E. 2018. Pet genomics medicine runs wild. Nature Publishing Group.

Nicholas FW. 2003. Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic acids research* **31**: 275-277.

Oh A, Pearce JW, Gandolfi B, Creighton EK, Suedmeyer WK, Selig M, Bosiack AP, Castaner LJ, Whiting RE, Belknap EB et al. 2017. Early-onset progressive retinal atrophy associated with an IQCB1 variant in African black-footed cats (Felis nigripes). *Scientific reports* **7**: 43918.

Online Mendelian Inheritance in Animals (OMIA). Sydney School of Veterinary Science, 03/12/2019. World Wide Web URL: http://omia.org/.

Ontiveros ES, Ueda Y, Harris SP, Stern JA, 99 Lives Consortium. 2018. Precision medicine validation: identifying the MYBPC 3 A31P variant with whole-genome sequencing in two Maine Coon cats with hypertrophic cardiomyopathy. *Journal of feline medicine and surgery*: 1098612X18816460.

Ottoni C, Van Neer W, De Cupere B, Daligault J, Guimaraes S, Peters J, Spassov N, Prendergast ME, Boivin N, Morales-Muñiz A. 2017. The palaeogenetics of cat dispersal in the ancient world. *Nature Ecology & Evolution* **1**: 0139.

Perez-Losada J, Wu D, DelRosario R, Balmain A, Mao JH. 2012. Allele-specific deletions in mouse tumors identify Fbxw7 as germline modifier of tumor susceptibility. *PLoS One* **7**: e31301.

Pontius JU, Mullikin JC, Smith DR, Agencourt Sequencing T, Lindblad-Toh K, Gnerre S, Clamp M, Chang J, Stephens R, Neelam B et al. 2007. Initial sequence and comparative analysis of the cat genome. *Genome Res* **17**: 1675-1689.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*: 201178.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130-135.

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333-i339.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.

Roversi G, Picinelli C, Bestetti I, Crippa M, Perotti D, Ciceri S, Saccheri F, Collini P, Poliani PL, Catania S et al. 2015. Constitutional de novo deletion of the FBXW7 gene in a patient with focal segmental glomerulosclerosis and multiple primitive tumors. *Sci Rep* **5**: 15454.

Shendure J, Findlay GM, Snyder MW. 2019. Genomic medicine–progress, pitfalls, and promise. *Cell* **177**: 45-57.

Small C, Bassham S, Catchen J, Amores A, Fuiten A, Brown R, Jones A, Cresko W. 2016. The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome biology* **17**: 258.

Smit A, Hubley R, Green P. 2013. 2013–2015. RepeatMasker Open-4.0.
Spycher M, Bauer A, Jagannathan V, Frizzi M, De Lucia M, Leeb T. 2018. A frameshift variant in the COL5A1 gene in a cat with Ehlers-Danlos syndrome. *Anim Genet* **49**: 641-644.
Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J et al. 2018. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**: 1161-1170.
Teng H, Zhang Y, Shi C, Mao F, Cai W, Lu L, Zhao F, Sun Z, Zhang J. 2017. Population Genomics Reveals Speciation and Introgression between Brown Norway Rats and Their Sibling Species. *Mol Biol Evol* **34**: 2214-2228.
Tetzlaff MT, Yu W, Li M, Zhang P, Finegold M, Mahon K, Harper JW, Schwartz RJ, Elledge SJ. 2004. Defective cardiovascular development and elevated cyclin E and Notch proteins in mice lacking the Fbw7 F-box protein. *Proc Natl Acad Sci U S A* **101**: 3338-3345.
The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.
The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.
Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M, Kitts P. 2013. Eukaryotic genome annotation pipeline. In *The NCBI Handbook [Internet] 2nd edition*. National Center for Biotechnology Information (US).
TICA. 2015. Munchkin. In http://wwwticaorg/cat-breeds/item/236.
Tsunematsu R, Nakayama K, Oike Y, Nishiyama M, Ishida N, Hatakeyama S, Bessho Y, Kageyama R, Suda T, Nakayama KI. 2004. Mouse Fbw7/Sel-10/Cdc4 is required for notch degradation during vascular development. *J Biol Chem* **279**: 9417-9423.
Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**: 11 10 11-33.
Van Neer W, Linseele V, Friedman R, De Cupere B. 2014. More evidence for cat taming at the Predynastic elite cemetery of Hierakonpolis (Upper Egypt). *Journal of Archaeological Science* **45**: 103-111.
Vigne J-D, Guilaine J, Debue K, Haye L, Gérard P. 2004. Early taming of the cat in Cyprus. *Science* **304**: 259-259.
Visser M, Weber KL, Lyons LA, Rincon G, Boothe DM, Merritt DA. 2019. Identification and quantification of domestic feline cytochrome P450 transcriptome across multiple tissues. *J Vet Pharmacol Ther* **42**: 7-15.
Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.
Wang P, Mazrier H, Caverly Rae J, Raj K, Giger U. 2018. A GNPTAB nonsense variant is associated with feline mucolipidosis II (I-cell disease). *BMC Vet Res* **14**: 416.

Wen Y, Li J, Wang L, Tie K, Magdalou J, Chen L, Wang H. 2014. UDP-glucose dehydrogenase modulates proteoglycan synthesis in articular chondrocytes: its possible involvement and regulation in osteoarthritis. *Arthritis Res Ther* **16**: 484.

Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, Tamburro C, Williams MS, Green ED. 2019. Genomic medicine for undiagnosed diseases. *The Lancet*.

Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, Dahdouli M, Rio Deiros D, Below JE, Salerno W et al. 2016. The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. *Genome Res* **26**: 1651-1662.

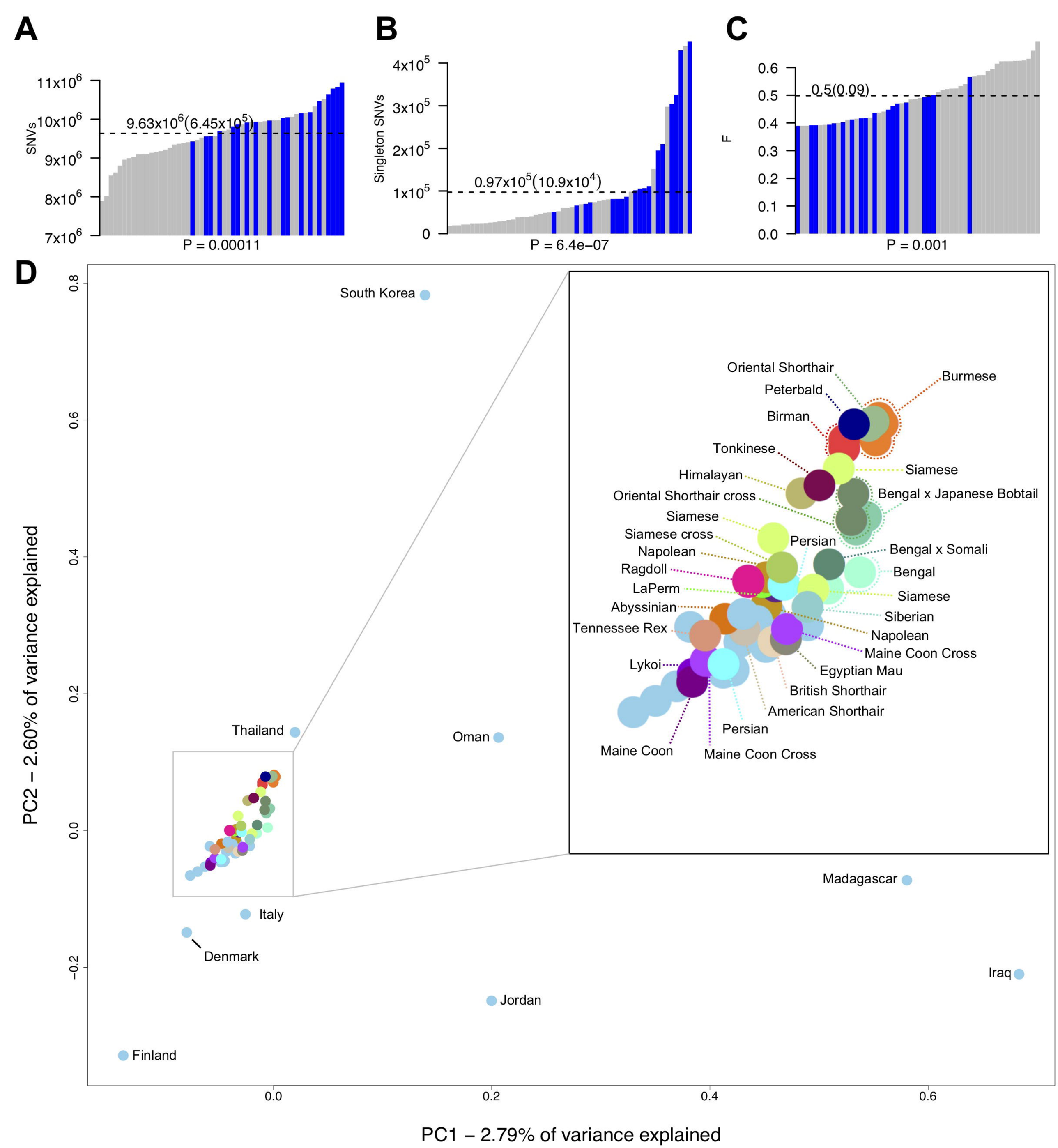
Yang J, Li WR, Lv FH, He SG, Tian SL, Peng WF, Sun YW, Zhao YX, Tu XL, Zhang M et al. 2016. Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. *Mol Biol Evol* **33**: 2576-2592.

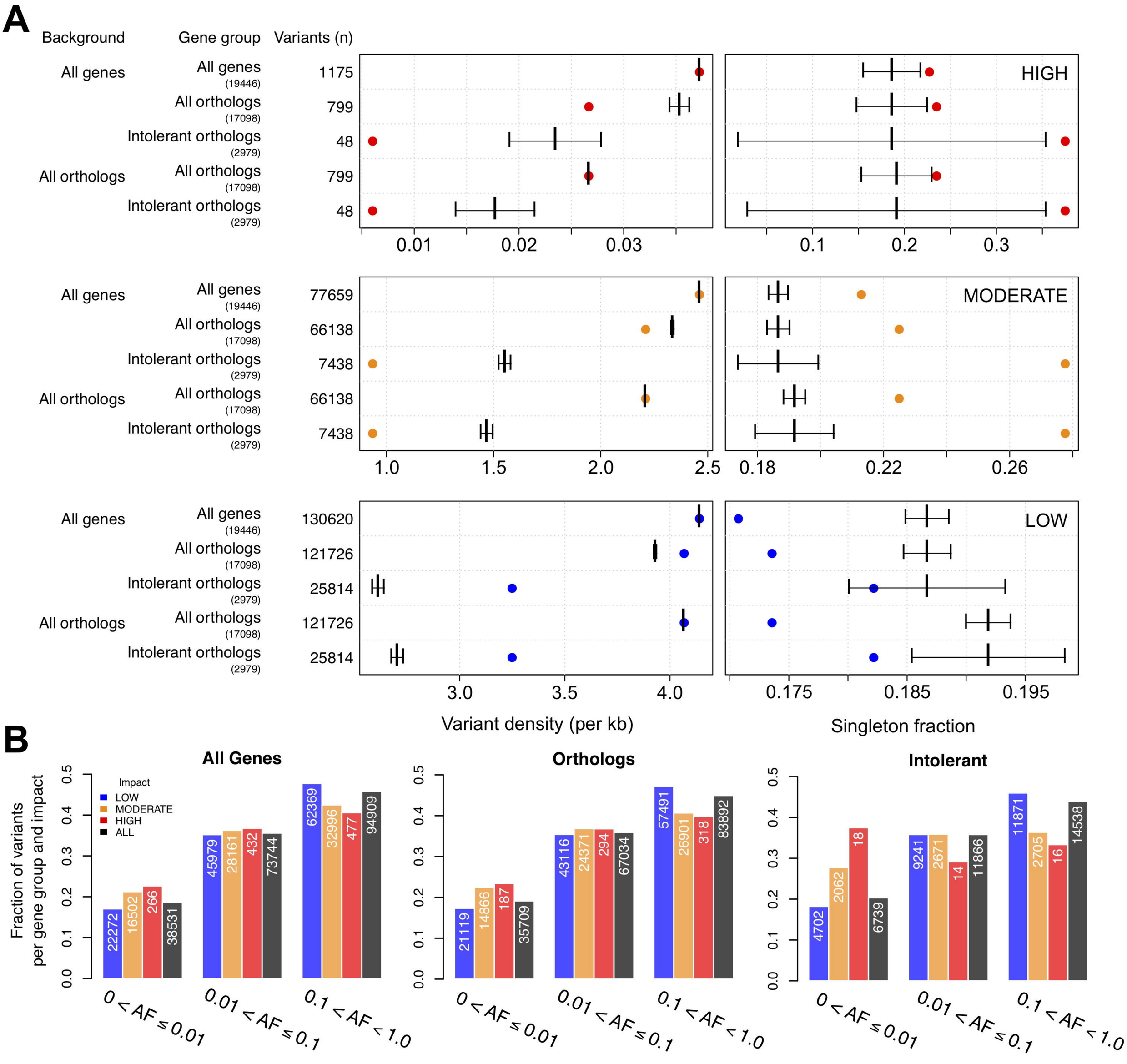
Yeh CH, Bellon M, Nicot C. 2018. FBXW7: a critical tumor suppressor of human cancers. *Mol Cancer* **17**: 115.

Yu Y, Grahn RA, Lyons LA. 2019a. Mocha tyrosinase variant: a new flavour of cat coat coloration. *Anim Genet* **50**: 182-186.

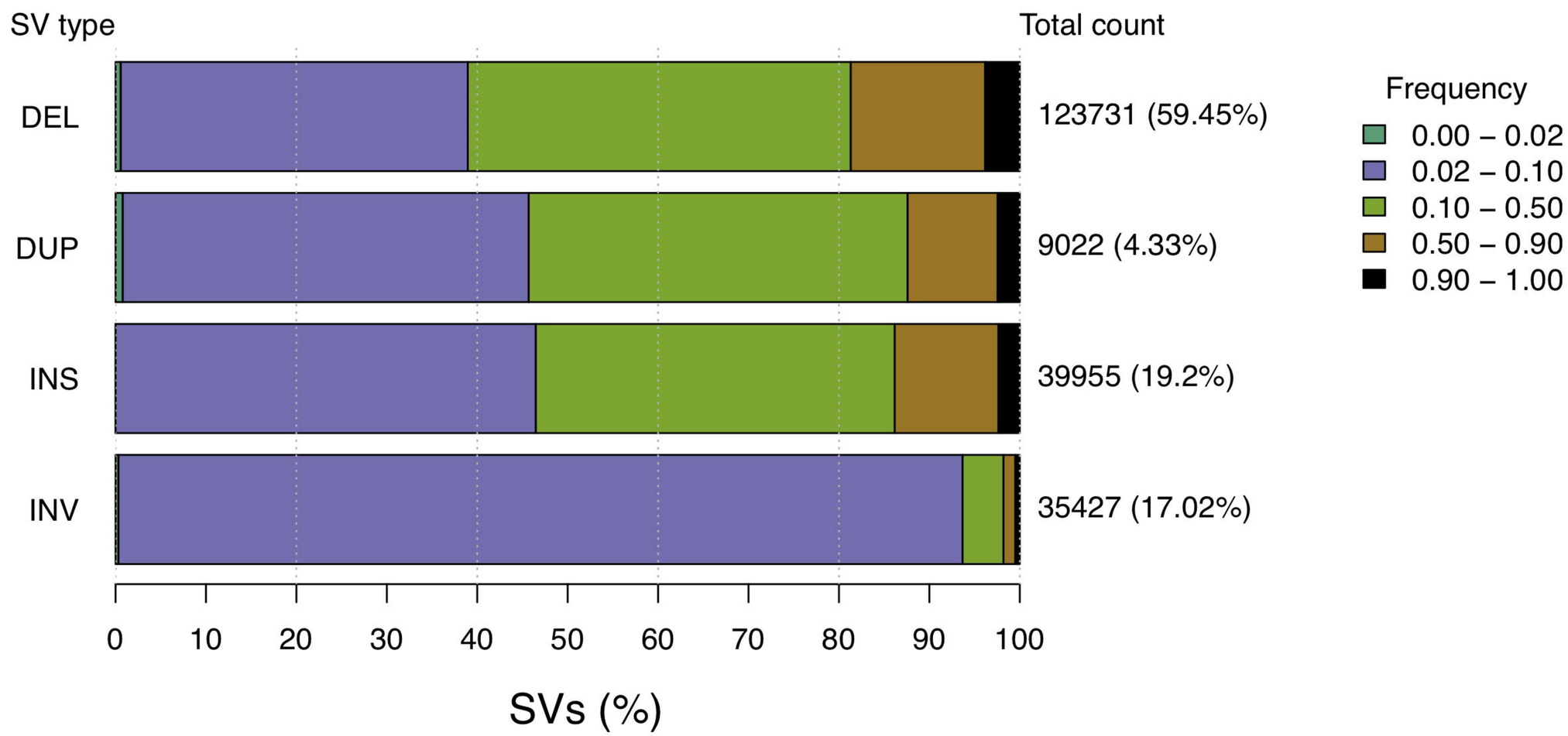
Yu Y, Shumway KL, Matheson JS, Edwards ME, Kline TL, Lyons LA. 2019b. Kidney and cystic volume imaging for disease presentation and progression in the cat autosomal dominant polycystic kidney disease large animal model. *BMC Nephrology* **20**: 259.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754-D761.

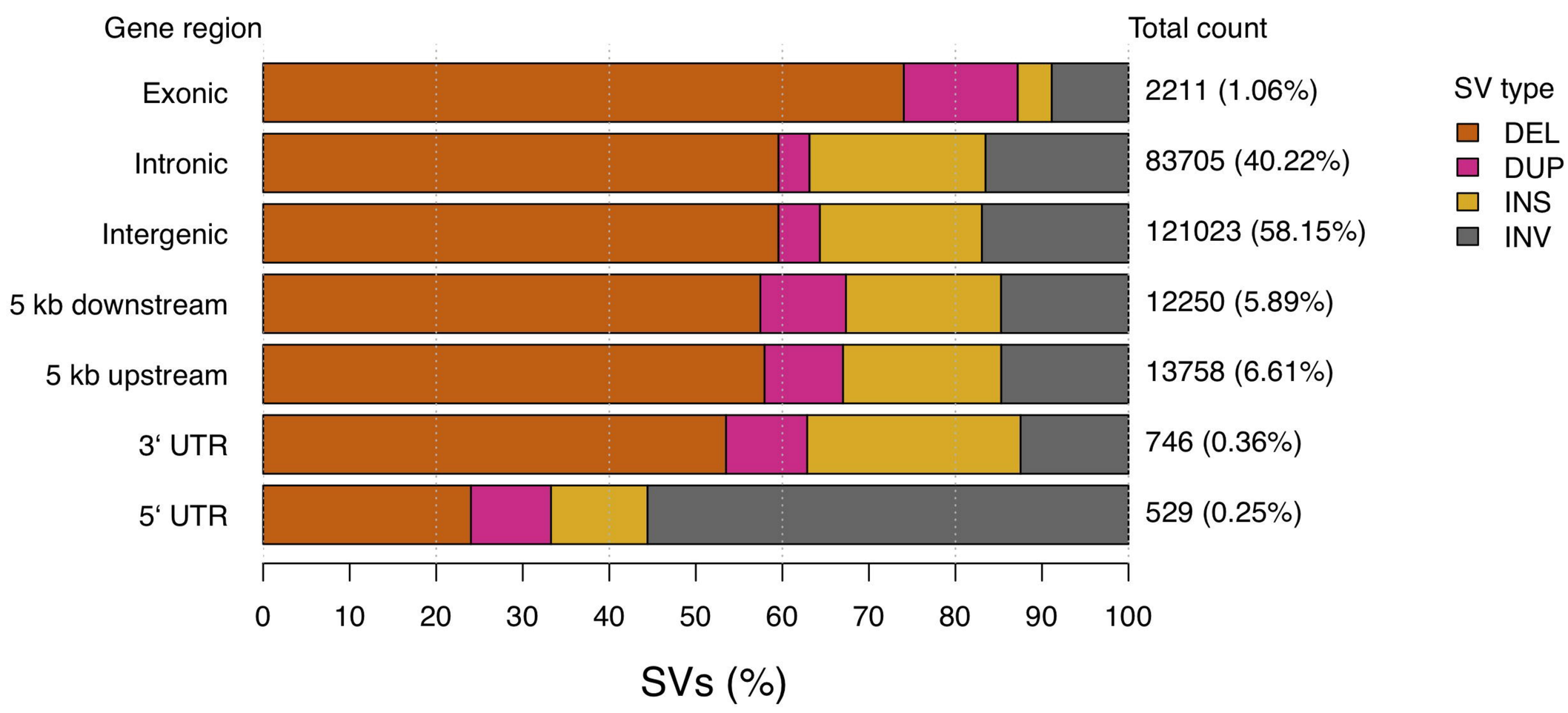

A
SNVs
11x10^6
10x10^6
9x10^6
8x10^6
7x10^6
9.63x10^6(6.45x10^5)
P = 0.00011
B
Singleton SNVs
4x10^5
3x10^5
2x10^5
1x10^5
0
0.97x10^5(10.9x10^4)
P = 6.4e−07
C
F
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.5(0.09)
P = 0.001
D
PC2 – 2.60% of variance explained
PC1 – 2.79% of variance explained
South Korea
Thailand
Oman
Madagascar
Italy
Denmark
Iraq
Jordan
Finland
Oriental Shorthair
Peterbald
Birman
Burmese
Tonkinese
Siamese
Himalayan
Oriental Shorthair cross
Bengal x Japanese Bobtail
Siamese
Siamese cross
Persian
Napolean
Bengal x Somali
Ragdoll
Bengal
LaPerm
Siamese
Abyssinian
Siberian
Tennessee Rex
Napolean
Maine Coon Cross
Lykoi
Egyptian Mau
British Shorthair
American Shorthair
Persian
Maine Coon
Maine Coon Cross

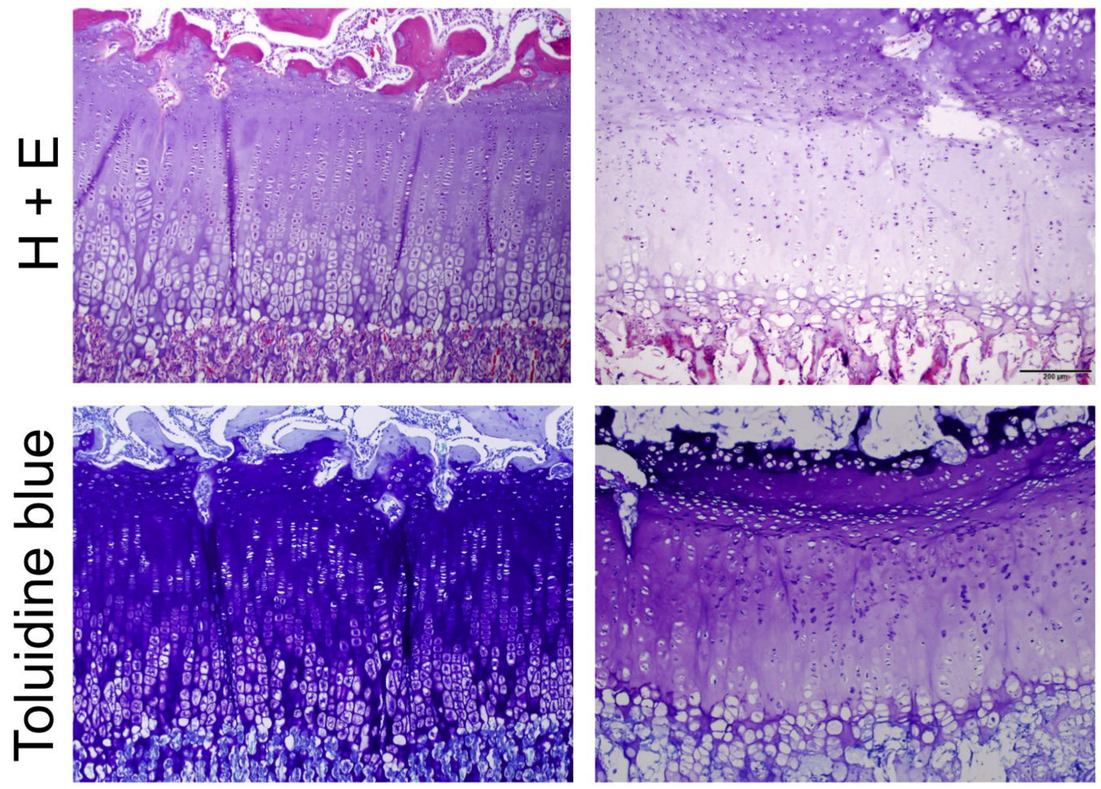

A
Background
Gene group
Variants (n)
All genes
All genes (19446)
All orthologs (17098)
Intolerant orthologs (2979)
All orthologs
HIGH
MODERATE
LOW
Variant density (per kb)
Singleton fraction
B
All Genes
Orthologs
Intolerant
Fraction of variants per gene group and impact
Impact
LOW
MODERATE
HIGH
ALL
0 < AF ≤ 0.01
0.01 < AF ≤ 0.1
0.1 < AF < 1.0

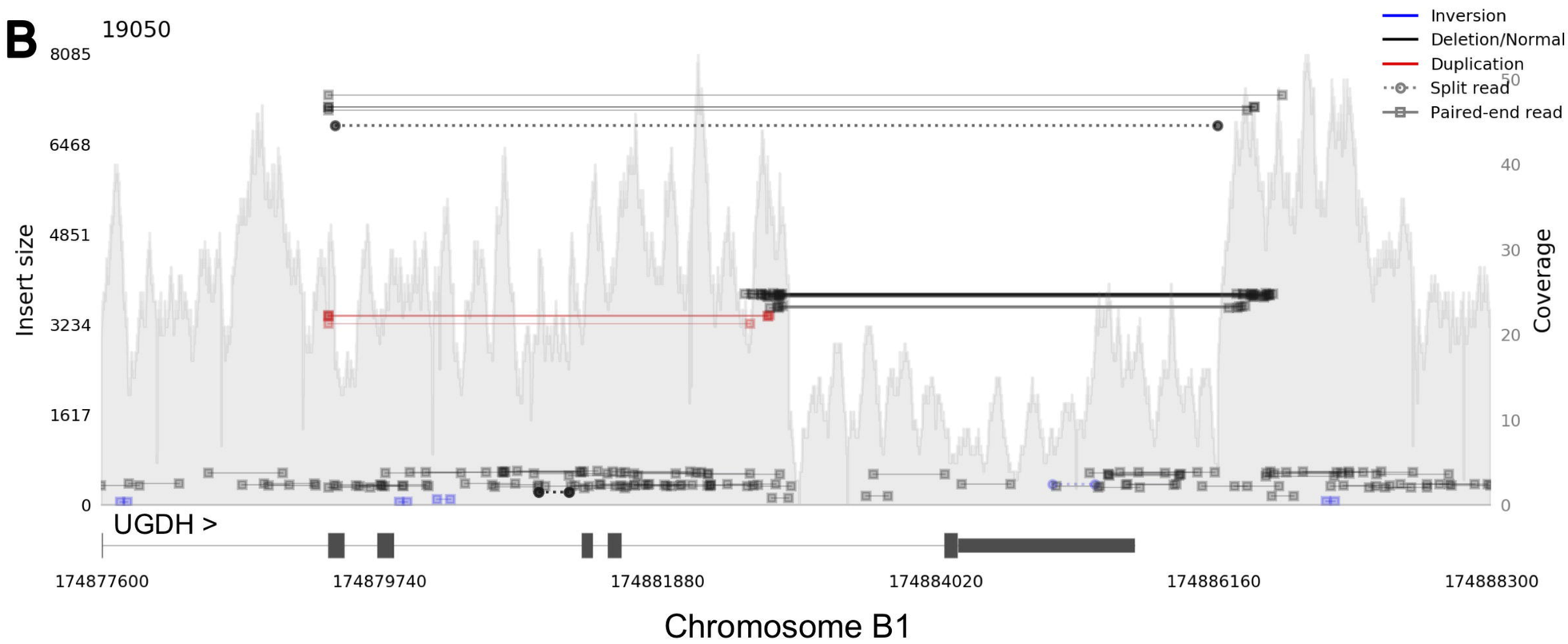
**A**

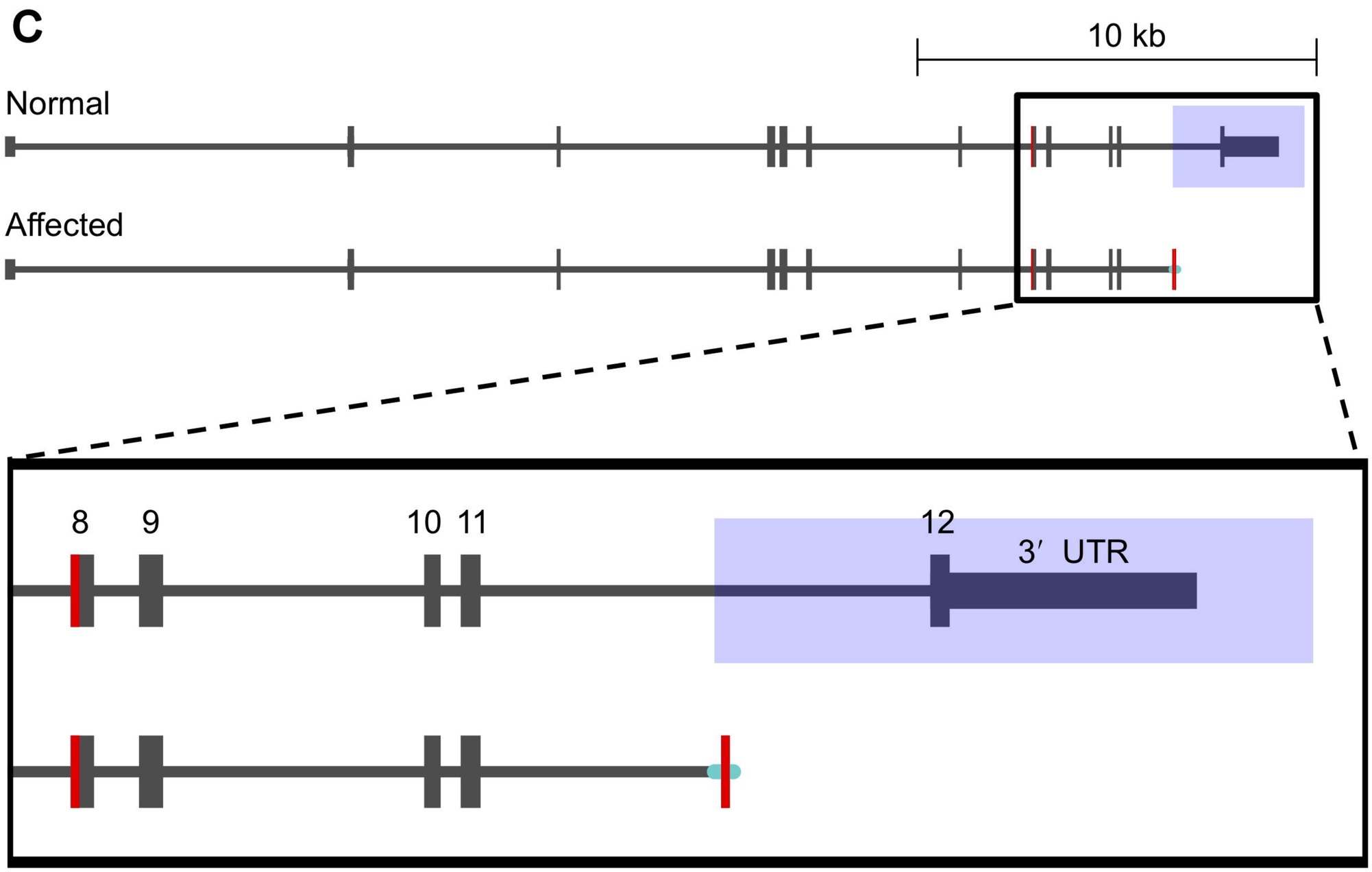

SV type
Total count
DEL
DUP
INS
INV
123731 (59.45%)
9022 (4.33%)
39955 (19.2%)
35427 (17.02%)
Frequency
0.00 – 0.02
0.02 – 0.10
0.10 – 0.50
0.50 – 0.90
0.90 – 1.00
0 10 20 30 40 50 60 70 80 90 100
SVs (%)

**B**


Gene region
Total count
Exonic
Intronic
Intergenic
5 kb downstream
5 kb upstream
3‘ UTR
5‘ UTR
2211 (1.06%)
83705 (40.22%)
121023 (58.15%)
12250 (5.89%)
13758 (6.61%)
746 (0.36%)
529 (0.25%)
SV type
DEL
DUP
INS
INV
0 10 20 30 40 50 60 70 80 90 100
SVs (%)


D
Normal
Dwarf
H + E
Toluidine blue


B
8085
19050
Inversion
Deletion/Normal
Duplication
Split read
Paired-end read
Insert size
6468
4851
3234
1617
0
Coverage
40
30
20
10
0
UGDH >
174877600
174879740
174881880
174884020
174886160
174888300
Chromosome B1


C
10 kb
Normal
Affected
8
9
10
11
12
3′ UTR