DISCOVERY OF CANINE CANCER-ASSOCIATED SOMATIC MUTATIONS IN WES, WGS, AND RNA-SEQ

by

KUN-LIN HO

(Under the Direction of Shaying Zhao)

ABSTRACT

Spontaneous cancers in dogs provide valuable yet underutilized models for cancer research. To enhance their utility, we comprehensively analyzed DNA sequencing data from 684 canine tumors using whole exome sequencing across over 35 breeds and 7 common cancer types. Our results demonstrated that the genetic landscape of canine cancers is predominantly driven by tumor type, with each cancer displaying distinct mutational patterns that mirror those in human counterparts. For example, canine mammary tumors had frequent PI3K pathway mutations, while osteosarcomas showed a high prevalence of TP53 mutations. We also found variable tumor mutation rates across cancer types, with higher rates in oral melanoma, osteosarcoma, and hemangiosarcoma but lower rates in mammary tumors and gliomas. Interestingly, mutation rates are consistently associated with TP53 but not PIK3CA mutations. In parallel, we developed an efficient pipeline to identify somatic mutations from tumor-only RNA-seq data by leveraging a large database of known human cancer mutations, variant allele frequencies, and machine learning. This approach reduced the need for matched normal samples while recapitulating expected mutation patterns, such as PIK3CA mutations in mammary tumors, and revealing consistent mutation rate patterns across canine cancer types. Our integrated analyses provide

optimized methods to unlock the genetics of spontaneous canine cancers for translational insights into human disease.

INDEX WORDS: Germline and somatic mutations ; spontaneous canine tumors ; whole exome sequencing (WES) ; whole genome sequencing (WGS) ; RNA-seq ; cancer-associated somatic mutations

DISCOVERY OF CANINE CANCER-ASSOCIATED SOMATIC MUTATIONS IN WES, WGS, AND RNA-SEQ

by

KUN-LIN HO

BS, Tamkang University, Taiwan, 2007

MS, Northeastern University, Boston, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2023

DISCOVERY OF CANINE CANCER-ASSOCIATED SOMATIC MUTATIONS IN WES, WGS, AND RNA-SEQ

by

KUN-LIN HO

Major Professor:	Shaying Zhao
Committee:	Kevin Dobbin
	Tianming Liu
	Stephen Tompkins
	Zhong-Ru Xie

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2023

DEDICATION

I dedicate this work to my beloved parents, Su-Chaio Chen and Wu-Hsiung Ho, and to my cherished wife, Pei-Hua Yu

ACKNOWLEDGEMENTS

I would like to extend my deep gratitude to my supervisor, Dr. Shaying Zhao, for her unwavering support and guidance throughout this research project. Under Dr. Zhao's guidance, I learned the importance of delving deeper into scientific questions, as well as how to approach them effectively. I also acquired the valuable skill of scrutinizing my findings, subjecting them to doubt, and seeking methods to validate my results. These lessons have been instrumental in shaping my research journey.

I would also like to express my sincere appreciation to my committee members: Dr. Kevin Dobbin, Dr. Tianming Liu, Dr. Stephen Tompkins, and Zhong-Ru Xie. Their constant support and constructive feedback have played a pivotal role in solidifying and enhancing the quality of my analyses.

I also want to thank my fellow lab members for their invaluable assistance and feedback. Their contributions have not only improved my computational skills but also deepened my understanding of various research projects. I am particularly grateful for our collaborative efforts on other research projects, where our professional teamwork skills allowed us to learn from one another and achieve success in our graduate studies.

In addition, I extend my gratitude to the administrative team of the Institute of Bioinformatics (IOB). Starting with the IOB Graduate Program Administrator, April Mosley, whose dedication and professionalism in promptly addressing my inquiries and providing assistance were invaluable. I would also like to thank the IOB Human Resources Manager, Sandra Gets, and Administrative Coordinator, Jara Lane Usherwood, for their timely and

efficient support. Each of them ensured that I received the necessary assistance to make my academic journey smoother and more efficient.

I am profoundly appreciative to all who contributed to my academic and research pursuits. This work would not have been possible without their collective efforts.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

LITERATURE REVIEW AND INTRODUCTION

INTRODUCTION OF CANCERS

Cancer is the leading cause of death worldwide as reported by the World Health Organization (WHO) and estimated 608,570 Americans died in 2021, corresponding 1600 deaths per day[1]. Therefore, cancer causes significant economic and personal and national health impact. The magnitude of these numbers is not only a sobering reflection of the profound individual suffering but also a striking reminder of the substantial economic and public health impact that cancer wields. The gravity of the cancer challenge underscores the urgency of research and innovation in understanding its complex biology, identifying effective treatments, and ultimately finding a cure. By exploring somatic mutations and their potential as biomarkers for early detection and personalized therapies, we aim to enhance the arsenal of tools available in the fight against this pervasive disease. In the following sections, we will further explore the multifaceted nature of cancer, its underlying genetic complexity, and the rationale behind the use of the canine model as a valuable resource in advancing our understanding and treatment of this formidable adversary.

CANCER BIOLOGY

Understanding the features of the cancer is a crucial step to develop effective methods for diagnosis and treatment. The defining features of cancer, often referred to as "hallmarks,"

encompass a range of physiological traits that enable malignant cells to thrive and evade the body's regulatory mechanisms[2]. These traits include:

1. Self-Sufficiency in Growth Signals[2]: Unlike normal cells that depend on external signals to trigger growth, cancer cells are capable of autonomously generating their own growth signals. This autonomy reduces their reliance on external growth stimuli and enables uncontrolled proliferation, which is typically recognized as one of the initial and defining characteristics by cancer researchers[2].

2. Insensitivity to Growth[2]-Inhibitory Signals[2]: In contrast to normal cells, cancer cells display resistance to signals that inhibit their growth. While healthy cells would receive these signals and cease proliferation, cancer cells remain unresponsive and keep proliferation.

3. Evasion of Programmed Cell Death[2]: Under normal circumstances, cells will undergo programmed cell death, known as apoptosis, when the cells have suffered severe damages or have outlived their utility. Cancer cells, however, have developed mechanisms to evade this critical self-destructive process, allowing them to persist, accumulate and promote uncontrolled cell proliferation.

4. Limitless Replicative Potential[2]: Healthy cells possess a built-in mechanism that limits their capacity to replicate—a process known as cellular senescence. In contrast, cancer cells evade this limit, enabling them to divide indefinitely and amass a growing tumor. This trait is also known as immortalization[3].

5. Sustained Angiogenesis[4]: For tumor growth to continue, cancer cells must establish a network of blood vessels that supply them with the nutrients and oxygen needed to thrive. This process,

called angiogenesis, is a hallmark feature of cancer and represents a key driver of tumor expansion.

6. Tissue Invasion & metastasis: Invasive behavior is a critical hallmark that empowers cancer cells to infiltrate neighboring tissues and even spread to distant sites in the body. This ability to invade and metastasize is a major contributor to the human cancer death[5].

The acquisition of hallmark features in cancer cells is intricately linked to genome instability and genetic mutations. Mutations can arise from various sources, including:

- Replication Errors: During cell division, errors may occur in the duplication of DNA, leading to mutations in the resulting daughter cells[6].
- DNA Damage[6]: Exogenous factors, such as exposure to chemicals, ultraviolet (UV) light, and ionizing radiation, can inflict damage to the DNA. Furthermore, endogenous factors, such as reactive oxygen species, aldehydes, or errors during mitosis, can also induce DNA damage[6]. These damaging events may trigger mutations in the affected DNA sequences, further fueling the evolution of the cancer cell's genetic profile.

Certain mutations in the genome provide growing advantage, leading to dominance in a local tissue environment.

MUTATIONS ASSOCIATED WITH CANCER AND NORMAL CELLS:

In the realm of genetic mutations, two types of mutations are discerned: germline mutations and the somatic mutations. Germline mutations occurs in sperm and eggs and they are heritable. Conversely, somatic mutations occur in other cell types and are non-heritable. Both

germline and somatic mutations have the potential to significantly reshape cellular function, providing a selective advantage to the affected cells, which, in turn, promotes their survival. This type of mutation is often referred as driver mutation[7], in contrast to passenger mutations, which have either no phenotypic consequences or biological effects.

In general, germline mutations are less frequently associated with cancer driver mutations, whereas somatic mutations are frequently linked to such driver mutations[8]. These driver mutations can manifest as gain-of-function alterations, such as the activation of oncogenes like KRAS and BRAF, which enhance cell growth and proliferation. Conversely, they can lead to loss-of-function mutations in tumor suppressor genes like TP53, BRCA1, and BRCA2, which normally serve to restrain cell growth and repair DNA damage.

Understanding the intricacies of these genetic mutations and their consequences is central to unraveling the complexities of cancer biology. Therefore, determining the somatic mutation profile of cancer could elucidate biomarkers for early detection or personalized therapies.

The progression of sequencing technology and analytical methodologies has opened the door to large-scale identification of somatic mutations across diverse cancer genomes. This advancement significantly bolsters the prospects of identifying crucial biomarkers for cancer treatment.

CANINE CANCER ANIMAL MODELS

A useful animal model is essential to accelerate the development of cancer treatment and enhance the understandings of tumor progression. Rodent models have been wildly used in preclinical studies and serve to replicate the disease process. However, crucial genetic molecular,

immunologic, environmental, and physiological differences may prohibit rodent models from serving as effective models for cancer treatment[9]. Therefore, many promising preclinical studies with rodent models failed to translate into clinical success[10-12]. Compared to rodents models, the dog model has several advantages.

1. Pet dogs are exposed to similar carcinogens and microbes and share the same environment as their caregivers: Dogs can spontaneously develop cancer that resembles human cancer without genetic manipulation, thereby better reflecting the etiology of cancer development observed in humans[13]. Basically, canine cancers are described in the same language as that of human and can be classified according to histologic and/or clinical staging systems used in human cancers[14].
2. Dogs have large populations and comparable cancer incidence rate as human[14]: Cancer is the most common cause of death in dogs, affecting approximately four million dogs per year. Data from the Animal Tumor Registry of Genoa estimated that the incidence of cancer in dogs ranged from 99.3 to 272.1 per 100,000 dogs[15]. These data are comparable to the estimated cancer incidence in humans reported by the National Cancer Institute SEER program.
3. Immune System Experience in Pet Dogs[13]: Pet dogs encountered various immunizations and infections prior to cancer development. This diverse immunological experience shapes their immune repertoire, making them more immunologically experienced than lab-raised rodents. Additionally, dogs often develop tumors spontaneously, allowing their immune systems extended time to recognize and respond to tumors before clinical diagnosis. This prolonged exposure to tumor antigens educates the canine immune system in a way that's unmatched by rodent models in the

laboratory, where tumors are artificially induced. This natural, extended interaction with tumor-associated antigens makes dogs valuable for studying immunotherapies in a context resembling human cancer development.

With these advantages, researchers can enhance the efficiency of cancer therapy development and significantly accelerate the transition from preclinical studies to human clinical trials. Ultimately, the use of dogs as a model system not only benefits cancer research but also holds the potential to improve the lives of both human and canine patients, underscoring the mutually beneficial relationship between medical advancements and our furry companions.

CHAPTER 2

CANINE TUMOR MUTATIONAL BURDEN IS CORRELATED WITH TP53 MUTATION ACROSS TUMOR TYPES AND BREEDS[1]

[1]Alsaihati, B.A., Ho, K.L., Watson, J., Feng, Y., Wang, T., Dobbin, K.K., and Zhao,S. (2021). Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. Nature Communication 12, 4670. 10.1038/s41467-021-24836-9.

ABSTRACT

Spontaneous canine cancers are valuable but relatively understudied and underutilized models. To enhance their usage, we reanalyze whole exome and genome sequencing data published for 684 cases of >7 common tumor types and >35 breeds, with rigorous quality control and breed validation. Our results indicate that canine tumor alteration landscape is tumor type-dependent, but likely breed-independent. Each tumor type harbors major pathway alterations also found in its human counterpart (e.g., PI3K in mammary tumor and p53 in osteosarcoma). Mammary tumor and glioma have lower tumor mutational burden (TMB) (median < 0.5 mutations per Mb), whereas oral melanoma, osteosarcoma and hemangiosarcoma have higher TMB (median ≥ 1 mutations per Mb). Across tumor types and breeds, TMB is associated with mutation of TP53 but not PIK3CA, the most mutated genes. Golden Retrievers harbor a TMB-associated and osteosarcoma-enriched mutation signature. Here, we provide a snapshot of canine mutations across major tumor types and breeds.

INTRODUCTION

Cancers in pet dogs arise spontaneously in animals that have intact immune systems and share the same environment as humans. Compared to traditional cancer models such as cell lines and rodents, these canine cancers more accurately emulate human cancers in etiology, complexity, heterogeneity, behavior, treatment and outcome. Hence, they have the potential to effectively bridge a current gap between preclinical studies and human clinical trials, accelerating bench-to-bedside translation[13,16,17]. As such, the National Cancer Institute (NCI) has recently issued programs targeting canine cancers. These include funding multi-institute immunotherapy trials in pet dogs[18] and a 5-year project to build the NCI Integrated Canine Data Commons[19], a database for canine data dissemination similar to the cancer genome atlas (TCGA) data portal. Private foundations are also funding canine studies, including the Vaccination Against Canine Cancer Study, a 5-year, $6 million trial to vaccinate 800 healthy dogs using tumor-specific neoantigens to determine if the vaccination will prevent or delay the onset of cancer[20].

However, current deficiencies create roadblocks to the effective use of canine cancers. This is clearly exemplified by sequence mutation, a hallmark of cancer[21]. Mutation landscape, burden and signature have all been extensively investigated in human cancer via pan-cancer studies[22-27]. However, to our knowledge, no pan-cancer research has been published for the dog, and fundamental questions remain unanswered. For example, does canine tumor mutation landscape match that of human cancer? Does canine tumor mutational burden (TMB) also vary significantly among cancer types, as it does in human cancers[22,23]?

The lack of pan-breed cancer study also leaves key questions unanswered. For example, Golden Retrievers are predisposed to the development of osteosarcoma, lymphoma and hemangiosarcoma; do the mutation landscape and TMB of Golden Retriever differ among these cancer types? Golden Retriever, Greyhound and Rottweiler dogs are all predisposed to osteosarcoma; do the mutation landscape and TMB of osteosarcoma differ among these breeds? Addressing these questions will significantly enhance the usage of >300 pure breeds of the dog in cancer research.

To answer these questions, we performed a pan-tumor and pan-breed study with matched tumor and normal samples of 684 cases, which represent over 7 common canine tumor types and over 35 popular breeds, with published whole exome sequencing (WES)[28-36] (654 cases) and/or whole genome sequencing (WGS) data[30,37] (86 cases). We performed comprehensive quality controls (QC), including breed validation, of these datasets. We then investigated somatic mutations, which include somatic base substitutions and small indels, as well as gene amplifications and deletions in 597 tumors from 591 cases with WES data passing our QC measures. Our results indicate that these alterations are tumor type-dependent, but mostly breed-independent. Across tumor types and breeds, TMB, defined as the number of somatic base substitutions and small indels per Mb callable coding sequence (CDS), is associated with mutation of *TP53* but not *PIK3CA*, the two most mutated genes. Finally, each tumor type harbors major pathway alterations that are also found in its human counterpart. Our study provides a snapshot of mutations across major tumor types and breeds in pet dogs.

RESULTS

Quality control (QC) of published canine sequencing data.

The WES dataset consists of 1,316 paired tumor and normal samples of 654 animals from 9 Bioprojects. These include 204 cases (408 samples) of mammary tumor[28,29], 56 cases (112 samples) of glioma[30], 61 cases (122 samples) of B-cell lymphoma[31], 39 cases (78 samples) of T-cell lymphoma[31], 65 cases (136 samples) of oral melanoma[32], 78 cases (156 samples) of osteosarcoma[33,34], 68 cases (138 samples) of hemangiosarcoma[35,36] and 83 cases (166 samples) of unclassified tumors. They represent over 35 breeds, including Golden Retriever (163 dogs), Maltese (69 dogs), Poodle (38 dogs), Boxer (36 dogs) and others.

One of the mammary tumor studies[28] provides the most comprehensive case information, with patient (e.g., age, sex, breed), histological subtype and limited clinical (e.g., tumor invasiveness, patient alive/death status) data. The osteosarcoma, lymphoma, glioma and hemangiosarcoma studies all have patient information, but lack clinical data. The oral melanoma study lacks patient information, including breed.

The WES data were generated by different groups, using different exome-capturing kits and Illumina sequencing machines. We hence performed a rigorous QC to ensure that data chosen from each study meet a set of quality standards before any integrative analysis.

For the sequencing amount, except for certain mammary and hemangiosarcoma sample sets, all datasets have a median of >50 million (M) read pairs per sample (Figure 2.1a). We excluded two samples with <5M read pairs from further analyses.

We then examined the mapping of read pairs to the canine reference genome[38]. Except for the glioma and one hemangiosarcoma datasets, all studies have >80% read pairs in nearly every sample uniquely and concordantly mapped to the genome, with the median close to or larger than 90% (Figure 2.1b). We excluded 9 samples with mapping rates <60%. Furthermore, except for glioma and hemangiosarcoma, nearly all samples have >70% reads (close to 90% for mammary and melanoma samples) with a mapping quality score of >30 (Figure 2.1c). For the target rate, all studies except two have, on average, >50% read pairs that are uniquely and concordantly mapped placed to the CDS regions, with the melanoma study and one mammary tumor study[28] achieving >60% (Figure 2.1d). We excluded three samples with target rates <30%. For the average mapped read coverage in CDS regions, except for a hemangiosarcoma dataset[36], all studies have reached a median of >70X (Figure 2.1e). We excluded 24 samples with coverage <30X. For the mapped read distribution in the target regions (which reflects sequencing randomness), we determined the deviation of each sample from its theoretical Poisson distribution (as a completely random sequencing process can be approximated by the Poisson distribution). The results indicate that one mammary tumor study[28] has the most random sequencing, closely followed by the oral melanoma study (Figure 2.1f). We excluded one sample which is a clear outlier (Figure 2.1f). After these steps, all samples have >10Mb callable bases in total in CDS regions (used for somatic base substitution and small indel discovery; see Methods) (Figure 2.1g).

To assess the tumor-normal sample-pairing accuracy, we used germline base substitution and small indel variants detected in each sample, assuming that correctly paired samples, compared to other samples in the same study, should share the most variants. We found a total of 24 mis-paired cases (Figure 2.1h), and excluded them from further analysis.

In summary, our QC analysis indicates that one of the mammary studies[28] and the oral melanoma study[32] have the highest sequence quality, and that the mammary study[28] has the most comprehensive case information. A total of 591 cases (597 tumors and 591 matching normal samples) have passed our QC measures, and were used for further analyses.

We also performed similar QC analyses on the WGS dataset, which consists of 172 paired tumor and normal samples from 86 animals with glioma (67 cases)[30], oral or ocular melanoma (4 cases)[37], or osteosarcoma (15 cases)[34]. Close to 30 breeds are covered, including Boxer (24 animals), Boston Terrier (11 animals) and others. We found 25 samples with a mapping rate <60% and 25 samples with a sequence coverage <30X, and excluded them from further analysis. Because of the small sample size (only 72 paired tumor and normal samples from 36 cases passed QC), we used the WGS dataset only for breed validation and non-coding mutation signature finding.

Breed-specific germline analysis for breed validation.

To assess the breed data accuracy, we focused on the 10 pure breeds in the WES dataset with each having ≥10 animals passing QC measures specified in Figure 2.1. We identified 5,363 breed-specific variants, defined as germline base substitutions and small indels that are unique to or enriched in one of these breeds. We then performed clustering analysis using the variant allele frequency (VAF) values of these variants in the normal samples of the animals. Our analysis validated the breeds of 385 dogs and corrected 5 dogs with breed error (3 Yorkshire Terriers reassigned to 2 Shih Tzus and 1 Schnauzer; 1 Maltese each reassigned to Shih Tzu and Yorkshire Terrier) (Figure 2.2). We also reclassified 5 dogs as "unknown", as they lack VAF patterns seen in any of the 10 pure breeds (Figure 2.2).

To corroborate our strategy, we first performed the same clustering analyses using the WGS dataset after QC. As shown in Figure 2.2, all 22 dogs (3 having WGS data only), whose reported breeds belong to one of the 10 pure breeds investigated above, were confirmed. Second, we divided the WES studies into discovery and validation sets based on their sample size. We identified breed-specific germline variants for 9 pure breeds with ≥10 animals per breed in the discovery set, with which we clustered dogs from both sets. The analysis confirmed 17 of 19 animals from the validation set, and reassigned the breed for the remaining 2 dogs. These results indicate that our approach is valid.

We repeated this analysis to attempt breed prediction for 107 cases in the WES dataset with no breed data (e.g., oral melanoma cases[32]). We were able to unambiguously assign breeds to 50 dogs (14 to Golden Retriever, 10 to Cocker Spaniel, 8 each to Boxer and Rottweiler, 4 each to Shih Tzu and Maltese, and 1 each to Yorkshire Terrier and Schnauzer) (Figure 2.2).

Lastly, we clustered all 626 animals with WES and/or WGS data passing QC (Figure 1), including 85 dogs with reported breeds not among the 10 pure breeds investigated (other breeds), as well as 24 dogs of mixed breed. We hypothesize that if our approach is valid, the vast majority of these dogs would not cluster with the 10 pure breed dogs. Our analysis classified 18 mixed breed dogs as "unknown" and reassigned the remaining 6 dogs to specific pure breeds (2 Maltese, 2 Schnauzer, 1 Rottweiler and 1 Shih Tzu). For 85 dogs of other breeds, the analysis classified 82 dogs as "unknown" and reassigned the breed for the remaining 3 dogs. All other dogs shared the same breed validation, correction, prediction and reclassification indicated in Figure 2.2. The results support our hypothesis, indicating that our approach is effective.

In summary, we discovered breed-specific germline variants for 10 breeds, with which we successfully validated 385 dogs, corrected 5 dogs and predicted 50 dogs in the WES dataset

for their breed assignment, as shown in Figure 2.2. These dogs were used for downstream breed-related analyses described later.

Alteration landscape varies with tumor types but not breeds examined.

For somatic mutations (i.e., base substitutions and small indels), we focused on the WES dataset, because of the large sample size (597 tumor-normal pairs from 591 cases after QC) and high sequencing coverage (Figure 2.1), and the CDS regions, which are more accurately annotated than other genomic regions. We assembled a mutation discovery pipeline that used sequence coverage, mutant allele frequency (MAF) and paired-read strand orientation[39] to reduce mutation artifacts (see Methods). This effectively reduces C>T artifacts originated from the fixation process in FFPE samples[40], as well as G>T artifacts arisen from 8-oxoG DNA oxidative damage[39] in frozen samples of certain studies.

We compared each mutation in each tumor between our study and the original publications, including the genomic coordinate and the actual mutation, which are published only for the mammary tumor[28] and oral melanoma[32] studies. For oral melanoma, we found a median overlap rate of 67% with 5-step filtering and of 59% with further paired-read strand orientation filtering. We manually examined >20 mutations detected only by our pipeline or in the original publications, and found that all appear to be valid base changes. Thus, the difference is likely due to variations in read cleaning, germline mutation filtering and artifact filtering. For mammary tumor, the overlap rate is lower (43%) due to different mutation calling software, as 66% overlap was achieved when we used MuTect2 as in the original publication[28].

We identified genes that harbor somatic non-synonymous base substitutions or small indels, as well as genes that are amplified or deleted, in each tumor. We then examined the alteration landscape (Figure 2.3a), which consists of these altered genes that can be detected at

≥0.8 power within a tumor type or a breed based on our sample size calculation. The study reveals unique alteration features for each canine tumor type (Figure 2.3a), many of which are consistent with individual tumor type findings[28-37,41].

Mammary tumors harbor frequent PI3K pathway alteration, with 50% of the tumors having at least one member gene altered (Figure 2.3a). The PIK3CA H1047R mutation is especially common, found in 26% of the tumors. However, another PIK3CA mutation hotspot, the E542/545 site, is intriguingly missing, differing from human breast cancer[42].

Oral melanoma and osteosarcoma both harbor frequent p53 pathway alteration (61%) (Figure 2.3a). However, the actual altered genes differ, with *TP53* mutated in 50% of osteosarcomas and *MDM2* amplified in 45% of oral melanomas (Figure 2.3a). Moreover, while deletion is common in osteosarcoma, amplification is frequent in oral melanoma. Indeed, *CDKN2A* is deleted in 22% of osteosarcomas and *CDK4* is amplified in 28% of oral melanomas, resulting in frequent cell cycle gene alteration in both tumor types.

Hemangiosarcoma has a *TP53* mutation frequency of 59%, the highest among the 7 tumor types (Figure 2.3a). *PIK3CA* is another frequently mutated gene, mutated in 31% of hemangiosarcomas. The most significantly mutated genes include *FBXW7* (encoding WNT signaling molecule) in B-cell lymphoma, *SATB1* (functioning in chromatin remodeling) in T-cell lymphoma, and *CALD1* (encoding an actin and myosin binding protein) in glioma (Figure 2.3a). However, they are less recurrent than *PIK3CA* mutation in mammary tumor or *TP53* mutation in hemangiosarcoma or osteosarcoma (Figure 2.3a).

In contrast to tumor type, the canine alteration landscape appears largely breed-independent among the breeds examined (Figure 2.3a). To statistically test this, we performed Fisher exact tests on the most recurrently altered genes (*TP53*, *PIK3CA* and *CDKN2A*) and

pathways (p53, PI3K, cell cycle and RTK/RAS) to achieve a larger power. Most of these alterations do not differ significantly in their enrichment or depletion levels among different breeds within the same tumor type, unlike the tumor type comparison (Figure 2.3b). For example, mammary tumors of Maltese, Shih Tzu and Yorkshire Terrier dogs all have frequent *PIK3CA* mutation and PI3K pathway alteration (Figure 2.3). However, various tumor types of Golden Retriever dogs differ significantly in these alterations (Figure 2.3b).

Canine TMB varies mostly among tumor types but not breeds.

We investigated TMB, defined as the number of somatic base substitutions and small indels per Mb callable CDS, in each of the 597 canine tumors of the WES dataset after sequence QC (Figure 2.1). To increase the accuracy, we first identified 1,564 retrogenes and other problematic genes (see Methods) in the current canine gene annotation database. We excluded these problematic genes from TMB calculations, as they harbor significantly more mutations compared to protein-coding genes (Figure 2.6a).

Resembling human cancer[23], TMB varies among these canine tumors, ranging from 0 to 36 (Figure 2.5). However, the overall TMB is low, with a median of 0.53. Hypermutation (TMB > 10) was found in 1.17% of canine tumors, and ultra-hypermutation (TMB > 100) was not detected in any tumors. Both are rarer compared to adult human tumors, of which 2.3% are hypermutated and 0.32% are ultra-hypermutated.

TMB varies among tumor types (Figure 2.6a). Canine mammary tumor, glioma and B-cell lymphoma have lower TMB, with a median range of 0.37-0.4, and are therefore classified as TMB-low (TMB-L) (Figure 2.6a). Canine T-cell lymphoma, oral melanoma, osteosarcoma and hemangiosarcoma have significantly higher TMB, with a median range of 0.81-1.08, and are

thus classified as TMB-high (TMB-H) (Figure 2.6a). Except for lymphomas (see Discussion), these findings are confirmed with different mutation discovery strategies.

As sequence coverage influences the sensitivity of somatic mutation discovery[43], we performed TMB comparison across tumor types controlling for sequence coverage (at 30-50x, 50-100x and ≥100x). The analysis confirms our original conclusion that TMB is tumor type-dependent.

Within the same tumor type, TMB appears to be similar among breeds, except for osteosarcoma where Golden Retrievers have significantly higher TMB than Rottweilers and Greyhounds (Figure 2.7a). We thus conclude that canine TMB primarily varies with tumor types, but not breeds for those examined (Figure 2.7a).

In general, canine TMB values are significantly lower than their human adult counterparts[26], and are more comparable to their pediatric counterparts[44,45] .

Canine TMB is correlated with *TP53* but not *PIK3CA* mutation.

TP53 is mutated in 16.7% of the 597 canine tumors, and is the most frequently mutated gene (Figure 2.5). Importantly, we observed a strong association between *TP53* mutation and TMB across tumor types (Figure 2.6b). This is clearly seen in canine hemangiosarcoma and osteosarcoma, both TMB-H (Figure 2.6a), and with *TP53* mutated in 59% and 50% of their tumors respectively (Figure 2.3a). The median TMB of osteosarcomas and hemangiosarcomas with mutant *TP53* is increased to 1.31 and 1.33 respectively, from 0.7 and 0.67 for the corresponding tumors with wild type *TP53* (Figure 2.6b). A clear association between TMB and *TP53* mutation is also noted across breeds (Figure 7b). Indeed, the median TMB increases with

TP53 mutation in Golden Retriever (0.53 to 1.4), Maltese (0.34 to 0.93), Greyhound (0.7 to 1.25) and Rottweiler (0.78 to 0.96) (Figure 2.7b).

PIK3CA is the second most frequently mutated gene, mutated in 16.4% of the tumors (Figure 2.5). However, in contrast to *TP53*, we did not observe a strong association ($P < 0.05$ and median fold change >1.5) between TMB and *PIK3CA* mutation in any tumor type or breed (Figures 2.6b, and 2.7b).

To unbiasedly screen the association between individual gene mutation and TMB, we studied all 104 genes that are mutated in ≥5 tumors in a tumor type or breed (which can be detected with a power >0.9). We determined the association within each tumor type as shown in Figure 2.6a, as well as within a breed after normalizing each TMB value with its tumor type median. In both analyses, *TP53* remains the most significant gene across most tumor types and breeds. The study also identified other genes with significant association within at least one tumor type or breed, including *ASPM*, which functions in mitotic spindle, and *SPEF2* and *FSIP2*, both related to spermatogenesis. Notably, many of these genes are mutually inclusive with *TP53* in mutation.

At the pathway level, TMB is consistently associated with p53 pathway alteration. Cell cycle is another pathway with the association found. Importantly, we also observed the association of TMB with *TP53* mutation, but not with *PIK3CA* mutation, in corresponding human adult or pediatric cancers (Figure 2.6c). Moreover, in breast cancer and pediatric cancer, *TP53* has the most significant association revealed by unbiased screen. These findings support the dog-human homology.

DISCUSSION

Taking advantage of public canine data, we have investigated 684 canine cases of over 7 tumor types and over 35 breeds that are common in dogs. To our knowledge, this represents the first pan-tumor and pan-breed study for the dog. We have built pipelines for comprehensive sequence QC, breed validation and artifact reduction in somatic mutation discovery. Importantly, our work answers several important questions regarding canine tumor mutation, leading to more precise use of the canine model in cancer research (e.g., tumor type, but not breed, should be a primary factor to consider in mutation-targeting therapy trials).

Canine tumor alteration landscape, TMB and *TP53*.

Our study indicates that the canine somatic alteration landscape is tumor type-dependent, but largely breed-independent, for the tumor types and breeds examined here. Each of the 7 canine tumor types harbors distinct gene mutations and copy number alterations. The difference is especially evident among adenoma/carcinoma, sarcoma and lymphoma. Moreover, the alteration landscape is similar among different breeds within the same tumor type, but differs among different tumor types from the same breed.

Canine TMB differs among adenoma/carcinoma, sarcoma and other tumor types; however, it generally does not vary with breed for those examined. Loss of function of TP53 is a potential reason. Canine osteosarcoma, hemangiosarcoma and oral melanoma harbor higher TMB, as well as frequent *TP53* mutation or *MDM2* amplification (which promotes TP53 protein degradation). In contrast, canine mammary tumor and glioma harbor infrequent *TP53* mutation and lower TMB. Moreover, *TP53* mutation is strongly associated with TMB across tumor types and breeds, a pattern not observed for *PIK3CA*, the second most frequently mutated gene after

TP53. We propose that these observations are related to the cells of origin and tumorigenesis mechanisms, as discussed below.

Mammary adenomas or carcinomas originate from epithelial cells. The establishment of epithelial cell apical-basolateral polarity decreases cell proliferation and invasiveness, acting as a potent tumor suppressor[46-48]. PIK3CA H1047R mutation, common in canine mammary tumors, increases cell stemness[49] and decreases epithelial cell polarity, leading to accelerated cell proliferation and tumorigenesis. However, even with accelerated cell proliferation, the cell cycle checkpoint is functional and DNA damage can be repaired. This leads to slower accumulation of mutations in the genome and lower TMB.

Hemangiosarcoma, osteosarcoma and oral melanoma arise from mesenchymal cells, which lack cell polarity and cell adhesion. Loss of function of TP53, due to either *TP53* mutation or *MDM2* amplification, leads to defective cell cycle checkpoints and accelerated cell cycle. As a result, fewer DNA damages are repaired and fewer DNA replication errors are corrected[50], leading to rapid accumulation of mutations in the genome and higher TMB.

In supporting the hypothesis above, we have noted a strong association between cell cycle gene alteration and TMB. However, further experimental and computational analyses are required to test this hypothesis.

Compared to B-cell lymphomas, T-cell lymphomas harbor more somatic base substitutions that have low MAF and are more random, resulting in a higher TMB using our main mutation discovery pipeline. More studies are required to understand this.

Dog-human comparison

Our pan-cancer study reveals dog-human homology in the alteration landscape. Each canine tumor type shares many of the major pathway and gene alterations with its human counterpart. However, certain differences are also noted. Different subtype composition could be one reason, e.g., more frequent *ERBB2* amplification in human breast cancer may be due to more prevalent Her2-enriched subtype in humans than in dogs. Moreover, genes can be altered via other mechanisms not examined here, including epigenetic and expression alterations. Hence, future canine subtyping and dog-human subtype comparison, along with more comprehensive alteration investigation, may further increase the dog-human homology.

The dog-human homology is also seen in TMB. First, the order of canine tumor types sorted by TMB (i.e., mammary tumor < glioma < lymphoma, etc.) is the same as that of the corresponding human cancer types. Second, across tumor types in both species, TMB is strongly associated with *TP53* mutation and p53 pathway alteration, but not with *PIK3CA* mutation. This may be related to cells of origin in both species, as discussed previously.

Canine TMB is overall lower than the corresponding human adult TMB, but comparable to pediatric TMB. Chronological age (in clock time) may be a factor, considering the dominance of the aging mutation signature (due to deamination of cytosine) in both species. Difference in subtype composition and driver mutations is another reason, which is clearly seen in glioma where *IHD1* mutation is frequent in humans but rare in dogs. Tumor progression stage could also be a factor, as most human adult tumors[26] used in the comparison are late stage tumors (nearly all of human breast tumors are invasive and 33% harbor *TP53* mutation, while only half of canine mammary tumors are invasive and <5% harbor *TP53* mutation). Further studies are needed to address this TMB difference and underlying reasons.

MATERIALS AND METHODS

Data collection

Canine data: Canine WES and WGS data were downloaded from the Sequence Read Archive (SRA) database, including PRJNA391455 (osteosarcoma), PRJNA489159 (mammary tumor), PRJEB12081 (oral melanoma), PRJNA579792 (glioma), PRJNA552034 (hemangiosarcoma), PRJNA247493 (lymphoma and unclassified) and others. We also obtained other information from relevant publications[28-37,41,51,52].

Canine genome canFam3.1 and gene annotation canFam3 1.99 GTF were downloaded from the Ensembl database. Known canine germline base substitutions and small indels (55, 447 and 895 total) were combined from 1) Ensembl canine dbSNP, canFam3; 2) the DoGSD database[53] and 3) a published study[54].

Human data: Mutated or amplified/deleted genes in human cancers were extracted from published studies, including 996 breast cancers[55,56], 86 high grade pediatric gliomas[45,57], 511 low grade adult gliomas[56], 37 diffuse large B-cell lymphomas[56,58], 42 T-cell lymphomas[59], 59 mucosal melanomas[60], 57 pediatric osteosarcomas[44], 46 adult osteosarcomas[44] and 48 angiosarcomas[61]. Human TMB values were derived from published adult[26,56,61] and pediatric[44,45] cancer studies. Curated canonical cancer pathway gene lists were obtained from a TCGA pan cancer study[62].

Canine read mapping

Canine sequence read pairs were mapped to the canine reference genome canFam3 using BWA-aln (version 07.17)[63]. Concordantly and uniquely mapped pairs were identified based on the flag values and TAG values (with XT: AU or XT: AM), and were used to calculate the

mapping rate of each sample. Such pairs with at least one read with ≥1bp overlapping a coding sequence (CDS) region of the canFam3 1.99 GTF annotation were used to calculate the CDS-targeting rate. Mapped read coverage was obtained using GATK (version 3.8.1)[64] DepthOfCoverage, with minimum mapping quality 10 and base quality 10. Sequencing randomness was assessed with the root mean square error (RMSE) between the actual read coverage distribution in target regions and the theoretical Poisson distribution, with λ set to the mean coverage of each sample.

Germline base substitution and small indel calling

Germline base substitutions and small indels were first called by applying GATK 3.8.1 HaplotypeCaller to the realigned bam files of individual tumor or normal samples (see Somatic mutation calling) with parameters of dontUseSoftClippedBases -stand_call_conf 20.0. Variants were then filtered with GATK VariantFiltration with parameters of FS > 30.0 and QD < 2.0. Furthermore, variants with total read coverage <10 were excluded. Only germline base substitutions and small indels that were detected in both tumor and normal samples of at least one case were used for further analyses below.

Tumor-normal sample pairing accuracy

For a given study, let T and N be the total number of germline base substitutions and small indels in a tumor and normal sample respectively, and S be the total number of those shared between T and N. The shared fraction between the tumor sample of case i (T_i) and the normal sample of case j (N_j) is given by $F_{i,j} = \frac{S_{i,j}}{\min(T_i, N_j)}$. When $i = j$, "self" fraction ($Self_i$) is obtained. When $i \neq j$, "nonself" fraction is obtained. For a given case i, its best non-self match

is identified by $Best\ nonself_i = \max(F_{i,j}, F_{j,i}),\ \forall\ j\ \in [1,n]\ and\ j \neq i$, where n is the total case number of the study. Thus, $Self_i - Best\ nonself_i$ is negative if and only if either the tumor or the normal sample of case i has a better match to a sample of a different case, which indicates tumor-normal sample pairing error for case i.

Breed validation and prediction (The whole algorithm is developed by Dr. Burair A. Alsaihati and I performed the task)

Variant allele frequency (VAF) of each of the 157,628 germline base substitution and small indel variants identified as previously described was calculated in each normal sample by $VAF = \frac{variant\ allele\ reads}{total\ reads}$. Each variant was classified as reference (VAF <0.2), non-reference (VAF ≥0.2), or not determined (ND) if total read coverage <10. Variants with ND in >20% samples were excluded, due to their poor coverage.

Only breeds with ≥10 dogs were used for breed-specific base substitution and small indel variant discovery. A variant is considered "breed-specific" if it is either breed-unique or breed-enriched. To be considered breed-unique, a variant must be: 1) non-reference in ≥5 dogs of the breed; 2) non-reference in ≥40% dogs of the breed; and 3) reference in all dogs with ≥10 read coverage of the remaining breeds. Breed-enriched variants were identified with Fisher exact tests between any two breeds using the reference and non-reference sample counts. To be considered breed-enriched, a variant must be 1) enriched in breed A and 2) not enriched in any breed that is not A, against every other breed at $P \leq 0.1$.

Identified breed-specific variants were used for breed validation or prediction. First, to reduce noise, a sample to be validated or predicted should have >80% of the combined breed-specific variant sites with ≥10 read coverage for VAF calculation. For those sites with <10 read

coverage, a random VAF value was assigned to each site. Breed validation was then achieved via standard hierarchical clustering with VAF values of breed-specific variants in the normal sample of each dog, as illustrated in Figure 2.2.

Somatic mutation calling

Somatic mutations include somatic base substitutions and small indels. Canine read-pair mapping, selection, duplicate-marking with Picard (version 2.16.0), realignment with GATK (version 3.8.1)[64] and other processing were performed as previously described[51,52].

MuTect (version 1.1.7)[65] was then used to detect somatic base substitutions, with a minimum base quality of 30, and filtering known canine germline base substitutions from previously described sources. Additional filtering steps were used to reduce artifacts. First, the results were subjected to a 5-step filtering process as described[32], which considers both total read coverage and mutant allele frequency (MAF). This effectively reduces artifacts with very low MAF including: 1) C>T artifacts originated from the fixation process in FFPE samples[40]; and 2) G>T artifacts arisen from 8-oxoG DNA oxidative damage[39] in frozen samples from specific studies. Second, the results were further filtered based on paired-read strand orientation bias[39]. Specifically, F1R2, where Illumina read 1 and read 2 respectively align to the forward strand (F1) and the reverse strand (R2) of the reference genome, and F2R1, the opposite of F1R2, were determined for each mutation. Then, Fisher exact tests were applied with F1R2 and F2R1 reference and mutant reads to identify and exclude mutations with significant orientation bias ($P \leq 0.05$). Furthermore, cutoffs of ≥4 in total and ≥5% being mutant reads for both F1R2 and F2R1 reads were applied for: 1) G>T and C>A mutations in Broad frozen tumors; and 2) for

C>T and G>A mutations in FFPE tumors of all studies, to further reduce paired-read strand orientation bias.

Somatic indels in CDS regions were discovered with Strelka[66] as described[32]. As expected, small indels account for only 5% of the mutations. Mutation annotation was performed with Annovar (version 2017Jul16) [67], using the canine annotation file described previously.

Tumor mutational burden (TMB)

TMB values were calculated by $TMB = \frac{total\ somatic\ base\ substitutions\ and\ small\ indels\ in\ CDS}{total\ callable\ bases\ in\ millions\ in\ CDS}$ for each case. Callable bases were identified with MuTect with the minimum base quality score set to 30.

Validation with other somatic mutation calling tools

To validate somatic mutation findings, other tools GATK4 MuTect2 (version 4.1.6)[64], Varscan2 (version 2.4.2)[68] and LoFreq (version 2.1.2)[69] were used as described[30]. Briefly, MuTect2 was run in the panel-of-normals (PON) mode, using paired normal samples that passed QC (n=591) to create the PON file. Canine germline mutation filtering and other parameters were used as described[30]. SomaticSeq (version 3.4.1)[70] was used to find consensus mutation callings among MuTect2, Varscan2 and LoFreq.

Canine retrogene and other problematic gene identification

Problematic genes in the canFam3 1.99 GTF annotation file were identified after excluding mitochondrial genes. Problematic genes are defined as genes that: 1) have only an

Ensembl ID and lack a gene symbol, name or other biologically meaningful description; and 2) consist of a single exon. A problematic gene is classified as a retrogene if its single exon arises from fusion of partial or complete exons of a protein-coding gene.

Somatic copy number alteration (CNA) identification (Dr. Tianfang Wang performed the copy number alteration)

VarScan (version 2.4.2)[71] was first applied on WES data of matched tumor and normal sample pairs. Then, CBS[72] (DNAcopy R package) was used to segment CDS regions, with 0.01 the significance level set to 0.01 for change point identification, and 10,000 permutations performed for p-value calculation. Segments with $\left|\log_2(\frac{T}{N})\right| > 1$ (T: tumor; N: normal) were considered CNAs and their overlapping genes were identified. Further selection was made by finding genes with CNAs also detected by another software, SEG[73]. SEG was run as previously described[41,74,75].

Significant alteration identification and cross-species comparison

Both MAF and sample recurrence were used to identify significant mutations and mutant genes (Figure 3). Fisher exact tests were performed to first identify individual mutations that have significantly higher MAF compared to the remaining mutations within a tumor. Among the identified mutations, two analyses were then performed. First, Fisher exact tests were used to find individual mutations that are significantly recurrent among the samples within a tumor type or breed. Such mutations could potentially be gain-of-function and genes harboring them may be oncogenes, e.g., PIK3CA H1047R. Second, to discover potential tumor suppressors, which

harbor loss-of-function mutations that could occur at different loci among tumors (e.g., *TP53*), genes that harbor significant mutations in any tumor were identified. Then, using these genes as the background, Fisher exact tests were performed to identify mutant genes that are significantly recurrent across the samples within a tumor type or breed.

Amplified/deleted genes within a tumor were identified via Z-tests at $q \leq 0.01$, as described previously. Fisher exact tests were then used to identify those amplified/deleted genes that are significantly recurrent among samples within a tumor type or breed.

Multiple testing correction with the Benjamini and Hochberg strategy[76] was applied on each Fisher exact test described above.

Enrichment of gene and pathway alterations in a tumor type or breed

Enrichment scores were determined by $-log_{10}(q)$ and with positive values indicating enrichment and negative values indicating depletion. Each q value was obtained from a Fisher exact test that compares the ratio of altered versus wild-type tumors of a tumor type or a breed to that of the remaining tumor types or breeds and after applying multiple test correction.

Association between gene mutation and TMB

For canine tumor type specific association, genes that are mutated in ≥ 5 tumors in a specific tumor type (each of the 7 canine tumor types, TMB-L and TMB-H) were selected for Wilcoxon tests, using TMB without normalization. For breed-specific association, genes that are mutated in ≥ 5 tumors within a specific breed were chosen, and Wilcoxon tests were conducted with normalized TMB values, given by $normalized\ TMB = \frac{TMB}{Tumor\ type\ TMB\ median}$. S1-high

tumors (defined as tumors with ≥15 S1 mutations) and unclassified tumors were excluded from all analyses. Separate association analysis was conducted in Golden Retriever S1-high tumors only. Human association studies were performed for genes that are mutated in ≥ 20 tumors overall and ≥ 5 tumors in a specific cancer type. TMB normalization was performed for cross cancer type association determination.

FIGURES

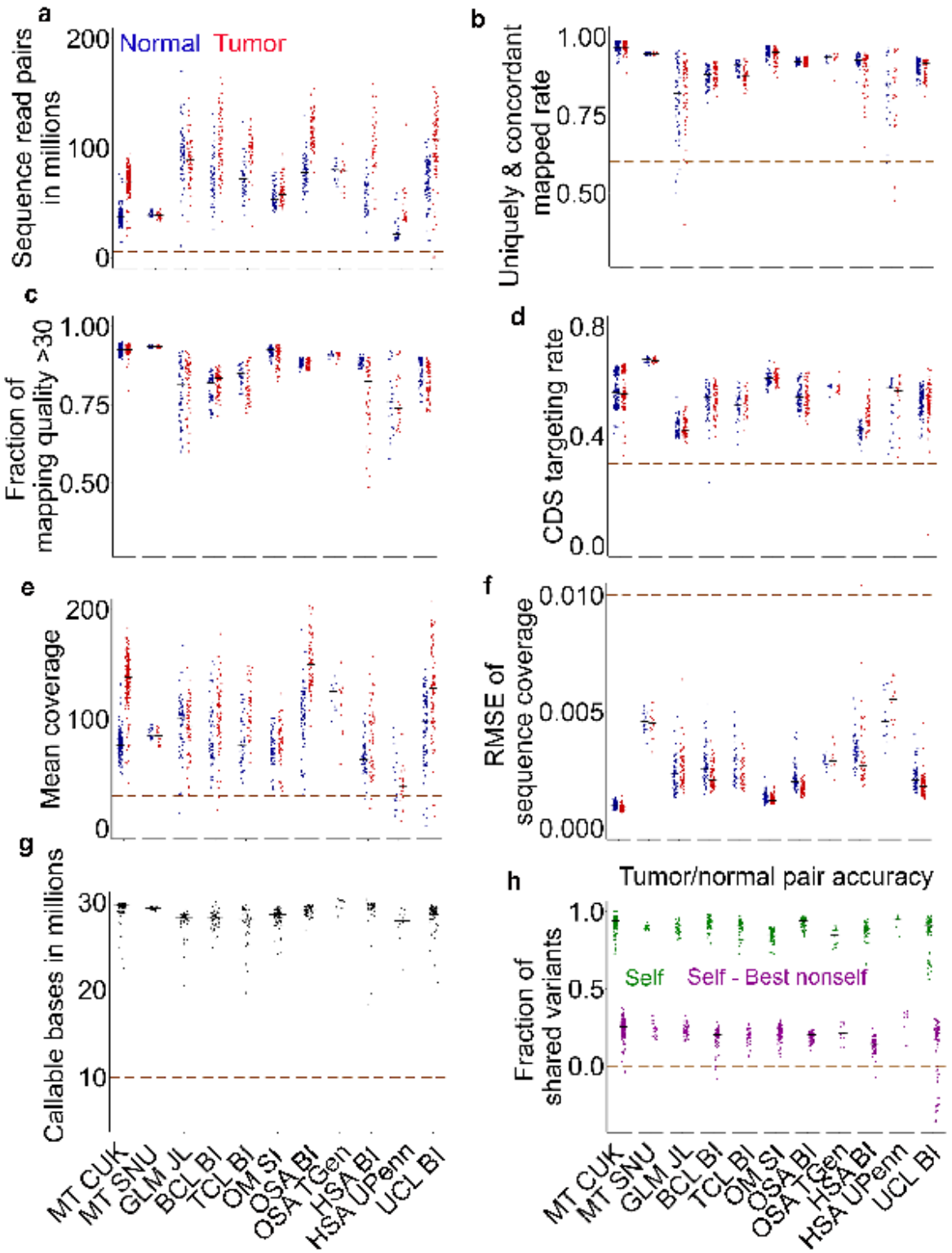

a
Sequence read pairs in millions
200
100
0
Normal Tumor
b
Uniquely & concordant mapped rate
1.00
0.75
0.50
c
Fraction of mapping quality >30
1.00
0.75
0.50
d
CDS targeting rate
0.8
0.4
0.0
e
Mean coverage
200
100
0
f
RMSE of sequence coverage
0.010
0.005
0.000
g
Callable bases in millions
30
20
10
h
Tumor/normal pair accuracy
Fraction of shared variants
1.0
0.5
0.0
Self
Self - Best nonself
MT CUK
MT SNU
GLM JL
BCL BI
TCL BI
OM SI
OSA BI
OSA TGen
HSA BI
HSA UPenn
UCL BI

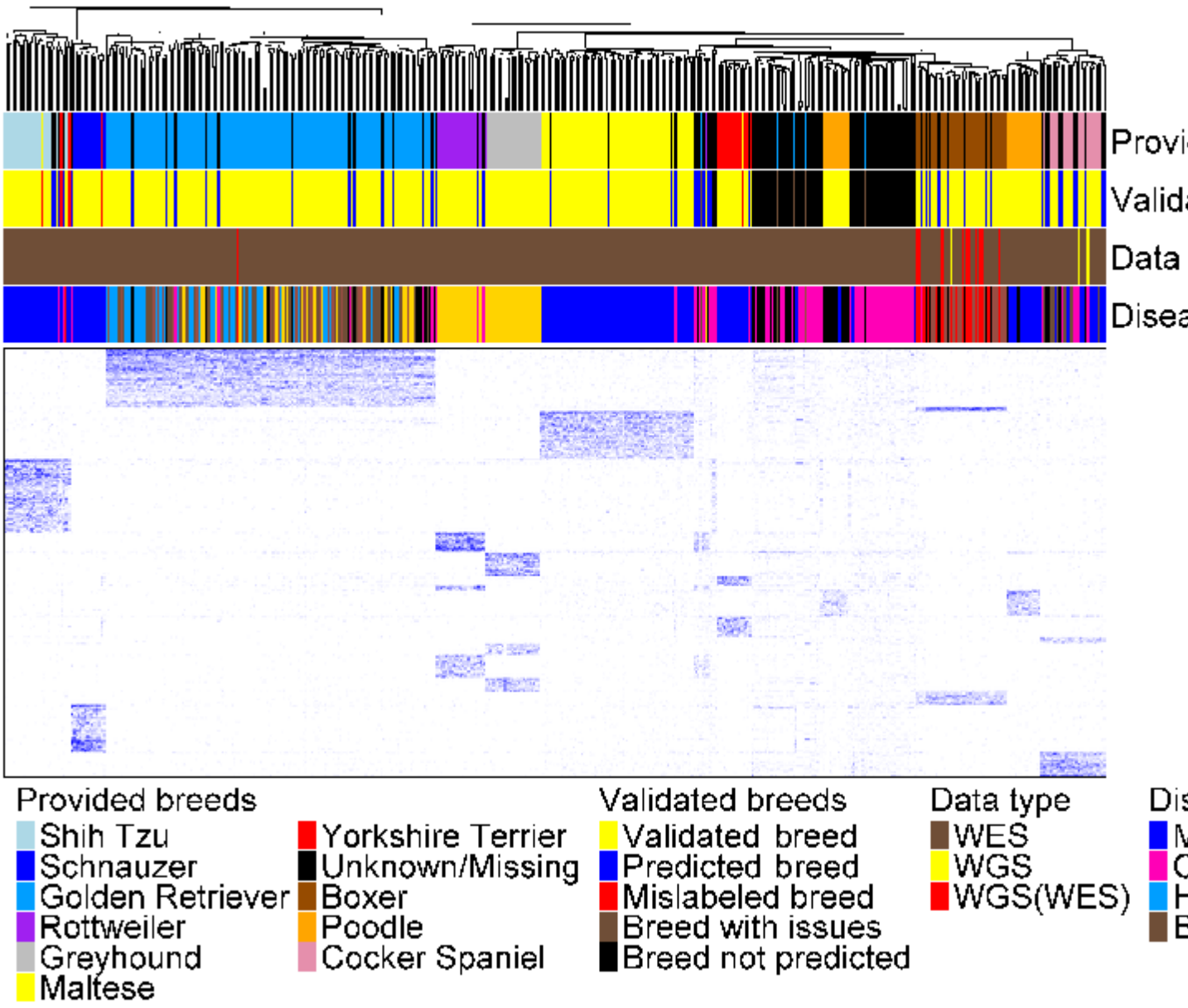
Figure 2.1. We performed a rigorous quality control (QC) of whole-exome sequencing (WES) data published for 654 canine cases.

a. Distributions of total read pairs per sample of the tumor and normal sample sets of each study. Each dot represents a sample, and the median is indicated by a black line. The dashed line specifies the QC cutoff. Each study is represented by the tumor type and the institute name. MT: mammary tumor; GLM: glioma; BCL: B-cell lymphoma; TCL: T-cell lymphoma; OM: oral melanoma; OSA: osteosarcoma; HSA: hemangiosarcoma; UCL: unclassified. CUK: Catholic University of Korea; SNU: Seoul National University; JL: Jackson Laboratory; SI: Sanger Institute; BI: Broad Institute; UPenn: University of Pennsylvania.

b-f. Distributions of per sample rate of read pairs that aligned concordantly and uniquely to the canFam3 reference genome (b), fraction of reads with mapping quality of ≥30 (c), CDS-targeting rate (the fraction of read pairs that align concordantly and uniquely to the canFam3 CDS regions) (d), mean read coverage in CDS regions (e) and root-mean-square error (RMSE) between the actual distribution and theoretical distribution (based on the Poisson distribution) of sequence coverage in CDS regions (f). Each plot is presented as described in a**.**

g. Distributions of the total number of callable bases per case, determined by MuTect.

h. Tumor-normal pairing accuracy. "Self" (in green) is the fraction of germline variants shared between the normal and tumor samples of a dog. "Best nonself" is the fraction of germline variants shared between a normal or tumor sample of one dog and its best matched sample from another dog. "Self – Best nonself" (in purple) indicates the difference, and a negative difference points to incorrect tumor-normal pairing.


Provi
Valida
Data
Disea
Provided breeds
Shih Tzu
Schnauzer
Golden Retriever
Rottweiler
Greyhound
Maltese
Yorkshire Terrier
Unknown/Missing
Boxer
Poodle
Cocker Spaniel
Validated breeds
Validated breed
Predicted breed
Mislabeled breed
Breed with issues
Breed not predicted
Data type
WES
WGS
WGS(WES)

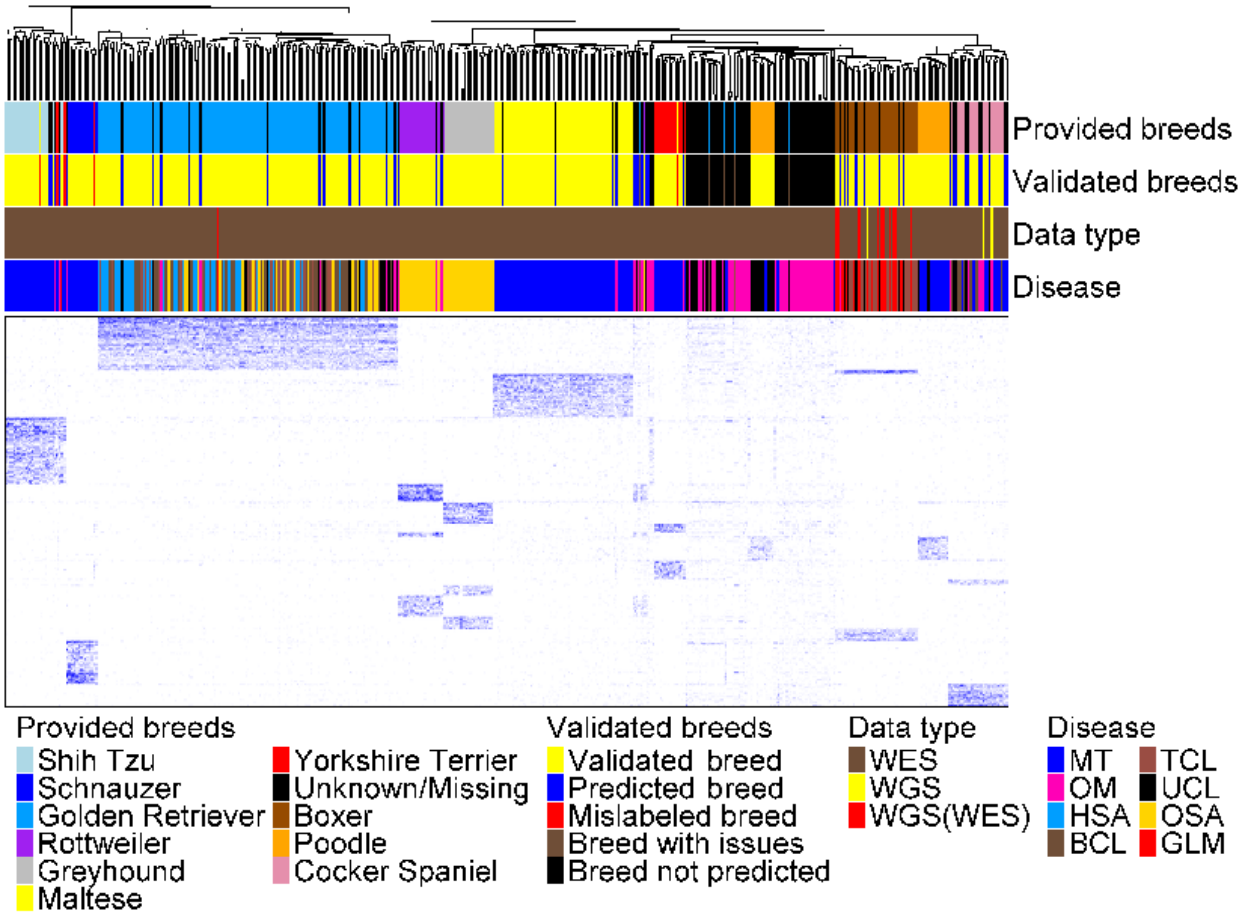


Figure 2.2. We conducted breed validation and prediction using breed-specific germline base substitutions and small indels discovered. The heatmap shows the clustering of 505 animals (398 dogs with breeds provided and 107 dogs with no breeds provided), using variant allele frequency (VAF) values of the 5,363 breed-specific germline base substitution and small indel variants in their normal samples. These variants were discovered with the WES dataset (see Methods). The WGS dataset was used for validation as specified in the "Data Type" bar, where "WGS(WES)" indicates that a dog has both WGS and WES data but only WGS data were used in the clustering analysis. The "Provided breeds" bar and the "Disease" bar respectively indicate the breed and tumor type of each dog provided by the source studies. The "Validated breeds" bar denotes the analysis outcome as specified, with "Unknown" representing dogs whose

provided breeds could not be validated or corrected, due to the lack of any specific VAF clustering patterns of the 10 pure breeds investigated. **(The breed prediction and validation algorithm is developed by Dr. Burair A. Alsaihati, and I performed the task)**

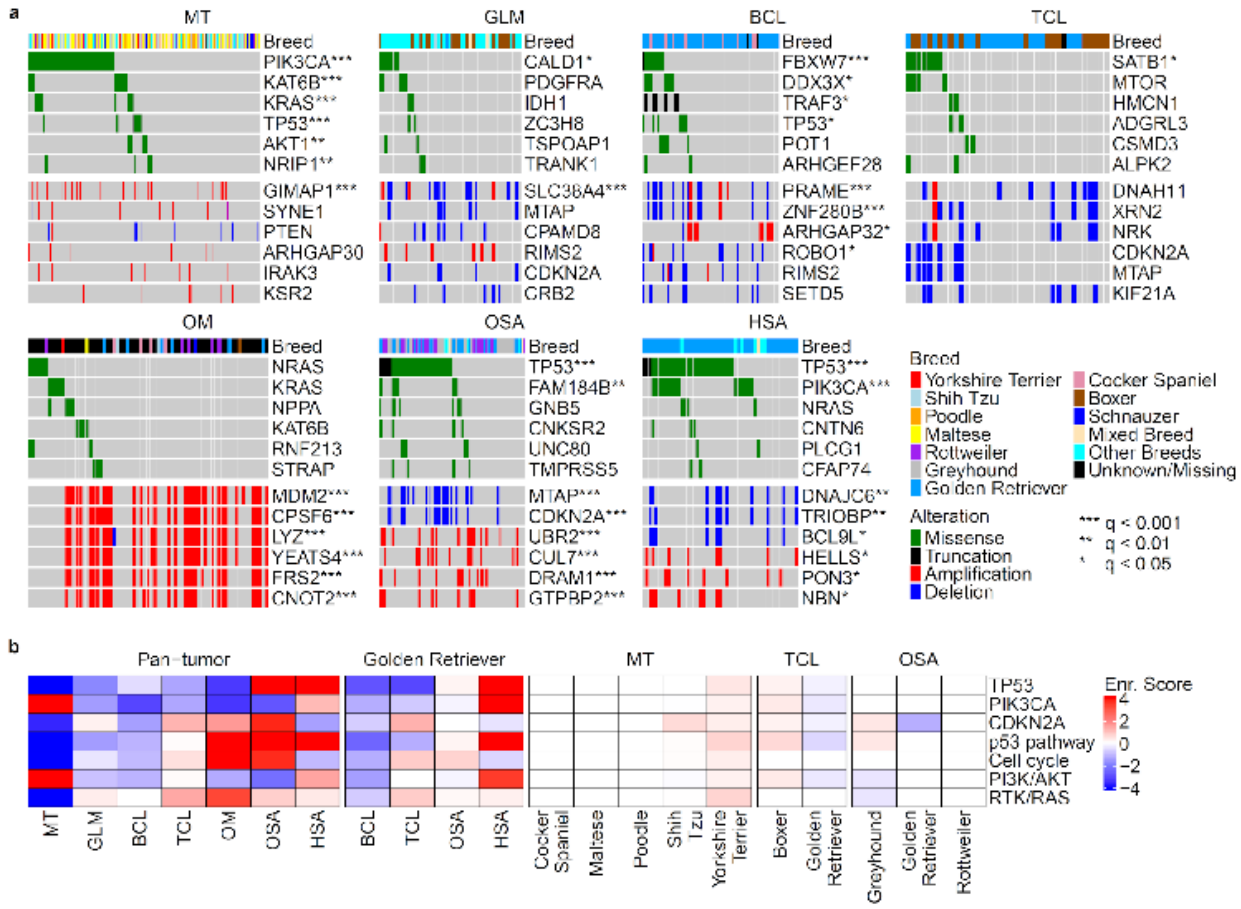


Figure 2.3. Canine tumor alteration landscape, consisting of genes recurrently mutated and/or amplified/deleted, varies with tumor types but not with breeds in general.

Oncoprints indicate top 6 most recurrently altered genes with nonsynonymous somatic base substitutions or small indels (top), or copy number alterations (CNAs) (bottom), in CDS regions in each tumor type indicated. Significant recurrence, identified by Fisher exact tests, are denoted by "*" as shown. The breed of each animal is specified in the breed bar.

a. Enrichment scores of the most recurrently altered genes and pathways, obtained via Fisher exact test q-values (see Method), in each tumor type of all breeds (left) and of Golden Retriever (middle), as well as in each breed with ≥10 dogs within a tumor type (right). **(The**

copy number alternation (CNA) were identified by Dr. Tianfang Wang, and I incorporated her data into this analysis).

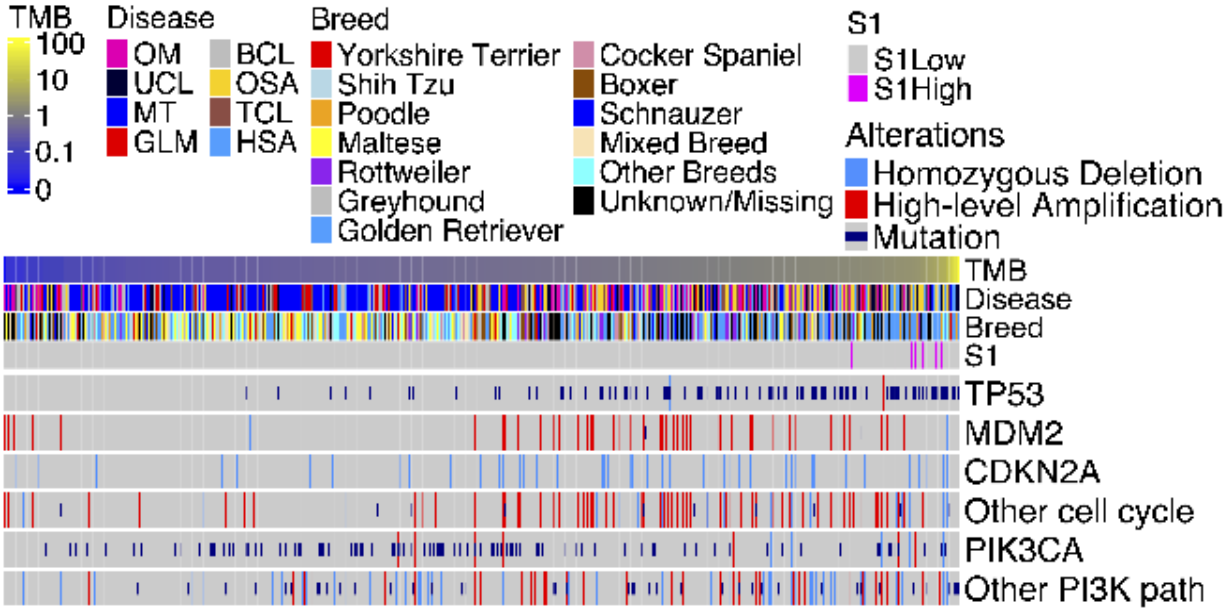


Figure 2.4. We investigated TMB and common alterations in each of the 597 tumors of over 7 tumor types and over 35 breeds. The tumors in the oncoprint were ordered from left to right by lowest to highest TMB. Seven tumor types as indicated in Figure 2.3 and unknown tumor types (UCL; see Figure 2.1) are included. Breeds shown include those validated, corrected, predicted, or unknown (with issue or not predicted) as shown in Figure 2.2, as well as other breeds, which are not validated due to small sample size, and mixed breeds. Top recurrent gene and pathway alterations are shown. **(I provided the mutation data, and Dr. Joshua Watson created the heatmap)**

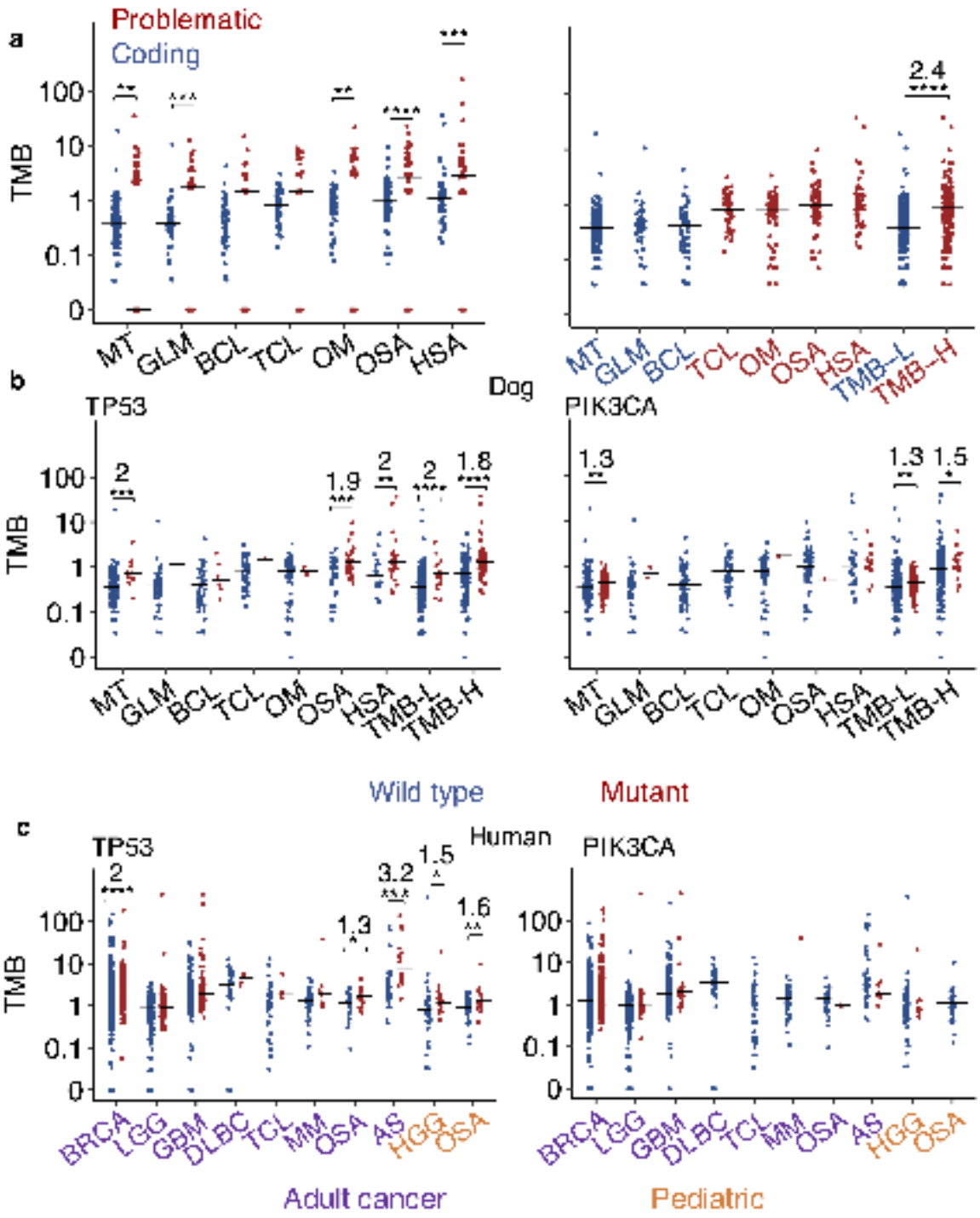


Figure 2.5. TMB varies among tumor types and is correlated with *TP53* mutation .

a. TMB distributions of each canine tumor type, ordered left to right from lowest to highest median values. The left plot shows that problematic genes (see Methods) have significantly

higher TMB than other genes, and thus were excluded from further analyses. The right plot indicates that canine tumors are classified into TMB-low (TMB-L) and -high (TMB-H). Wilcoxon tests were conducted to examine the TMB difference between two groups indicated, with *, **, *** and **** representing $p<0.05$, <0.01, <0.001 and <0.0001 respectively. For significant comparisons, the fold change in median TMB is also indicated.

b & c. TMB distributions of cases with wild type (blue) or mutant (red) *TP53* or *PIK3CA* within each canine (b) or human (c) tumor type. For tumor types with both wild type and mutant groups having ≥5 tumors, Wilcoxon tests were conducted to determine the significance of the association between TMB and *TP53* or *PI3KCA*, with p-values and fold changes shown as described in a. LGG: low grade glioma; GBM: glioblastoma; HGG: high grade glioma. **(Dr. Joshua Watson collected and analyzed the human mutation data).**

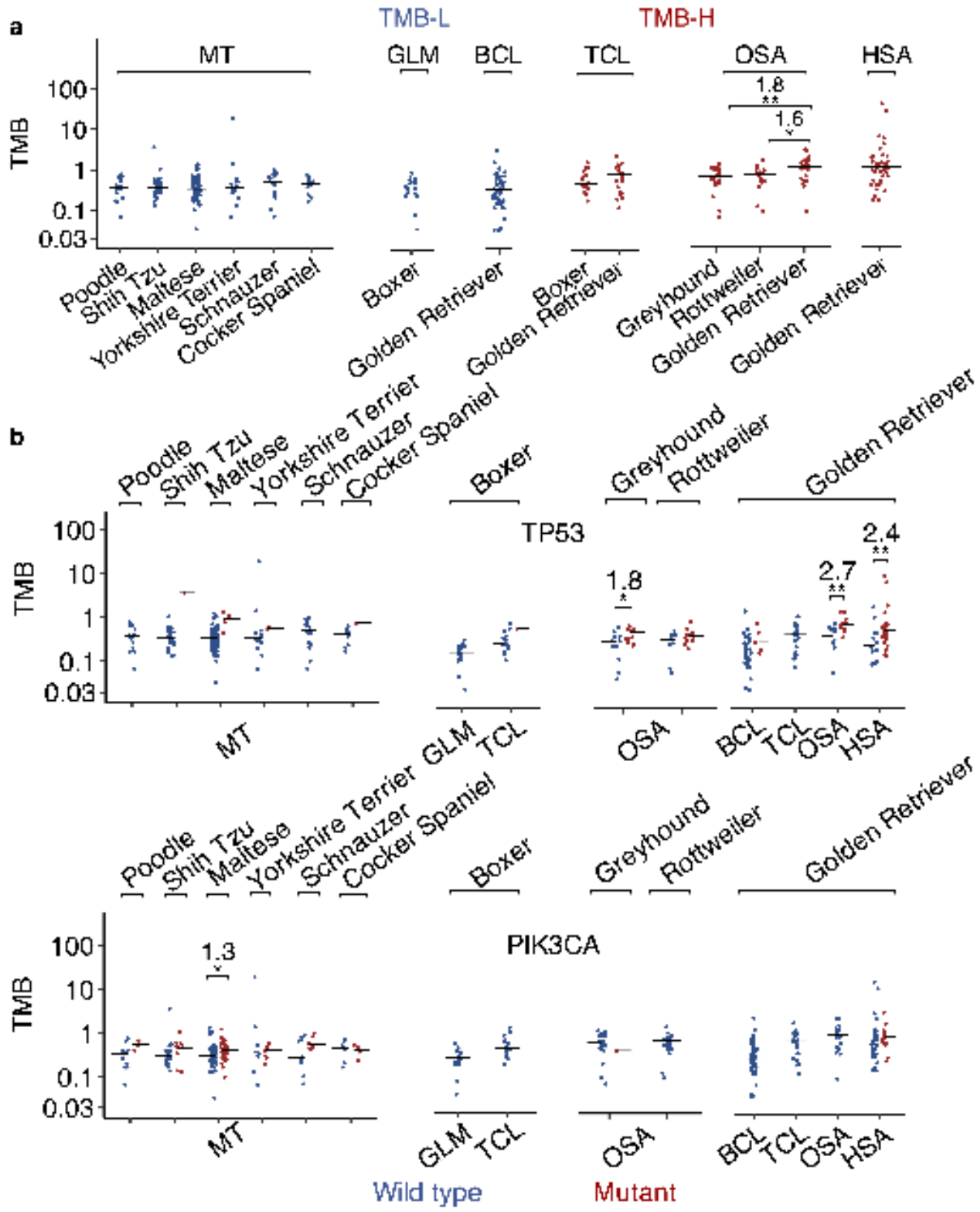


Figure 2.6. TMB is largely independent of breeds (I provided the mutation data, and Dr. Joshua Watson created the scatter plot) .

a. TMB distributions of cases grouped by tumor type and then breed. Only groups with ≥10 tumors are shown, plotted as described in Figure 2.6a.

b. TMB distributions of tumors grouped by breed, tumor type, and finally *TP53* (top) or *PIK3CA* (bottom) mutation status. Only groups with *TP53* (or *PIK3CA*) wild type and mutant combined tumors of ≥10 are shown.

CHAPTER 3

AN EFFICIENT PIPELINE TO IDENTIFY CANINE CANCER-SOMATIC MUTATIONS USING TUMOR-ONLY RNA-SEQ DATA[1]

[1]Ho, Kun-Lin., Zhao, Shaying. To be submitted to *Scientific Reports*

ABSTRACT

The dog is an important spontaneous cancer model. Expressed mutation identification is needed for studies such as tumor-specific neoantigen discovery. It can be achieved using RNA-seq data, which however is a challenge, especially for the dog. To address this, we developed a pipeline that uses variant allele frequency distribution pattens and massive human somatic mutations known for germline - somatic mutation discrimination. We applied our pipeline to RNA-seq data of ~400 canine tumor samples sequenced by various groups. The pipeline known cancer genes and mutations, as well as mutation landscape, and mutation burden that largely agree with those found by paired-tumor and normal whole exome sequencing. The results indicate our pipeline is effective.

INTRODUCTION

Somatic mutations present in the cancer genome can arise from various mechanisms, including DNA replication errors, exposure to exogenous or endogenous mutagens, enzymatic modifications of DNA, or deficiencies in DNA repair processes, ultimately leading to alterations in gene function[6,21]. The identification of these mutations within tumor samples holds immense clinical significance. Specifically, when these mutations accumulate in "cancer-driver" genes, they have the potential to trigger the development of cancer, promote cancer progression, or confer resistance to therapeutic interventions[8]. Therefore, the detection of these mutations is imperative for the development of effective targeted cancer therapies.

More recently, somatic mutations within a tumor type have been served to estimate the tumor mutational burden (TMB), and TMB can be served as a promising proxy for neoantigen load[77]. TMB is defined as the count of non-synonymous mutations detected in tumor DNA and has been established as an independent marker for assessing patient responses to immune checkpoint inhibitor therapy (ICI), as well as for predicting patient survival, encompassing both treated and treatment-naive individuals[78].

Traditionally, the detection of somatic mutations involves the utilization of whole exome or genome sequencing of paired tumor-normal DNA samples[79,80]. Matched-normal DNA samples are indispensable for distinguishing between somatic mutations exclusive to the tumor sample and germline variants shared across all cells within an individual. Recent advancements in the human cancer research of somatic mutation identification have led to the development of 'tumor-only' pipelines in WES and RNA-seq to detect somatic mutations[80,81]. The identification of expressed mutations is critical for advancing cancer immunotherapy[82,83]. Detecting mutations

in RNA-seq data can facilitate the identification of expressed mutations, a crucial aspect in identifying neoantigens.

Conversely, while dogs serve as a valuable animal model in cancer research for humans, their resources are comparatively limited. Identifying somatic mutations in RNA-seq data can be challenging, especially in the absence of matched normal samples, and this challenge is amplified in the context of canine studies. Furthermore, the identification of a matched normal RNA sample is hindered by the distinct expression patterns and profiles between normal and tumor samples. A recent study has devised a method to identify somatic mutations in whole exome sequencing without the need for matched normal DNA samples in the canine, but this study primarily focused on a limited set of genes and did not systematically identify somatic mutations in other genes[84].

In this study, we present a novel pipeline designed for the detection of cancer-associated somatic mutations in canine using "tumor-only" RNA-seq samples, named "TOSMIC," which stands for Tumor-Only Somatic Mutation Identification in Canine. Our approach involves the utilization of variant allele frequency (VAF), human somatic mutation databases, and machine learning techniques to facilitate the identification of cancer-associated somatic mutations. This method is particularly valuable for laboratories generating their own RNA-seq data from tumor samples, enabling them to assess whether the genes of interest harbor disease-driving onco-mutations.

RESULTS

RNA-seq QC:

The RNA-seq dataset consists of 376 tumor samples of 376 animals from 9 BioProjects (Table S1). These include 158 mammary tumors (MTs)[28,85], 71 oral melanomas (OMs), 52 osteosarcomas (OSAs)[34], 39 gliomas (GLMs) , 23 hemangiosarcomas (HSAs)[36,86], and 11 bladder tumors (BLAs), 11 urothelial bladder tumors(UBs) and 11 prostate tumors (PROs) [87,88] (Table S1). We also collected the patient (e.g., age, sex and breed), histological subtype and clinical (e.g., tumor invasiveness and patient alive/death status) data if published for each Bioporject at the SRA database or in the literature. The RNA-seq data were generated by different groups. We hence performed a rigorous QC to ensure that data chosen from each study meet a set of quality standards before any integrative analysis. For the sequencing amount, all datasets have a median between 20 million (M) and 80 M read pairs per sample (Fig. 3.1a ; Table S1). All samples have >5 M read pairs and no samples were excluded. We then examined the mapping of read pairs to the canine reference genome[34]. All studies have >60% read pairs in nearly every sample uniquely and concordantly mapped to the genome, with the median close to or larger than 80% (Fig. 3.1b). We excluded 2 samples with mapping rates <60% (Table S1). Furthermore, all but two samples have >60% reads with a mapping quality score of >30 (Fig 3.1c), and we excluded the two samples with low mapping quality (Table S1). For the target rate, all but 5 samples have, on average, >50% read pairs that are uniquely and concordantly mapped to the CDS regions (Fig. 3.1d). We also examined the overall distribution of the expression levels of all ~200,000 protein-coding genes in each sample, and find one significant any clear outliners (Fig. 3.1e).

Somatic mutation discovery pipeline

We compared the RNA-seq reads to the canine reference genome, and detected sequence changes of 14,939,660 total (211,453 unique) in 367 samples that pass the QC measures. These changes however consist of somatic mutations, germline mutations, and artifacts from sequencing and other experimental errors. To narrow down somatic mutations, we developed a pipeline that maximally utilizes existing resources for both dogs and humans (Fig. 3.2).

We first used already known canine germline mutations, ~9 M in total, to filter out germline mutations. This reduces the sequence changes to 269,771 total (88,824 unique), 98% reduction (58% unique) (Fig. 3.2). We then used the same strategy as described[32,89] that combines read coverage, variant allele frequency (VAF), This further reduce the sequence changes to 98,444 total (29,070 unique), 64 % reduction (67 % unique) (Fig. 3.2). We next use known human somatic mutations (~4.3 M) to divide the sequence changes to two groups. One group consists of 7,433 total (1,883 unique) changes that match human somatic mutations (Fig. 3.2), referred to as "Human" hereafter. The remaining changes, 91,011 total (27,187 unique), consists of the other group (Fig. 3.2), referred to as "Remained". We then used the VAFs of the sequence changes to distinguish between somatic and germline mutations. The rationale is that VAFs of homozygous and heterozygous germline mutations in most samples should cluster near 1.0 and 0.5 respectively, whereas VAFs of somatic mutations should randomly distribute. To more accurately apply this concept, we divided the sequence changes in the Remain group into those detected in $\geq$ 10 samples, 5-9 samples, or <5 samples (Fig. 3.2). For the first two subgroups, we classified changes with VAF in 0.4-0.6 $\geq$ 50% or 40% samples respectively as germline mutations (Fig. 3.2), based on the VAF distribution of known canine germline

mutations (Fig. S3.1a-b). For the third subgroup (<5 samples), we classified mutations with VAFs in 0.4-0.6 as germline (Fig. 3.2). After filtering these germline mutations, we combined all three subgroups and identified those mutations with $VAF \geq 0.9$ in $\geq 20\%$ samples as germline mutations (Fig. 3.2), again based on the known germline mutation distribution (Fig. S3.2c). These analyses reduce sequence change by 67 % (57 % unique), 51 % (52 % unique), and 40 % (36 % unique) for the three subgroups respectively (Fig. 3.2). The Human group contains fewer sequence changes, and we only filtered those with $VAF \geq 0.9$ in $\geq 20\%$ samples. This reduces 7,433 total (1,883 unique) sequence changes to 7,048 total (1,815 unique), only 5% reduction (Fig. 3.2). We lastly used mutation recurrency to filter germline mutations, assuming that germline mutations are more recurrent than somatic mutations. Based on the most recurrent somatic mutations (e.g., PIK3CA H1047R in canine mammary tumor) from our previous work[89], we classified sequence changes that were detected in $\geq 30\%$ samples within a specific tumor type (Fig. S2d) or $\geq 15\%$ samples across all tumor types (Fig. S2e) as germline mutation. Furthermore, we use 5363 breed-specific variants identified as describe [32,89] to filter germline mutations, This analysis reduces sequence changes to 53,072 total (36,667 unique) (Fig. 3.2 and Table S2). Finally, we apply a machine learning model to do final filtering, leading to 8402 total (3704 unique) as somatic mutation candidates.

Compare mutation identified in RNA-seq with mutations identified in WES-normal tumor pair.

WES data has been made available for both tumor and normal samples within the mammary cohort[28]. We leveraged this dataset to assess the performance of our pipeline before

integrating the machine learning model. Our analysis involved a comparison of somatic mutations identified through WES analysis[90] and those detected by our RNA-seq pipeline in each dog, as presented in Figure 3.2. We found that a variable proportion of mutations, ranging from 0.8% to 74% with a median of 11%, were exclusively identified in the WES analysis (Figure 3.3a). Conversely, only a small fraction of mutations, ranging from 0% to 20% with a median of 2%, were identified by both approaches, while our RNA-seq pipeline unveiled a significantly higher number of mutations (Figure 3.3c). Considering that gene expression levels influence mutation detection, we investigated the gene expression profiles of these mutations. Interestingly, we observed that mutations exclusively identified in the WES analysis exhibited lower expression levels, while those shared between the two methods displayed higher expression levels. To obtain a deeper understanding of the mutations exclusively identified in our RNA-seq pipeline, we individually examined each of them and validated their presence using WES data from matched normal-tumor pairs (See Methods). Our findings are summarized in Figures 3.3d and 3.3e. Analyzing the variant allele frequency (VAF) distribution within each category revealed that germline mutations exhibited a higher VAF (median ~0.45), whereas Wild-Type (WT) and RNA-editing mutations displayed a lower VAF (median ~0.25). In summary, except for a small portion of mutations classified as somatic (0.7%), the majority of mutations exclusively identified in our RNA-seq pipeline non-somatic mutations, with 26% being germline, 34% WT, and 39% RNA-editing mutations. This outcome demonstrated the reliability of mutations identified through WES of normal-tumor pairs but highlighting the constraints and challenges associated with mutation identification in RNA-seq data, which may be influenced by potential experimental artifacts or the absence of matched normal samples.

Machine learning filtering

Our manual comparison of WES and RNA-seq mutations provided valuable insights into mutation identification in RNA-seq. However, manual filtering of these mutations is not practical without matching tumor-normal WES samples. To address this challenge, we harnessed the power of a machine learning model to enhance the filtration of potential artifacts resulting from differences in sequencing techniques.

To effectively train the machine learning model, we leveraged various classes of mutations from different sources, including mutations validated through WES analysis (refer to Fig 3.3) and the human somatic mutation database (see also the "Methods" section). Our aim was to evaluate the model's effectiveness. Therefore, we selected mutations that passed the variant allele frequency (VAF) filtering from the "remain" category (refer to Fig 3.2) as our validation dataset.

The effectiveness of our VAF filtering step may be influenced by the frequency of mutations occurring in different samples. Additionally, the effectiveness of our VAF filtering step is expected to improve as mutations occur in more samples due to the utilization of a more precise VAF distribution cut-off. The machine learning model captured this trend, with 37% of the mutations predicted as somatic mutations when mutations were found in ≥10 samples. This percentage decreased to 23% for mutations found in 5-9 samples and further decreased for mutations found in <5 samples. Additionally, non-somatic mutations, such as wild type (WT) and RNA-editing mutations, exhibited a decreasing frequency (Fig 3.4b). The VAF distributions of these mutations were consistent with the patterns observed in the training data, with higher

VAF in germline mutations and low VAF in mutations classified as WT and RNA-editing (Fig 3.4c).

To further evaluate the effectiveness of our mutation identification pipeline with the machine learning model, we utilized variants from normal samples in our mammary tumor dataset. These variants from normal samples were presumed to have a lower likelihood of being cancer-associated somatic mutations and, consequently, were expected to be less frequently included in the final selection of somatic mutation candidates. As demonstrated in Figure 3.4d, the overlap ratio between variants from RNA-seq normal samples and variants in the total tumor samples decreased from 3% to 0.5% throughout the whole pipeline filtering steps. This reduction suggests that our RNA-seq mutation identification method effectively eliminates non-somatic mutations.

Variant allele frequency (VAF) distribution

We investigated the Variant Allele Frequency (VAF) distribution in three contexts: our pipeline without additional filtering (Pipeline only) (Fig. 3.5a), our pipeline with manually curated data filtering (Pipeline + MT RNA-seq + WES) (Fig. 3.5b), and our pipeline with the integration of a machine learning model (Pipeline + MT RNA-seq + WES+ ML) (Fig. 3.5c), focusing on mutations classified as 'remained' category. Mutations identified in our pipeline without additional filtering displayed high VAF (Fig. 3.5a), and the incorporation of manual curation data can further refinement by reducing mutations with high VAF (Fig. 3.5b). Additionally, our pipeline, complemented by the machine learning model, effectively eliminated high VAF mutations present in ≥10, 5-9, or <5 samples (Fig. 3.5c), notably reducing the occurrence of homozygous germline mutations.

When comparing these findings with somatic mutations identified in paired tumor-normal WES studies (Fig. 3.5d), somatic mutations in RNA-seq data typically demonstrate higher VAF values, which could be attributed to their increased expression levels in tumor samples. However, it's important to highlight that somatic mutations consistently exhibit VAF values below 0.4.

Identification of frequently mutated genes using our pipeline

We identified genes harboring somatic non-nonsynonymous base substitutions in each tumor using different cancer-associated somatic mutation identification methods, namely, a pipeline without manual curation of data and machine learning filtering (left panel), a pipeline with manually curated data but without machine learning filtering (middle panel), and a pipeline with manually curated data and machine learning filtering (right panel) (Fig. 3.6).

To assess the effectiveness of manually curated data and machine learning methods in identifying cancer-associated somatic mutations in canine cancers, we subsequently examined the alteration landscape across these three methods. Each method revealed unique alteration features specific to different canine tumor types.

For instance, in mammary tumors, we observed a frequent PIK3CA alteration, with 35% of the tumors showing mutations in this gene according to all three methods (Fig. 3.6, top panel). This finding aligns with previously published results for canine mammary tumors[16]. TP53 emerged as the most frequently mutated gene in osteosarcoma, and hemangiosarcoma across all

three methods. This consistency with previously reported findings[16] suggests the robustness of our methods in capturing this feature.

In gliomas, PIK3CA and PDGFRA were identified as among the top three frequently mutated genes by our machine learning-based pipeline, consistent with published canine mutation results[16,31]. Additionally, in hemangiosarcoma, PIK3CA was identified as a frequently mutated gene by all three methods, aligning with published results[16,37,79].

While our three methods effectively capture the essential mutation landscape in line with published results, our machine learning-based pipeline excels at eliminating non-somatic mutations. For instance, IGFBP7 and OTUB1 ranked among the top ten frequently mutated genes in four and three out of five tumor types, and even when manually curated RNA-WES comparison mutation data was included, SOD3 and MRPS2 were among the top ten frequently mutated genes in four and two out of five tumor types, respectively. Such recurrent mutations across different tumor types are less likely to be cancer-associated somatic mutations and can be effectively filtered out by our machine learning-based pipeline, highlighting its effectiveness

Identify tumor mutation burden with different tumor types using our pipeline

To further evaluate our pipeline, we examined the Tumor Mutation Burden (TMB), defined as the count of somatic base substitutions and small indels per mega base callable coding sequences, within a cohort of 367 canine tumors from RNA-seq data after sequence quality control (Fig. 3.1a). To enhance precision, we initially excluded mutations occurring in 1,564 retrogenes, previously identified as genes with significantly higher mutation rates compared to protein-coding genes[90]. Furthermore, for improved TMB accuracy assessment, we included only

mutations classified as “human” and “remained detected in at least 10 samples” within our analysis pipeline (Fig. 3.2). Among all tumor types, hemangiosarcoma (HSA) and Bladder tumor (BLAs) exhibited the highest TMB, aligning with prior studies in canine cancer and human cancer[23,90], which have also suggested high TMB in these tumor types. This consistency was observed across all three analytical methods we employed. Osteosarcoma (OSA) was similarly identified as a high TMB tumor type in previous canine cancer studies[90], and our pipeline, whether with or without machine learning filtering, effectively captured this characteristic. Conversely, mammary tumors (MT) were consistently recognized as low TMB tumor types in both human and canine cancer research[23,90], and our pipeline with curated data or machine learning filtering, reflected this with MT exhibiting the lowest TMB. Notably, with the exception of oral Melanomas (OM) (see Discussion), our cancer associated somatic mutation identification pipeline, especially when machine learning filtering, demonstrating its ability to capture TMB characteristics consistent with existing knowledge.

DISCUSSION

Challenges arise when attempting to directly identify somatic mutations from RNA-seq data, primarily due to the requirement for mutations to be sufficiently expressed. While this prerequisite may reduce sensitivity in specific downstream analyses, it could enhance sensitivity in other aspects. Specifically, expressed mutations harbor the potential to serve as neoantigens, potentially triggering an immune response, a critical facet of cancer research[83].

In this study, we introduce TOSMIC, a novel computational pipeline tailored for the identification of somatic mutations in canine RNA-seq data, reducing the need for matched-normal samples. Our pipeline incorporates variant allele frequency (VAF) filtering, human somatic mutation databases, and machine learning techniques to discriminate somatic mutations from non-somatic variants. Significantly, our pipeline integrates vital elements, including manually curated data derived from mutation comparisons between RNA-seq and whole exome sequencing (WES), as well as human somatic mutation databases.

Our pipeline underwent extensive testing through a range of methodologies, encompassing the assessment of somatic and non-somatic mutation overlap ratios, the identification of frequently mutated genes across diverse tumor types, and the quantification of tumor mutational burden (TMB). In direct comparison with published findings, our pipeline exhibited consistent trends in frequently mutated genes and TMB estimation when analyzed alongside tumor-normal paired DNA samples. These results underscore the pipeline's efficacy in accurately filtering non-somatic mutations from RNA-seq data.

Of particular interest, we identified plenty mutations exclusive to the RNA-seq samples, which are classified as neither somatic nor germline mutations. Our endeavors to validate these RNA-seq-exclusive mutations via DNA whole exome sequencing unveiled the potential presence of experimental artifacts in the RNA-seq sample preparation process, which could elucidate the occurrence of these mutations in RNA-seq but not in DNA samples (i.e., wild type in the DNA samples). Additionally, while accounting for experimental artifacts, the presence of mutations classified as RNA-editing warrants attention, as they may play a role in DNA-RNA transcription processes or exert biological effects on tumor progression.

Canine tumor alteration landscape and TMB

Our pipeline can capture the essential mutation landscape and TMB features as previous published results. Osteosarcoma, hemangiosarcoma, and oral melanoma are considered higher TMB tumor type. Osteosarcoma and oral melanoma originate from mesenchymal cells that lack cell polarity and adhesion. Loss of TP53 function, whether due to TP53 mutation or MDM2 amplification, disrupts the cell cycle checkpoints and accelerates the cell cycle. This, in turn, leads to less effective repair of DNA damage and a reduced correction of DNA replication errors, ultimately causing a rapid accumulation of mutations in the genome and a higher TMB. However, our pipeline didn't identify oral melanoma as high TMB tumor type. One possible explanation for this could be oral melanoma tend to have their copy number alternation such as MDM2 amplification then single point mutations as demonstrated in our previous published results. TP53 and NRAS are common mutations in oral melanoma, but we didn't identify prominent TP53 and NRAS mutations in this dataset. Moreover, the pipeline can't identify the copy number variants so we can't confidently point out what happens in this oral melanoma samples. Further experimental and computational analyses are necessary to validate this phenomenon conclusively.

Limitations of the Pipeline:

Our pipeline effectively identifies cancer-associated single nucleotide variants and small indels of somatic mutations from RNA-seq data. However, it is essential to acknowledge that copy number alterations (CNAs) in DNA sequences represent another crucial aspect of cancer

genomics. CNAs involve variations in the number of copies of specific DNA segments within the genome, encompassing deletions (loss of genetic material) and amplifications (duplication of genetic material)[91]. These alterations are well-documented for their substantial impact on cancer initiation and progression, particularly in sarcoma, liposarcoma, and glioblastoma[92]. Unfortunately, our mutation identification pipeline does not account for CNAs, which means that it may overlook significant mutations associated with cancer development and progression.

One of the strategies employed in our pipeline is the use of Variant Allele Frequency (VAF) to distinguish somatic mutations from germline mutations. VAFs of homozygous and heterozygous germline mutations in most samples are expected to cluster around 1.0 and 0.5, respectively, while VAFs of somatic mutations should exhibit a more random distribution. However, it's important to note that significant CNAs have the potential to shift the VAF distribution. As a result, our VAF filtering steps may not always provide the level of accuracy we anticipate, which is a limitation that warrants consideration.

MATERIALS AND METHODS

Data collection:

Canine RNA-seq were downloaded from the Sequence Read Archive (SRA) database, including PRJNA489087 (mammary tumor), PRJNA749900 (oral melanoma), PRJNA525883(osteosarcoma), PRJNA579792 (glioma), PRJNA562916 (hemangiosarcoma), and others. Other information was obtained from relevant publications of these studies. Canine genome canFam3.1 and gene annotation canFam3 1.99 GTF were downloaded from the Ensembl

database. Known canine germline base substitutions and small indels (91,918,943 total) were combined from (1) Ensembl canine dbSNP, canFam3; (2) the DoGSD database[53] and (3) a published study[93].

Canine RNA-seq data quality control (QC) and processing:

RNA-seq read pairs were mapped to the canine reference genome canFam3 using HISAT2 (version 2.21)[94] . Concordantly and uniquely mapped pairs were identified and were used to calculate the mapping rate of each sample. Such pairs with at least one read with ≥1 bp overlapping a coding sequence (CDS) region of the canFam3 1.99 GTF annotation were used to calculate the CDS-targeting rate. Quality control of canine RNA-seq data was performed as described[87]. First, MultiQC (version 1.5) was used to examine GC content and duplicate level. Second, the distributions of per sample read-pair total amount, mapping quality, and CDS targeting rate were examined to identify and exclude samples that fail to meet the cutoffs. A total of 9 canine RNA-seq samples failed the QC and were excluded from further analysis (Figures 2.1a-e). For each sample that passed QC measures, Subread (version 2.0.0)[95] was used to identify read pairs that are uniquely and, for paired-end RNA-seq, concordantly mapped to the exonic regions of the canFam3 1.99 GTF annotation, the sum of which yields raw RNA-seq counts. Cufflinks (version 2.2.0)[96] were used to calculated FPKM (fragments per kilobase of exon per million mapped) value of each gene in each sample, which was then converted to TPM (transcript per million).

Mutation calling from GATK:

Base substitutions and small indels from tumor samples were first called by applying GATK[97] 3.8.1 HaplotypeCaller with parameters of dontUseSoftClippedBases -stand_call_conf 20.0. Variants were then filtered with GATK VariantFiltration with parameters of FS > 30.0 and QD< 2.0.

Tumor only RNA-seq somatic mutation pipeline:

The tumor only RNA-seq somatic mutation identification pipeline, as illustrated in Figure 2 and below:

Step1 : 367 RNA-seq samples from 9 bioprojects in Sequence Read Archive (SRA) were mapped to Canfam3 with STAR[98] 2.6.1 using STAR 2-pass procedure with variants detection using HaplotypeCaller[97] (GATK 3.8.1).

Step2: Variants were filtered with known germline variants databases.

Step3: Variants were subjected to a 4-step filtering process to filter sequencing errors as described except the steps (5)[32] and divided the remaining variants into the following two categories: (1) variants found in human somatic mutation databases (COSMIC V95[99] and C-bioportal[100] (2021/7/27), 4,290,320 somatic mutations total) after human dog mutation translation, and (2) variants not found in human (hereafter referred to as remained).

Step4: Variants in different categories are subject to different filtering steps based on variant allele frequency (VAF).

Step5: Variants found in ≥ 40% samples of one tumor type or variants found in ≥ 30% samples of ≥ 2 tumor types or found in ≥ 15% of total samples were filtered out.

Step6: Variants found in breed-specific variants database were filtered out.

Step7: Variants were filtered out with a machine learning model.

Human dog mutation translation:

The human and canine protein sequences were derived from the Ensembl annotations of the human (hg37, Release 103), and dog (CanFam3.1, Release 99) and protein sequences were aligned using Clustal-Omega[101]. The corresponding positions between the two species were identified by the protein alignments. Human somatic mutations were derived from cBioPortal (2021/7/27), and COMSMIC (V95). Variants of canine that can be found in the orthologous human somatic mutation database were classified as mutation matched the human somatic databases.

Compare mutations in RNA-seq and WES / manually curated data:

We compared mutations identified from RNA-seq and WES within the same sample and categorized each mutation as either HQ (High Quality) or LQ (Low Quality) for RNA-seq and WES to represent the quality of the mutation. For RNA-seq, if a mutation had alternative reads ≥

3, total reads ≥ 20, and a Variant Allele Frequency (VAF) > 0.1, we labeled it as HQ RNA. In the case of WES, a mutation was labeled as HQ WES if it had alternative reads ≥ 3 and total reads ≥ 20 in both the normal and tumor samples. Mutations that did not meet these criteria were labeled as LQ variants. To classify mutations as somatic, germline, wild-type (WT), or RNA editing, we used mutations identified in WES to verify(refer to Figure 3.3e). Specifically, if the base in the normal sample matched the reference base (N==R), or if the base in the tumor sample matched the reference base (T==R), we considered it as WT (wild-type). A somatic mutation was defined as the base in the normal sample matching the reference base, but differing from the base in the tumor sample. A germline mutation was defined as a base differing from the reference base in both the normal and tumor samples, with the base in the normal sample matching the tumor sample. In the WT category, if the variants called in the RNA-seq were of high quality, we classified them as RNA editing.

Model classification:

The training sets were derived as followed and shown as Figure 3.5a.

Somatic :

1. Mutations shared in both RNA-seq and WES.
2. Somatic mutations reported in human data after pipeline filtering (without machine learning filtering).
3. Manually curated data (as illustrated in Figure 3e).

Germline:

1. Mutations in the remained category found in ≥10 samples but filtered by VAF filtering step.
2. Manually curated data (as illustrated in Figure 3e).

WT/RNA-editing:

1. Manually curated data (as illustrated in Figure 3e).

These mutations were then divided to 5 pairs of training and validation sets with 80% and 20 % in each group, respectively. A random forest classifier with 50 tress was applied on each of the training sets, using the somatic, germline, WT/RNA-editing labels. The resulting model was then tested on the corresponding validation set, and the precision and recall were calculated per variant to evaluate the model's performance.

Tumor mutational burden (TMB)

TMB values were calculated by $TMB = \frac{total\ somatic\ base\ substitutions + small\ indels\ in\ CDS}{total\ callable\ bases\ in\ millions\ in\ CDS}$ for each sample. Somatic base substitutions and small indels were selected from mutation identified as human category and remained category that found in ≥ 10 samples. Callable bases were identified with GATK 3.8.1 with the minimum base quality score set to 10.

Figures

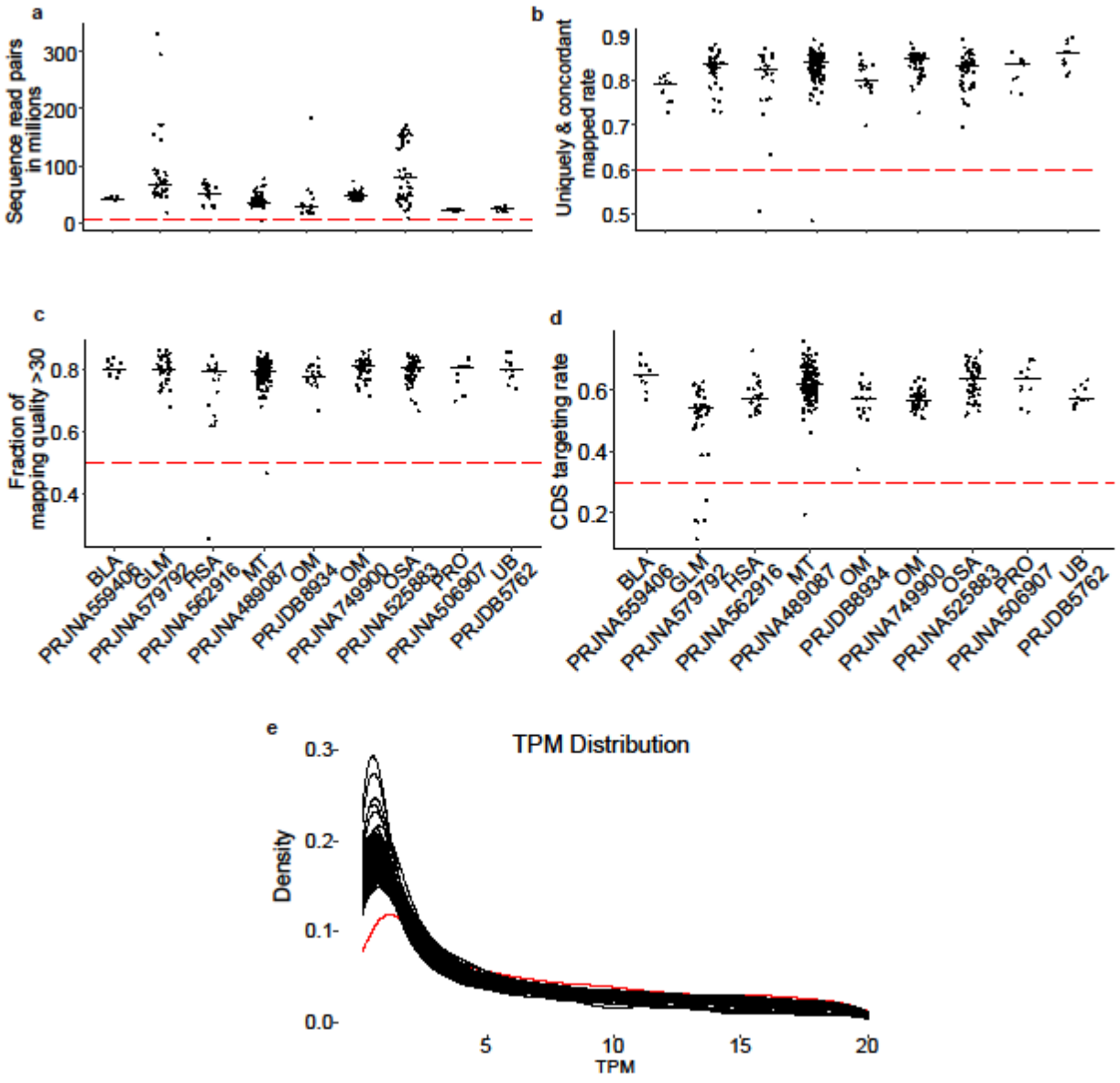


Figure 3.1. We performed a rigorous quality control (QC) of RNA-seq data published for 376 canine tumor samples

a. Distributions of total read pairs per sample of the tumor of each study. Each dot represents a sample, and the median is indicated by a black line. The dashed line specifies the QC cutoff. Each study is represented by the tumor type and the bioproject. BLA: bladder tumor; MT:

mammary tumor; GLM: glioma; OM: oral melanoma; OSA: osteosarcoma; HSA: hemangiosarcoma; UB: urothelial bladder tumors.

b-d. Distributions of per sample rate of read pairs that aligned concordantly and uniquely to the canFam3 reference genome (b), fractions of reads with mapping quality of ≥30 (c), CDS targeting rate (the fraction of read pairs that align concordantly and uniquely to the canFam3 CDS regions) (d).

e. Distribution of TPM values for 376 tumor samples; the red line highlights the sample with an outlier in the TPM distribution.

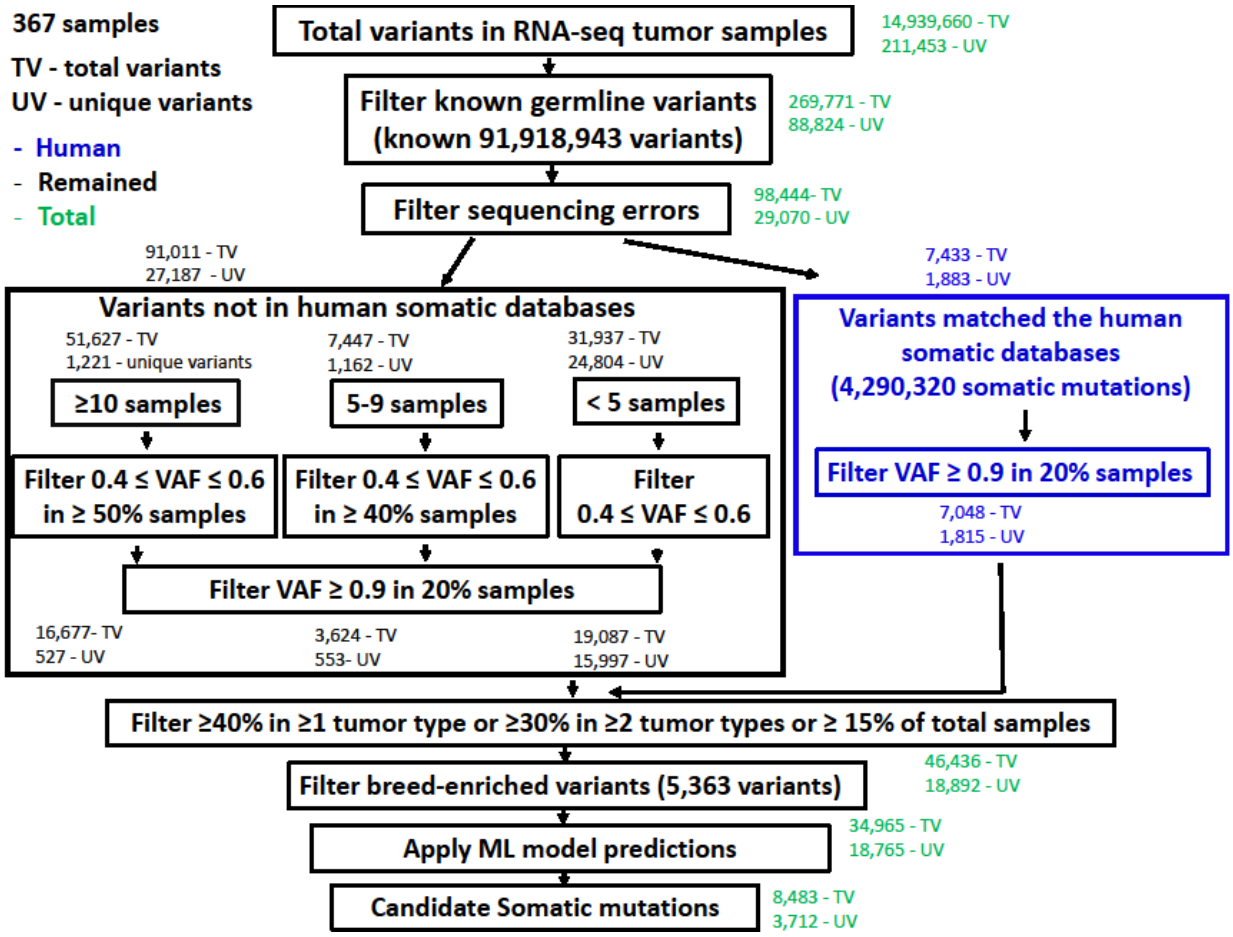
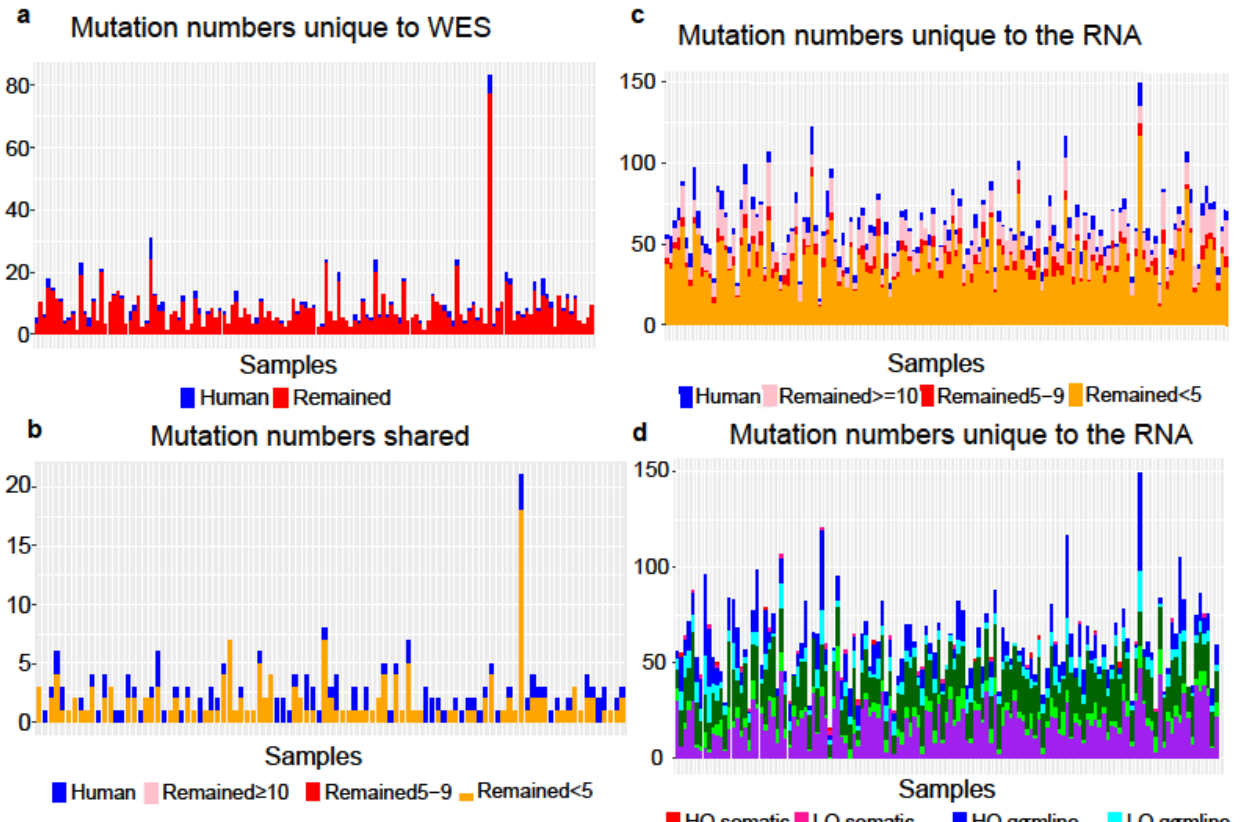


Figure 3.2**. The somatic mutation identification pipeline in tumor-only RNA-seq pipeline uses known germline database, VAF distribution, sample recurrency filtering, breed specific variants, and machine learning model to filter out germline mutations.** For each step in the pipeline, total variants (TV) and unique variants (UV) are indicated. The blue box indicates the variants can be identified in the human somatic mutation databases, and the black is not identified in the human somatic mutation databases. The green number shows the combined variant numbers.



e Summary of manually curated data

WES Category	Sequence quality	Label	N
N!=R,T!=R N==T	HQ RNA-seq, HQ WES	HQ germline	1407
N!=R,T!=R N==T	HQ RNA-seq, LQ WES	LQ germline	101
N!=R,T!=R N==T	LQ RNA-seq, HQ WES	LQ germline	591
N==R,T==R	HQ RNA-seq, HQ WES	HQ RNA editing	763
N==R,T==R	HQ RNA-seq, LQ WES	LQ RNA editing	2489
N==R,T!=R	HQ RNA-seq, HQ WES	HQ somatic	16
N!=R,T==R	HQ RNA-seq, HQ WES	HQ somatic	3
N!=R,T==R	HQ RNA-seq, LQ WES	LQ somatic	5
N==R,T!=R	LQ RNA-seq, HQ WES	LQ somatic	2
N==R,T!=R	LQ RNA-seq, LQ WES	LQ somatic	8
N==R,T!=R	HQ RNA-seq, LQ WES	LQ somatic	26
N==R,T==R	LQ RNA-seq, HQ WES	LQ WT	1199
N==R,T==R	LQ RNA-seq, LQ WES	LQ WT	1589

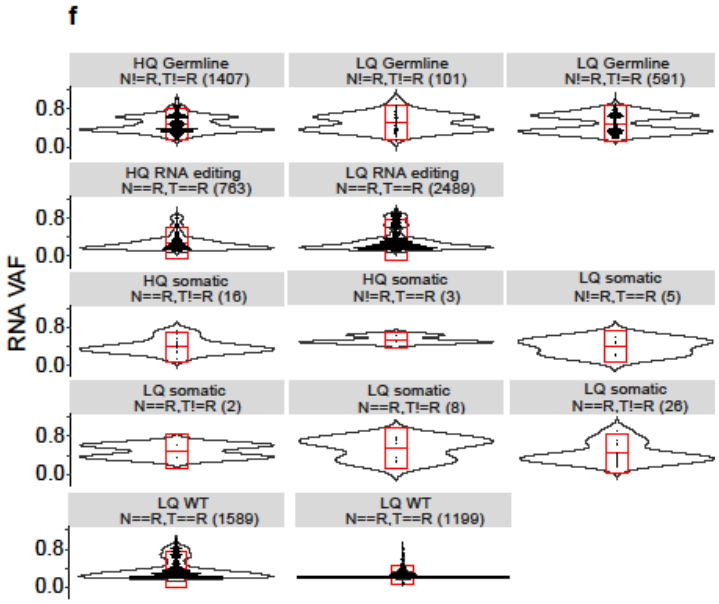


Figure 3.3**. Comparison of Mutations Identified in RNA-seq and WES.** This figure illustrates the comparison of mutations identified through RNA-seq with our pipeline without machine

learning model and whole-exome sequencing (WES). For each mutation in each sample, the genomic coordinate and the actual mutation were compared. Each mutation count is divided into different categories with different color.

a. The mutation counts only found in normal-tumor pairs WES.
b. The mutation counts shared between RNA-seq and WES.
c. The mutation counts exclusively detected in RNA-seq after filtering (no machine learning model).
d. The mutation counts found in RNA-seq only and classified through WES normal-tumor pairs samples. See also Fig. c.
e. A summary of mutations found exclusively in RNA-seq (from Fig. d).
f. Visualization of Variant Allele Frequency (VAF) distribution for each category identified in Fig. e.

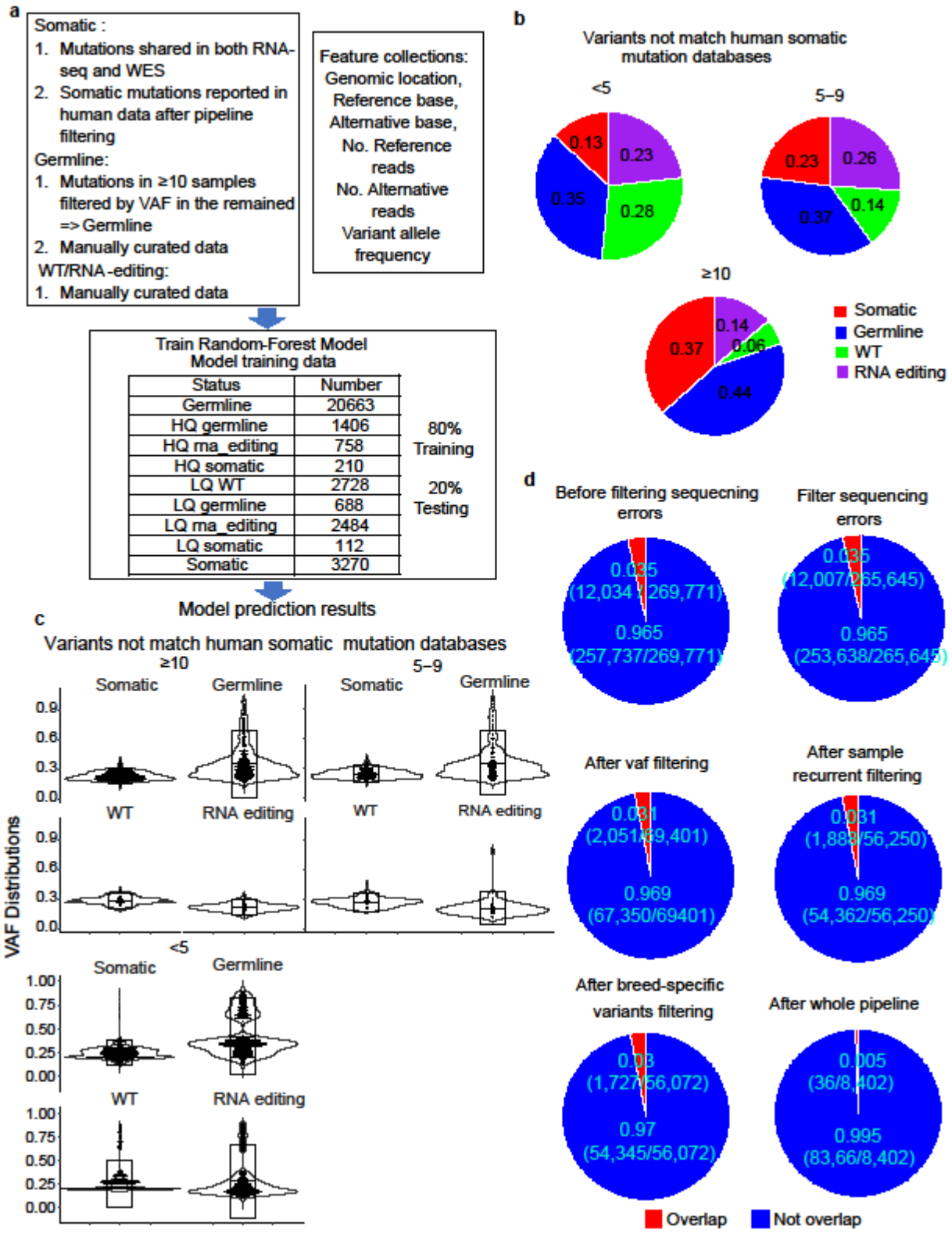


Figure 3.4. Machine Learning Assisted Filtering of Non-Somatic Mutations: This figure

demonstrates the effectiveness of our pipeline, combined with a machine learning model, in filtering out non-somatic mutations and capture somatic mutations.

a. An overview of the training process for our random forest classifier. The model is trained using a 5-fold cross-validation approach with a training set consisting of features for variants of each category (see Methods for details).
b. Ratios of model-predicted results from variants not matched in the human somatic mutation database. Mutations are categorized based on the number of samples in which they were identified: "<5" for mutations found in less than 5 samples, "5-9" for those found in 5 to 9 samples, and "≥10" for those found in 10 or more samples.
c. VAF distributions of model-predicted results from variants not matched in the human somatic mutation database. Somatic mutations typically exhibit lower VAF values, whereas RNA-editing events often have even lower VAF values, which may be attributed to potential sequence errors during the RNA-seq process.
d. Overlapped ratios between germline variants from matched normal samples and total variants identified in the tumor sample at each step in the pipeline. See also Figure2.

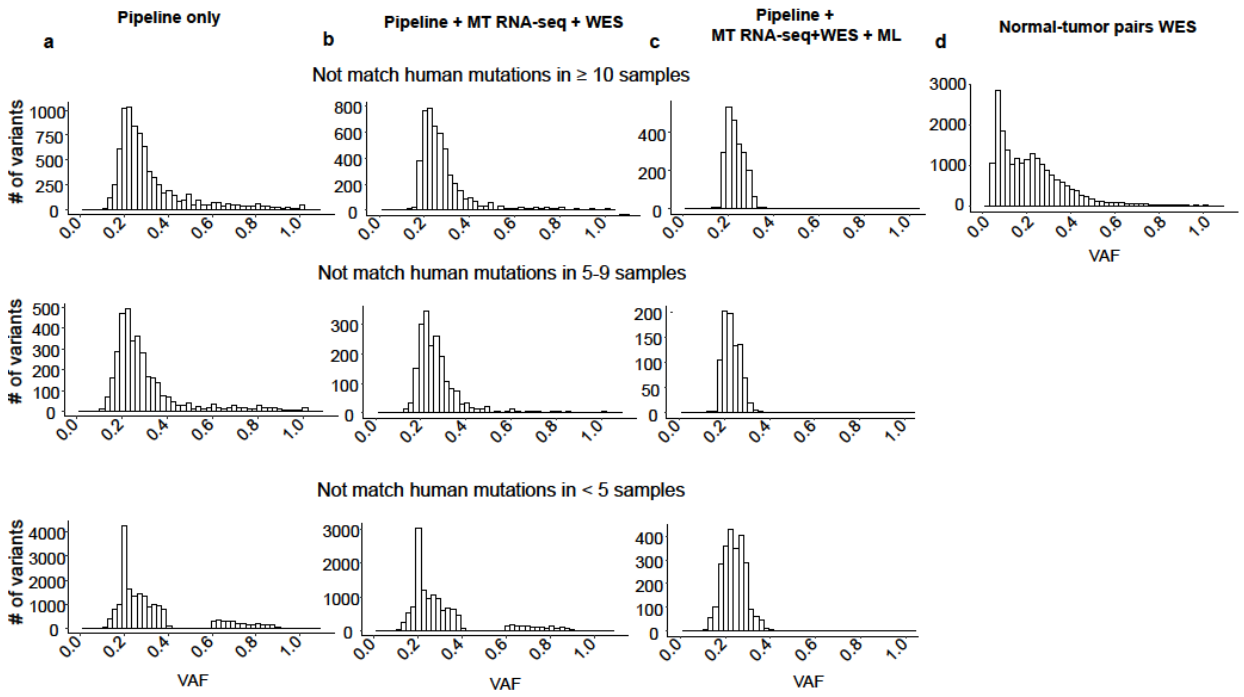


Figure 3.5. Efficient Removal of Germline Mutations by the Pipeline and Machine Learning Classifier. This figure illustrates how our pipeline, combined with a machine learning classifier, effectively eliminates germline mutations. Each column represents a method to visualize the VAF distribution of the mutations.

a, b, c: The top panel in each of the three sub-figures represents variants not found in the human somatic mutation database but found in more than 10 samples. The middle panel represents variants found in 5-9 samples. The bottom panel represents variants found in fewer than 5 samples.

d. The VAF distribution of the mutations identified in tumor-normal pairs WES sequences.

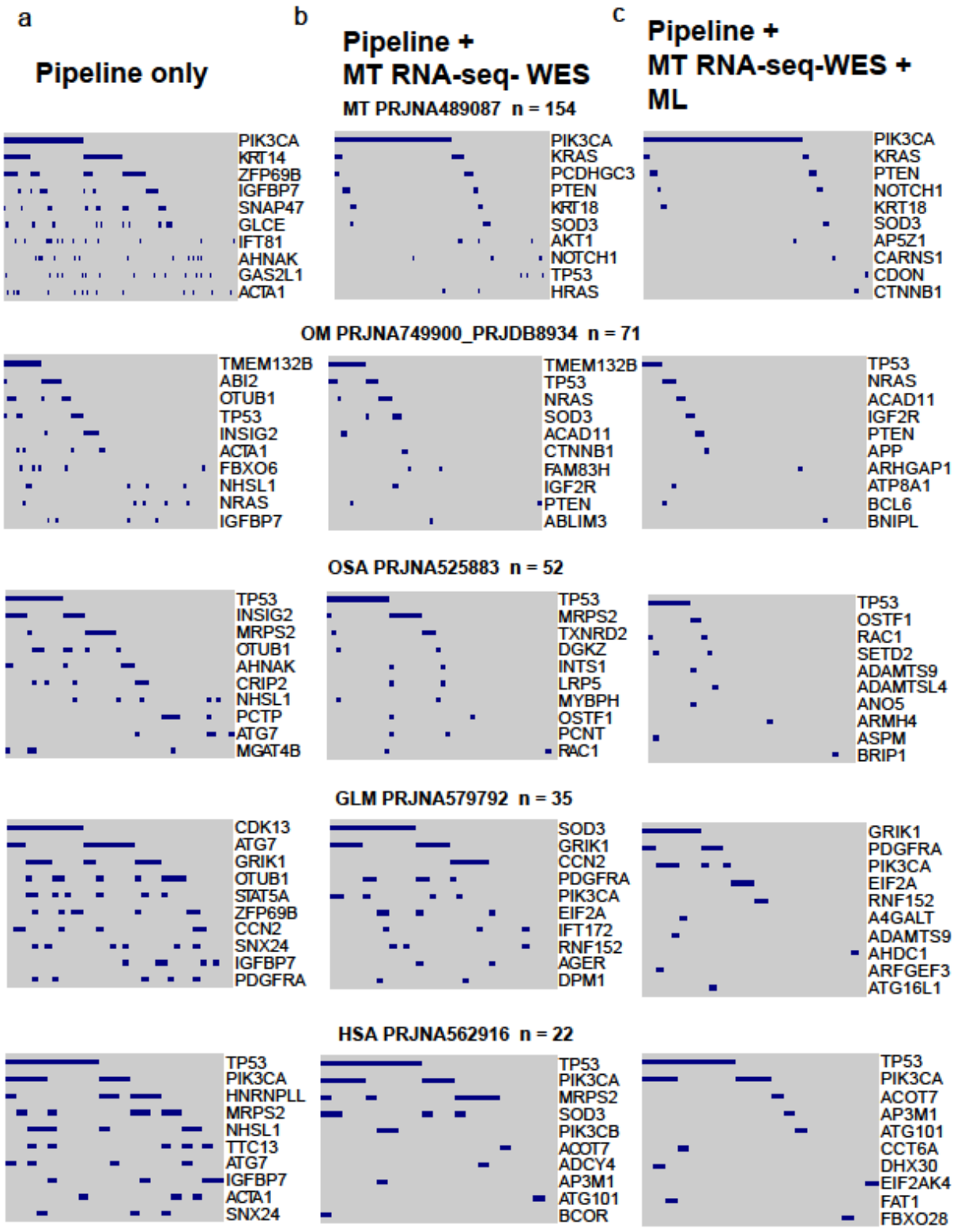


Figure 3.6. Capturing Recurrently Mutated Genes with Our Pipeline and Machine Learning. Our pipeline with a machine learning model can capture most frequent mutated gene with a pattern

similar to published results. Oncoprints indicate the top ten most recurrently altered genes with nonsynonymous somatic base substitutions in CDS regions in each tumor type indicated. Somatic base substitutions and small indels were selected from mutation in human and remained that found in ≥ 10 samples. Plots shown from left to the right:

a. Mutations after our pipeline filtering (without machine learning).
b. Mutations after our pipeline and manual data curation.
c. Mutations after our pipeline combined with manually curated data and a machine learning model filtering.

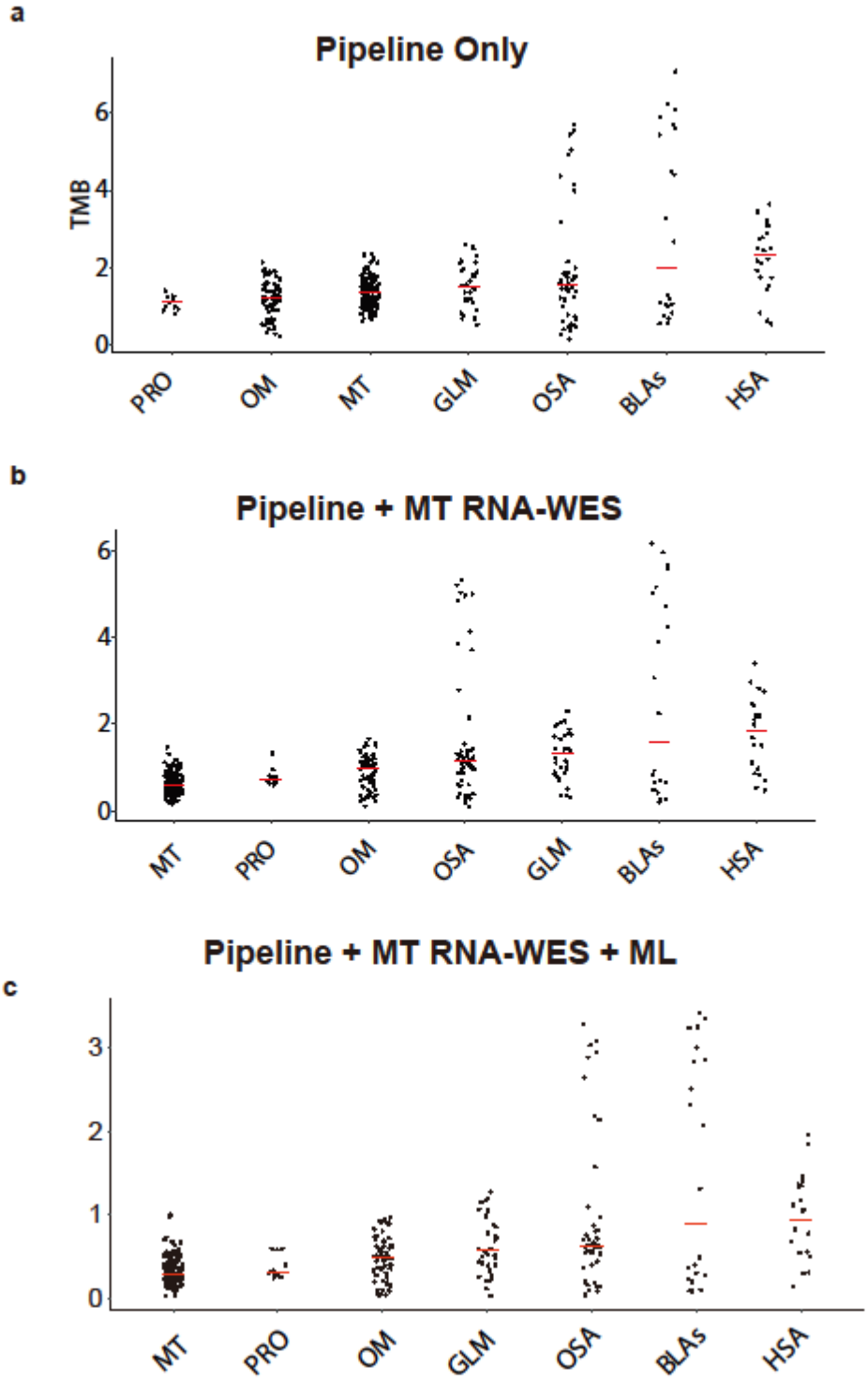

a
Pipeline Only
TMB
0
2
4
6
PRO
OM
MT
GLM
OSA
BLAs
HSA
b
Pipeline + MT RNA-WES
0
2
4
6
MT
PRO
OM
OSA
GLM
BLAs
HSA
c
Pipeline + MT RNA-WES + ML
0
1
2
3
MT
PRO
OM
GLM
OSA
BLAs
HSA

Figure 7. Variation of TMB in Canine Tumor Types and Model Performance. Figure 7 illustrates the variation in TMB across different canine tumor types. Red bar indicated the median TMB value. The TMB distributions for each tumor type are ordered from lowest to highest median values, as follows: (a) TMB after applying our pipeline without machine learning and manual curated data filtering, (b) TMB after combining our pipeline with manual data curation, and (c) TMB after combining our pipeline with manual data curation and machine learning model filtering. Our machine learning model effectively captures patterns similar with published results.

CHAPTER 4

CONCLUSION AND POTENTIAL IMPACT OF THE STUDY

My dissertation research involves conducting comprehensive analyses of somatic mutations in canine cancers, utilizing whole exome/genome sequencing and RNA sequencing data. These methodologies collectively offer crucial insights into the mutational landscapes prevalent in major canine tumor types and breeds. Additionally, the research includes the development of computational tools designed to efficiently analyze sequencing data and identify cancer-associated mutations in canine.

Within these studies, the exploration of mutation landscapes across canine cancers facilitates the improved design of clinical trials testing mutation-targeted therapies in pet dogs. This approach allows for the selection of drugs tailored to specific genomic profiles, as mutation patterns are found to be breed-independent, enabling the focus on tumor types rather than breeds. This expansion of eligible populations enhances the relevance and applicability of the findings.

Moreover, the identification of common hotspot mutations in both pet dogs and humans provides dogs as valuable animal models for studying cancer development mechanisms. Clinical trials designed around these common mutations can benefit both human and canine cancer research.

While whole-exome (WES) and whole-genome sequencing (WGS) are useful in identifying cancer-associated mutations, the consideration of gene expression levels becomes is essential for identifying potential biomarkers or neoantigens for cancer treatment or prevention.

However, WES and WGS do not capture gene expression, making the development of TOSMIC—an innovative tool designed to capture cancer-associated mutations with guaranteed expression—a valuable complement. TOSMIC enhances our ability to identify potential biomarkers or neoantigen for cancer diagnosis and treatment.

In summary, by integrating DNA and RNA sequencing data across diverse canine cancers, these studies provide critical resources to accelerate use of pet dogs in cancer research. The computational and biological insights gained can ultimately inform therapeutic development to benefit both canine and human patients.

Table

BioProject	Total_sample	Read pairs not same length	Total read pairs < 5M	Unique mapped rate <0.6	Fraction of Mapping quality < 0.5	CDS targeting rate < 0.3	TPM_distribution_fail_qc	Pass_QC
PRJDB5762	11	0	0	0	0	0	0	11
PRJDB8934	20	0	0	0	0	0	0	20
PRJNA489087	158	0	1	1	0	1	1	154
PRJNA506907	11	0	0	0	0	0	0	11
PRJNA525883	52	0	0	0	0	0	0	52
PRJNA559406	11	0	0	0	0	0	0	11
PRJNA562916	23	0	0	1	0	0	0	22
PRJNA579792	39	0	0	0	0	4	0	35
PRJNA749900	51	0	0	0	0	0	0	51

Table S1: RNA-seq data QC summary.

REFERENCES

1 Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA Cancer J Clin* **71**, 7-33 (2021). https://doi.org:10.3322/caac.21654

2 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000). https://doi.org:10.1016/s0092-8674(00)81683-9

3 Wright, W. E., Pereira-Smith, O. M. & Shay, J. W. Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts. *Mol Cell Biol* **9**, 3088-3092 (1989). https://doi.org:10.1128/mcb.9.7.3088-3092.1989

4 Hanahan, D. & Folkman, J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* **86**, 353-364 (1996). https://doi.org:10.1016/s0092-8674(00)80108-7

5 Sporn, M. B. The war on cancer. *Lancet* **347**, 1377-1381 (1996). https://doi.org:10.1016/s0140-6736(96)91015-6

6 Hoeijmakers, J. H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366-374 (2001). https://doi.org:10.1038/35077232

7 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724 (2009). https://doi.org:10.1038/nature07943

8 Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-1489 (2015). https://doi.org:10.1126/science.aab4082

9 Beura, L. K. *et al.* Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532**, 512-516 (2016). https://doi.org:10.1038/nature17655

10 Seok, J. *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A* **110**, 3507-3512 (2013). https://doi.org:10.1073/pnas.1222878110

11 Mak, I. W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res* **6**, 114-118 (2014).

12 Schuh, J. C. Trials, tribulations, and trends in tumor modeling in mice. *Toxicol Pathol* **32 Suppl 1**, 53-66 (2004). https://doi.org:10.1080/01926230490424770

13 Dow, S. A Role for Dogs in Advancing Cancer Immunotherapy Research. *Front Immunol* **10**, 2935 (2019). https://doi.org:10.3389/fimmu.2019.02935

14 Gardner, H. L., Fenger, J. M. & London, C. A. Dogs as a Model for Cancer. *Annu Rev Anim Biosci* **4**, 199-222 (2016). https://doi.org:10.1146/annurev-animal-022114-110911

15 Merlo, D. F. *et al.* Cancer incidence in pet dogs: findings of the Animal Tumor Registry of Genoa, Italy. *J Vet Intern Med* **22**, 976-984 (2008). https://doi.org:10.1111/j.1939-1676.2008.0133.x

16 Thamm, D. H. Canine Cancer: Strategies in Experimental Therapeutics. *Front Oncol* **9**, 1257 (2019). https://doi.org:10.3389/fonc.2019.01257

17 Somarelli, J. A. *et al.* Improving Cancer Drug Discovery by Studying Cancer across the Tree of Life. *Mol Biol Evol* **37**, 11-17 (2020). https://doi.org:10.1093/molbev/msz254

18 NCI. NCI-Funded Canine Immunotherapy Trials Network Treats Pet Dogs to Study Cancers Common to Humans (2019).

19 NCI. Integrated Canine Data Commons (ICDC). (2020).

20 CNN. These dogs are getting a cancer vaccine. If it works, humans could be next. (2019).

21 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011). https://doi.org:10.1016/j.cell.2011.02.013
22 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020). https://doi.org:10.1038/s41586-020-1943-3
23 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013). https://doi.org:10.1038/nature12477
24 Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371-376 (2018). https://doi.org:10.1038/nature25795
25 Grobner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321-327 (2018). https://doi.org:10.1038/nature25480
26 Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830 e814 (2018). https://doi.org:10.1016/j.immuni.2018.03.023
27 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020). https://doi.org:10.1038/s41586-020-1969-6
28 Kim, T. M. *et al.* Cross-species oncogenic signatures of breast cancer in canine mammary tumors. *Nat Commun* **11**, 3616 (2020). https://doi.org:10.1038/s41467-020-17458-0
29 Lee, K. H., Hwang, H. J., Noh, H. J., Shin, T. J. & Cho, J. Y. Somatic Mutation of PIK3CA (H1047R) Is a Common Driver Mutation Hotspot in Canine Mammary Tumors as Well as Human Breast Cancers. *Cancers* **11** (2019). https://doi.org:ARTN 2006 10.3390/cancers11122006
30 Amin, S. B. *et al.* Comparative Molecular Life History of Spontaneous Canine and Human Gliomas. *Cancer Cell* **37**, 243-257 e247 (2020). https://doi.org:10.1016/j.ccell.2020.01.004
31 Elvers, I. *et al.* Exome sequencing of lymphomas from three dog breeds reveals somatic mutation patterns reflecting genetic background. *Genome research* **25**, 1634-1645 (2015).
32 Wong, K. *et al.* Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. *Nat Commun* **10**, 353 (2019). https://doi.org:10.1038/s41467-018-08081-1
33 Sakthikumar, S. *et al.* SETD2 is recurrently mutated in whole-exome sequenced canine osteosarcoma. *Cancer research* **78**, 3421-3431 (2018).
34 Gardner, H. L. *et al.* Canine osteosarcoma genome sequencing identifies recurrent mutations in DMD and the histone methyltransferase gene SETD2. *Commun Biol* **2**, 266 (2019). https://doi.org:10.1038/s42003-019-0487-2
35 Megquier, K. *et al.* Comparative genomics reveals shared mutational landscape in canine hemangiosarcoma and human angiosarcoma. *Molecular Cancer Research* **17**, 2410-2421 (2019).
36 Wang, G. *et al.* Molecular subtypes in canine hemangiosarcoma reveal similarities with human angiosarcoma. *PLoS One* **15**, e0229728 (2020). https://doi.org:10.1371/journal.pone.0229728
37 Hendricks, W. P. D. *et al.* Somatic inactivating PTPRJ mutations and dysregulated pathways identified in canine malignant melanoma by integrated comparative genomic analysis. *PLoS genetics* **14**, e1007589 (2018).

https://doi.org:10.1371/journal.pgen.1007589
38 Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
39 Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research* **41** (2013). https://doi.org:ARTN e67
10.1093/nar/gks1443
40 Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin Chem* **61**, 64-71 (2015). https://doi.org:10.1373/clinchem.2014.223040
41 Liu, D. *et al.* Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. *Cancer research* (2014). https://doi.org:10.1158/0008-5472.CAN-14-0392
42 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012). https://doi.org:10.1038/nature11412
43 Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6** (2015). https://doi.org:ARTN 10001
10.1038/ncomms10001
44 Behjati, S. *et al.* Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. *Nature Communications* **8**, 15936 (2017). https://doi.org:10.1038/ncomms15936
45 Wu, G. *et al.* The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nature Genetics* **46**, 444-450 (2014). https://doi.org:10.1038/ng.2938
46 Li, Y. *et al.* Cancer driver candidate genes AVL9, DENND5A and NUPL1 contribute to MDCK cystogenesis. *Oncoscience* **1**, 854 (2014).
47 Tianfang Wang, S.-H. K., Xiao Peng, Severine Urdy, Zefu Lu, Robert J. Schmitz, Stephen Dalton, Keith E. Mostov, Shaying Zhao. A Qualitative Change in the Transcriptome Occurs after the First Cell Cycle and Coincides with Lumen Establishment during MDCKII Cystogenesis. *iScience* **23** (2020). https://doi.org:https://doi.org/10.1016/j.isci.2020.101629.
48 Tang, J. *et al.* Cancer driver-passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer. *Oncogene* **33**, 814-822 (2014). https://doi.org:10.1038/onc.2013.17
49 Koren, S. *et al.* PIK3CA H1047R induces multipotency and multi-lineage mammary tumours. *Nature* **525**, 114-118 (2015).
50 Campbell, B. B. *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* **171**, 1042-+ (2017). https://doi.org:10.1016/j.cell.2017.09.048
51 Wang, J. *et al.* Proliferative and Invasive Colorectal Tumors in Pet Dogs Provide Unique Insights into Human Colorectal Cancer. *Cancers (Basel)* **10** (2018). https://doi.org:10.3390/cancers10090330
52 Wang, J. *et al.* Collaborating genomic, transcriptomic and microbiomic alterations lead to canine extreme intestinal polyposis. *Oncotarget* **9**, 29162-29179 (2018).

https://doi.org:10.18632/oncotarget.25646
53 Bai, B. *et al.* DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res* **43**, D777-783 (2015). https://doi.org:10.1093/nar/gku1174
54 Plassais, J. *et al.* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature communications* **10**, 1-14 (2019).
55 Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690-705 e699 (2018). https://doi.org:10.1016/j.ccell.2018.03.014
56 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1-pl1 (2013). https://doi.org:10.1126/scisignal.2004088
57 Mackay, A. *et al.* Molecular, Pathological, Radiological, and Immune Profiling of Non-brainstem Pediatric High-Grade Glioma from the HERBY Phase II Randomized Trial. *Cancer Cell* **33**, 829-842 e825 (2018). https://doi.org:10.1016/j.ccell.2018.04.004
58 Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e296 (2018). https://doi.org:10.1016/j.cell.2018.03.022
59 da Silva Almeida, A. C. *et al.* The mutational landscape of cutaneous T cell lymphoma and Sezary syndrome. *Nat Genet* **47**, 1465-1470 (2015). https://doi.org:10.1038/ng.3442
60 Zhou, R. *et al.* Analysis of mucosal melanoma whole-genome landscapes reveals clinically relevant genomic aberrations. *Clinical Cancer Research* **25**, 3548-3560 (2019).
61 Painter, C. A. *et al.* The Angiosarcoma Project: enabling genomic and clinical discoveries in a rare cancer through patient-partnered research. *Nature medicine* **26**, 181-187 (2020).
62 Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337 e310 (2018). https://doi.org:10.1016/j.cell.2018.03.035
63 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
64 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
65 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219 (2013).
66 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811-1817 (2012). https://doi.org:10.1093/bioinformatics/bts271
67 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-e164 (2010). https://doi.org:10.1093/nar/gkq603
68 Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-477 (2013). https://doi.org:10.1016/j.cell.2013.09.034
69 Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40**, 11189-11201 (2012). https://doi.org:10.1093/nar/gks918
70 Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol* **16**, 197 (2015). https://doi.org:10.1186/s13059-015-0758-2

71 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576 (2012).
72 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004). https://doi.org:DOI 10.1093/biostatistics/kxh008
73 Zhang, M. *et al.* SEG - A Software Program for Finding Somatic Copy Number Alterations in Whole Genome Sequencing Data of Cancer. *Comput Struct Biotechnol J* **16**, 335-341 (2018). https://doi.org:10.1016/j.csbj.2018.09.001
74 Tang, J. *et al.* Copy number abnormalities in sporadic canine colorectal cancers. *Genome Res* **20**, 341-350 (2010). https://doi.org:10.1101/gr.092726.109
75 Liu, D. *et al.* Canine spontaneous head and neck squamous cell carcinomas represent their human counterparts at the molecular level. *PLoS Genet* **11**, e1005277 (2015). https://doi.org:10.1371/journal.pgen.1005277
76 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* **57**, 289-300 (1995). https://doi.org:DOI 10.1111/j.2517-6161.1995.tb02031.x
77 Jardim, D. L., Goodman, A., de Melo Gagliato, D. & Kurzrock, R. The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer Cell* **39**, 154-173 (2021). https://doi.org:10.1016/j.ccell.2020.10.001
78 Luksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517-520 (2017). https://doi.org:10.1038/nature24473
79 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594 (2018). https://doi.org:10.1038/s41592-018-0051-x
80 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013). https://doi.org:10.1038/nbt.2514
81 Katzir, R., Rudberg, N. & Yizhak, K. Estimating tumor mutational burden from RNA-sequencing without a matched-normal sample. *Nat Commun* **13**, 3092 (2022). https://doi.org:10.1038/s41467-022-30753-2
82 Jiang, T. *et al.* Tumor neoantigens: from basic research to clinical applications. *J Hematol Oncol* **12**, 93 (2019). https://doi.org:10.1186/s13045-019-0787-5
83 Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818-834 e813 (2020). https://doi.org:10.1016/j.cell.2020.09.015
84 Rodrigues, L. *et al.* Shared hotspot mutations in oncogenes position dogs as an unparalleled comparative model for precision therapeutics. *Sci Rep* **13**, 10935 (2023). https://doi.org:10.1038/s41598-023-37505-2
85 Lee, K. H., Hwang, H. J., Noh, H. J., Shin, T. J. & Cho, J. Y. Somatic Mutation of PIK3CA (H1047R) Is a Common Driver Mutation Hotspot in Canine Mammary Tumors as Well as Human Breast Cancers. *Cancers (Basel)* **11** (2019). https://doi.org:10.3390/cancers11122006
86 Megquier, K. *et al.* Comparative Genomics Reveals Shared Mutational Landscape in Canine Hemangiosarcoma and Human Angiosarcoma. *Mol Cancer Res* **17**, 2410-2421

(2019). https://doi.org:10.1158/1541-7786.MCR-19-0221
87 Watson, J. *et al.* Human basal-like breast cancer is represented by one of the two mammary tumor subtypes in dogs. *Breast Cancer Res* **25**, 114 (2023). https://doi.org:10.1186/s13058-023-01705-5
88 Thiemeyer, H. *et al.* Suitability of ultrasound-guided fine-needle aspiration biopsy for transcriptome sequencing of the canine prostate. *Sci Rep* **9**, 13216 (2019). https://doi.org:10.1038/s41598-019-49271-1
89 Alsaihati, B. A. *et al.* Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. *Nature Communications* **12** (2021). https://doi.org:ARTN 4670
10.1038/s41467-021-24836-9
90 Alsaihati, B. A. *et al.* Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. *Nat Commun* **12**, 4670 (2021). https://doi.org:10.1038/s41467-021-24836-9
91 Shlien, A. & Malkin, D. Copy number variations and cancer. *Genome Med* **1**, 62 (2009). https://doi.org:10.1186/gm62
92 Shetty, S. *et al.* MDM2 amplification in malignant Brenner tumors may play a role in progression to malignancy and aid in separation from urothelial and other ovarian carcinomas. *Hum Pathol* **117**, 42-50 (2021). https://doi.org:10.1016/j.humpath.2021.08.001
93 Plassais, J. *et al.* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun* **10**, 1489 (2019). https://doi.org:10.1038/s41467-019-09373-w
94 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019). https://doi.org:10.1038/s41587-019-0201-4
95 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**, e108 (2013). https://doi.org:10.1093/nar/gkt214
96 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010). https://doi.org:10.1038/nbt.1621
97 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010). https://doi.org:10.1101/gr.107524.110
98 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013). https://doi.org:10.1093/bioinformatics/bts635
99 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-811 (2015). https://doi.org:10.1093/nar/gku1075
100 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-404 (2012). https://doi.org:10.1158/2159-8290.CD-12-0095
101 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence

alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011). https://doi.org:10.1038/msb.2011.75