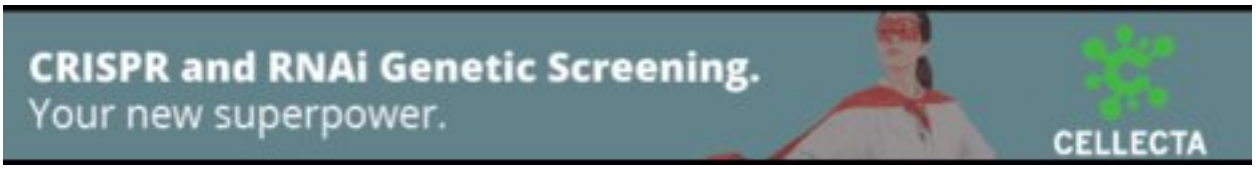
Modeling molecular development of breast cancer in canine mammary tumors

Kiley Graim, Dmitriy Gorenshteyn, David G Robinson, et al.

Genome Res. published online December 23, 2020
Access the most recent version at doi:10.1101/gr.256388.119



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

Authors: Kiley Graim[1,2], Dmitriy Gorenshteyn[2,3], David G. Robinson[2,3], Nicholas J. Carriero[1], James A. Cahill[4], Rumela Chakrabarti[5], Michael H. Goldschmidt[6], Amy C. Durham[6], Julien Funk[1], John D. Storey[2,7], Vessela N. Kristensen[8], Chandra L. Theesfeld[2*], Karin U. Sorenmo[5*], Olga G. Troyanskaya[1,2,9*]

Affiliations:
[1]Flatiron Institute, Simons Foundation, New York, New York, USA
[2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA
[3]Graduate Program in Quantitative and Computational Biology, Princeton University, Princeton, New Jersey, USA
[4]Laboratory of the Neurogenetics of Language, Rockefeller University, New York, New York, USA
[5]Department of Biomedical Sciences and the Penn Vet Cancer Center, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[6]Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[7]Center for Statistics and Machine Learning, Princeton University, Princeton, New Jersey, USA
[8]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway
[9]Department of Computer Science, Princeton University, Princeton, New Jersey, USA
*Co-Corresponding Authors:
ogt@cs.princeton.edu
karins@vet.upenn.edu
chandrat@princeton.edu

Keywords: Genomics, dog, canine, breast cancer

Abstract

Understanding the changes in diverse molecular pathways underlying the development of breast tumors is critical for improving diagnosis, treatment, and drug development. Here, we used RNA-profiling of canine mammary tumors (CMTs) coupled with a robust analysis framework to model molecular changes in human breast cancer. Our study leveraged a key advantage of the canine model, the frequent presence of multiple naturally occurring tumors at diagnosis, thus providing samples spanning normal tissue, benign and malignant tumors from each patient. We demonstrated human breast cancer signals, at both expression and mutation level, are evident in CMTs. Profiling multiple tumors per patient enabled by the CMT model allowed us to resolve statistically robust transcription patterns and biological pathways specific to malignant tumors versus those arising in benign tumors or shared with normal tissues. We demonstrated that multiple-histological-samples per patient is necessary to effectively capture these progression-related signatures, and that carcinoma-specific signatures are predictive of survival for human breast cancer patients. To catalyze and support similar analyses and use of the CMT model by other biomedical researchers, we provide FREYA, a robust data processing pipeline and statistical analyses framework.

Introduction

While there has been extensive progress in the field of breast cancer research, our understanding of the process of tumorigenesis remains incomplete (Bombonati and Sgroi 2011, Yates et al. 2017, Karagiannis et al. 2017, Harbeck et al. 2019). Studies of tumor progression in humans generally rely on disparate patient samples, with inter-individual genetic variability obscuring the

molecular progression signal (Crawford and Oleksiak 2007; Storey et al. 2007; Hughes et al. 2015). *In vitro* approaches using human cell lines have been used to control for this sample heterogeneity; however, they are not fully reflective of in vivo tumor progression, including the effects of the microenvironment and the immune system (Gillet et al. 2011; Stein et al. 2004). The (*in vivo*) murine model of breast cancer has proven incredibly useful in deciphering cancer mechanisms, however, it requires experimental modification of the host via genetic modification (transgenic mice) or the transplantation of foreign tissue (xenografts) (Rangarajan and Weinberg 2003; Boone et al. 2015), which alters the tumor dynamics (Ben-David et al. 2017).

Canine mammary tumor (CMT) is a promising emerging model for studying naturally occurring breast tumors (Klopfleisch et al. 2011; Liu et al. 2014; Pinho et al. 2012). CMTs and human breast cancer (BRCA) have similar histopathological profiles, including incidence rates, relationship with age and body mass index, hormonal influence, and clinical presentation as demonstrated in many clinical and smaller scale studies (Cekanova and Rathore 2014; Paoloni and Khanna 2008; Rowell et al. 2011; Kol et al. 2015; Kristiansen et al. 2016). Canine simple carcinomas share especially strong similarities with human breast cancer in terms of both histological and genetic features (Liu et al. 2014). Additionally, BRCA and CMT share chromosomal abnormalities such as copy number variations in several key breast cancer marker genes like *MYC* and *PTEN* (Borge et al. 2015). A significant advantage of the canine model is the high incidence of multiple naturally occurring tumors in the same patient (Sorenmo et al. 2009), which are rarely possible to biopsy in humans but common in canines since they have five pairs of mammary glands and often limited clinical monitoring. Thus, it is possible to design studies that overcome the effect of inter-individual genetic variability by assaying multiple naturally occurring tumor samples from a single patient, something that is rarely possible to

biopsy in humans (Toole et al. 2014). As such, the canine model provides a powerful complement to both laboratory mice and clinical human studies in studying breast cancer *in vivo*. Furthermore, discoveries with therapeutic potential made in CMTs can lead to rapid translational and clinical studies (LeBlanc et al. 2016).

In this study, we map the molecular signals underlying the similarities and differences between normal tissue and benign and malignant tumors. We find that the molecular signals in CMTs broadly reflect molecular changes in human breast cancer, including PAM50 molecular subtypes of clinical significance, and genetic cancer signatures. We then move beyond traditional normal vs. malignant comparisons to leverage the multiple-mammary-tumor nature of the canine model, where dogs can simultaneously present three types of samples found in canine tumor development: non-neoplastic mammary gland tissues (normal), benign/pre-malignant and malignant. We consider the three types of CMT samples in each of our patients as progression-ordered groups, enabling us for the first time to identify distinct signatures of gene expression reflective of progression from normal to benign to malignant. These signatures are relevant to human cancer biology; in TCGA and METABRIC, human breast cancers with stronger CMT carcinoma progression signature have significantly worse survival than patients with weaker carcinoma progression signature. Throughout, our analysis is driven by a robust statistical framework we developed for the molecular analysis of CMT -omic data, including a turn-key computational analytic pipeline (FREYA, **FR**amework for **E**xpression anal**Y**sis **A**cross species) tailored to dog and dog-human cancer comparisons that we make available to all researchers to promote naturally occurring CMTs as a model for human breast cancer. Altogether, our comprehensive genomic characterization demonstrates that CMTs are a powerful translational

model of BRCA, providing insights that inform our understanding of tumor development and treatment in both humans and dogs.

Results

To study the development of tumors from normal tissue to carcinoma we collected 89 mammary tissue samples (26 normal, 41 benign, 22 malignant) from 16 dogs of diverse breeds being treated through the Penn Vet Shelter Canine Mammary Tumor Program (see **Methods, Fig. 1**). The multiple independent primary tumors typical of CMT (required in this study design) present a unique window into tumor progression (**Supplemental Fig. S1A**). The presence of multiple independent lesions at different stages in the same individual (independence determined via systematic analysis of tumor mutational profiles with phylogenetic analysis, **Supplemental Fig. S1B, Methods**) allows us to identify molecular signals specific to each stage of tumorigenesis. There are many types of canine carcinoma and while on a semantic level the tubular carcinoma may be most similar to human cancer, on the molecular level these relationships are largely unexplored. To provide the broadest analysis of carcinoma-specific signals we included all available carcinoma samples (**Supplemental Table S1**). Using RNA sequencing, we generated genome-wide gene expression profiles (~13k genes) and called somatic mutations for each sample**.** We developed a robust analytical framework, FREYA, to detect and interpret the molecular signals in CMTs to facilitate translational BRCA research (**Fig. 1**, freya.flatironinstitute.org).

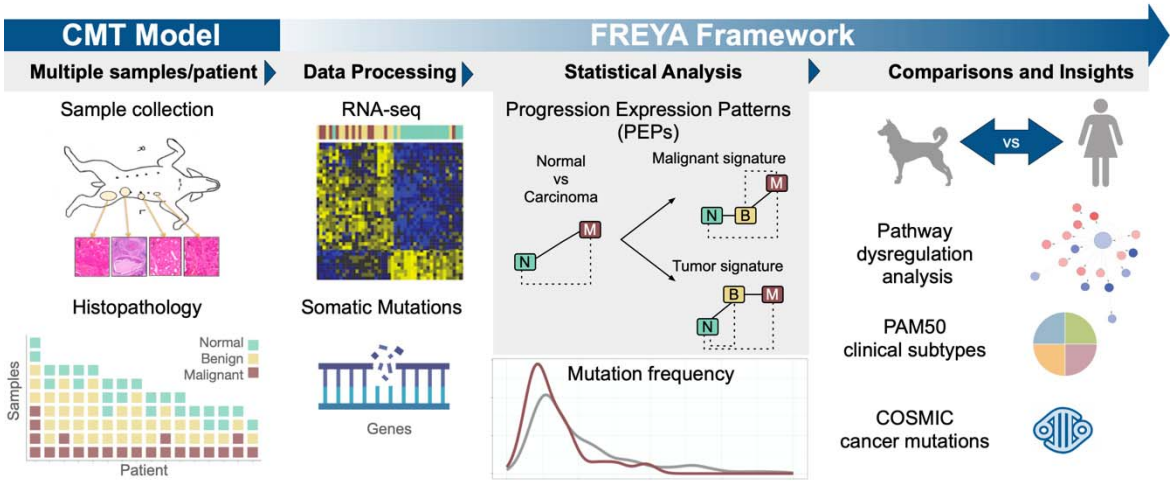


Fig. 1. Multiple CMTs per patient model enables discovery of carcinoma-specific processes that inform human BRCA. Left panel (CMT Model): Tissue samples were collected and annotated for each of the 89 samples from 16 canine patients. For study inclusion, each patient was required to provide a minimum of one sample (represented by colored blocks) from each histological group: normal (green), benign (yellow), and malignant (red). Many of the dogs have multiple samples of different tumor histologies. Right Panel (FREYA framework): We developed the FREYA framework to study tumor development. Using FREYA, we analyzed multiple primary tumors per patient with RNA and mutation profiling and developed a statistical framework to determine differences in gene expression between normal, benign, and malignant samples, and compared CMTs molecular signals to human breast cancer.

Molecular and cancer subtype similarities between canine and human tumors

As a first step, we assessed global cancer signals in malignant CMTs compared to normal tissue by identifying differentially expressed genes (FDR < 0.05, **Supplemental Table S2**). We found that these genes form four major modules in the genome-scale mammary epithelial functional network, wherein connections between genes reflect close interactions and participation in

pathways and processes specific to the tissue (Greene et al. 2015) (**Fig. 2A**). These clusters are characterized by distinct enrichment signatures indicating diverse dysregulated hallmark cancer processes **(Fig. 2A)**, including DNA repair and cell cycle regulation (module 1, including genes such as *BRCA1*, *BRCA2*, and *CDKN1A*), apoptotic signaling, response to hormone, immune functions, and response to stress in the endoplasmic reticulum (module 2, including genes *PIK3CA*, *FOXA1* and *MAPK1*), responses to hypoxia that enable tumor formation including angiogenesis and cell migration (module 3, including genes *GATA3*, *HIF1A*, and *VEGFA*), and immune function and hormone signaling (module 4, including *ARMC6*). These results indicate that cancer signals in CMT are broadly similar to common human cancer signals.

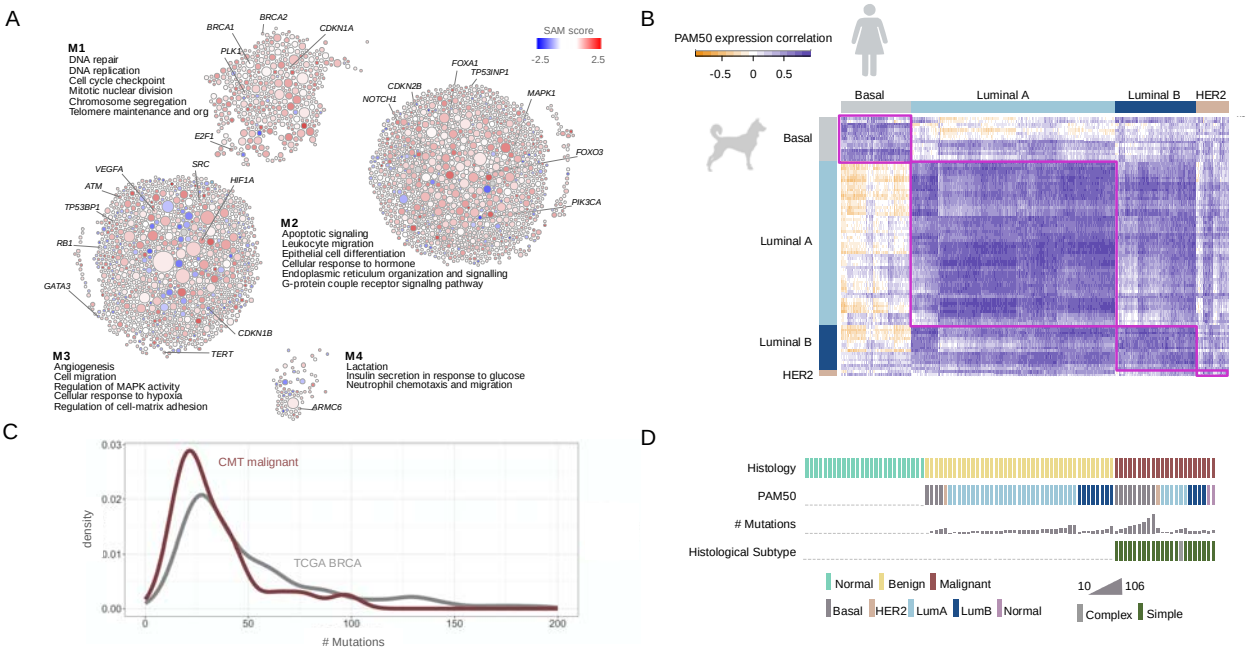


Fig. 2. Cancer hallmark processes found in CMT transcriptional programs. **(A)** Biological processes showing differential gene expression between normal and carcinoma samples (**Supplemental Table S2**) were identified by network-based enrichment method at humanbase.flatironinstitute.org. Differentially expressed genes were clustered using a shared-

nearest-neighbor-based community-finding algorithm to identify distinct modules of tightly connected genes (Krishnan et al. 2016) within the mammary epithelium functional network (Greene et al. 2015). Gene Ontology (GO) enrichment was performed on each module and representative significant processes are displayed (entire list in **Supplemental Table S3**). Circles are genes and the size of the circle indicates the sum of connections in the graph. Gene expression values (SAM scores) are overlaid. Red indicates increased expression in carcinoma and blue indicates decreased expression in carcinoma. COSMIC cancer census genes are indicated in each module (M1-M4). **(B)** Human PAM50 intrinsic subtype signals are found in CMTs. Each bar represents the number of samples predicted for each PAM50 subtype, human or canine. Predictions for CMT samples were based on gene expression programs using a classifier trained on human BRCA samples and PAM50 subtype gene expression signature data. 98% of human samples were correctly predicted, reflecting the accuracy of the predictor. **(C)** Density plot showing the genome-wide number of mutations per tumor sample in human (grey) and canine (maroon). **(D)** Oncoprint showing histology, predicted PAM50 subtype, number of mutations, and histologic subtype (simple/complex) for each sample in the cohort.

We next examined the representation of signals from human PAM50 intrinsic molecular BRCA subtypes within CMTs (Parker et al. 2009). Using a multinomial elastic net regression model we trained on known human PAM50 subtype samples from TCGA, CMT samples were predicted to be one of these types (**Fig. 2B, Supplemental Table S1**). Correlations between dog and human samples of the same subtype are significantly higher (p-val 8.65×10^{-43}) compared to different subtypes, showing that dog is reflective of the human intrinsic subtypes. Following the same protocols used to define PAM50 subtypes in human, we performed unsupervised clustering of

the CMT tumor samples (see **Methods, Supplemental Table S1, Supplemental Fig. S2**). Unsupervised CMT clusters are weakly correlated with their predicted PAM50 subtypes based on the human model (p-value = 0.047) and as well as histology (p-value = 0.024). This clustering (**Supplemental Fig. S2A**) and PCA analysis (**Supplemental Fig. S2B**) also both point to molecular heterogeneity among the canine tumor samples of similar histology, which is similar but stronger than that observed in human samples (Sorlie et al. 2003, The Cancer Genome Atlas 2012). As in previous human reports (Sorlie et al. 2003), unsupervised CMT clustering does not simply recapitulate hormone receptor status (**Supplemental Fig. S2C**) although cluster 3 samples are enriched in Basal tumors and have low hormone receptor expression levels. Thus, these naturally occurring CMTs display cancer dysregulations resembling human cancers at both a global level of transcriptional changes (**Fig. 2A**) and at the level of specific changes characteristic of clinical subtypes of breast cancer (**Fig. 2B, Supplemental Fig. S2**).

Canine mammary tumors harbor human cancer-implicated mutations

To date, only targeted small-scale sequencing studies have examined CMT mutations. We extended our analysis of the molecular signals in CMT by analyzing the whole-transcriptome somatic mutations (see **Methods, Fig. 2C**) in the CMT tumor samples and compared them to human cancer mutations. For most genes, read depth was sufficient for mutation calling (see **Methods, Supplemental Fig. S3**), but some mutations could be missed due to limitations of RNA-seq-based mutation calling, including low expression levels, allele-specific expression, and intron-side splice site variants that exomes would miss.

We observed 1904 mutations in 524 genes, of which 226 mutations fall in genes belonging to the COSMIC catalog of human cancer-related genes (~600 total genes; Tate et al. 2019). Four of the top thirty recurrently mutated genes (**Supplemental Fig. S4**) are COSMIC genes (*B2M*, *CTNNB1*, *EML4*, *FGFR1*) and twenty-two mutations in these genes are SnpEff predicted high impact mutations (stop-gain or frameshift), making them candidate driver mutations. We also observed mutations in many genes involved in human breast cancer, including *TP53*, *PIK3CA*, *NOTCH1*, *GATA3*, *FLNA*, *CDKN1B*, and *BAP1* (**Supplemental Table S4**) (Tate et al. 2019). The CMT mutation landscape is similar to that of human breast cancer, with most tumors harboring fewer than 75 mutations and a small subset of highly mutated tumors (**Fig. 2C**, **Supplemental Table S4**; Bailey et al. 2018). Among CMT, predicted Basal tumors have significantly higher somatic SNV burden compared to all other predicted PAM50 tumor types (p-value < $1x10^{-16}$, Spearman's rho, **Fig 2D**), consistent with human breast cancer. Overall, human cancer genes (COSMIC) are significantly more likely to be mutated in CMT samples than non-cancer genes (p-value = $8.037\ x10^{-15}$), with breast cancer genes specifically enriched in CMT samples versus general cancer genes (p-value=$7.81\ x10^{-12}$), indicating strong similarities between BRCA and CMT at the mutational level.

Extraction of mutational signatures can indicate mutational processes driving tumorigenesis. We examined the patterns of base substitutions in the CMTs and found, as expected, more transitions than transversions across all tumors (**Supplemental Fig. S5**) with the exception of patient 11. Six of the seven sequenced tumors from patient 11 have more transversions, suggesting distinct mutagenesis processes in this patient. Deamination of cytidine by APOBEC enzymes can lead to

C→T transitions; these APOBEC signature mutations are significantly associated with BRCA and correlate with increased somatic SNV burden and clinically aggressive features (Burns et al. 2013, Harris 2015, Takahashi et al. 2020).

Capture and Characterization of Progression Expression Patterns (PEPs)

Despite breakthroughs in characterization of breast cancer subtypes, targeted therapy development, and great strides in patient outcomes, the precise mechanisms and processes mediating invasiveness and malignancy are not yet fully characterized (Yates et al. 2017; Karagiannis et al. 2017). The presence of multiple histologies per patient in CMTs can confer sensitivity to detect altered pathways specific to malignant tumors that might not be identified in paired normal/carcinoma comparisons in human. To leverage this aspect of the canine model and discover tumor-stage-specific dysregulations, we analyzed signatures specific to each of the three epithelial tissue groups: normal/non-neoplastic (normal, normal with atypia, duct ectasia, hyperplasia); benign (simple adenoma and complex adenoma); carcinoma (simple carcinoma, *in situ* carcinoma, carcinoma in a mixed tumor). More specifically, to identify genes driving the differences between these histologic types, we performed differential gene expression analysis (see **Methods**) for each pairwise combination of histologic categories and identified genes with expression signatures that are significantly different (FDR < 5%) in at least two of the paired comparisons. We then systematically identified genes specific to histologic types by using both the significance of changes and direction of the expression change (**Fig. 3**). We refer to these signatures as Progression Expression Patterns (PEPs, see **Supplemental Table S5** for full list of genes).

In order to empirically assess the advantages of the CMT-based study for detection of tumor-relevant signals, we compared this multiple tumor types per patient design to the traditional tumor vs. normal design by subsampling patient samples in our dataset. We then assessed how much signal was lost in each case by assessing the ability of each subsampled study, with equitable sample sizes, to identify PEPs discoverable in the full dataset. PEPs identified using three histologies were consistently able to more closely recapitulate the PEPs generated with the full dataset than PEPs identified using just two histological groups but the same number of samples (Wilcoxon rank sum p-values 6.8 $x10^{-15}$ Tumor PEP and 1.4 $x10^{-08}$ Carcinoma PEP, **Supplemental Fig. S6**, **Methods**). Specifically for the carcinoma PEP, the simulation using the two histologies design shows no correlation with the carcinoma PEP generated with the full dataset, underlining the importance of having three stages of tumor development to discover malignant-specific processes and underscoring the power of the canine model to detect signals with smaller numbers of samples.

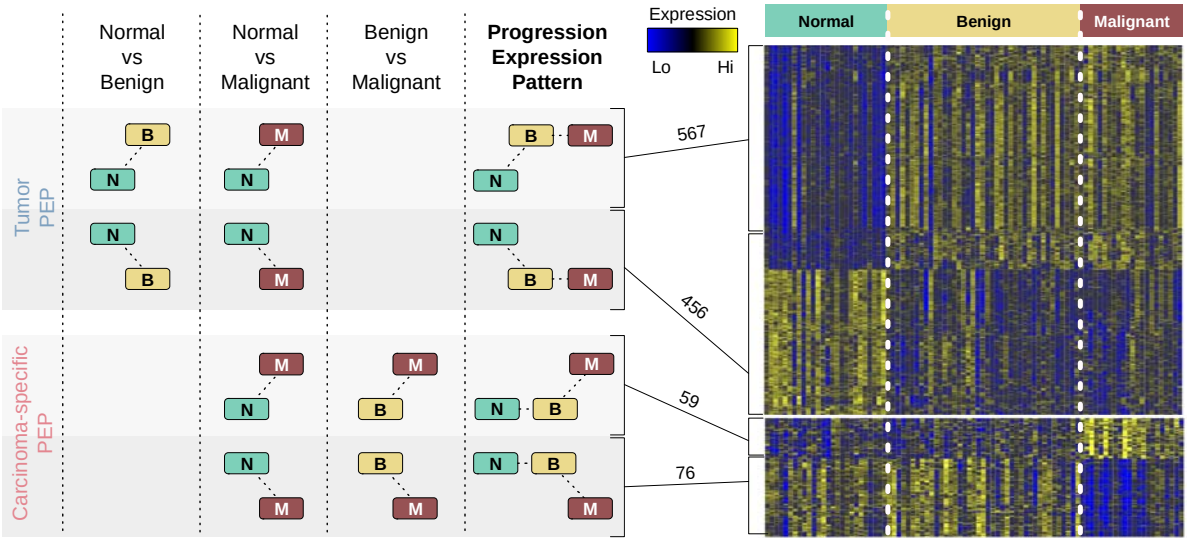


Fig. 3. Identification of Progression Expression Patterns. Progression Expression Patterns (PEPs) are identified using differential expression analysis between histological groups (top, Tumor PEP, 1023 genes; bottom, Carcinoma-specific PEP, 136 genes). The diagrams illustrate how each PEP pattern is defined. For example, Tumor PEP includes genes up-regulated in tumors (significantly differentially expressed both between normal and benign samples and between normal and malignant samples). The heatmap shows the expression patterns for these genes, with patterns divided into up- and down-regulated (e.g. Tumor PEP includes 567 genes significantly up-regulated in tumors and 456 genes significantly down-regulated in tumors).

Resolving carcinoma-specific processes versus those altered at the benign transition

We explored processes and pathways represented in each PEP using Gene Ontology term enrichment analysis. The Tumor PEP represents genes whose expression is changed concordantly in both benign tumors and malignant carcinomas, while the Carcinoma PEP consists of genes whose expression is uniquely altered in carcinomas, but not significantly altered in benign tumors relative to normals. Tumor PEP genes are significantly enriched for a number of known human tumor-associated pathways, including control of cell cycle transitions, DNA repair pathways, regulation of MAP kinase activity, regulation of adaptive immune response, and mammary gland epithelial cell proliferation (**Supplemental Table S3**). The Carcinoma PEP (**Fig. 3, Supplemental Table S5**) is significantly enriched for known breast cancer processes, including negative regulation of apoptosis, regulation of epithelial cell differentiation, and lipid metabolic processes (**Supplemental Table S3**). Included in the Carcinoma PEP are a number of genes implicated in breast cancer aggression and metastasis, such as *PDGFB*, *GATA3*, and *SMO* (Jansson et al. 2018; Benvenuto et al. 2016). This suggests

that whereas Tumor PEP genes are associated with cancer processes, Carcinoma PEP genes are associated with cancer aggression.

The availability of multiple tumor types in CMTs presents a unique opportunity to resolve those pathways that are dysregulated between normal tissues and all tumors versus the pathways specific to the carcinoma transition. To accomplish this, we compared the biological processes enriched in genes identified in the traditional Normal vs Carcinoma differential expression analysis (many hallmark tumor processes **(Figs. 2A & 4)** to those processes found enriched in either the Tumor or Carcinoma PEPs (**Methods**, **Fig. 4A, Supplemental Table S3**)). We found that apoptotic processes were represented in the Normal-Carcinoma comparison, genes involved in negative regulation of *internal* apoptotic signaling were distinctly represented in the Carcinoma signature, while genes relating to response of cells to *external* death cues were enriched in the Tumor signature. This could reflect that tumors in general are antagonized by the immune system, yet the more aggressive carcinomas are expressing genes that are turning off their ability to die in response to internal apoptotic signaling pathways, thus contributing to malignancy (French and Tschopp 2002, Fernald and Kurokawa 2013, Ashkenazi 2015, Mantovani et al. 2019). Processes uniquely enriched in carcinomas also included calcium signaling and homeostasis and regulation of lipid biosynthesis, which may be linked to managing endoplasmic reticulum and membrane stresses that can drive malignancy (Urra et al. 2016). Thus, comparing multiple canine tumor types effectively distinguishes processes linked to specific stages of tumor development and identifies pathways that are unique to aggressive tumors.

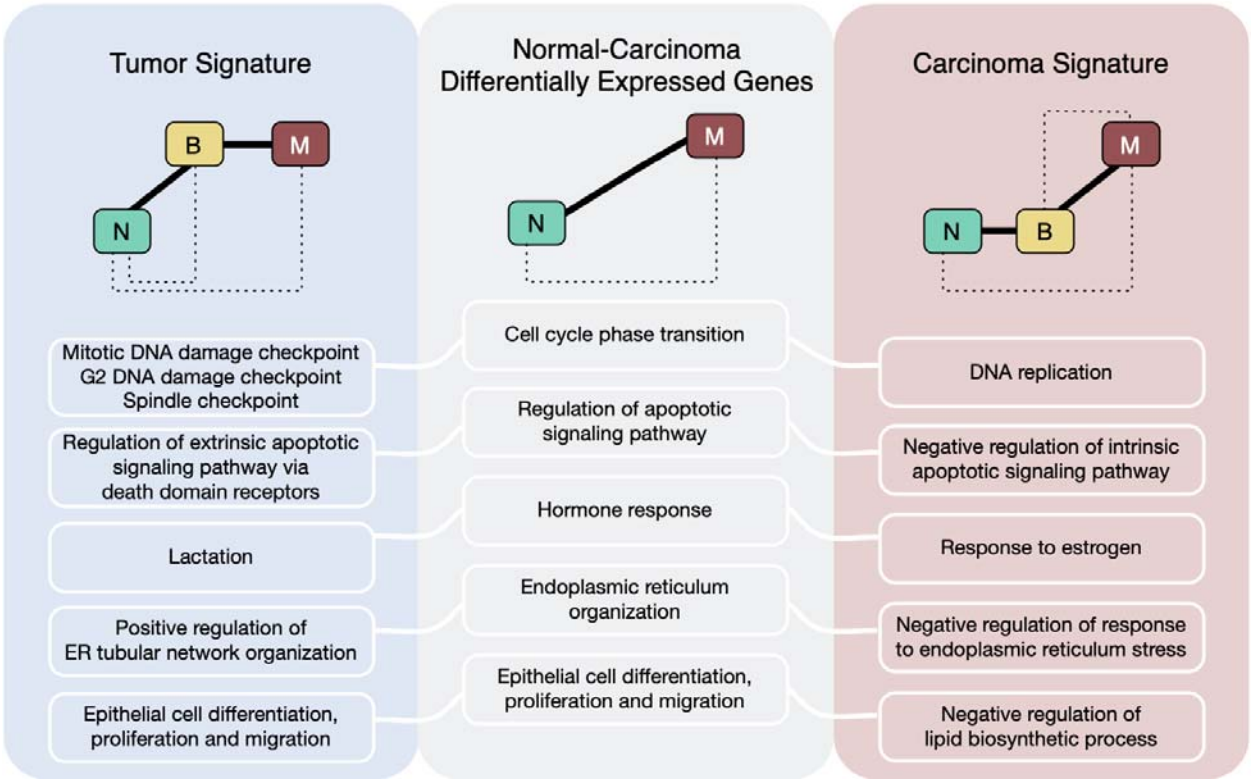


Fig. 4. Resolution of cancer hallmark processes by PEPs to discern malignancy-specific processes. Genes differentially expressed between Normal and Carcinoma samples (as in traditional gene expression analysis) exhibit Tumor or Carcinoma-specific signatures. This experimental design stratifies tumor processes into those specific to malignant tumors (carcinoma-specific pattern) and those that are perturbed in both benign and malignant tumors (tumor-specific pattern). Five representative example GO terms from each pattern are shown (see **Supplemental Table S3** for a complete list).

Carcinoma PEP signature is predictive of survival in human breast cancer

To understand how the Carcinoma PEP (which delineates malignant-specific tumor signals) relates to human tumors, we investigated whether this group of genes is predictive of clinical survival in breast cancer patients. Indeed, we found that in TCGA BRCA and METABRIC

samples, levels and direction of expression change of Carcinoma PEP is predictive of human survival: patients expressing the weakest Carcinoma PEP signature have significantly better outcomes (all data, Peto-Peto p=0.0038 TCGA and p=0.0058 METABRIC, **Supplemental Fig. S7**), and this goes beyond reflecting PAM50 subtypes. While PAM50 subtype and Carcinoma PEP strength correlate (for example most Basals have strong Carcinoma PEP signalling), there is a significant difference in survival within subtype. In the most prevalent subtype, Luminals, there is a survival difference relative to the Carcinoma PEP scores (**Fig 5A-B**: Luminal A, TCGA p-value 0.004, METABRIC p-value 0.0048; Luminal B, TCGA p-value 0.004, METABRIC p-value 0.0048). Thus, the Carcinoma PEP signature has clinical relevance in human breast cancer, underscoring the utility of the canine model for capturing molecular signatures associated with human breast cancer.

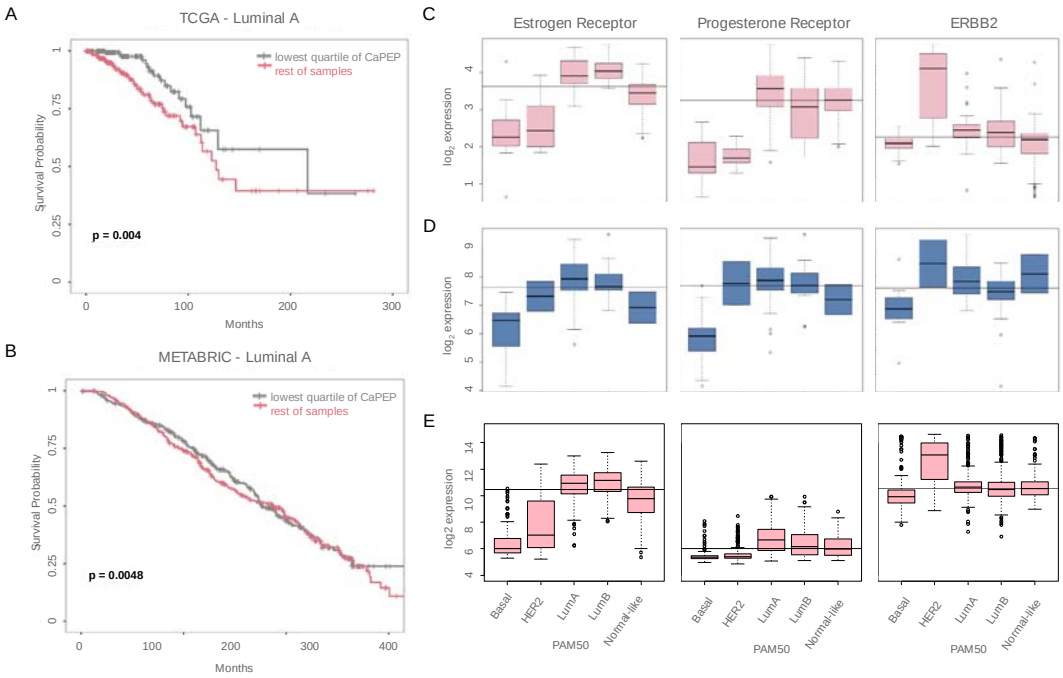


Fig. 5. Dog PEP signature is predictive of survival in human breast cancer. (A-B) Kaplan-Meier plots showing patients with breast cancers bearing strongest Carcinoma PEP signal have worse outcomes in two independent human breast cancer cohorts: (A) TCGA BRCA and (B) METABRIC. (C-E) Dogs, like humans, have strong hormone receptor expression signaling differences between PAM50 subtypes. Estrogen receptor (*ESR1*), progesterone receptor (*PGR*) and *Erb-B2* receptor (*ERBB2*) expression within each PAM50 subtype shown for (C) TCGA BRCA, (D) this CMT data and (E) METABRIC. Horizontal lines across each graph indicate median receptor expression across the entire cohort.

Discussion

In this study we presented a novel statistical approach for studying the mechanism of tumorigenesis by leveraging the multiple naturally-occurring samples per patient features of the canine mammary tumor model to define processes and pathways that are dysregulated between normal tissue, benign, and malignant tumors. We characterize the genome-wide landscape of molecular signals in CMT, both at the transcriptional and mutational level, demonstrating that many hallmark human breast cancer processes, including cell migration, cell cycle checkpoints, and apoptotic signaling, (see **Fig. 2A** and **Supplemental Table S3**) and molecular subtypes of human breast cancer are reflected in CMTs. In addition to demonstrating the molecular similarities between CMT and human breast cancer and providing a computational framework to facilitate using CMT as an effective model for BRCA, our analysis distills malignant-specific signals from overall tumor-associated signals. We show that these cancer dysregulations and aggressive biology captured by the canine carcinoma PEP are relevant to the dysregulations in human breast cancer, with weakest Carcinoma PEP signature correlating with significantly

increased patient survival. This information is distinct from that captured by predicted PAM50 subtype status -- the Carcinoma PEP is able to stratify survival within Luminal A patients, presenting an important perspective on factors that make this prevalent breast cancer subtype more dangerous. This underscores the potential of the CMT model for molecular studies of human breast cancer.

A critical challenge to translational studies in model organisms is in effective analysis that integrates the findings with human biology. To promote comparative oncology studies that leverage the multiple-samples-per patient approach afforded with CMTs, we developed and are sharing with all biomedical researchers a novel analytical framework, FREYA. FREYA is a computational suite which enables any researcher to perform analyses in this manuscript, from data processing to human cancer comparison to figure generation, using either the provided CMT data or user-provided data. FREYA is available at freya.flatironinstitute.org.

Altogether, our study demonstrates the relatedness of human breast cancer and canine mammary tumors at the molecular level as well as the utility of the CMT model for discerning signals that are obscured in other model systems. While our understanding of human breast carcinogenesis remains incomplete, understanding the progression from benign to malignant and identifying its molecular signature is key both to understand breast carcinogenesis and to identify targets for cancer prevention and therapy. The molecular signals in CMT that we identify indicate the canine model offers unique opportunities to fill gaps in our understanding of human breast tumorigenesis and provide a comprehensive canine breast cancer model that captures the major variables (predicted PAM50 subtypes within CMT do have the characteristic hormone receptor

expression (**Fig. 5C-E**)), hallmarks (tumor genetics, microenvironment, hormonal effect and immune function), and their interactions. In such, CMT as a model of human breast cancer provides a powerful complement to both human clinical and *in vitro* studies as well as model organism studies (e.g. in mouse). Insights from CMTs can be used to direct future mechanistic studies in other model systems, and the CMT model offers unique opportunities for expedited clinical trials of therapies, availability of material for isolation of breast cancer stem cells, and analyses of tumor evolution at both the level of mutations and associated transcriptional programs.

Materials and Methods

Ethics Statement

Animal work was approved by the University of Pennsylvania Institutional Animal Care and Use Committee (IACUC), listed as protocol 804298, principal investigator Karin Sorenmo, and titled "Molecular Evaluation of Canine Mammary Tumors."

Experimental Design

Dogs have a high incidence of multiple primary tumors, making it possible to study mammary tumor progression without the effects of inter-individual genetic variability. We created a pipeline to map the genomic landscape of CMTs, then compared them to BRCA. We showed that the multiple-diagnoses-per-patient experimental design was essential for capturing progression-related patterns of expression. Our analyses identified pathways and processes dysregulated in CMTs parallel to those altered in BRCA. We demonstrated that CMT mutation profiles recapitulated those seen in BRCA. CMT has the potential of being a uniquely impactful

model integrating transcriptional and other -omics data in a model organism that can bridge mechanistic studies in mouse/rat and human clinical data.

Sample Gathering

Tumor samples were collected from naturally occurring mammary tumors within sexually intact dogs treated through the Penn Vet Shelter Canine Mammary Tumor program. All dogs underwent routine clinical staging (including mapping and measuring of all tumors as well as thoracic imaging) followed by surgical removal of the affected glands. All tissues were processed immediately after removal. Two parallel small incisional sections were collected from each tumor as well as sections from visually normal mammary tissue; one section was flash frozen in liquid nitrogen and stored in -80F freezer and the adjacent section was fixed in formalin for routine H&E staining and histopathological evaluation. For mixed carcinomas, the carcinoma portion of the tumor was extracted for sequencing. In addition, the whole tumor was also evaluated histopathologically. A standard published classification system on canine mammary gland tumors was used to classify and grade all tumors/tissues by board certified Veterinary Pathologists (Goldschmidt and Durham) (Goldschmidt et al. 2011).

Sample Processing

Tissue samples were cryo-pulverized then homogenized using a rotor-stator in TRIzol (Cat. No.15596-026, Invitrogen). The lysate was further homogenized using a Qiashredder spin column. mRNA was extracted using the Qiagen RNeasy Kit. The sequencing libraries were prepared at the Princeton University Genomics Core Facility using the PrepX mRNA Library Protocol for the Apollo324 System (Wafergen). Sequencing was performed using the Illumina

HiSeq 2000 platform. This resulted in 5.041 billion mapped reads across 89 samples, with 56 million mapped reads on average per sample.

RNA-seq processing

RNA-seq reads were mapped to the CanFam 3.1 genome assembly (Ensembl release 91; Zerbino et al. 2018) using the HISAT2 aligner (Kim et al. 2019), after which assemblies were filtered using FastQC (Andrews 2010). DEXseq-Count (Anders et al. 2012; Reyes et al. 2013) was used to construct read counts for each gene in this combined transcriptome assembly. The resulting counts matrix was normalized using TMM (Robinson and Oshlack 2010). We then regressed out the effect of the individual and row-centered the resulting data in order to remove breed bias. This was necessary because of the high heterogeneity between dog breeds.

Variant calling and identification of somatic mutations

CMT mutations were called following the GATK Best Calling Practices for RNA-seq pipeline (software.broadinstitute.org/gatk/documentation/article.php?id=3891; DePristo et al. 2011; Van der Auwera et al. 2013). We made mutation calls in all genes and the average read depth of 164-fold coverage (**Supplemental Fig. S3** surpassed the threshold for maximum mutation call confidence as demonstrated by Sun et al. (Sun et al. 2017). Read depth was calculated using BEDTools coverage (BEDTools2 version v2.29.2; Quinlan and Hall 2010). We further filtered the variants by comparing each variable site in the tumor sample to the normal samples from that same individual, and discarded sites where tumor and normal samples matched. In cases where normal samples from the same individual had different genotype calls, we required that tumor samples differ from all normal tissue samples in order to call a mutation. We also excluded

genotype calls with quality scores less than 40 and calls generated from less than 4-fold coverage. Additionally, genotype calls annotated with HIGH and INTERMEDIATE functional effect scores (SnpEff (Cingolani et al. 2012)) were retained and used in downstream analyses. In order to compare human and CMT mutation rates, overall mutation counts for each TCGA BRCA sample were downloaded on June 17, 2019 from cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018. Due to uneven read coverage in RNA-seq data, there are limitations in RNA-seq mutation calling, for example, indels in low expressed genes such as tumor suppressors may not be identified due to a lack of coverage in the region, so that there is not enough data to make a high confidence call.

Phylogenetic Analysis

An identity-by-descent phylogeny with proportional branch length (**Supplemental Fig. S1**) was generated using SNPRelate (Manichaikul et al. 2010) and all SNPs called by FREYA. All samples, including normals, were included in this analysis. Pairwise similarity scores were calculated for all sample pairs, such that similarity of mutations in samples *a* and *b* (m_a, m_b) is:

$$\frac{m_a \cap m_b}{\min\left(|m_a|, |m_b|\right)}$$

Moderate and high impact mutations in named genes, called as described in Methods section *Variant calling and identification of somatic mutations*, were included in the similarity score calculation. For each pair of tumors, the similarity score is the fraction of mutations in the lesser mutated sample that are mutated in both samples, creating a similarity score ranging from 0 to 1.

Canine Pairwise Differential Expression Analysis

Pairwise differential expression comparisons were performed between normal and adenoma, normal and carcinoma, and adenoma and carcinoma samples. Differential expression testing was performed using edgeR (McCarthy et al. 2012), using a negative binomial generalized linear model explaining expression based on histology while controlling for individuals. Samples were normalized using weighted trimmed mean of M-values (TMM) (Robinson and Oshlack 2010). Genes with more than one count per million in at least 30 samples were used for this analysis. False discovery rate control was performed using the q-value method (Storey and Tibshirani 2003) on each comparison.

GO Enrichment

We identified enriched processes in differentially expressed genes in the normal-carcinoma comparison (FDR<0.05, **Supplemental Table S3**, **Fig. 2A**) and in each PEP (**Supplemental Table S5, Fig. 4**) using the Functional Module Detection query at humanbase.flatironinstitute.org. Each gene list was clustered using a shared-nearest-neighbor–based community-finding algorithm to identify distinct modules of tightly connected genes (Krishnan et al. 2016) within the mammary epithelium functional network (Greene et al. 2015). GO enrichment was performed on each module.

Unsupervised CMT Clustering

In order to identify the presence of molecular subtypes within the dog samples, we performed unsupervised clustering of the samples using the intrinsic analysis described previously in (Parker et al. 2009, Sorlie et al. 2003), to identify genes with low variability in expression within paired (tumor/normal) samples from the same patient but high variability across tumors from

different patients. Genes with a low ratio of within-dog variance versus between-dog variance, those below one standard deviation of the mean ratio, were defined as 'intrinsic genes' and used in the unsupervised clustering. Benign and malignant tumors were clustered based on those 2076 intrinsic genes (**Supplemental Table S1, Supplemental Fig. S2**). 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.

Progression Expression Profile Identification

Results of the pairwise differential expression analysis between the 3 histologies were used to identify the expression patterns (see **Fig. 3**). Each pattern is characterized as having 2 of the 3 comparisons showing differential expression, using a cutoff q-value below 0.05. Specifically, classification of genes to the appropriate pattern is determined using the following criteria:

Tumor-Specific: A gene differentially expressed in both the normal-adenoma and the normal-carcinoma comparisons; the sign of the change is the same for both comparisons.

Carcinoma-specific: A gene differentially expressed in both the normal-carcinoma and the adenoma-carcinoma comparisons; the sign of the change is the same for both comparisons.

PAM50 Subtypes

In order to identify the presence of human PAM50 molecular subtypes within the dog samples (normal, benign, and malignant), we combined the expression data from 89 dog and the 981 human TCGA tumor samples with PAM50 annotations, subset to the 42 PAM50 genes present in the dog samples (canine orthologs of some genes are currently unknown), then removed species

batch effects using sva (Leek et al. 2019). An elastic net (R package glmnet) (Friedman 2010) was trained to predict PAM50 subtype based on the human data and applied to samples for both species. 98% of human samples were correctly predicted (in accordance with the sample label in the TCGA dataset). R version 3.6.1 (2019) was used. CMT samples were predicted to be all four PAM50 subtypes and in similar ratios seen in humans. PAM50 subtype correlation statistic was calculated by calculating the Wilcoxon rank sum test p-value for each subtype (e.g. LumA dog samples vs LumA human samples compared to LumA dog samples vs non-LumA human samples; p-values are: LumA $<2.2 \times 10^{-16}$, HER2 .36, Basal $<2.2 \times 10^{-16}$, LumB $<2.2 \times 10^{-16}$). We then calculated a joint statistic using the Fisher combined probability test (R package metap) (Dewey 2020).

Simulation

The inclusion of three histologies in this study enabled us to define PEPs relevant to development of malignant tumors. We used a simulation to compare the accuracy of three vs two histology per patient setup for generating PEP signatures. For this analysis we subsampled the full dataset both for the three histologies per patient setup and the two histologies for patient setup and ran the entire PEP derivation from FREYA on each subsampled dataset. Since the two subsampled datasets in each random rerun are controlled for sample size, if there is a significant difference in how well they recapitulate the original PEP signature, it is driven by the difference in histologies.

To generate a dataset where only two histologic categories are represented (as is typical in normal versus tumor design), the 16 patients in our study were randomly split into 2 groups,

where one group contains pairs of normal and benign samples and the other contains normal and carcinoma samples, for a total of 32 samples. To generate a dataset where three histologic categories are represented, random sub-sampling of 10 patients was performed, where one of each of the three histology groups were randomly selected (30 samples per simulation total). Using these two subsample groups, we repeated the steps used to generate the PEPs and compare the maximum q-values characterizing the PEPs from simulations (300 simulations per experimental design). Q-values from these simulations were compared to the actual study in this paper using the respective pattern to calculate Spearman's correlation (**Supplemental Fig. S6**). The three-histology approach significantly outperformed the paired approach (Wilcoxon rank-sum test; p-values 6.8×10^{-15} Tumor PEP and 1.4×10^{-08} Carcinoma PEP) when using a comparable number of samples.

Comparison to Human Breast Cancer (BRCA) Data

RSEM normalized, $\log_2$-scaled RNA-seq data from two human breast cancer cohorts, TCGA BRCA (Hoadley et al. 2018) and METABRIC (Pereira et al. 2016), were obtained from via cBioPortal (cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018 and cbioportal.org/study/summary?id=brca_metabric). We identified orthologs between dog and human using BioMart (Kasprzyk 2011); only one-to-one mappings were used for this analysis. Differential expression of normal-malignant CMT samples was calculated on the $\log_2$ scaled data using siggenes (Schwender 2019). The list of known cancer-related genes, oncogenes, and tumor suppressors was taken from COSMIC (Tate et al. 2019). The representation of alternate haplotypes in hg38, absent in hg19, do not affect the findings of our study.

Carcinoma PEP Signature in Human Breast Cancer

In order to identify the significance of the PEPs in human breast cancer, we projected the Carcinoma PEP into TCGA BRCA and METABRIC data (see above). We first subtracted the median normal expression levels from the malignant gene expression values in each dog's samples and used the sum of positive and negative differences to designate each PEP gene as positive or negative (increased or decreased expression, respectively). We then generated signature scores for each human sample, calculating the sum of all PEP genes for which the gene was in the top of the human expression value ranges for positive PEP genes and in the bottom quartile for negative PEP genes. We then divided the signature scores into 4 groups of equal size and applied the Peto-Peto significance test.

We assigned each Carcinoma PEP gene *g* to the positive or negative signature set by subtracting the median expression levels in normal samples (e_{dn}) from the median expression levels in malignant samples (e_{dm}), within each dog *d*. Genes are grouped into *G*+ (up) and *G*- (down) determined by the ratio of dogs for which the direction of change in expression from tumor to normal is positive or negative:

$$\begin{cases} \frac{1}{|D|}\sum_{d\in D} 1_{e_{mdg}>e_{ndg}} > \frac{1}{2} & g \in G^{+} \\ else & g \in G^{-} \end{cases}$$

Given these groups, we then calculate Carcinoma PEP signature scores for each human sample, such that for each Carcinoma PEP gene *g*, the signature is considered present for that patient if

expression levels e_g are in either the top or bottom quartiles (Q_1, Q_4), depending on the direction of the expression change in CMT samples:

$$\sum_{g \in G^+} I_{e_{pg} > Q_3} + \sum_{g \in G^-} I_{e_{pg} < Q_1}$$

such that *G*+ and *G*- are Carcinoma PEP genes that follow a positive or negative direction of change within the dogs. Carcinoma PEP signature scores were assigned to all human samples in the TCGA BRCA and METABRIC cohorts and subtype analysis with Luminal A and Luminal B samples was performed in parallel.

FREYA Statistical Framework

The FREYA framework (freya.flatironinstitute.org) described here generates expression and mutation profiles from raw sequence data, then runs all analyses described in this manuscript on that data (with the exception of HumanBase functional network module detection, https://humanbase.flatironinstitute.org/, SNV substitution profiles, and phylogenetic analysis). No installation is necessary; a button click within the GitHub repository will automatically build an interactive docker image containing FREYA. Users have the option of passing unprocessed sequencer data to FREYA's DataPrep module or providing their own pre-processed data. Alternatively, we provide a version of FREYA optimized for a cluster environment. All versions of FREYA can be run with user-provided data.

Software Availability

Computer code underlying this statistical approach is available at freya.flatironinstitute.org and in the Supplemental Code file. To help with reproducibility and to encourage use of our statistical framework, we provide version information for each tool as well as parameters settings in the README and in the automated pipeline script.

Data Access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE136197.

Acknowledgments

General: We thank the Princeton University Genomics Core facility for the library construction and sequencing services, and members of the Sorenmo and Troyanskaya labs for discussions and editorial help, in particular Rachel Sealfon. We also want to thank the Flatiron Institute's Scientific Computing Core for helping package the analysis pipeline. **Funding:** Funding for this work was provided by the Puppy Up (2 Million Dogs) Foundation. The Penn Vet Shelter Canine Mammary Tumor Program made this work possible through the acquisition of tumor tissues and clinical data used in this study. **Author contributions:** KG, CLT, DG, DGR, KUS, and OGT designed the studies. KUS provided the clinical care, collected clinical data and performed tumor tissue sampling. KG, DG, DGR, RC, JAC, MHG, ACD, NJC, JF performed experiments and analyses. JDS contributed statistical aspects of the analyses. VK

provided expert feedback. KG, DG, CLT, and OGT wrote and edited the manuscript; All authors reviewed and approved the manuscript.

Disclosure Declaration

Competing interests: The authors have no competing interests to declare.

References

Abolhassani A, Riazi GH, Azizi E, Amanpour S, Muhammadnejad S, Haddadi M, Zekri A, Shirkoohi R. 2014. FGF10: type III epithelial mesenchymal transition and invasion in breast cancer cell lines. *J of cancer*. **5**: 537-547.

Alexa A and Rahnenfuhrer J. 2019. topGO: Enrichment Analysis for Gene Ontology. R package version 2.38.1. doi:10.18129/B9.bioc.topGO

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Nat precedings*. 1-1.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

Ashkenazi A. 2015. Targeting the extrinsic apoptotic pathway in cancer: lessons learned and future directions. *The J of clin investigation*. **125**: 487-489.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy□Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr protoc bioinf*. **43**: 483-492.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. 2018. Comprehensive characterization of cancer driver genes and mutations. *Cell*. **173**: 371-385.

Ben-David U, Ha G, Tseng YY, Greenwald NF, Oh C, Shih J, McFarland JM, Wong B, Boehm JS, Beroukhim R, et al. 2017. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat genet*. **49**: 1567-1575.

Benvenuto M, Masuelli L, De Smaele E, Fantini M, Mattera R, Cucchi D, Bonanno E, Di Stefano E, Frajese GV, Orlandi A, et al. 2016. In vitro and in vivo inhibition of breast cancer cell growth by targeting the Hedgehog/GLI pathway with SMO (GDC-0449) or GLI (GANT-61) inhibitors. *Oncotarget*. **7**: 9250-9270.

Bombonati A and Sgroi DC. 2011. The molecular pathology of breast cancer progression. *The J of pathology*. **223**: 308-318.

Boone JD, Dobbin ZC, Straughn Jr JM, Buchsbaum DJ. 2015. Ovarian and cer-vical cancer patient derived xenografts: The past, present, and future. *Gyn Onc*. **138**: 486-491.

Borge KS, Nord S, Van Loo P, Lingjærde OC, Gunnes G, Alnæs GI, Solvang HK, Lüders T, Kristensen VN, Børresen-Dale AL, et al. 2015. Canine mammary tumours are affected by frequent copy number aberrations, including ampli cation of MYC and loss of PTEN. *PLoS One*. **10**: doi:10.1371/journal.pone.0126371

Broad Institute TCGA Genome Data Analysis Center. 2016. Analysis Overview for Breast Invasive Carcinoma (Primary solid tumor cohort) - 10 January 2016. Broad Institute of MIT and Harvard. doi:10.7908/C1V40TJ9

Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, et al. 2013. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*. **494**: 366-70. doi: 10.1038/nature11881.

Cekanova M and Rathore K. 2014. Animal models and therapeutic molecular targets of cancer: utility and limitations. *Drug des, dev & ther*. **8**: 1911-1922.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, and Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. **6**: 80-92.

Consortium GO. 2019. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. **47**: D330-D338.

Crawford DL and Oleksiak MF. 2007. The biological importance of measuring individual variation. *J of Experimental Biology*. **210**: 1613-1621.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. 2012. The genomic and transcriptomic archi-tecture of 2,000 breast tumours reveals novel subgroups. *Nature*. **486**: 346-352.

Davis T, Loos B, Engelbrecht AM. 2014. AHNAK: the giant jack of all trades. *Cellular signalling*. **26**: 2683-2693.

De Craene B and Berx G. 2013. Regulatory networks de ning EMT during cancer initiation and progression. *Nat rev cancer*. **13**: 97-110.

De Francesco EM, Pellegrino M, Santolla MF, Lappano R, Ricchio E, Abonante S, Maggiolini M. 2014. GPER mediates activation of HIF1 /VEGF signaling by estrogens. *Cancer res*. **74**: 4053-4064.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat genet*. **43**: 491-498.

Dewey M (2020). *metap: meta-analysis of significance values*. R package version 1.4.

Donnem T, Al-Saad S, Al-Shibli K, Busund LT, Bremnes RM. 2010. Co-expression of PDGF-B and VEGFR-3 strongly correlates with lymph node metastasis and poor survival in non-small-cell lung cancer. *Annals of oncology*. **21**: 223-231.

Fernald K and Kurokawa M. 2013. Evading apoptosis in cancer. *Trends in cell biology*. **23**: 620-633.

French LE and Tschopp J. 2002. Defective death receptor signaling as a cause of tumor immune escape. *Seminars in cancer biology*. **12**: 51-55.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22. http://www.jstatsoft.org/v33/i01/.

Gillet JP, Calcagno AM, Varma S, Marino M, Green LJ, Vora MI, Patel C, Orina JN, Eliseeva TA, Singal V, et al. 2011. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc Natl Acad Sci*. **108**: 18708-18713.

Goldschmidt M, Peña L, Rasotto R, and Zappulli V. 2011. Classification and grading of canine mammary tumors. *Vet pathology*. **48**: 117-131.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nat genet*. **47**: 569-576.

Hanahan D and Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell*. **144**: 646-674.

Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, Ruddy K, Tsang J, and Cardoso F. 2019. Breast cancer (Primer). *Nature Reviews: Disease Primers.*

Harris RS. 2015. Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer. *Breast Cancer Res*. **17**: 8. doi: 10.1186/s13058-014-0498-3.

Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F, et al. 2012. Subtype and pathway specific responses to anti-cancer compounds in breast cancer. *Proc Natl Acad Sci*. **109**: 2724-2729.

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. 2018. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*. **173**: 291-304.

Hughes DA, Kircher M, He Z, Guo S, Fairbrother GL, Moreno CS, Khaitovich P, Stoneking M. 2015. Evaluating intra-and inter-individual variation in the human placental transcriptome. *Genome biology*. **16**: 54.

Jansson S, Aaltonen K, Bendahl PO, Falck AK, Karlsson M, Pietras K, Ryden L. 2018. The PDGF pathway in breast cancer is linked to tumour aggressiveness, triple-negative subtype and early recurrence. *Breast cancer res & treatment*. **169**: 231-241.

Karagiannis GS, Pastoriza JM, Wang Y, Harney AS, Entenberg D, Pignatelli J, Sharma VP, Xue EA, Cheng E, D'Alfonso TM, et al. 2017. Neoadjuvant chemotherapy induces breast cancer metastasis through a TMEM-mediated mechanism. *Sci transl med*. **9**: eaan0026. doi:10.1126/scitranslmed.aan0026

Kasprzyk A. 2011. BioMart: driving a paradigm change in biological data management. *Database*. **2011**: doi:10.1093/database/bar049

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat biotech*. **37**: 907-915.

Klopfleisch R, Von Euler H, Sarli G, Pinho S, Gartner F, Gruber A. 2011. Molecular carcinogenesis of canine mammary tumors: news from an old disease. *Vet pathology*. **48**: 98-116.

Kol A, Arzi B, Athanasiou KA, Farmer DL, Nolta JA, Rebhun RB, Chen X, Griffiths LG, Verstraete FJ, Murphy CJ, et al. 2015. Companion animals: Translational scientist's new best friends. *Sci transl med*. **7**: 308ps21-308ps21. doi:10.1126/scitranslmed.aaa9116

Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG. 2016. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat neuroscience*. **19**: 1454-1462.

Kristiansen V, Pena L, Diez Cordova L, Illera J, Skjerve E, Breen A, Cofone M, Langeland M, Teige J, Goldschmidt M, et al. 2016. Effect of ovariohysterectomy at the time of tumor removal in dogs with mammary carcinomas: a randomized controlled trial. *J of vet internal med*. **30**: 230-241.

LeBlanc AK, Breen M, Choyke P, Dewhirst M, Fan TM, Gustafson DL, Helman LJ, Kastan MB, Knapp DW, Levin WJ, et al. 2016. Perspectives from man's best friend:

National Academy of Medicine's Workshop on Comparative Oncology. *Sci transl med*. **8**: 324ps5-324ps5.

Lee I, Sohn M, Lim H, Yoon S, Oh H, Shin S, Shin J, Oh S, Kim J, Lee D, et al. 2014. Ahnak functions as a tumor suppressor via modulation of TGF /Smad signaling pathway. *Oncogene*. **33**: 4675-4684.

Leek JT, Johnson WE, Parker HS, Fertig EJ, Ja e AE, Storey JD, Zhang Y, Torres LC. 2019. sva: Surrogate Variable Analysis. R package version 3.34.0. doi:10.18129/B9.bioc.sva

Liu D, Xiong H, Ellis AE, Northrup NC, Rodriguez CO, O'Regan RM, Dalton S, and Zhao S. 2014. Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. *Cancer res*. **74**: 5045-5056.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, and Chen WM. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics. **26**: 2867–2873.

Mantovani F, Collavin L, Del Sal G. 2019. Mutant p53 as a guardian of the cancer cell. *Cell Death & Differentiation*. **26**: 199-212.

McCarthy DJ, Chen Y, and Smyth GK. 2012. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic acids res*. **40**: 4288-4297.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat genet*. **34**: 267-273.

The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature*. **490**: 61-70.

Paoloni M and Khanna C. 2008. Translation of new cancer treatments from pet dogs to humans. *Nat Rev Cancer*. **8**: 147-156.

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J of clin onc*. **27**: 1160-1167.

Pereira B, Chin SF, Rueda OM, Vollan HKM, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, et al. 2016. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat comm*. **7**: 1-16.

Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, Pollack JR,Ross DT, Johnsen H, Akslen LA, et al. 2000. Molecular portraits of human breast tumours. *Nature*. **406**: 747–752.

Pinho SS, Carvalho S, Cabral J, Reis CA, Gartner F. 2012. Canine tumors: a spontaneous animal model of human carcinogenesis. *Translational Research*. **159**: 165-172.

Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**: 841-842.

R Core Team. 2019. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria https://www.R-project.org.

Rangarajan A and Weinberg RA. 2003. Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nat rev Cancer*. **3**: 952-959.

Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. 2013. Drift and conservation of differential exon usage across tissues in primate species. *Proc Natl Acad Sci.* **110**: 15377-15382.

Robinson MD and Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*. **11**: R25. doi.org/10.1186/gb-2010-11-3-r25

Rowell JL, McCarthy DO, Alvarez CE. 2011. Dog models of naturally occurring cancer. *Trends Mol Med*. **17**: 380-388.

Schito L, Rey S, Tafani M, Zhang H, Wong CCL, Russo A, Russo MA, Semenza GL. 2012. Hypoxia-inducible factor 1-dependent expression of platelet-derived growth

factor B promotes lymphatic metastasis of hypoxic breast cancer cells. *Proc Natl Acad Sci*. **109**: E2707-E2716.

Schwender H. 2019. siggenes: Multiple Testing using SAM and Efron's Empirical Bayes Approaches. R package version 1.60.0. doi:10.18129/B9.bioc.siggenes

Shi H, Fu C, Wang W, Li Y, Du S, Cao R, Chen J, Sun D, Zhang Z, Wang X, et al. 2014. The FGF-1-specific single-chain antibody scFv1C9 selectively inhibits breast cancer tumour growth and metastasis. *J Cell Mol Med*. **18**: 2061-2070.

Silva TA, Smuczek B, Valadão IC, Dzik LM, Iglesia RP, Cruz MC, Zelanis A, de Siqueira AS, Serrano SM, Goldberg GS, et al. 2016. AHNAK enables mammary carcinoma cells to produce extracellular vesicles that increase neighboring fibroblast cell motility. *Oncotarget*. **7**: 49998–50016.

Sorenmo KU, Kristiansen VM, Cofone MA, Shofer FS, Breen A, Langeland M, Mongil CM, Grondahl AM, Teige J, Goldschmidt MH. 2009. Canine mammary gland tumours; a histological continuum from benign to malignant; clinical and histopathological evidence. *Vet Comp Oncol*. **7**: 162-172.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron J, Nobel A, Deng S, Johnsen H,Pesich R, Geisler S, et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences. **100**: 8418–8423.

Stein WD, Litman T, Fojo T, Bates SE. 2004. A Serial Analysis of Gene Expression (SAGE) Database Analysis of Chemosensitivity: Comparing Solid Tumors with Cell Lines and Comparing Solid Tumors from Different Tissue Origins. *Cancer Res*. **64**: 2805-2816.

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. 2007. Gene-expression variation within and among human populations. *Am J Hum Genet* **80**: 502-509.

Storey JD and Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci.* **100**: 9440-9445.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* **102**: 15545-15550.

Sun Z, Bhagwate A, Prodduturi N, Yang P, Kocher JPA. 2017. Indel detection from RNA-seq data: Tool evaluation and strategies for accurate detection of actionable mutations. *Briefings in bioinf*. **18**: 973-983.

Takahashi H, Asaoka M, Yan L, Rashid OM, Oshi M, Ishikawa T, Nagahashi M, Takabe K. 2020. Biologically Aggressive phenotype and Anti-cancer immunity counterbalance in Breast cancer with High Mutation Rate. *Scientific reports.* **10**: 1-13.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. **47**: D941-D947.

Toole MJ, Kidwell KM, Van Poznak C. 2014. Oncotype Dx results in multiple primary breast cancers. *Breast cancer basic & clin res.* **8**: 1-6.

Urra H, Dufey E, Avril T, Chevet E, Hetz C. 2016. Endoplasmic Reticulum Stress and the Hallmarks of Cancer. *Trends in cancer*. **2**: 252-262.

Weissmueller S, Manchado E, Saborowski M, Morris IV JP, Wagenblast E, Davis CA, Moon SH, Pfister NT, Tschaharganeh DF, Kitzing T, et al. 2014. Mutant p53 drives pancreatic cancer metastasis through cell-autonomous PDGF receptor signaling. *Cell*. **157**: 382-394.

Yates LR, Knappskog S, Wedge D, Farmery JH, Gonzalez S, Martincorena I, Alexandrov LB, Van Loo P, Haugland HK, Lilleng PK, et al. 2017. Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer cell*. **32**: 169-184.e7.

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

Zhang Y, Zhao Y, Li L, Shen Y, Cai X, Zhang X, Ye L. 2013. The oncoprotein HBXIP upregulates PDGFB via activating transcription factor Sp1 to promote the proliferation of breast cancer cells. *Biochem & biophys res comm*. **434**: 305-310.

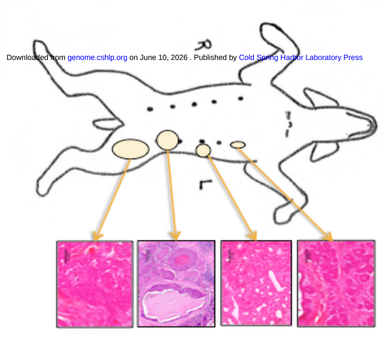
Figure 1

CMT Model

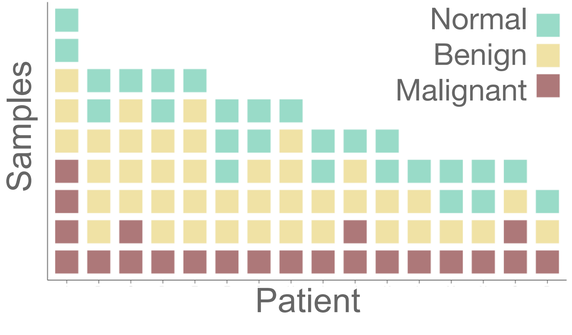
FREYA Framework

Multiple samples/patient

Sample collection

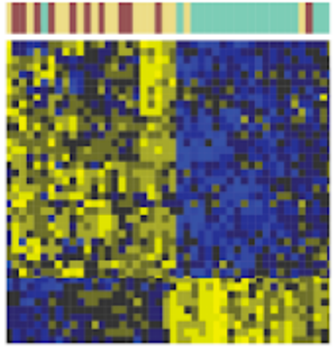


Histopathology

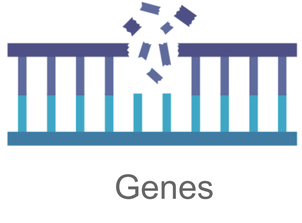


Data Processing

RNA-seq

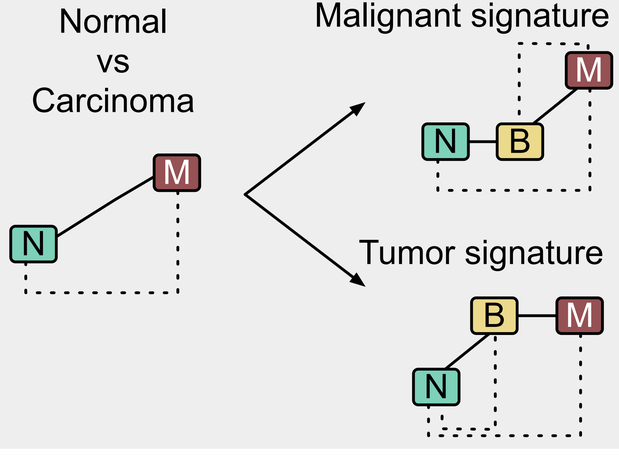


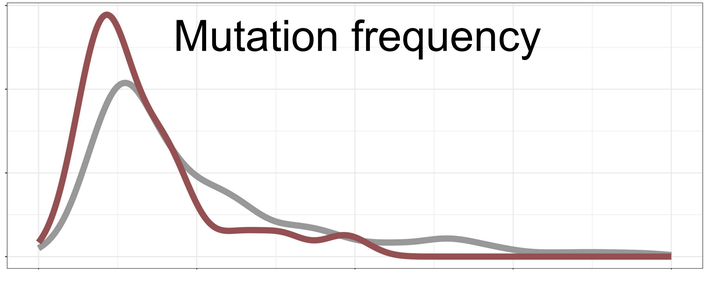
Somatic Mutations



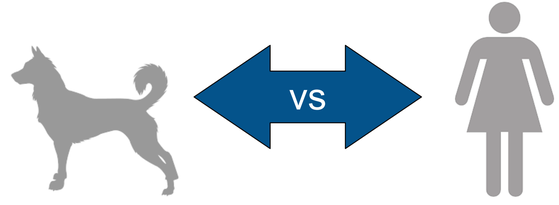
Statistical Analysis

Progression Expression Patterns (PEPs)

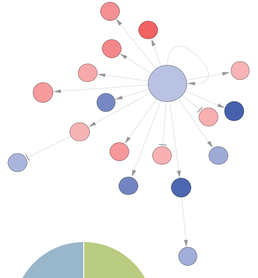




Comparisons and Insights



Pathway dysregulation analysis

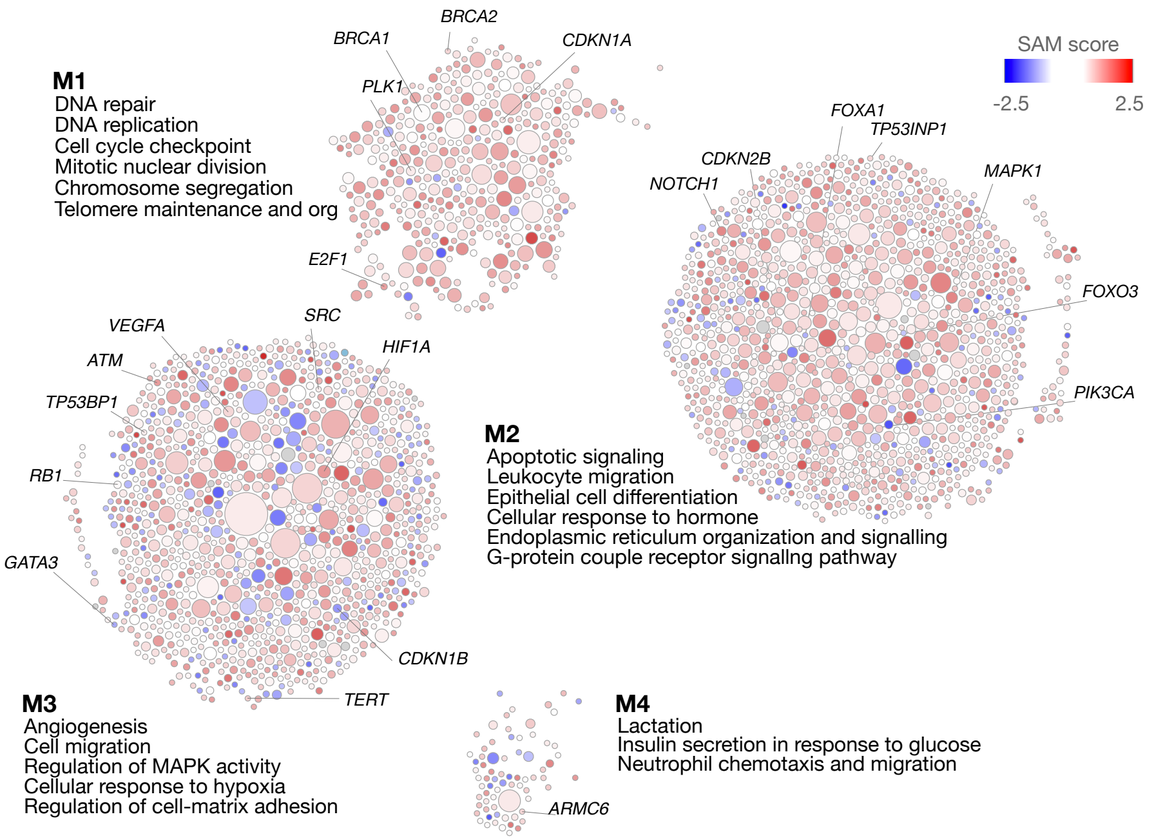


PAM50 clinical subtypes

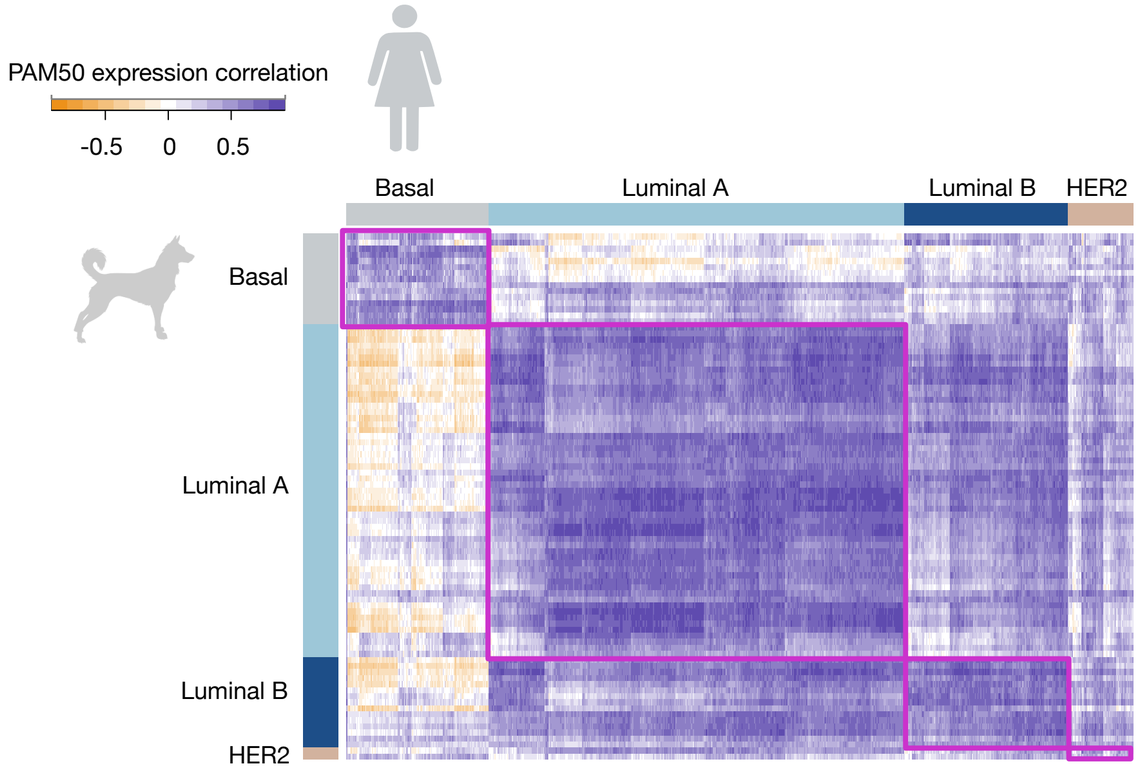
COSMIC cancer mutations

Figure 2

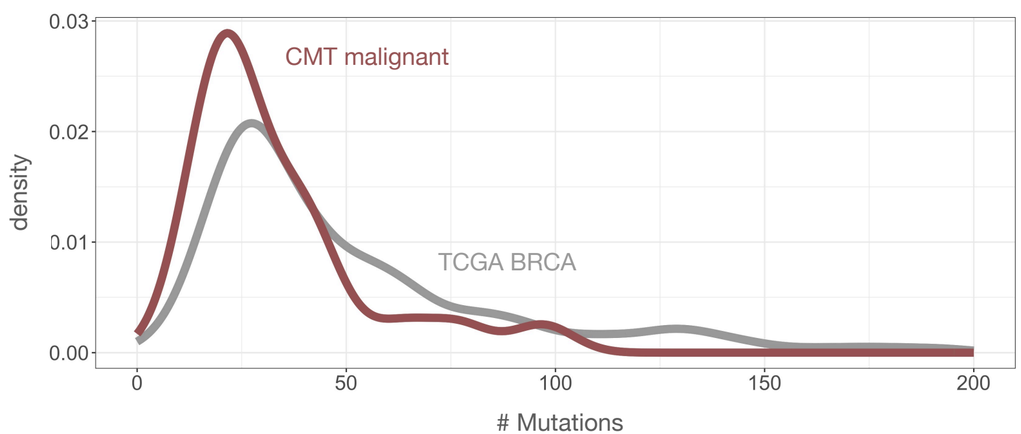
A



B



C



D

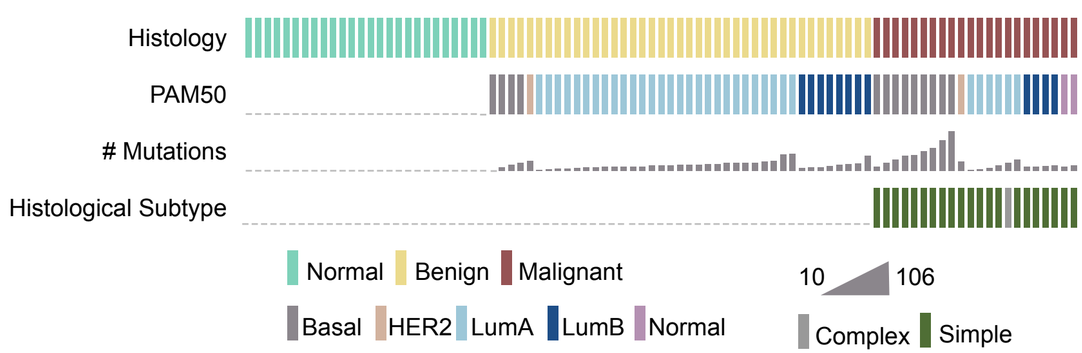


Figure 3

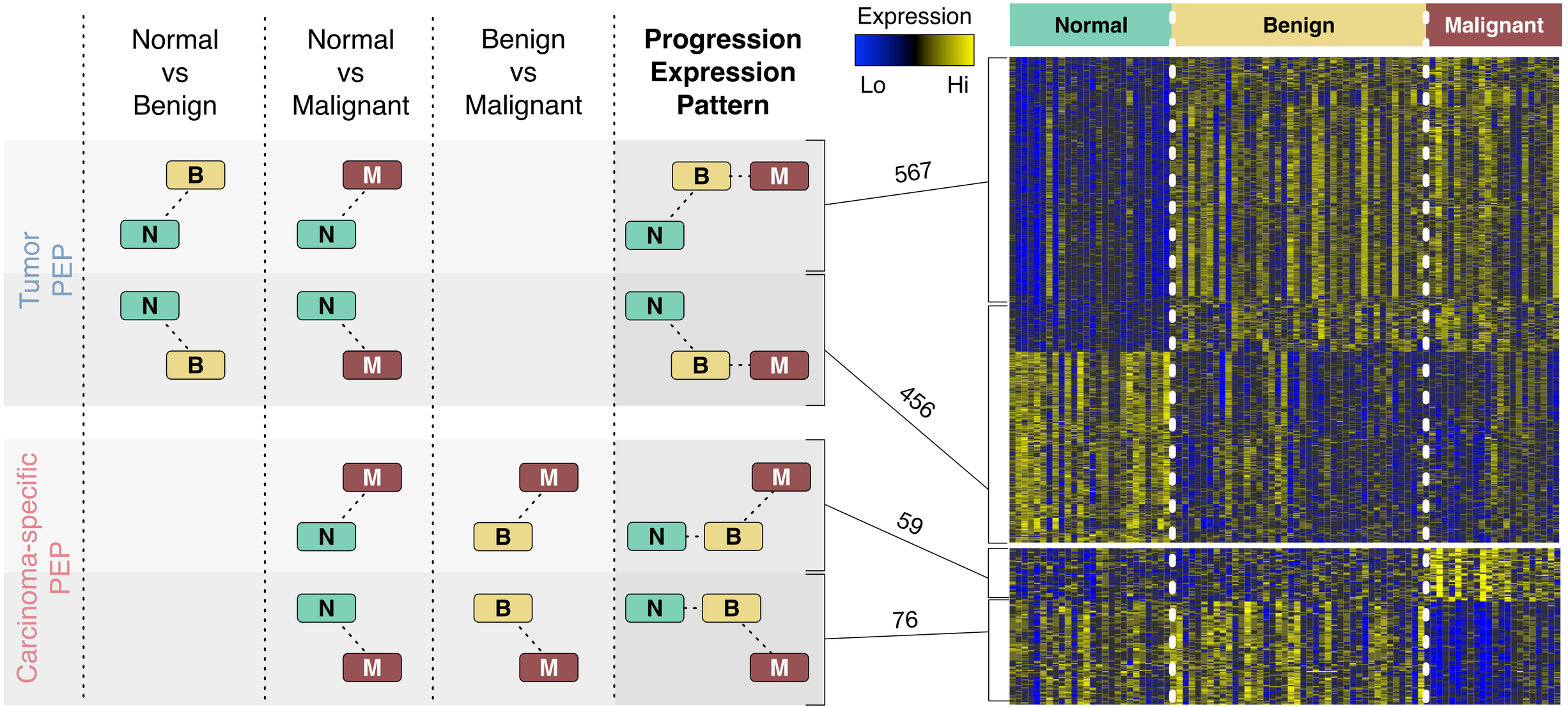


Figure 4

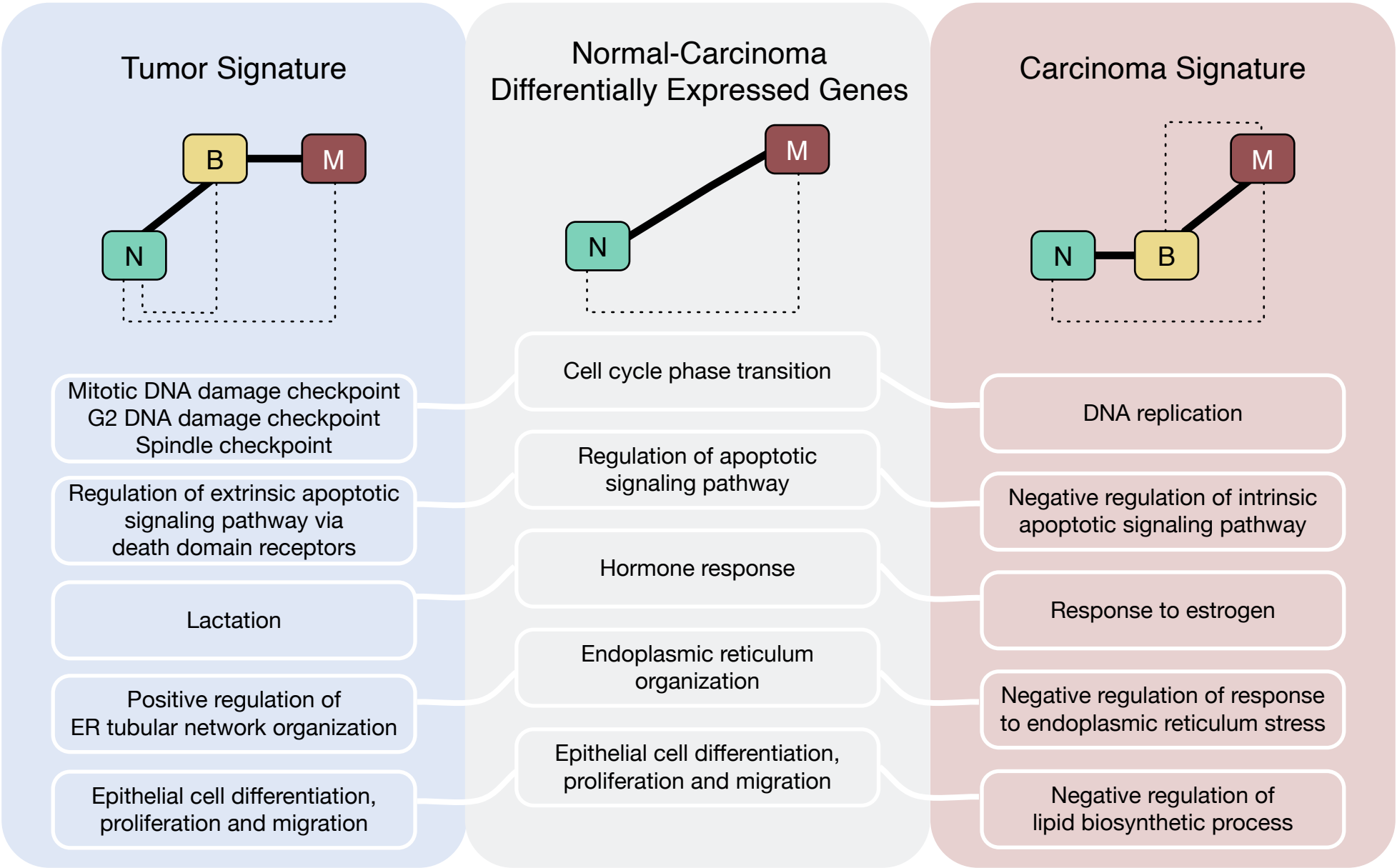

Tumor Signature
B
M
N
Mitotic DNA damage checkpoint
G2 DNA damage checkpoint
Spindle checkpoint
Regulation of extrinsic apoptotic signaling pathway via death domain receptors
Lactation
Positive regulation of ER tubular network organization
Epithelial cell differentiation, proliferation and migration
Normal-Carcinoma Differentially Expressed Genes
M
N
Cell cycle phase transition
Regulation of apoptotic signaling pathway
Hormone response
Endoplasmic reticulum organization
Epithelial cell differentiation, proliferation and migration
Carcinoma Signature
M
N
B
DNA replication
Negative regulation of intrinsic apoptotic signaling pathway
Response to estrogen
Negative regulation of response to endoplasmic reticulum stress
Negative regulation of lipid biosynthetic process

Figure 5