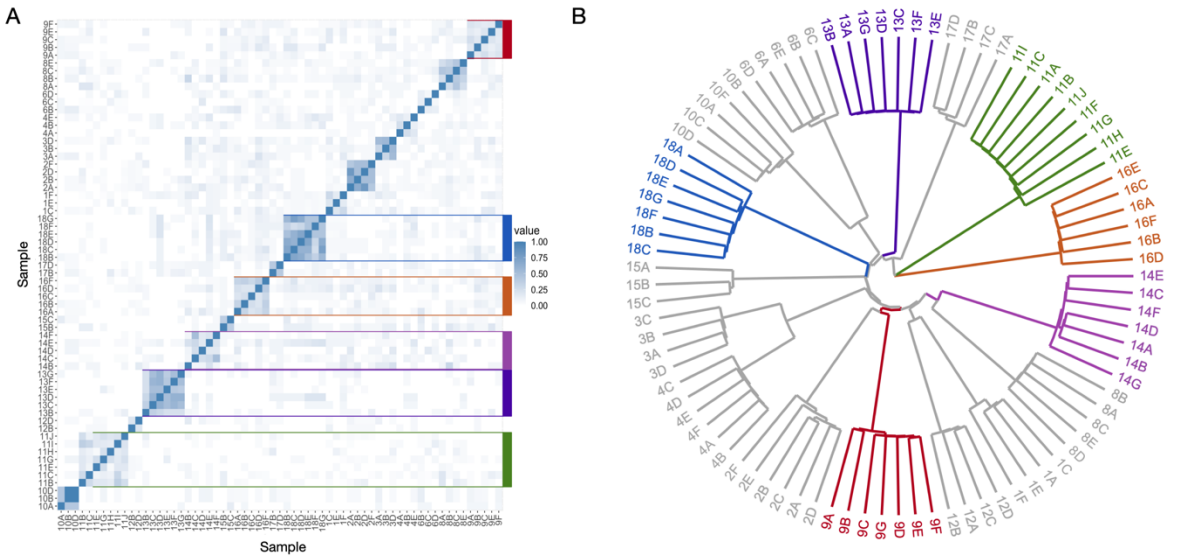
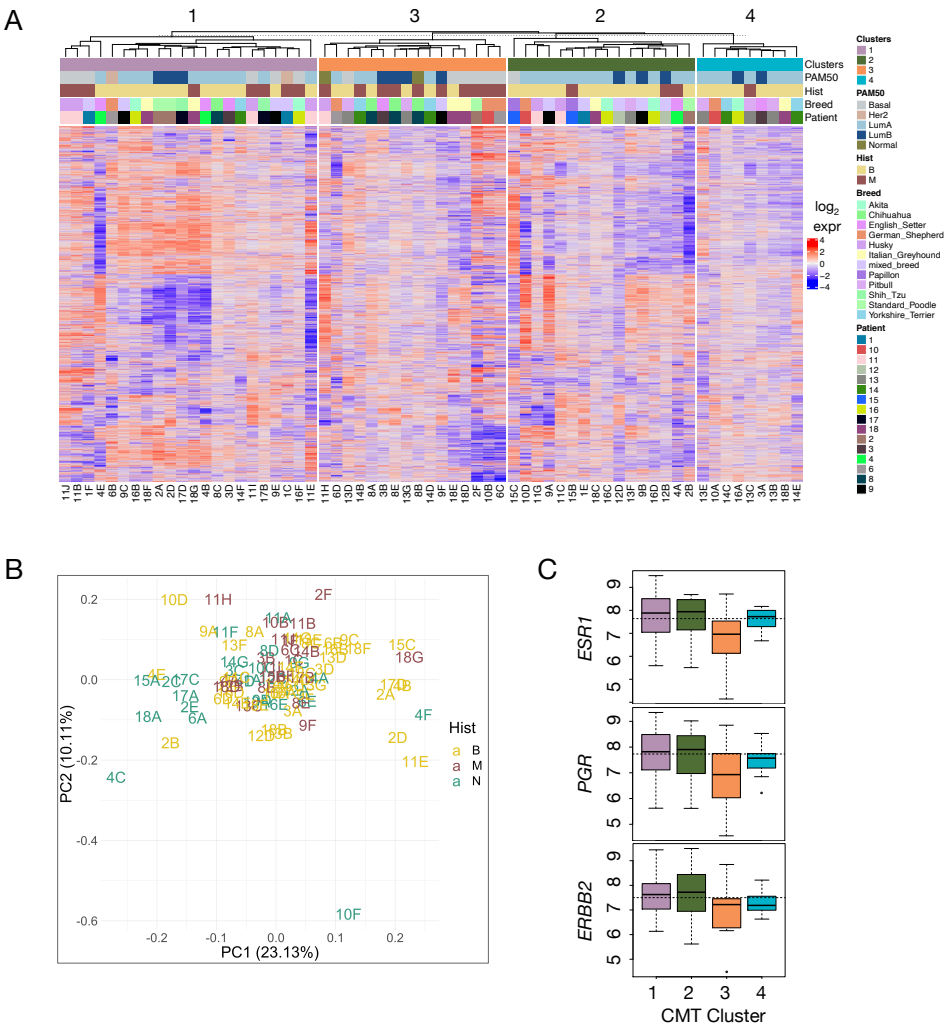
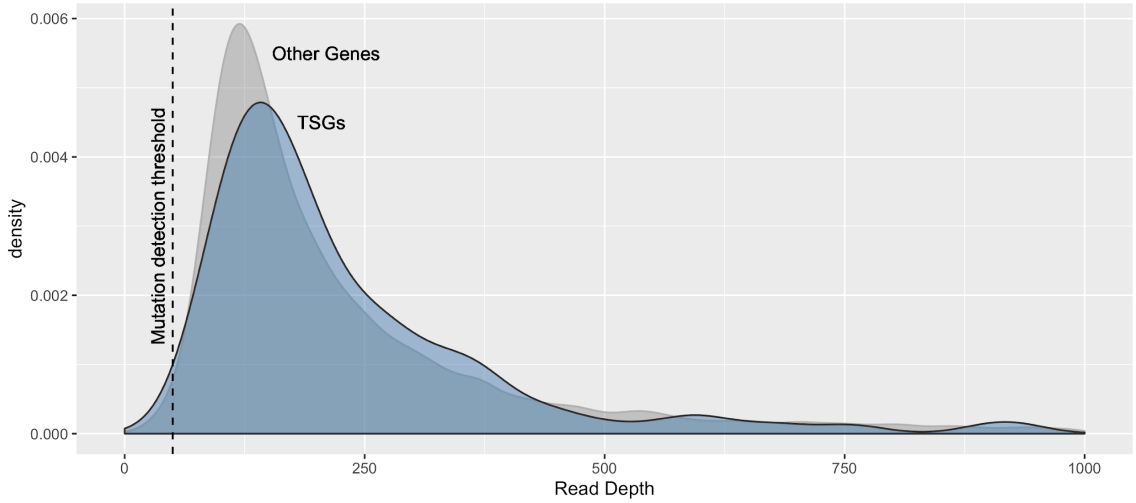
SUPPLEMENTAL MATERIALS FOR
Title: Modeling molecular development of breast cancer in canine mammary tumors



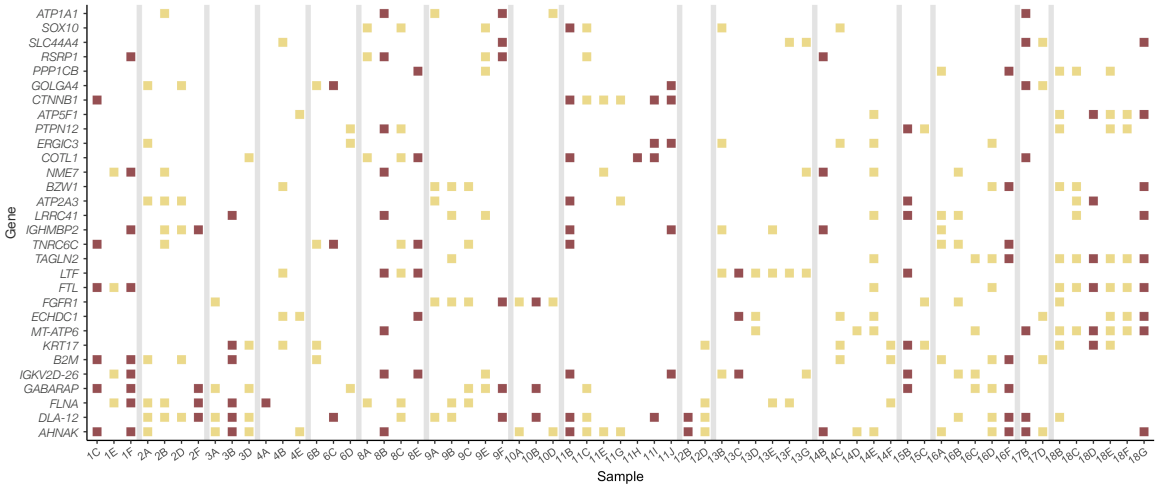
Supplemental Fig. S1: Mutational Independence of Tumors. (A) Mutational similarity scores between all pairs of tumors (see Methods). Samples outlined in different colors are dogs with at least 5 tumor samples. (B) Phylogeny with proportional branch lengths, calculated using SNPRelate (Manichaikul et al. 2010) and the SNPs called by FREYA for all samples (including normals). Branches are colored by dog to match panel A.



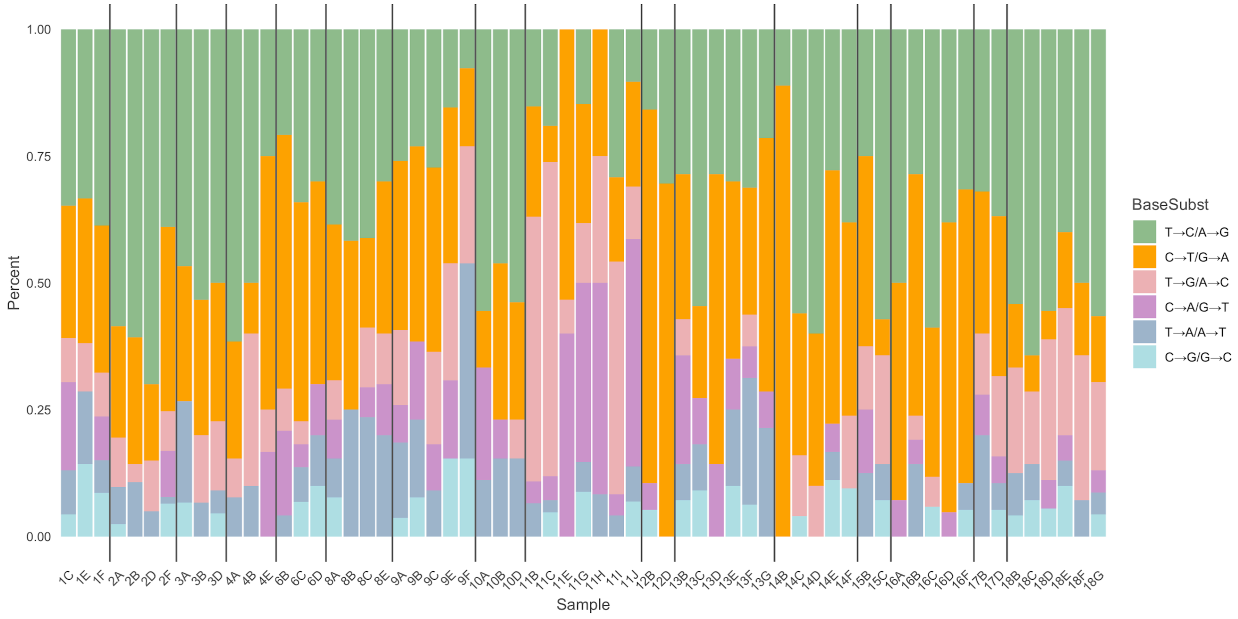
Supplemental Fig. S2: Unsupervised Clustering of CMT Data Identifies 4 Subtypes: (A) Heatmap highlighting expression patterns across the CMT clusters, predicted PAM50 subtypes, patient, breed. Heatmap shows median-centered expression for all genes in the CMT data. Different tumors from a dog end up in different clusters, indicating distinct gene expression programs in tumors from the same patient. We note the ten canine tubular carcinomas are not uniform, and in unsupervised clustering they fall into three separate clusters, reinforcing the importance of sampling diverse histologies to identify those with shared molecular hallmarks. (B) PCA of the complete expression data, colored by histology. (C) Boxplots of hormone receptor expression levels in each cluster.



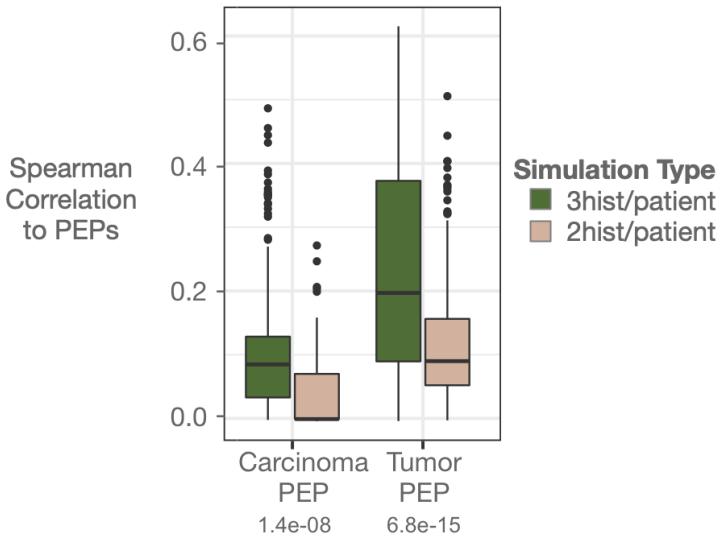
Supplemental Fig. S3: Median read depth per gene Median read depth per gene in our CMT dataset (across all samples) for COSMIC tumor suppressor genes and for all other genes in our data. The dotted vertical line shows the threshold beyond which RNA-seq calls are concordant with exome sequencing calls (threshold indicates 50-fold depth). Note that nearly all genes are measured above that threshold in our dataset.



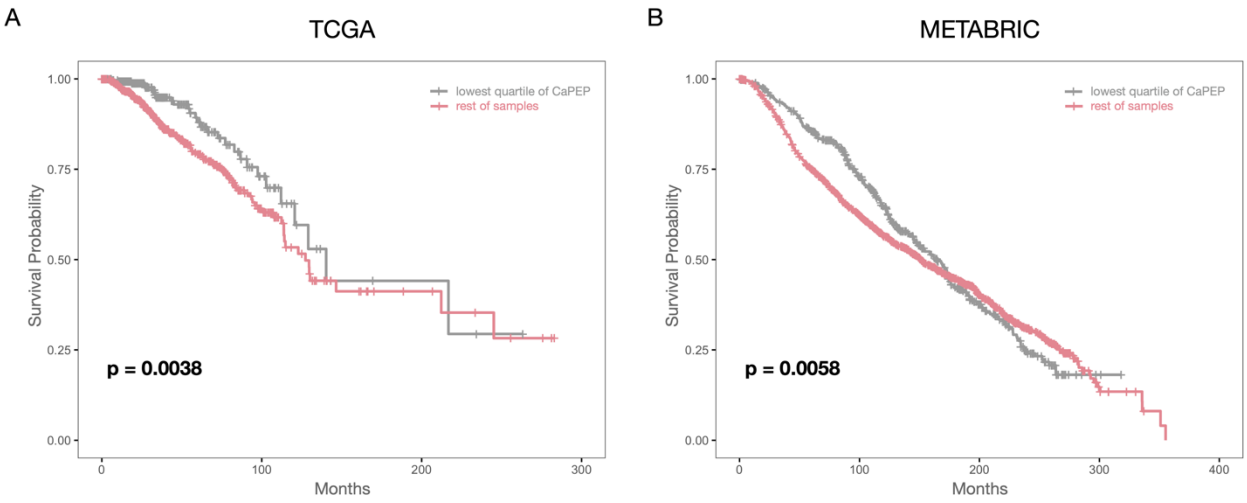
Supplemental Fig. S4: Mutational Recurrence Across Samples. The 30 most commonly mutated genes in the study are often mutated across patients, colored by histology (benign=yellow, malignant=red). Of the 30, four are human COSMIC genes (*B2M*, *CTNNB1*, *EML4*, *FGFR1*).



Supplemental Fig. S5: Base transitions C→T and T→C dominate in most patients, as expected. The ratio of base substitutions in each CMT tumor sample, with vertical lines separating the samples by dog of origin. Transitions are the more common mutation in most tumors, as expected by comparison with human tumors. Notable exceptions are patient 11 (more transversions than transitions) and patient 12 (almost exclusively transitions, despite only slightly lower than median mutation count), suggesting distinct mechanisms of mutagenesis in these dogs compared to the others.



Supplemental Fig. S6: Three histologies per patient drives discovery of malignant-specific processes: Results of 300 simulation runs of two subsampling approaches from the original data compared using the Spearman's correlation of each PEP to the complete run. Using samples from three histologic categories from each patient significantly outperforms runs with only two histologic groups per patient (all histologic groups represented in sample sets; Wilcoxon rank sum p-values $6.8x10^{-15}$ Tumor PEP and $1.4x10^{-08}$ Carcinoma PEP) indicating that PEP signals derived from three histology set-up cannot be identified using a two-histology study design.



Supplemental Fig. S7. Dog PEP signature is predictive of survival in human breast cancer. Kaplan-Meier plots showing patients with breast cancers bearing strongest Carcinoma PEP signal have worse outcomes in two independent human breast cancer cohorts: (A) TCGA BRCA and (B) METABRIC.