Breast Cancer Research

RESEARCH

Open Access

Human basal-like breast cancer is represented by one of the two mammary tumor subtypes in dogs

Joshua Watson[1], Tianfang Wang[2], Kun-Lin Ho[1], Yuan Feng[1], Tanakamol Mahawan[3], Kevin K. Dobbin[4] and Shaying Zhao[1,2*]

Abstract

Background About 20% of breast cancers in humans are basal-like, a subtype that is often triple-negative and difficult to treat. An effective translational model for basal-like breast cancer is currently lacking and urgently needed. To determine whether spontaneous mammary tumors in pet dogs could meet this need, we subtyped canine mammary tumors and evaluated the dog–human molecular homology at the subtype level.

Methods We subtyped 236 canine mammary tumors from 3 studies by applying various subtyping strategies on their RNA-seq data. We then performed PAM50 classification with canine tumors alone, as well as with canine tumors combined with human breast tumors. We identified feature genes for human BLBC and luminal A subtypes via machine learning and used these genes to repeat canine-alone and cross-species tumor classifications. We investigated differential gene expression, signature gene set enrichment, expression association, mutational landscape, and other features for dog–human subtype comparison.

Results Our independent genome-wide subtyping consistently identified two molecularly distinct subtypes among the canine tumors. One subtype is mostly basal-like and clusters with human BLBC in cross-species PAM50 and feature gene classifications, while the other subtype does not cluster with any human breast cancer subtype. Furthermore, the canine basal-like subtype recaptures key molecular features (e.g., cell cycle gene upregulation, TP53 mutation) and gene expression patterns that characterize human BLBC. It is enriched in histological subtypes that match human breast cancer, unlike the other canine subtype. However, about 33% of canine basal-like tumors are estrogen receptor negative (ER–) and progesterone receptor positive (PR+), which is rare in human breast cancer. Further analysis reveals that these ER–PR+ canine tumors harbor additional basal-like features, including upregulation of genes of interferon-*γ* response and of the Wnt-pluripotency pathway. Interestingly, we observed an association of *PGR* expression with gene silencing in all canine tumors and with the expression of T cell exhaustion markers (e.g., *PDCD1*) in ER–PR+ canine tumors.

Conclusions We identify a canine mammary tumor subtype that molecularly resembles human BLBC overall and thus could serve as a vital translational model of this devastating breast cancer subtype. Our study also sheds light on the dog–human difference in the mammary tumor histology and the hormonal cycle.

*Correspondence:
Shaying Zhao
szhao@uga.edu
Full list of author information is available at the end of the article

Keywords Dog–human comparison, Canine mammary tumor subtyping, RNA-seq, Basal-like breast cancer, Estrogen receptor, Progesterone receptor, Prolactin receptor, Wnt signaling, Interferon-γ response, PAM50 subtyping, Machine learning

Background

Human breast cancer is heterogeneous, consisting of well-established molecularly distinct subtypes [1–10]. One of these subtypes is basal-like breast cancer (BLBC; human BLBC will be referred to as hBLBC hereafter), which makes up roughly 15–20% of human breast cancers and has the worst prognosis of all subtypes [1–10]. About 70% of hBLBCs are triple negative, expressing neither estrogen receptor (ER) nor progesterone receptor (PR) and without HER2 amplification or overexpression [1–10]. These cancers also tend to have increased rates of cell proliferation and metastasis [1–10]. All these highlight the need for an effective translational model for hBLBC, which is critically missing at present [11–13].

Mammary cancers in pet dogs naturally occur in animals with an intact immune system [13, 14], overcoming many limitations of traditional cancer models such as cell lines and genetically modified rodent models. These canine cancers more accurately emulate human breast cancers in etiology, complexity, heterogeneity, behavior, treatment, and outcome [13–16]. They are also common in bitches, with an annual incidence rate estimated at 198 per 100,000 [17], which is comparable to the rate of 125 per 100,000 for breast cancer in women in the USA [18]. Mammary cancer is especially common in bitches that are not spayed or are spayed after the second estrus, with the risk of malignant tumor development expected at 26% [17]. Thus, canine mammary tumors have the potential to serve as a much-needed translational model of hBLBC, effectively bridging a current gap between preclinical models and human clinical trials to accelerate bench-to-bedside translation.

The effective use of the canine model is, however, complicated by issues including the dog–human hormonal cycle difference, e.g., the luteal phase lasts ~ 14 days for humans but ~ 2 months for dogs. Another difference is histology. About 50% of canine mammary tumors are complex or mixed, with multiple cell lineages (e.g., epithelial and myoepithelial cells) proliferating [14, 19, 20]. These histologies (e.g., adenomyoepithelioma) are, however, are very rare (< 1%) in human breast cancers [14, 21, 22]. It remains unknown how these differences shape the molecular homology and difference between canine and human mammary tumors.

The same as human breast cancers, spontaneous canine mammary cancers are heterogeneous and consist of distinct subtypes [19, 20]. Thus, subtype-level dog–human comparison is needed to evaluate the dog–human homology. Canine mammary cancers have been histopathologically and clinically subtyped, as well as molecularly subtyped with immunohistochemical markers established for human breast cancer (anti-ER, PR, HER2, -CK 5/6 and -CK14) [19, 20, 23, 24]. However, to our knowledge, canine mammary tumors have not been independently subtyped by genome-wide molecular studies.

For dog–human comparison, our previous study indicates that one histological subtype, simple carcinoma, molecularly resembles hBLBC in cross-species PAM50 classification with human and canine tumors [14]. However, this study is limited by its small sample size. Another group has performed PAM50 classification on the RNA-seq data recently published for 154 canine mammary tumors [25] and reported a higher homology between canine and human luminal A tumors than between canine and human basal-like tumors [26]. This study, however, did not perform cross-species PAM50 or other classifications to directly compare human and canine tumors. Moreover, many of the canine luminal A tumors are complex and mixed tumors, histologically differing from the vast majority of human luminal A tumors [14, 21, 22].

To address these discrepancies and deficiencies, we set out to independently subtype canine mammary tumors using RNA-seq data of 236 tumors [14, 25, 27] and then perform dog–human comparison at the subtype level, as described below.

Materials and methods

Data collection

Canine RNA-seq data were downloaded from the Sequence Read Archive (SRA) database, including data of 154 mammary tumors from PRJNA489087 (excluding 4 metastatic osteosarcomas and fibrosarcomas) [25] and of 63 mammary tumors from PRJNA561580 [27]. RNA-seq data of 25 mammary tumor samples sequenced in house [14] were also included (PRJNA203086 and PRJNA912710). Human breast cancer RNA-seq data were downloaded from the National Cancer Institute (NCI) Genomic Data Commons (GDC) database, and the PAM50 classification of these cancers was obtained from the cBioportal database [28]. Gene expression microarray data of two canine mammary tumor studies (GSE20718 and GSE22516) and one human breast cancer

(GSE20685) were downloaded from the Gene Expression Omnibus (GEO) database [29–31] and processed with affyPackage [32]. 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. Canine genome canFam4 (GSD 1.0) and the NCBI RefSeq gene annotation (NCBI RefSeq Curated and Predicted subsets from the NCBI *Canis lupus familiaris* Annotation Release 106, 2021-01-11) were obtained from the UCSC Genome Brower site. Canine mutation data and tumor mutation burden (TMB) values from whole exome sequencing analysis were obtained from a previous publication [33].

Canine sample collection and RNA-seq

Fresh-frozen (FF) canine tissues and spontaneous tumors were obtained from the canine tissue archive bank at Ohio State University, as previously described [34]. Samples were collected from client-owned dogs that developed the disease spontaneously, under the guidelines of the Institutional Animal Care and Use Committee for use of residual diagnostic specimens and with owner informed consent. The case information was provided by the tissue bank. The research received the ethical approval from the Institutional Animal Care and Use Committee.

Cryosectioning of FF tissues, H&E staining, and cryomicrodissection was performed as described [34] to enrich tumor cells for tumor samples. Genomic DNA and RNA were extracted from the dissected tissues using the AllPrep DNA/RNA Mini Kit (cat. no. 80204) from QIAGEN (Germantown, MD, USA). Only samples with a 260/280 ratio of ~ 2.0 (RNA) and showing no degradation and other contaminations were subjected to further quality control with qRT-PCR analysis with a panel of genes as previously described [34, 35]. RNA-seq libraries were constructed using KAPA Stranded mRNA- Seq Kit. The samples were subjected to 75- or 125-bp paired-end sequencing using Illumina HiSeq 2500 or NextSeq 500 at Georgia Genomics Facility.

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

Canine RNA-seq data were processed as described [34–36]. Briefly, RNA-seq read pairs were mapped to the canine reference genome canFam3 using HISAT2 (version 2.21) [37]. Concordantly (for paired-end RNA-seq data only) 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 [36]. First, MultiQC [38] (version 1.5) was used to examine GC content and duplicate level. Base quality distribution before and after Trimmomatic trimming was also examined. 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 6 canine RNA-seq samples from PRJNA561580 failed the QC and were excluded from further analysis (Additional file 1: Fig. S1A–E).

For each sample that passed QC measures, Subread (version 2.0.0) [39] was used to identify read pairs that are uniquely and, for paired-end RNA-seq, concordantly mapped to the CDS regions of the canFam3 1.99 GTF annotation, the sum of which yields raw RNA-seq counts. Cufflinks version 2.2.0 [40] was used to calculate the 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). For studies that combine RNA-seq data from multiple sources, comBat [41] was applied to correct batch effect for TPM values, and ComBat-seq [42] was used to correct batch effect for mapped RNA-seq read count values.

RNA-seq reads were also mapped to the canFam4 genome using HISAT2 (version 2.21), and the FPKM and TPM value of each gene was then calculated using Cufflinks version 2.2.0 and the NCBI RefSeq annotation described previously.

ER, PR, and HER2 status

Samples with an *ESR1* or *PGR* expression level of $FPKM \leq 1$ and $FPKM > 1$ were classified as ER or PR negative and positive, respectively. Samples with an *ERBB2* expression level of $FPKM \leq 35$ and $FPKM > 35$ were classified as HER2 not enriched and enriched, respectively.

Canine mammary tumor subtyping

A gene was selected if it has an official gene name associated with its Ensembl gene ID in the canFam3 1.99 GTF annotation and is expressed (with $FPKM \geq 1$ in at least one sample across a cohort or study). This yields 13,416 genes in the discovery set (paired-end RNA-seq data) [14, 25] and 13,608 genes in the validation set (single-end RNA-seq data) [27] (see Results). The NMF R package [43] was then applied on all of these selected genes, as well as on the top 5000, 2000, 1000, and 500 most variable genes among them, with 30 runs for the rank determination. These analyses consistently divided 143 out of 179 samples of the discovery set into two subtypes. For validation, K-means clustering, consensus clustering, and hierarchical clustering via R packages stats,

ConsensusClusterPlus, and pvclust, respectively [44–46], were used to subtype the 143 samples using the top 10% most variable genes. The same process was repeated to subtype the samples in the validation set.

PAM50 classification of canine and human tumors

A total of 43 canine homologues of the 50 PAM50 genes were identified in the canFam3 1.99 annotation file. These 43 genes were then used to perform PAM50 classification with canine samples alone, as well as with canine and human combined samples, for RNA-seq studies [14, 25, 27]. For microarray studies [30, 31], 40 of the PAM50 genes were identified based on the probes and used to classify canine samples alone and canine-human combined samples. The PAM50 subtypes of human breast cancer samples were downloaded from the cBioportal database or relevant publications. For canine and human combined sample PAM50 clustering for RNA-seq studies, 60 human tumors were randomly sampled from the Cancer Genome Atlas (TCGA) breast cancer study for each of the luminal A, luminal B, hBLBC, and HER2-enriched subtypes. These samples, along with all 27 human normal-like tumors, were merged with all 143 subtyped canine tumors for PAM50 classification analysis. The clustering dendrogram was then cut at the minimum number of clusters that maximally separate hBLBC tumors from hLumA tumors using the R package dendextend [47]. The number of tumors of each canine or human subtype in each cluster was counted. If a cluster contains the majority of tumors of a human subtype as well as the majority of a canine subtype, the human and canine subtypes were considered matched. This process was repeated 100 times to ensure each human tumor was sampled at least once.

Multidimensional scaling on the Euclidean distance matrix was performed for each of the 100 random samplings, from which the Mahalanobis distance between the centers of any two subtypes was calculated using the R package 'GenAlgo' v2.2.0 [46].

The above process was repeated with the 49 PAM50 genes identified in the canFam4 NCBI RefSeq annotation described previously.

The above process was repeated for microarray studies [29–31], except that only 40 canine homologues of the 50 PAM50 genes were identified, and 30 tumors per subtype were randomly sampled from the human dataset [29].

Differentially expressed (DE) genes and gene set enrichment analysis (GSEA)

DESeq2 [48] was used to identify DE genes between subtypes or subgroups. Genes with ≥ 2 fold change in read count and the Benjamini–Hochberg (BH)-adjusted $p \leq 0.05$ were considered differentially expressed. Enriched functions of DE genes were investigated with the GSEA [49] and DAVID [50] web tools. Pathway and signature gene sets were acquired from previous publications [34, 35, 51]. Single-sample GSEA (ssGSEA v. 10.1.0) was performed using GenePattern [52].

Machine learning for feature gene selection

The analysis was performed using R version 4.2.2 as follows. First, TCGA human breast cancers and canine mammary tumors were combined, and batch effect correction was applied using Combat [53] on the expression values ($\log_2(\text{TPM}+1)$) of the entire gene set. Second, with the batch-corrected human breast cancer data, DE genes were identified between hBLBC or hLumA and another PAM50 subtype with adjusted $p \leq 0.05$. This yielded 8034, 6958, and 4783 DE genes between hBLBC and each subtype of hLumA, Luminal B, and HER2-enriched, respectively, as well as 5843 and 5697 DE genes between hLumA and each subtype of luminal B and HER2-enriched, respectively. Third, human breast cancer samples were divided into an 80% training set, used for model development and feature selection, and a 20% test set, used for model evaluation. Lastly, Boruta [54] (a wrapper built around the random forest classification algorithm) was applied on the training set iteratively 50 times to identify the most relevant genes among each DE gene set that separates the two subtypes compared, with a frequency cutoff of 50 times. Combining identified genes from all subtype comparisons yielded 115 feature genes for hBLBC and 130 feature genes for hLumA. These genes were used to classify canine tumor samples alone, as well as canine and human combined samples, as described in PAM50 classification.

Correlation and other statistical analyses

Genes with $FPKM > 1$ in at least 10% of the samples of a subtype were chosen for *ESR1*,*PGR*, or *PRLR* (encoding the prolactin receptor) correlation analysis. *PGR* was excluded from hBLBC as < 10% of samples have an $FPKM > 1$. Positively or negatively correlated genes were defined as those with BH-adjusted $p < 0.05$ and correlation coefficient $|R| > 0.3$ for both Pearson and Spearman correlation analysis. The software enrichR [55] was used to identify transcription factors targeting each group of significantly correlated genes. Wilcoxon rank-sum tests were used for statistical comparison between subtypes or subgroups.

Results

Canine mammary tumors consist of two distinct subtypes

We first performed nonnegative matrix factorization (NMF) [43], a widely used subtyping strategy, to subtype 179 canine mammary tumors with paired-end

RNA-seq data (the discovery set), after combining 154 tumors sequenced by Kim et al. [25], and 25 tumors sequenced by us [14] followed by batch correction. NMF subtyping was repeated with all 13,416 genes that are expressed in at least one tumor, as well as with the top 5000, 2000, 1000, and 500 most variable genes among the 13,416 expressed genes. The analysis consistently clustered 143 of 179 tumors into two subtypes (Fig. 1; Additional file 2: Table S1). To validate this, we also subtyped these tumors using other popular strategies, including K-means, consensus clustering, and hierarchical clustering via multiscale bootstrap resampling [44, 45] (Additional file 1: Fig. S2A–C). These

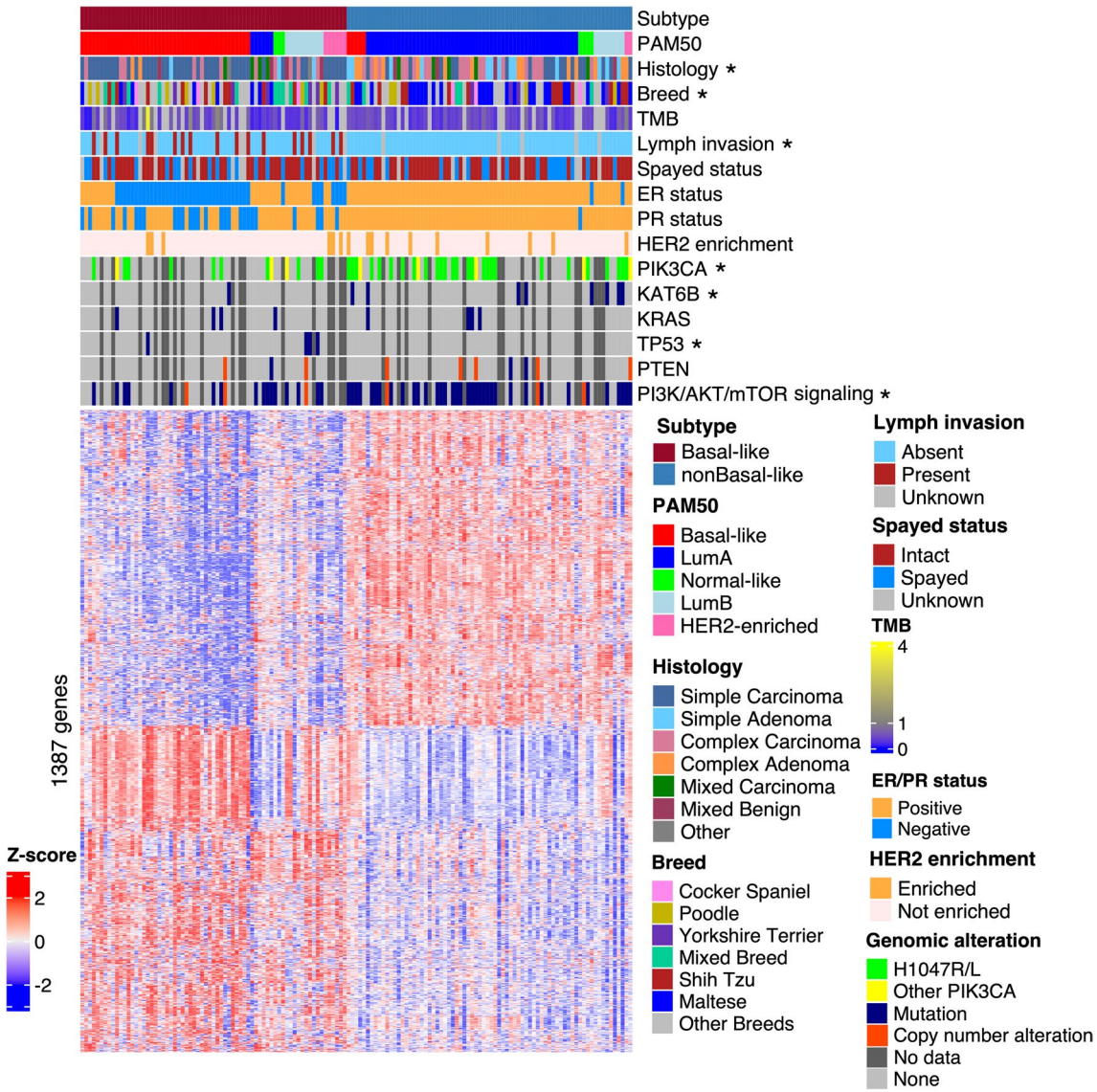

Fig. 1 Two molecularly distinct subtypes of canine mammary tumors were identified. Heatmap of the row scaled $\log_2$(TPM) values of 1387 metagenes from 143 canine mammary tumors (columns), ordered from left to right by subtypes, PAM50 classification, and the *ESR1* expression level from high to low. The metagenes (rows), identified from the NMF analysis (see Methods), are ordered by hierarchical clustering. Lymph invasion is defined as having tumor cells in peritumoral lymphatic vessels and/or regional lymph nodes. Tumor mutation burden (TMB), defined as the number of somatic base substitutions and small indels per megabase (Mb) of callable coding sequence, is obtained from a previous publication [33]. For ER/PR status, a tumor is considered "negative" or "positive" if its *ESR1*/*PGR* has a FPKM value of ≤ 1 or > 1, respectively. For HER2 enrichment, a sample is considered "not enriched" or "enriched" if its *ERBB2* has a FPKM value of ≤ 35 or > 35, respectively. "Other PIK3CA" represents all non-H1047 coding mutations in PIK3CA. "No data" represents samples with no mutation data. Annotation row titles marked with a "*" indicate a significant ($p \leq 0.05$) difference in enrichment between the subtypes

strategies consistently identified the same two subtypes as the NMF approach (Additional file 1: Fig. S2D).

We then performed the same subtyping analyses on the other large study of canine mammary tumors ($n=57$) (the validation set), whose RNA-seq data are single-end [27], differing from the discovery set. These tumors also clustered into two subtypes, consistent with the discovery set (Additional file 1: Fig. S2E–F).

The two subtypes differ in tumor histology and invasiveness. One subtype is significantly ($p = 0.007$) enriched in simple adenomas/carcinomas (where only one cell lineage proliferates prominently), while the other subtype is enriched in complex or mixed adenomas/carcinomas ($p < 0.001$) (where more than one cell lineages are proliferating prominently) (Fig. 1; Additional file 2: Table S1) [14, 19, 25]. Moreover, the simple adenomas/carcinomas-enriched subtype contains significantly ($p < 10^{-6}$) more cases with lymph node invasion (Fig. 1; Additional file 2: Table S1). Interestingly, the other subtype contains significantly ($p = 0.0014$) more Maltese dogs (Fig. 1; Additional file 2: Table S1).

The two subtypes display distinct molecular features. Among the top most mutated genes in this cohort [33], one subtype is enriched in *PIK3CA* hotspot mutation H1047R/L ($p = 0.0034$) and *KAT6B* mutation ($p = 0.036$), while the other subtype (simple adenomas/carcinomas-enriched and with more lymph node invasion) is enriched in *TP53* mutation ($p = 0.043$) (Fig. 1; Additional file 2: Table S1). However, we did not observe any significant difference in *KRAS* mutation and TMB between the two subtypes (Fig. 1; Additional file 2: Table S1). For pathways, the two subtypes differ in mutations and copy number alterations of genes in PI3K signaling ($p = 0.0039$) (Fig. 1; Additional file 2: Table S1), the most altered pathway in canine mammary tumors [33].

Other molecular differences between the two subtypes include PAM50 classification, ER and PR expression status, and others, and will be described in more details below.

Canine and human basal-like tumors cluster together in PAM50 classification

We performed PAM50 classification [56, 57] on the 179 canine tumors from the discovery set. About 74% of the tumors were classified as either basal-like ($n=62$, 35%) or luminal A ($n=71$, 39%) (Fig. 2A; Table 1). Importantly, 90% of the basal-like tumors belong to one of the two subtypes shown in Fig. 1, while 90% of the luminal A tumors belong to the other subtype. For this reason and reasons described below, the two subtypes shown in Fig. 1 are named canine basal-like mammary tumor (cBLMT) and canine non-basal-like mammary tumor (cNBLMT), respectively.

To quantitatively assess the canine-human homology at the subtype level, we performed cross-species PAM50 classification as described [14]. Briefly, we randomly sampled 60 human tumors from each of the luminal A, luminal B, HER2-enriched, and basal-like subtypes from the TCGA RNA-seq study [2, 3]. These, along with all 27 normal-like tumors in TCGA, amount to 267 human tumors covering all five intrinsic subtypes. We then performed PAM50 clustering on these human tumors together with all 143 subtyped canine tumors shown in Fig. 1. This analysis was repeated 100 times, ensuring that each TCGA tumor was sampled at least once.

In 66 of 100 random sampling analyses, cBLMTs and hBLBCs clustered together and away from tumors of any other canine or human subtype (Fig. 2B–C). To the contrary, in 87 of 100 random sampling analyses, cNBLMTs, the other canine subtype, did not cluster with tumors of any human subtype (Fig. 2B–C). These

(See figure on next page.)

Fig. 2 PAM50 and feature gene classifications group canine and human basal-like tumors together, but separates canine non-basal-like tumors from human tumors. **A**. PAM50 classification of 143 subtyped canine mammary tumors. The heatmap shows hierarchical clustering of the canine tumors using the row-scaled $\log_2$(TPM) values of 43 out of the 50 PAM50 genes. The bars indicate the canine subtype from Fig. 1, as well as the PAM50 subtype and confidence score of each tumor. **B**. An example of cross-species PAM50 classification. All 143 subtyped canine tumors, together with 267 human tumors (60 tumors per subtype of luminal **A** (LumA), luminal **B** (LumB), basal-like, or HER2-enriched randomly sampled from TCGA database, along with all 27 normal-like tumors from TCGA), were subjected to PAM50 classification. The dendrogram is colored to indicate the minimum number of clusters that maximally separate the hLumA tumors from hBLBC tumors. This cross-species PAM50 classification was repeated 100 times. **C**. Bar plot showing the number of random samplings in which human and canine basal-like tumors clustered together or separately, compared to those of human luminal **A** and canine non-basal-like tumors. The *p*-value is based on Fisher's exact test. **D**. Multidimensional scaling plot of the cross-species PAM50 classification shown in **B**. Each dot represents a tumor from a subtype specified by the color as indicated in the legend. **E**. Violin plot indicating the distribution of the Mahalanobis distances between the centers of two subtypes on the multidimensional scaled plot **D** of each of the 100 cross-species PAM50 classifications **B** achieved via random samplings (see Methods). The p-values were obtained from Wilcoxon tests. **F.** Classification of 143 subtyped canine mammary tumors using 115 feature genes for hBLBC and 130 feature genes for hLumA identified via machine learning. **G**. Violin plot indicating the distribution of the Mahalanobis distances as described in **E**, using the 115 hBLBC feature genes or 130 hLumA feature genes for cross-species classification, instead of the PAM50 genes. The *p*-values were obtained from Wilcoxon tests.

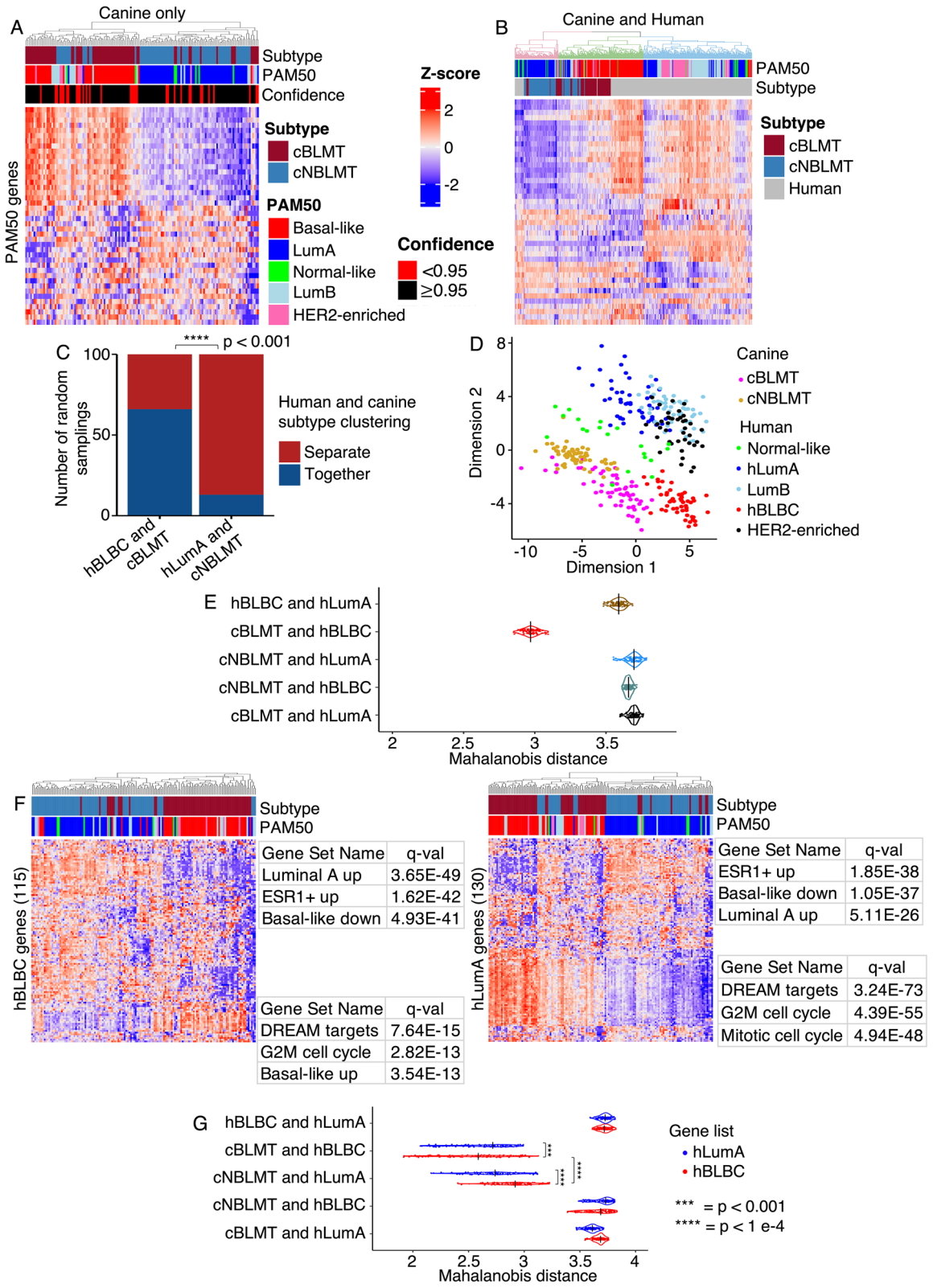

Fig. 2 (See legend on previous page.)

Table 1 Features of canine mammary tumor subtypes and subgroups

Subtype	cBLMT[1]				cNBLMT[2]
Subgroup	ER−PR−	ER−PR+	ER+PR+	ER+PR−	96% ER+PR+
Basal-like	17	18	6	3	5
Luminal A	0	0	4	2	55
Luminal B	2	1	6	1	8
HER2-enriched	1	3	2	0	2
Normal-like	0	1	2	0	4
Enriched functions of upregulated genes in subgroup	Hedgehog signaling; EMT; Wnt/β-catenin signaling	Interferon-γ response; IL6/JAK/STAT3 Signaling; Wnt & pluripotency	Interferon-γ response; Hedgehog signaling	ND[3]	
Histological & clinical features	Simple adenoma/carcinoma enriched; tumor with lymph node invasion enriched				Complex or mixed tumors enriched
Enriched functions of upregulated genes in subtype	Cell cycle; proliferation; MYC target; Wnt signaling; EMT; basal-like upregulation				Ion transport; ESR1 targets; basal-like downregulation
Mutation enrichment	*TP53* mutation				PIK3CA H1047R/L; *KAT6B* mutation; PI3K-AKT-mTOR pathway alteration
Cross-species PAM50	Clustered with hBLBC				Not clustered with any human breast cancer subtype

[1] cBLMT Canine basal-like mammary tumor

[2] cNBLMT canine non-basal-like mammary tumor

[3] cBLMT ER+PR- subgroup has a small sample size and is not investigated

observations are supported by multidimensional scaling of each cross-species PAM50 classification, as shown by the example provided in Fig. 2D. We then calculated the Mahalanobis distance [46] between the centers of canine and/or human subtypes for each of the 100 random sampling analyses. The distributions clearly indicate that the Mahalanobis distances between hBLBC and cBLMT are significantly shorter than those between hBLBC and human luminal A (hLumA) or cNBLMT, as well as those between hLumA and cBLMT or cNBLMT (Fig. 2E; Additional file 3: Table S2). These results support that cBLMT molecularly resembles hBLBC, but cNBLMT molecularly differs from hLumA.

The PAM50 classification analysis described above was performed with canFam3, for which we were able to identify only 43 PAM50 genes (see Methods). To determine whether the missing genes compromise the conclusions, we repeated the canine-only and cross-species PAM50 classifications by using canFam4, for which we identified 49 PAM50 genes (Additional file 3: Table S2). For canine-only classification, the subtype assignment between canFam3 and canFam4 agrees at 87% for luminal A and HER2-enriched, 78% for luminal B, and 71% for basal-like and normal-like (Additional file 3: Table S2). Importantly, cross-species classification with canFam 4 also reveals a significantly higher homology between cBLMT and hBLBC than between cNBLMT and hLumA (Additional file 1: Fig. S3A–E), consistent with the canFam3 analysis (Figs. 2A–E).

To validate this finding, we attempted to conduct the same analyses on the 57 tumors from the validation set [27]. PAM50 analysis of these canine tumors alone indeed classified a majority of the tumors as basal-like ($n=21$) or luminal A ($n=17$) (Additional file 1: Fig. S3F), consistent with the discovery set (Fig. 2A). Moreover, many of the basal-like canine tumors have a PAM50 gene expression pattern that closely matches hBLBCs (Additional file 1: Fig. S3G). However, likely due to having single-end RNA-seq data, these canine tumors could not co-cluster with human breast tumors (whose RNA-seq data are paired-end) in cross-species PAM50 classification even after batch correction.

We next performed the same analyses on the gene expression microarray data of two canine studies ($n=27$; $n=13$) and of one human study ($n=327$) [29–31]. Consistent with the RNA-seq analysis described above, canine-only PAM50 clustering classified most of these canine tumors as either basal-like or luminal A (Additional file 1: Fig. S3H). Moreover, cross-species PAM50 classification clustered cBLMTs and hBLBCs together in a majority of random sampling analyses (Additional file 1: Fig. S3H–I), further supporting the molecular homology between cBLMT and hBLBC.

Feature genes identified by machine learning support cross-species PAM50 classification results

The PAM50 genes used in our cross-species classification (Figs. 2A–E) were established in the past using human breast cancer studies [56, 57]. To avoid any biases from using this predefined gene set, we relied on machine learning to objectively identify genes whose expression patterns characterize the hBLBC or hLumA subtype. Briefly, among thousands of DE genes found between hBLBC/hLumA and other PAM50 subtypes, the machine learning model consistently picks up 115 genes for hBLBC and 130 genes for hLumA that differentiate the subtype from other PAM50 subtypes (Additional file 1: Fig. S3J–O; Additional file 3: Table S2). Importantly, the 115 hBLBC genes place cBLMTs significantly closer to hBLBCs, compared to cNBLMTs with hLumA tumors grouped by the 130 hLumA genes (Figs. 2F–G). This indicates a significantly closer homology between cBLMT and hBLBC, compared to between cNBLMT and hLumA, consistent with cross-species PAM50 analysis (Figs. 2A–E).

The cBLMT subtype captures key molecular features of hBLBC

We identified differentially expressed (DE) genes between cBLMT and cNBLMT. The 1123 genes upregulated in cBLMT are significantly enriched in cell cycle (e.g., DREAM targets, G2M checkpoint) and other functions that characterize hBLBC [3], as well as in genes that are known to be upregulated in hBLBC [51] (Fig. 3A; Table 1; Additional file 4: Table S3). Conversely, the 497 genes downregulated in cBLMT are significantly enriched in genes that are known to be downregulated in hBLBC [51], as well as in functions including ion transport and *ESR1* targets (Fig. 3A; Table 1; Additional file 4: Table S3). We conducted the same analysis with the validation set and observed similar findings (Additional file 1: Fig. S4A).

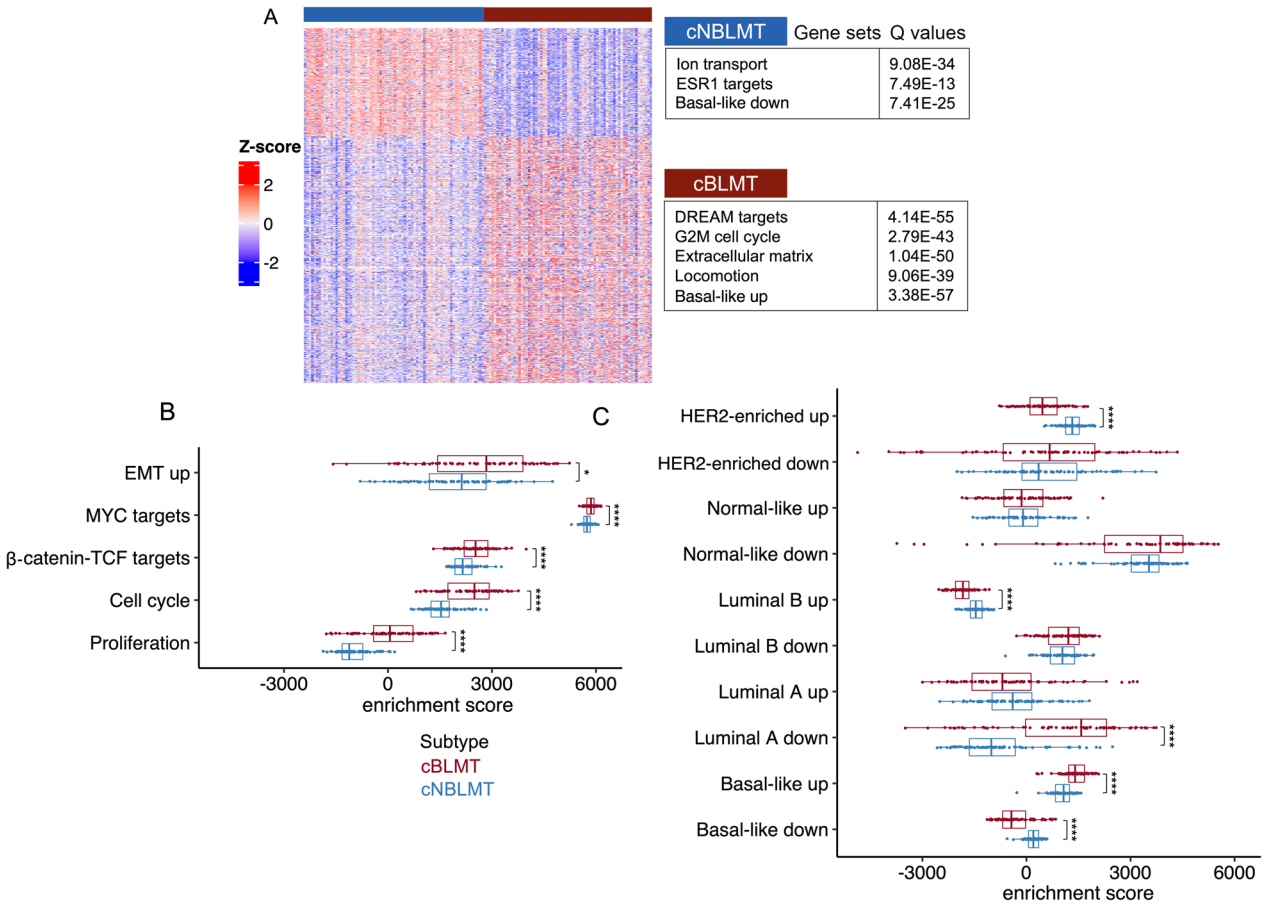

Fig. 3 Differentially expressed (DE) gene analysis indicates the enrichment of hBLBC signatures in cBLMT. **A**. Heatmap of the row scaled $\log_2$(TPM) values of 1,620 DE genes between cBLMT and cNBLMT samples, identified with an expression fold change of > 2 and a BH-adjusted *p*-value of < 0.01 for each DE gene (see Methods). The enriched functions among each DE gene group are indicated. **B** & **C** Distribution of single sample gene set enrichment analysis (ssGSEA) scores of canine tumors with hBLBC signature gene sets **B**, as well as with gene sets with expression patterns characterizing each human breast cancer subtype [51] shown **C**. P-values are from Wilcoxon tests. *: $p < 0.05$; ****: $p < 0.0001$

We then performed single-sample gene set enrichment analysis (ssGSEA) with the signature gene sets activated in hBLBC and gene sets known to be up- or downregulated in each human breast cancer subtype [3, 51, 52]. This analysis indicates cell cycle, cell proliferation, MYC targets, β-catenin-TCF targets, and epithelial–mesenchymal transition (EMT) genes are all significantly more upregulated in cBLMTs, compared to cNBLMTs (Fig. 3B). Moreover, the gene set known to be highly expressed in hBLBC is significantly upregulated in cBLMTs, while the gene set known to be lowly expressed in hBLBC is significantly downregulated in cBLMTs [51] (Fig. 3C). This pattern is, however, not observed in cNBLMTs, as the gene set known to highly expressed in hLumA tumors [51] does not show significant upregulation in cNBLMTs (Fig. 3C). These results are largely supported by findings with the validation set (Additional file 1: Figs. S4B–C). The analyses indicate that cBLMT captures key molecular features of hBLBC examined, while cNBLMT fails to do so with those of hLumA.

The cBLMT subtype contains ER−PR−, ER−PR+, and ER+ PR+ tumors

hBLBCs consist of approximately 70% ER-PR- (expressing neither ER nor PR) and 30% ER+PR− tumors, with ER−PR+and ER+PR+tumors being extremely rare, based on the *ESR1* and *PGR* transcript abundance levels (Figs. 4A–B). However, among cBLMTs, ER−PR+and ER+PR+tumors are significantly more frequent, accounting for 33% and 29%, respectively, while ER−PR− and ER+PR− tumors only make up 29% and 9%, respectively (Fig. 4C–D; Table 1). Notably, about 78% of ER−PR+cBLMTs were classified as basal-like in PAM50 analysis, similar to the percentage for ER−PR− cBLMT

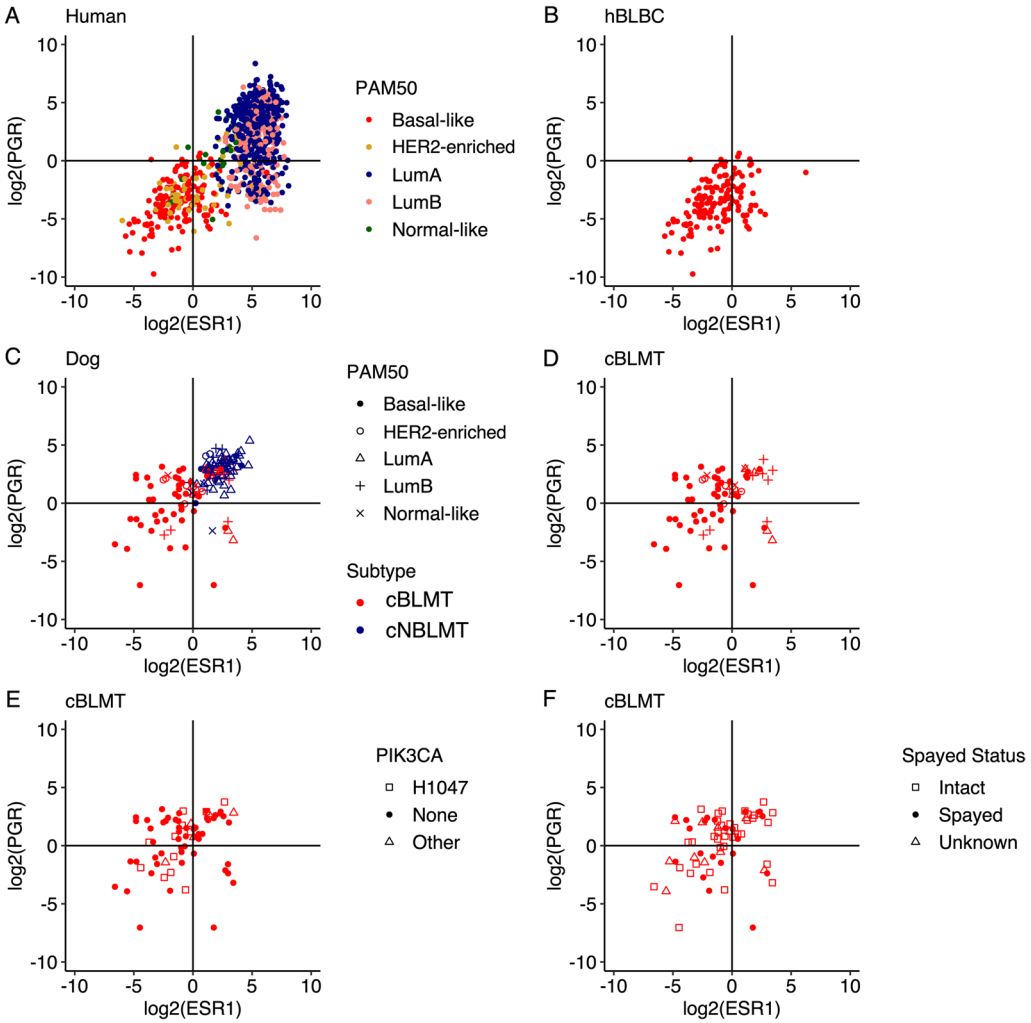

Fig. 4 cBLMT contains more ER-PR+tumors than hBLBC. Scatter plots show the $\log_2$(FPKM) values of *ESR1* and *PGR* for all TCGA human breast cancers with a PAM50 subtype ($n=827$) **A**, hBLBCs ($n=140$) **B**, both cBLMTs and cNBLMTs ($n=143$) **C**, as well as cBLMTs ($n=69$) with the PAM50 subtype **D**, *PIK3CA* mutation status **E**, dog's spayed status **F** indicated

(85%) (Table 1; Additional file 5: Table S4). PAM50 classification also categorized 30% of ER+PR+cBLMTs as basal-like, a proportion significantly higher than in cNBLMTs (7%) (96% of cNBLMTs are ER+PR+) (Table 1; Additional file 5: Table S4). The observation of higher *PGR* expression in cBLMTs and canine tumors than in human breast tumors (Fig. 4A–D) is supported by analysis using canFam4 (Additional file 1: Fig. S5A–B), as well as by other canine and human studies [27, 30, 31, 58] (Additional file 1: Fig. S5C–F). We noted no significant differences in the dog's spayed status or the *PIK3CA* mutation status among the ER±PR±cBLMT subgroups (Fig. 4E–F), indicating that the higher *PGR* expression is unlikely to be associated with either factor.

All cBLMT subgroups capture key molecular features of hBLBC

We compared each of ER−PR−, ER−PR+, and ER+PR+cBLMT subgroups to cNBLMT (the ER+PR− subgroup contains only 6 samples and was thus excluded from the analysis; see Table 1). We found that the DE genes are enriched in functions that characterize hBLBC [3, 7, 59] (Figs. 5A–C; Table 1). Briefly, cell cycle and Wnt signaling are enriched among upregulated genes in each cBLMT subgroup, and the enrichment is especially significant in ER−PR− and ER−PR+cBLMTs (Figs. 5A–C; Additional file 6: Table S5). Signatures for impaired BRCA2 function and p53 signaling, both known features of hBLBC [3], are also enriched among upregulated genes in ER−PR− and ER−PR+cBLMTs (Figs. 5A–C; Table 1; Additional file 6: Table S5).

Metabolic reprogramming is frequent in cancers, and we investigated 44 reprogrammed metabolic pathways reported in human cancers [60] (Additional file 6: Table S5). The analysis indicates that retinol metabolism is downregulated, while keratan sulfate synthesis is upregulated in all cBLMT, cNBLMT, hBLBC and hLumA (Additional file 1: Fig. S6). Notably, serine synthesis and purine de novo synthesis, two pathways known to increase stemness (hence basal-like features) in breast cancer cells [61, 62], are upregulated in both hBLBCs and cBLMTs (no significant difference was observed among the subgroups) (Additional file 1: Fig. S6). The analysis provides another piece of data supporting the cBLMT-hBLBC homology.

ER−PR+ cBLMTs harbor upregulated INF-γ response genes but downregulated *IFNG*

We performed DE analysis between cBLMT subgroups and found additional hBLBC characteristics [2, 3, 7, 63, 64] specific to each subgroup (Figs. 5D–F; Table 1). Compared to ER+PR+and ER−PR+cBLMTs, upregulated genes in ER−PR− cBLMTs are significantly enriched in functions such as EMT (Fig. 5D–E). Meanwhile, upregulated genes in ER+PR+and ER−PR+cBLMTs are enriched in functions including interferon-gamma (INF-γ) response (Fig. 5D–E). Other notable findings include that Wnt-signaling-initiated pluripotency genes and IL6/JAK/STAT3 signaling genes are significantly upregulated in ER−PR+cBLMTs, compared to ER-PR- cBLMTs (Fig. 5D–E). Hedgehog signaling genes are significantly upregulated in both ER−PR− and ER+PR+cBLMTs, compared to ER−PR+cBLMTs (Fig. 5D–E).

Interestingly, while the INF-γ response genes are upregulated in both ER+PR+and ER−PR+cBLMTs (Fig. 5D–E), the INF-γ gene (*IFNG*) itself is significantly downregulated in ER−PR+cBLMTs than in ER+PR+cBLMTs (Fig. 6A). This indicates that T cell exhaustion may occur in ER−PR+cBLMTs [63]. To investigate this possibility, we examined the expression of 8 known T cell exhaustion markers [65], but did not find a significant difference among the three subgroups (Fig. 6A).

As 6 out of 8 T cell exhaustion markers, including *PDCD1* (encoding PD-1), express higher in ER−PR+ and/or ER+PR+cBLMTs (Fig. 6A), we examined the association of each marker with *PGR* or *ESR1* in expression. We found that four markers, including *PDCD1*, *HAVCR2*, *CTLA4*, and *TIGIT*, have a significant positive association with *PGR* in ER−PR+cBLMTs (Fig. 6B; Additional file 1: Fig. S7A). No such associations were found for *PGR* in other cBLMT subgroups or the cNBLMT subtype, or for *ESR1* in any cBLMT subgroup or cNBLMT (Figs. 6B–C; Additional file 1: Fig. S7A–B).

PGR expression is associated with gene silencing

To further understand PR in canine tumors, we investigated genes that correlate with *PGR* or *ESR1* in transcript abundance in both human and canine tumors. The same as in hLumA tumors, over 1000 genes were found to be positively correlated with *ESR1* in cNBLMTs (Fig. 7A). Importantly, both sets of genes are enriched in the same functions, including cell cycle (Fig. 7A; Additional file 7: Table S6). For *PGR*, about 2.7 times as many (776 versus 289) positively correlated genes were identified for cNBLMT than hLumA (Fig. 7A). Moreover, while the *PGR*-correlated 289 genes in hLumA tumors are enriched in largely the same functions as those of the *ESR1*-correlated genes, the 776 *PGR*-correlated genes in cNBLMTs are enriched in interferon-γ response (Fig. 7A; Additional file 7: Table S6).

Fewer positively correlated genes with *ESR1* or *PGR* were identified in basal-like tumors in both species. For *ESR1*, we found about 400 genes in cBLMTs, which are enriched in functions including fatty acid metabolisms, oxidative phosphorylation, and SUZ12 targets, and 82

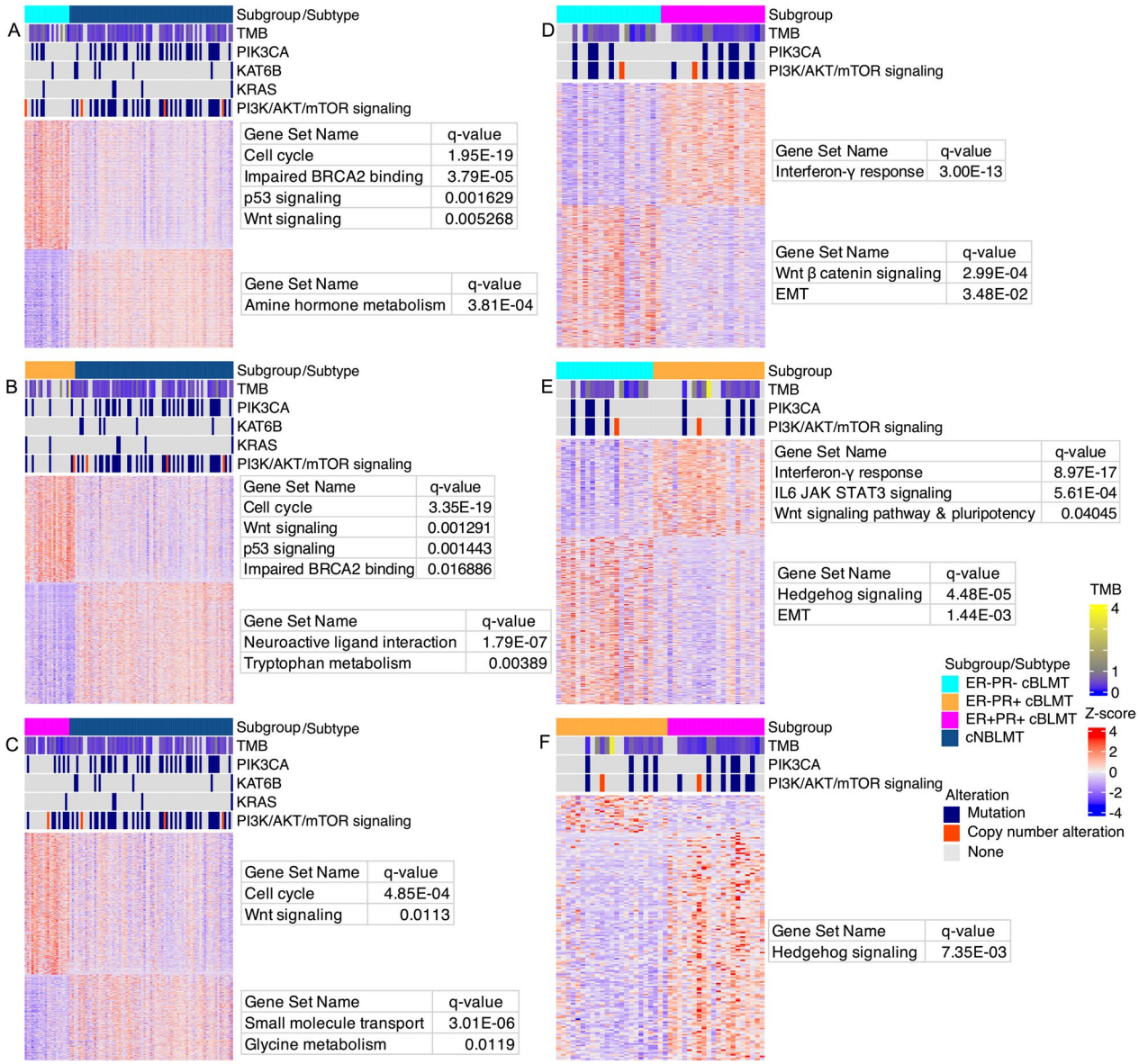

Fig. 5 Basal-like features are maintained in all cBLMT subgroups, especially in ER−PR+ and ER−PR− cBLMTs. **A**–**C**. Heatmaps showing the DE genes identified between cNBLMT and each of the ER−PR− **A**, ER−PR+ **B**, and ER + PR + **C** cBLMT subgroups, with a BH-adjusted *p*-value of < 0.05 and an expression-fold change of > 2. The heatmaps are presented as described in Fig. 3A, along with TMB and the most mutated genes indicated as described in Fig. 1. A tumor was classified ER+ or PR+ if its *ESR1* or *PGR* gene has a FPKM value of > 1, respectively; otherwise, the tumor was classified ER− or PR−. **D**–**F**. Heatmaps of DE genes identified between ER ± PR ± cBLMT subgroups, presented as described in **A**–**C**

genes in hBLBCs, which are not enriched in any specific functions (Fig. 7A; Additional file 7: Table S6). For *PGR*, while no genes were found in hBLBC, about 300 genes in cBLMT were identified and are enriched in interferon-*γ* response and GATA1 targets (Fig. 7A; Additional file 7: Table S6).

Negatively correlated genes show a larger dog–human difference, especially for the basal-like subtype and *PGR*. For *ESR1* in non-basal-like subtypes, 394 genes in hLumA tumors, enriched in functions including EMT, KRAS signaling, and SUZ12 targets, and 38 genes in cNBLMTs, enriched in functions such as oxidative phosphorylation, were identified (Fig. 7B; Additional file 7: Table S6). However, in basal-like subtypes, only 2 genes were negatively correlated with *ESR1* in hBLBCs, compared to 58 in cBLMTs that are enriched in cell cycle-related functions such as mitotic spindle and G2M checkpoints (Fig. 7B; Additional file 7: Table S6). For *PGR*, while no genes were

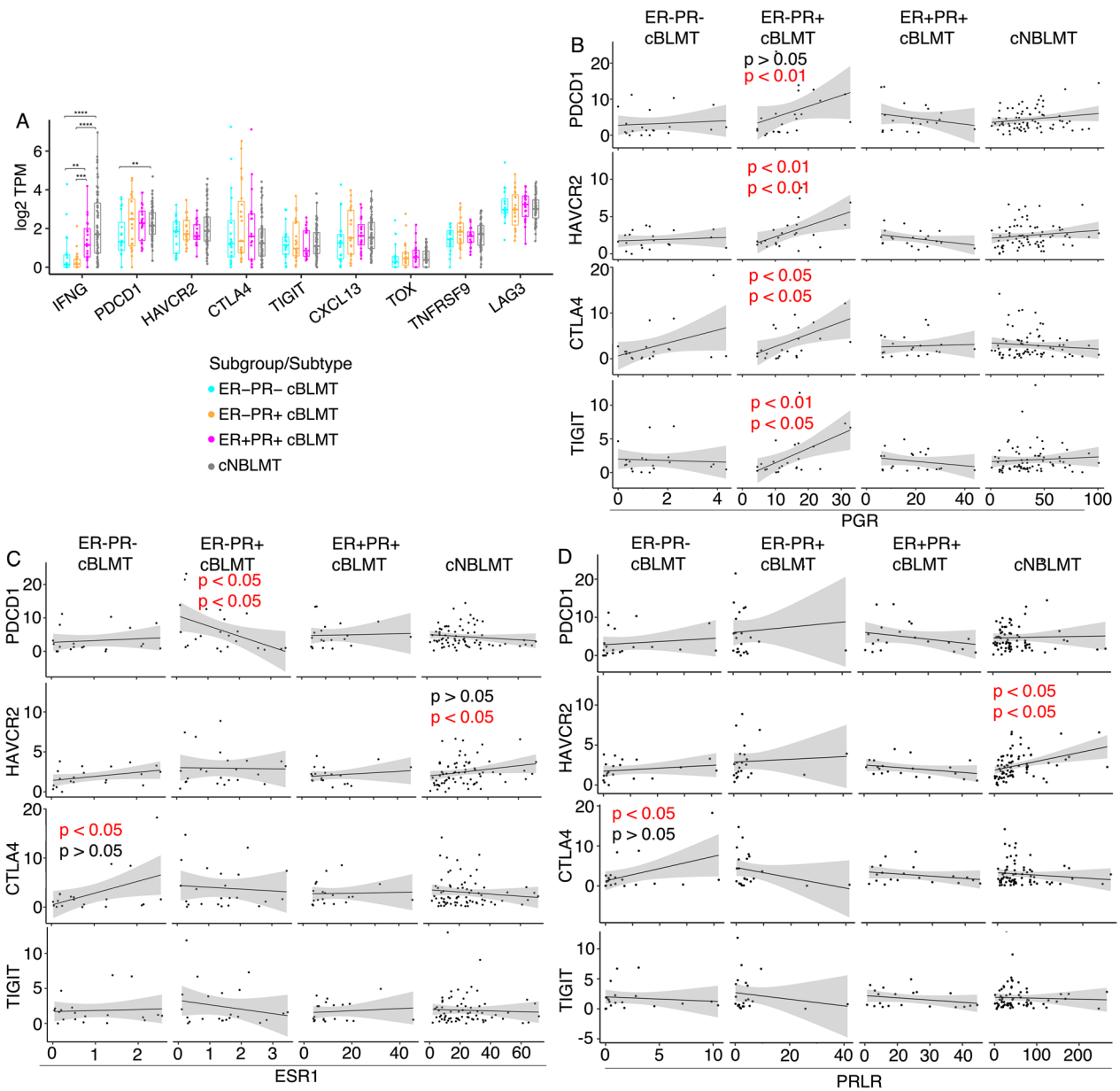

Fig. 6 *PGR* correlates with several T cell exhaustion signature genes in mRNA expression in ER−PR+ cBLMTs. **A** Dot plots of $\log_2$(TPM) values of *IFNG* and 8 canonical T cell exhaustion marker genes in each cBLMT subgroup and cNBLMT. P-values are from Wilcoxon tests. **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$. **B**–**D** Pearson (top) and Spearman (bottom) correlation analysis between *PGR* **B**, *ESR1* **C**, or *PRLR* **D** and *PDCD1*, *HAVCR2*, *CTLA4*, or *TIGIT* in mRNA expression in each subgroup and subtype shown. Only significant Pearson and/or Spearman correlations have p-values indicated

identified in either hLumA or hBLBC tumors, 142 genes in cNBLMT, enriched in functions such as hypoxia and glycolysis, and 71 genes in cBLMT, not enriched in any specific functions, were found (Fig. 7B; Additional file 7: Table S6).

PRLR has different features from PGR

Prolactin (also called luteotropin) is described as the third hormone in breast cancer [66]. We hence examined the expression of the prolactin gene *PRL* and the prolactin receptor gene *PRLR*. We found that *PRL* is essentially not expressed in any canine tumors ($FPKM < 1$) (Additional file 1: Fig. S7D). *PRLR*, however, is expressed in many canine tumors ($1 < FPKM < 60$) (Additional file 1: Fig. S7E). Among the canine subgroups/subtypes, the expression level of *PRLR* is the lowest in ER−PR− cBLMT tumors and increases by approximately 3-, 24-, and 44-fold

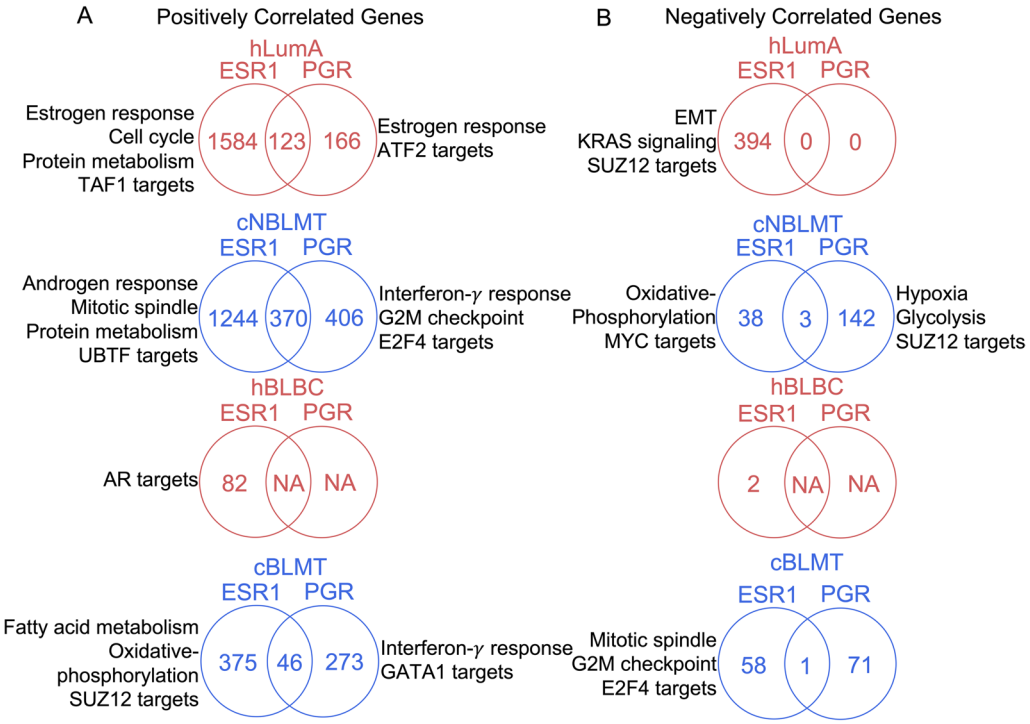

Fig. 7 *PGR* is associated with gene silencing in canine tumors but not in human tumors. **A** Venn diagrams indicating the number of genes that are positively correlated with *ESR1* and/or *PGR* in mRNA expression in each human or canine subtype specified. These genes were identified as those having correlation coefficient $R > 0.3$ and BH-adjusted $p < 0.05$ in both Pearson and Spearman correlation analyses. Also indicated are the top enriched functions of genes that are correlated with only *ESR1* or *PGR*. **B** Venn diagrams for genes negatively correlated with *ESR1* and/or *PGR*, identified with $R < -0.3$ and other cutoffs and presented as described in **A**

in ER−PR+cBLMT, ER+PR+cBLMT, and cNBLMT (> 90% of which are ER+PR+) samples, respectively (Additional file 1: Fig. S7E). Accordingly, *PRLR* expression is significantly positively correlated with *ESR1* and *PGR* expression in these canine tumors (Additional file 1: Fig. S7F).

Resembling canine mammary tumors, *PRL* is largely not expressed while *PRLR* is expressed in human breast cancers (Additional file 1: Fig. S7D–E). However, for the matched subtypes and subgroups, *PRLR* is expressed higher in human tumors than in canine tumors (Additional file 1: Fig. S7D–E). *PRLR* has lower positive correlation with *ESR1* and *PGR* in human tumors compared to canine tumors (Additional file 1: Fig. S7G).

To determine whether *PRLR* is correlated with T cell exhaustion markers, we repeated the same analysis with *ESR1* and *PGR* as shown in Figs. 6B–C. We found no significant correlations for *PRLR*, unlike *PGR* (Fig. 6D; Additional file 1: Fig. S7C). Furthermore, we observed no clear association of *PRLR* with gene silencing in cBLMT, differing from *PGR* (Additional file 1: Fig. S8).

Discussion

Taking advantage of the recently published RNA-seq data for hundreds of canine mammary tumor cases, we performed, to our knowledge, the first genome-wide and independent (not using any known biomarkers) subtyping of this cancer common in bitches that are intact or spayed late. The study identifies two subtypes and further shows that one subtype molecularly resembles hBLBC, while the other subtype appears not to match any human breast cancer subtypes. This conclusion is consistent with our previous study [14], but differs from a recent publication reporting that canine and human luminal A tumors have more molecular homology than canine and human basal-like tumors [26]. While our canine-only PAM50 analysis also classifies most tumors of one canine subtype as luminal A, the same as the study by Bergholtz et al. [26], our cross-species PAM50 analysis clearly separates canine luminal A tumors from hLumA tumors, unlike basal-like tumors (as such, we named the two canine subtypes as basal-like, cBLMT, and non-basal-like, cNBLMT) (note that Bergholtz et al. [26] did not perform cross-species PAM50 classification). Importantly, our conclusion is supported by PAM50-independent

classification using hBLBC and hLumA feature genes identified by machine learning. We further show that cBLMTs capture key molecular features and expression patterns of hBLBCs, whereas cNBLMTs fail to do the same with hLumA tumors. Our results are supported by the histology of the canine subtypes. cBLMTs are enriched in simple carcinomas and simple adenomas, where only one cell lineage prominently proliferates. cNBLMTs, however, are enriched in complex or mixed carcinomas and adenomas, where multiple cell lineages (e.g., epithelial cells and myoepithelial cells) proliferate [14, 19, 20]. Complex or mixed tumors (e.g., adenomyoepithelioma) are very rare in human breast cancer [14, 21, 22]; thus, cNBLMTs do not match hLumA tumors histologically.

Our findings have significant biological implications. The strong molecular homology between hBLBC and cBLMT revealed by this study provides the fundamental justification to use our powerful dog–human comparison strategy [67] for cancer driver—passenger discrimination, a central aim of cancer research [68], for hBLBC. Briefly, hBLBCs are characterized with extensive copy number alterations, and harbor thousands of amplified or deleted genes [3, 4], some being drivers (cancer-causative) and many others being passengers (not cancer-causative). Studies by us [33] and others [69] have also detected hundreds of amplified or deleted genes in canine mammary tumors. In our dog–human comparison strategy [67], drivers are those genes amplified/deleted in both hBLBC and cBLMT, while passengers are those genes amplified/deleted only in hBLBC and located near a dog–human genomic rearrangement breakpoint. Our strategy has been successfully applied to colorectal cancer [67], and its application to hBLBC will equally yield new insights into hBLBC carcinogenesis.

Our findings have significant clinical implications. About 70% of hBLBCs are triple negative (ER-, PR-, and HER2 not enriched). Thus, hormone therapy and anti-HER2 drugs are often not applicable, and hBLBC tends to have the poorest prognosis among the PAM50 subtypes. Effective treatments of hBLBCs are hence urgently needed. Our study shows that cBLMTs can serve as a much-needed translational model to speed up anti-hBLBC drug discovery. Due to the large dog population (> 80 million pet dogs in the US alone), more similar body size between humans and dogs, more relaxed FDA regulation, and, most importantly, the strong dog–human molecular homology, cBLMTs can be a stepping stone to bridge a current gap between preclinical research (mostly using cell lines and rodents) and human clinical trials, accelerating bench-to-bedside translation. The hBLBC driver discovery, as described above, can also yield targets for hBLBC intervention.

Although cBLMT co-clusters with hBLBC in PAM50 classification and captures the key molecular features and gene expression patterns of hBLBC, cBLMT contains ER−PR+ and ER+PR+tumors, which are rare in hBLBC. Moreover, while ER+PR− tumors are common in hBLBC and other breast cancer subtypes, ER−PR+ tumors are nearly nonexistent in any human breast cancer subtype. This is because that in human breast cells, PR is induced by ER [70, 71], and thus, without ER, PR will not be expressed. One notable difference between the estrous cycle in bitches and the menstrual cycle in women is the luteal phase, which lasts 14 days for humans but 2 months for dogs. As the canine mammary glands are constantly exposed to a high level of progesterone during the luteal phase [72, 73], it is possible that the *PGR* gene is still actively transcribed after the *ESR1* gene is silenced in dogs, resulting in the ER−PR+ tumors. More studies are needed to investigate this possibility.

Despite the difference in the PR expression status, ER+PR+ and ER−PR+ cBLMTs capture key molecular features of hBLBCs, the same as ER−PR− cBLMTs. These include upregulation of cell cycle genes and Wnt signaling. Upregulation of cell cycle genes could lead to high cell proliferation, a molecular characteristic of hBLBC [3]. Activated Wnt signaling is also a well-known feature of hBLBC [7, 59]. For example, WNT5B, one of the major Wnt signaling molecules, is known to drive the hBLBC phenotype, both in vitro and in vivo, by activating both canonical and non-canonical Wnt signaling [74]. *WNT5B* is upregulated in both ER−PR+ and ER+PR+cBLMTs. Wnt signaling-driving pluripotency genes are upregulated in ER-PR+cBLMTs, which likely further drives the basal-like features of these tumors [59].

One notable finding from our study is that in ER−PR+cBLMTs, interferon-γ response genes are upregulated, but the interferon-γ gene (*IFNG*) itself is downregulated. Moreover, ER-PR+cBLMTs appear to express more of the immune checkpoint genes *PDCD1* (encoding PD-1) and *CTLA4*, and only in these tumors, *PGR* is positively correlated with *PDCD1* and *CTLA4*. These results are consistent with T cell exhaustion, where T cells are hypofunctional [75]. T cell exhaustion occurs in many human cancers, including hBLBCs [63] and presents challenges and opportunities in cancer immunotherapy [75]. Due to the small sample size, we cannot conclude definitively that T cell exhaustion indeed occurs in cBLMTs. Once this possibility is validated with further studies, cBLMTs could be a valuable model to investigate the relationship among progesterone, PR, and T cell exhaustion. Importantly, cBLMTs may be good models to test novel immunotherapies targeting T cell exhaustion [75].

Our study reveals that unlike *ESR1*, *PGR* is associated with gene silencing in canine mammary tumors. However, the silenced genes appear to be random and not enriched in any particular functions, especially in ER−PR+ cBLMTs. Interestingly, we find that many of the *ESR1* or *PGR*-correlated genes are targets of SUZ12, a component of polycomb repressive complex 2 (PRC2) that primarily methylates lysine 27 of histone H3 (e.g., H3K27me3, a marker of transcriptionally silent chromatin). Further studies are needed to determine whether PRC2 is responsible for *PGR*-associated gene silencing.

PRLR, encoding the receptor for prolactin (which is also called luteotropin, and has been described as the third hormone in breast cancer [66]) resembles *ESR1* more than *PGR* in our study. For example, *PRLR* is not associated with T cell exhaustion markers and gene silencing. However, more studies are needed to understand the role of *PRLR* and prolactin in canine mammary tumors, including their protein expression levels among the subtypes and subgroups.

Conclusions

We identify two molecular subtypes in spontaneous canine mammary tumors. One subtype, cBLMT, molecularly and histologically resembles hBLBC, a breast cancer subtype that lacks an effective treatment and has the worst clinical outcomes. The other subtype, cNBLMT, appears not to match any human breast cancer subtype molecularly and histologically. While cBLMTs also consist of ER−PR+and ER+PR+ tumors (which may be related to the long luteal phase of the estrous cycle in dogs), a difference from hBLBC, we note that these tumors capture the key molecular features of hBLBCs, the same as ER−PR− cBLMTs. Thus, cBLMTs could serve as a much-needed spontaneous animal model for hBLBC, filling a critical gap in breast cancer research. Moreover, while much more studies are needed, ER−PR+cBLMTs may provide a valuable system to study T cell exhaustion, as well as estrogen/ER-independent roles of progesterone and PR in gene silencing.

Abbreviations

NMF	Non-negative matrix factorization
GEO	Gene Expression Omnibus
NCI	National Cancer Institute
GDC	Genomic Data Commons
TCGA	The Cancer Genome Atlas
cBLMT	Canine basal-like mammary tumors
cNBLMT	Canine non-basal-like mammary tumors
ER	Estrogen receptor
PR	Progesterone receptor
PRLR	Prolactin receptor
PRL	Prolactin
TPM	Transcripts per million
FPKM	Fragments per kilobase of exon per million mapped
QC	Quality control
CDS	Coding sequence
TMB	Tumor mutation burden
hBLBC	Human basal-like breast cancer
hLumA	Human luminal A breast cancer
EMT	Epithelial–mesenchymal transition
DE	Differentially expressed
PRC2	Polycomb repressive complex 2

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-023-01705-5.

Additional file 1 Fig. S1. RNA-seq quality control. **Fig. S2.** Validation of canine mammary tumor subtyping results shown in Fig. 1 using different strategies and data set. **Fig. S3.** Dog-alone and cross-species PAM50 classification using canFam4 and its annotation; hBLBC- and hLumA feature gene identification via machine learning. **Fig. S4.** Differentially expressed (DE) gene analysis indicates the enrichment of hBLBC signatures in cBLMT of the validation set. **Fig. S5.** Canine tumors, especially cBLMTs, express PGR more abundantly than hBLBCs. **Fig. S6.** Purine de novo synthesis and serine synthesis are more activated in cBLMTs and hBLBCs, compared to normal mammary tissues. **Fig. S7.** PGR, but not ESR1 or PRLR, correlates with several T-cell exhaustion signature genes in mRNA expression in ER-PR+ cBLMTs; PRL and PRLR expression in canine and human tumors. **Fig. S8.** PRLR positively correlates with many genes in mRNA expression, but is not associated with gene silencing in cBLMT

Additional file 2 Table S1, An file containing sample metadata and expression data used for Figs. 1, Additional file 1: Fig. S1, and S2.

Additional file 3 Table S2, An file containing PAM50 genes, PAM50 classification, cross-species PAM50 analysis data, and ML-identified feature genes used for Figs. 2 and Additional file 1: Fig. S3.

Additional file 4 Table S3, An file containing DE genes, enriched functions, and ssGSEA scores used for Figs. 3 and Additional file 1: Fig. S4.

Additional file 5 Table S4 An file containing ESR1 and PGR expression data of human and canine tumors used for Figs. 4 and Additional file 1: Fig. S5.

Additional file 6 Table S5 An file containing DE genes and enriched functions and pathways, and the metabolic signature genes for each pathway used for Figs. 5 and Additional file 1: Fig. S6.

Additional file 7 Table S6 An file containing ESR1/PGR/PRLR-correlated genes used for Figs. 7 and Additional file 1: Fig. S8.

Additional file 8 TPM data used for NMF, clustering, and other analyses of this manuscript.

Acknowledgements

We thank Ms. Jin Qian for her contribution to the study; Dr. Holly Borghese and Dr. William C. Kisseberth for their help in sample collection; the Georgia Advance Computing Resource Center (GACRC) for providing the computing power for this work; Dr. Michael Skaro for advice on figure design; and the authors who published canine and human data used in this work.

Author contributions

SZ, TF, KD, and JW conceived the experiment design. JW and SZ wrote the manuscript. KH performed all QC analysis and contributed to the machine learning study. YF performed DE analysis and plotted Fig. 3A. KD provided advice on statistical analyses. TF performed initial NMF and PAM50 classification. TM developed the machine learning pipeline. JW conducted all other analyses.

Funding

This research was funded by NCI R01 CA252713 and R01 CA182093. The Ohio State University College of Veterinary Medicine Biospecimen Repository, one of the sample sources, is supported by NCATS UL1TR001070 and NCI P30CA016058.

Availability of data and materials
The RNA-seq data generated from this study are submitted to the SRA database at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA912710. All other RNA-seq and microarray data sets analyzed in this study are obtained from the SRA and GEO databases at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA489087/, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA203086/, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA561580, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20718, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22516, and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse20685. TPM data used in this manuscript are included as a compressed zip file in Additional file 8.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Bioinformatics, University of Georgia, 120 E Green Street, Athens, GA 30602, USA. [2]Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, 120 E Green Street, Athens, GA 30602, USA. [3]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK. [4]Department of Biostatistics, University of Georgia, Athens, GA 30602, USA.

Received: 28 February 2023 Accepted: 31 August 2023
Published online: 03 October 2023

References

1. Polyak K. Heterogeneity in breast cancer. J Clin Invest. 2011;121(10):3786–8.
2. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163(2):506–19.
3. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
4. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
5. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, et al. The somatic mutation profiles of 2433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.
6. Fu Z, Chen S, Zhu Y, Zhang D, Xie P, Jiao Q, Chi J, Xu S, Xue Y, Lu X, Song X. Proteolytic regulation of CD73 by TRIM21 orchestrates tumor immunogenicity. Sci Adv. 2023;9(1):eadd6626.
7. Sharma M, Castro-Piedras I, Rodgers AD, Pruitt K. Genomic profiling of DVL-1 and its nuclear role as a transcriptional regulator in triple negative breast cancer. Genes Cancer. 2021;12:77–95.
8. Liao L, Zhang YL, Deng L, Chen C, Ma XY, Andriani L, Yang SY, Hu SY, Zhang FL, Shao ZM, et al. Protein phosphatase 1 subunit PPP1R14B stabilizes STMN1 to promote progression and paclitaxel resistance in triple-negative breast cancer. Cancer Res. 2023;83(3):471–84.
9. Reddy TP, Rosato RR, Li X, Moulder S, Piwnica-Worms H, Chang JC. A comprehensive overview of metaplastic breast cancer: clinical features and molecular aberrations. Breast Cancer Res. 2020;22(1):121.
10. Bu W, Liu ZY, Jiang WY, Nagi C, Huang SX, Edwards DP, Jo E, Mo QX, Creighton CJ, Hilsenbeck SG, et al. Mammary precancerous stem and non-stem cells evolve into cancers of distinct subtypes. Can Res. 2019;79(1):61–71.
11. Dow S. A role for dogs in advancing cancer immunotherapy research. Front Immunol. 2019;10:2935.
12. Zeng L, Li W, Chen CS. Breast cancer animal models and applications. Zool Res. 2020;41(5):477–94.
13. Kwon JY, Moskwa N, Kang W, Fan TM, Lee C. Canine as a comparative and translational model for human mammary tumor. J Breast Cancer. 2023;26(1):1.
14. Liu D, Xiong H, Ellis AE, Northrup NC, Rodriguez CO Jr, O'Regan RM, Dalton S, Zhao S. Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. Cancer Res. 2014;74(18):5045–56.
15. Gray M, Meehan J, Martinez-Perez C, Kay C, Turnbull AK, Morrison LR, Pang LY, Argyle D. Naturally-occurring canine mammary tumors as a translational model for human breast cancer. Front Oncol. 2020;10:617.
16. Thamm DH. Canine cancer: strategies in experimental therapeutics. Front Oncol. 2019;9:1257.
17. Meuten DJ. Tumors in domestic animals. 4th ed. Ames, Iowa: Iowa State University Press; 2002.
18. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin. 2012;62(1):10–29.
19. Goldschmidt M, Pena L, Rasotto R, Zappulli V. Classification and grading of canine mammary tumors. Vet Pathol. 2011;48(1):117–31.
20. Nakagaki KY, Nunes MM, Garcia APV, Nunes FC, Schmitt F, Cassali GD. Solid carcinoma of the canine mammary gland: a histological type or tumour cell arrangement? J Comp Pathol. 2022;190:1–12.
21. Tan PH, Ellis IO. Myoepithelial and epithelial-myoepithelial, mesenchymal and fibroepithelial breast lesions: updates from the WHO classification of tumours of the breast 2012. J Clin Pathol. 2013;66(6):465–70.
22. Hayes MM. Adenomyoepithelioma of the breast: a review stressing its propensity for malignant transformation. J Clin Pathol. 2011;64(6):477–84.
23. Sassi F, Benazzi C, Castellani G, Sarli G. Molecular-based tumour subtypes of canine mammary carcinomas assessed by immunohistochemistry. BMC Vet Res. 2010;6:5.
24. Sorenmo KU, Kristiansen VM, Cofone MA, Shofer FS, Breen AM, Langeland M, Mongil CM, Grondahl AM, Teige J, Goldschmidt MH. Canine mammary gland tumours; a histological continuum from benign to malignant; clinical and histopathological evidence. Vet Comp Oncol. 2009;7(3):162–72.
25. Kim KK, Seung BJ, Kim D, Park HM, Lee S, Song DW, Lee G, Cheong JH, Nam H, Sur JH, et al. Whole-exome and whole-transcriptome sequencing of canine mammary gland tumors. Sci Data. 2019;6(1):147.
26. Bergholtz H, Lien T, Lingaas F, Sorlie T. Comparative analysis of the molecular subtype landscape in canine and human mammary gland tumors. J Mammary Gland Biol Neoplasia. 2022;27(2):171–83.
27. Graim K, Gorenshteyn D, Robinson DG, Carriero NJ, Cahill JA, Chakrabarti R, Goldschmidt MH, Durham AC, Funk J, Storey JD, Kristensen VN. Modeling molecular development of breast cancer in canine mammary tumors. Genome Res. 2021;31(2):337–47.
28. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.
29. Kao KJ, Chang KM, Hsu HC, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer. 2011;11:143.
30. Klopfleisch R, Lenze D, Hummel M, Gruber AD. The metastatic cascade is reflected in the transcriptome of metastatic canine mammary carcinomas. Vet J. 2011;190(2):236–43.
31. Klopfleisch R, Lenze D, Hummel M, Gruber AD. Metastatic canine mammary carcinomas can be identified by a gene expression profile that partly overlaps with human breast cancer profiles. BMC Cancer. 2010;10:618.
32. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.
33. Alsaihati BA, Ho KL, Watson J, Feng Y, Wang T, Dobbin KK, Zhao S. Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. Nat Commun. 2021;12(1):4670.
34. Wang J, Wang T, Sun Y, Feng Y, Kisseberth WC, Henry CJ, Mok I, Lana SE, Dobbin K, Northrup N, et al. Proliferative and invasive colorectal tumors in pet dogs provide unique insights into human colorectal cancer. Cancers. 2018;10(9):330.

35. Wang T, Kwon SH, Peng X, Urdy S, Lu Z, Schmitz RJ, Dalton S, Mostov KE, Zhao S. A qualitative change in the transcriptome occurs after the first cell cycle and coincides with lumen establishment during MDCKII cystogenesis. iScience. 2020;23(10):101629.
36. Feng Y, Hess PR, Tompkins SM, Hildebrand WH, Zhao S. A Kmer-based paired-end read de novo assembler and genotyper for canine MHC class I genotyping. iScience. 2023. https://doi.org/10.1016/j.isci.2023.105996.
37. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
38. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
39. Liao Y, Smyth GK, Shi W. Feature counts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
40. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.
41. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.
42. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2(3):lqaa078.
43. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinform. 2010;11:367.
44. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2.
45. Wilkerson MD, Hayes DN. Consensus cluster plus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572–3.
46. R Core Team R, Team RC: R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2020. In.; 2021.
47. Galili T. Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015;31(22):3718–20.
48. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
49. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
50. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216-21.
51. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG, Foekens JA, Martens JW. Subtypes of breast cancer show preferential site of relapse. Cancer Res. 2008;68(9):3108–14.
52. Kuehn H, Liberzon A, Reich M, Mesirov JP. Using GenePattern for gene expression analysis. Curr Protoc Bioinform. 2008;7:7–12.
53. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
54. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw. 2010;36(11):1–13.
55. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, et al. Gene set knowledge discovery with enrichr. Curr Protoc. 2021;1(3):e90.
56. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7.
57. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52.
58. Mohr A, Luder Ripoli F, Hammer SC, Willenbrock S, Hewicker-Trautwein M, Kielbowicz Z, Murua Escobar H, Nolte I. Hormone receptor expression analyses in neoplastic and non-neoplastic canine mammary tissue by a bead based multiplex branched DNA assay: a gene expression study in fresh frozen and formalin-fixed, paraffin-embedded samples. PLoS ONE. 2016;11(9):e0163311.
59. Khramtsov AI, Khramtsova GF, Tretiakova M, Huo DZ, Olopade OI, Goss KH. Wnt/beta-catenin pathway activation is enriched in basal-like breast cancers and predicts poor outcome. Am J Pathol. 2010;176(6):2911–20.
60. Sun HY, Zhou Y, Skaro MF, Wu YR, Qu ZX, Mao FL, Zhao SW, Xu Y. Metabolic reprogramming in cancer is induced to increase proton production. Can Res. 2020;80(5):1143–55.
61. Samanta D, Semenza GL. Serine synthesis helps hypoxic cancer stem cells regulate redox. Cancer Res. 2016;76(22):6458–62.
62. Lv Y, Wang X, Li X, Xu G, Bai Y, Wu J, Piao Y, Shi Y, Xiang R, Wang L. Nucleotide de novo synthesis increases breast cancer stemness and metastasis via cGMP-PKG-MAPK signaling pathway. PLoS Biol. 2020;18(11):e3000872.
63. Guo L, Cao C, Goswami S, Huang X, Ma L, Guo Y, Yang B, Li T, Chi Y, Zhang X, et al. Tumoral PD-1hiCD8+ T cells are partially exhausted and predict favorable outcome in triple-negative breast cancer. Clin Sci. 2020;134(7):711–26.
64. Xu S, Feng Y, Zhao S. Proteins with evolutionarily hypervariable domains are associated with immune response and better survival of basal-like breast cancer patients. Comput Struct Biotechnol J. 2019;1(17):430–40.
65. Zheng L, Qin S, Si W, Wang A, Xing B, Gao R, Ren X, Wang L, Wu X, Zhang J, et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. Science. 2021;374(6574):abe6474.
66. Schuler LA, O'Leary KA. Prolactin: the third hormone in breast cancer. Front Endocrinol. 2022;13:910978.
67. Tang J, Li Y, Lyon K, Camps J, Dalton S, Ried T, Zhao S. Cancer driver-passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer. Oncogene. 2014;33(7):814–22.
68. Haber DA, Settleman J. Cancer: drivers and passengers. Nature. 2007;446(7132):145–6.
69. Borge KS, Nord S, Van Loo P, Lingjaerde OC, Gunnes G, Alnaes GI, Solvang HK, Luders T, Kristensen VN, Borresen-Dale AL, et al. Canine mammary tumours are affected by frequent copy number aberrations, including amplification of MYC and loss of PTEN. PLoS ONE. 2015;10(5):e0126371.
70. Lapidus RG, Nass SJ, Davidson NE. The loss of estrogen and progesterone receptor gene expression in human breast cancer. J Mammary Gland Biol Neoplasia. 1998;3(1):85–94.
71. Burstein HJ. Systemic therapy for estrogen receptor-positive, HER2-negative breast cancer. N Engl J Med. 2020;383(26):2557–70.
72. Concannon PW. Reproductive cycles of the domestic bitch. Anim Reprod Sci. 2011;124(3–4):200–10.
73. Sleeckx N, de Rooster H, Veldhuis Kroeze EJ, Van Ginneken C, Van Brantegem L. Canine mammary tumours, an overview. Reprod Domest Anim. 2011;46(6):1112–31.
74. Jiang S, Zhang M, Zhang Y, Zhou W, Zhu T, Ruan Q, Chen H, Fang J, Zhou F, Sun J, et al. WNT5B governs the phenotype of basal-like breast cancer by activating WNT signaling. Cell Commun Signal. 2019;17(1):109.
75. Chow A, Perica K, Klebanoff CA, Wolchok JD. Clinical implications of T cell exhaustion for cancer immunotherapy. Nat Rev Clin Oncol. 2022;19(12):775–90.

Publisher's Note