Supplementary Materials and Methods

Comparative Genomics Reveals Shared Mutational Landscape in Canine Hemangiosarcoma and Human Angiosarcoma

DNA extraction

FFPE tumor samples were macrodissected by microtome to select for regions of high tumor cell density, and tumor DNA was prepared using the Qiagen QIAamp DNA FFPE Tissue Kit. DNA from frozen tumor tissue and germline DNA from whole blood were prepared using the Qiagen DNeasy Blood and Tissue Kit. The DNA was then eluted into nuclease-free purified water. DNA concentration was determined using the NanoDrop microvolume spectrophotometer (ThermoFisher Scientific) and/or the Quant-iT PicoGreen system (Invitrogen).

Library construction

DNA from each tumor and normal sample was diluted in Tris-EDTA (TE) buffer for sonic fragmentation. Samples were fragmented to a target size of 500bp using a Covaris ultrasonicator (Covaris). Fragments were cleaned and subject to size selection using Agencourt AMPure XP magnetic beads (Beckman Coulter). Select samples were visualized using the Agilent BioAnalyzer to check the distribution of fragment sizes. The Kapa Hyper Prep Kit was used for library construction (Kapa Biosystems). Briefly, size-selected DNA fragments were subject to end-repair and A-Tailing reactions, followed by attachment of adaptors in preparation for molecular barcoding. Fragments were purified using Agencourt AMPure XP magnetic beads between reactions. NEBNext Multiplex Oligos for Illumina (New England BioLabs) were then used to barcode the individual libraries following product guidelines.

Amplification of samples

The 66 overamplified samples had significantly lower library complexity than the 28 samples amplified using the recommended number of PCR cycles (mean number of unique molecules 195,124,255 in standard libraries, 112,315,681 in overamplified libraries; $p = 1.8 \times 10^{-5}$; **Table S2**). Although the mean number of total mutations called in the overamplified samples was lower than in the standard samples (mean mutational burden per tumor/normal pair = 46.3 vs 64.2), this did not reach statistical significance ($p_{t\text{-test}} = 0.14$). (For this calculation, an outlier with 346 mutations was removed, and two pairs which were a combination of overamplified and standard library were removed.) Thus, in the overamplified samples, we may have been underpowered for variant discovery, leading to a more conservative set of mutations. We note that all significantly mutated genes discovered during data analysis were mutated in both standard and overamplified samples, with the sole exception of *ENSCAFG00000017407*, which was observed in the overamplified samples only (**Figure S2**).

Exome capture

Custom blocking oligonucleotides were designed complementary to the barcode sequences and synthesized by Integrated DNA Technologies (IDT). Briefly, the amplified libraries were incubated with the SeqCap EZ probes, blocking oligos, and Roche Developer Reagent (in place of human Cot-1 DNA), for 60 hours in a thermocycler (Eppendorf). Captured DNA was then recovered and washed using SeqCap-EZ Capture Beads. The captured library was amplified via ligation-mediated (LM)-PCR for the recommended 14 cycles, and the amplified captured library washed using AMPure XP magnetic beads. qPCR using primers for specific targeted regions and a negative control region that was not targeted was performed to test enrichment of a subset of the captured libraries using the LightCycler 480 instrument (Roche).

Somatic variant calling

We called variants in the GATK3 MuTect2 and GATK4 Mutect2 versions, with the addition of the *--dontUseSoftClippedBases* option. Prior to using this setting, we saw a large number of artifactual indels being called in our FFPE samples. These indels were being called with the only support for the variant being the ends of soft-clipped reads. A similar artifact has been reported in WGA TCGA data (1). We found no significant difference in the total number of mutations ($p_{t\text{-}test} = 0.75$) or percent of indels ($p = 0.95$) between frozen ($n = 17$) and FFPE ($n = 30$) samples in the final somatic mutation call set.

Variants for the panel of normals were called as recommended in the GATK3 workflow, using *--artifactDetectionMode* in MuTect2, and the calls from the normal samples were merged using CombineVariants, keeping any variant that was called in two or more dogs. All preprocessing was performed using GATK version 3.6.0 (2). BQSR was performed using a set of 19,112,082 known canine SNP positions drawn from multiple sources, including SNPs discovered by the Lindblad-Toh (3,4) and Axelsson labs (5), those included on the Affymetrix Axiom Canine HD array, and those contained in the DoGSD database (6).

Using GATK4 (version 4.beta.3), we again used the default Mutect2 parameters, with the addition of the *--dontUseSoftClippedBases* option. We then applied the FilterMutectCalls tool to this set of variant calls, using default cutoffs, with the exception of increasing the stringency of the median read position filter to ten, and specifying a unique alternate allele read count of four. The FilterByOrientationBias tool was also applied, using the “G,T” setting (for oxidation artifacts) for all samples, and the “C,T” setting for FFPE-preserved samples.

Somatic copy number aberration calling

Somatic copy number alterations in tumors compared to the matched normal were called in the exome data using VarScan2 (7) followed by circular binary segmentation to translate intensity measurements into regions. Recurrent SCNAs were then identified using Gistic2 (8) with default options and a cutoff threshold of 10000 for “max seg.”

RNA-sequencing

Total RNA was isolated from tissue samples using the TriPure Isolation Reagent (Roche Applied Science), and the RNeasy Mini Kit (Qiagen) was used for clean-up according to the manufacturer's instructions. Briefly, the TruSeq RNA sample preparation kit (Illumina) and a HiSeq 2000 or 2500 sequencing system (Illumina) were used to generate Illumina sequencing libraries. Each sample was sequenced to a targeted depth of approximately 20 – 80 million paired-end reads with mate-pair distance of 50 bp. Primary analysis and demultiplexing were performed using CASAVA software version 1.8.2 (Illumina) to verify the quality of the sequence data. The end result of the CASAVA workflow was demultiplexed into FASTQ files for analysis. Bioanalyzer quality control and RNA-seq were performed at the University of Minnesota Genomics Center (UMGC) or at the Broad Institute. Tumor purity estimates were made based on histologic examination of tumor slides, as well as from the RNA-seq data using the program ESTIMATE (9).

RNA-seq validation and variant identification

RNASEQ data was mapped using the STAR-Mapper (10) with STAR-FUSION mapping settings (11) to the CanFam3.1 genome. Bam files generated by STAR were sorted and indexed using Samtools (12). A pipeline was developed to identify the bases present at locations defined as somatic mutations in the tumor-normal exome calls. Starting from a file containing somatic mutation locations and a file containing a list of bam file location, the pipeline used Samtools functions to identify the bases present at each location at each file and return the total number of reads observed as the first number. Then the numbers of A:T:C:G:N:<>:+:-:* were returned. As the mpileup only returns values that are different from the expected base at the given position, only alternate reads were reported. For validation of a variant within the same individual, a minimum of 3 reads of the alternate allele found in the exome study were required. For identification of variants in additional individuals, at least 3 reads of the alternate allele representing greater than 10% of the reads present at that location were required.

Construction of lollipop plots

Canine variants were lifted over to the human genome (hg19) using the UCSC LiftOver tool (13). Plots were created using the MutationMapper function at cBioPortal. Six canine TP53 variants were not plotted due to mismatch of the reference allele in the canine and human genomes.

References

1. MuTect2 Insertion Artifacts [Internet]. National Cancer Institute Genomic Data Commons. 2016 [cited 2018 Sep 28]. Available from: https://gdc.cancer.gov/content/mutect2-insertion-artifacts

2. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1–33.

3. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438:803–19.

4. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011;7:e1002316.

5. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013;495:360–4.

6. Bai B, Zhao W-M, Tang B-X, Wang Y-Q, Wang L, Zhang Z, et al. DoGSD: the dog and wolf genome SNP database. Nucleic Acids Res. 2015;43:D777–83.

7. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.

8. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.

9. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

11. Haas BJ, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq [Internet]. bioRxiv. 2017 [cited 2019 Jul 6]. page 120295. Available from: https://www.biorxiv.org/content/early/2017/03/24/120295

12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

13. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.