수의학박사학위논문

오믹스 데이터를 이용한 개와 사람의 바이오마커 비교연구

Comparative Studies on Human and Canine Mammary Carcinoma Biomarkers by Omics Data Analysis

2021년 2월

서울대학교 대학원

수의학과 수의생명과학 전공

(수의생화학)

오믹스 데이터를 이용한 개와 사람의 바이오마커 비교연구

지도 교수 조 제 열

이 논문을 수의학박사학위논문으로 제출함
2020년 11월

서울대학교 대학원
수의학과 수의생명과학 전공 (수의생화학)
박 형 민

박형민의 수의학박사 학위논문을 인준함
2020년 12월

위 원 장 이 항 (인)

부위원장 조 제열 (인)

위 원 이 소 영 (인)

위 원 한 규 동 (인)

위 원 김 평 환 (인)

Comparative Studies on Human and Canine Mammary Carcinoma Biomarkers by Omics Data Analysis

Under the supervision of Professor Je-Yoel Cho

DISSERTATION

Presented in Partial Fulfillment of the Requirement for the Degree of DOCTOR OF PHILOSOPHY

By

Hyoung-Min Park

Major in Veterinary Biomedical Sciences
(Veterinary Biochemistry)
Department of Veterinary Medicine
The Graduate School
Seoul National University

February 2021

Abstract

Comparative Studies on Human and Canine Mammary Carcinoma Biomarkers by Omics Data Analysis

Breast cancer (BC), known as mammary gland carcinoma (MGC), is one of the most frequently diagnosed malignancies among women and canines. Despite the countless efforts to fully understand and overcome such cancer-related anomalies, various subtypes originating from specific regions of the mammary organ generates infrequent yet menacing malignancies. Comparative medicinal approach has emerged as a powerful method to approach human BC research on a different perspective. Together with various omics technologies, the paradigm for BC treatment has become shifting toward evidence-based large-scale discovery studies which leads to biomarkers specifically expressed in distinct BC subtypes. The incorporation of diverse omics data spreading from next generation sequencing (NGS) assembled epigenetic transcripts to mass spectrometry (MS) derived proteomics stands as a solution for breast malignancy differential diagnosis and drug target discovery. The research is divided into three chapters for detailed description.

CHAPTER Ⅰ describes sequenced RNA-seq data from ten pairs of canine mammary gland carcinoma (MGC) and matching adjacent normal tissues to identify canine MGC-associated transcriptomic signatures. Breast cancer (BC) and MGC is the most frequently diagnosed and leading cause of cancer-related mortality in both women and canines. To better understand both canine MGC- and human BC-specific genes which express similar transcriptomic profiles, we sequenced RNAs obtained from eight pairs of carcinomas and adjacent normal tissues in dogs. By comprehensive transcriptome analysis, 351 differentially expressed genes (DEGs) were identified in overall canine MGCs. Based on the DEGs, comparative analysis revealed correlation existing among the three histological subtypes of canine MGC (ductal, simple, and complex) and four molecular subtypes of human BC (HER2+, ER+, ER & HER2+, and TNBC). Eight DEGs shared by all three subtypes of canine MGCs had been previously reported as cancer-associated genes in human studies. Gene ontology (GO) and pathway analyses using the identified DEGs revealed that the biological processes of cell proliferation, adhesion, and inflammatory responses are enriched in up-regulated MGC DEGs. In contrast, fatty acid homeostasis and transcription regulation involved in cell fate commitment were down-regulated in MGC DEGs. Moreover, correlations are demonstrated between upstream promoter transcripts and DEGs. Canine MGC- and subtype-enriched gene expression allows us to better understand both human BC and canine MGC,

yielding new insight into the development of biomarkers and targets for both diseases. The resemblance in transcriptomic profiles will present canines as a suitable comparative model for MGC studies and its application to human BC.

CHAPTER Ⅱ focuses on the identification and treatment specific to a BC subtype. Among many types of BCs, triple-negative breast cancer (TNBC) has the worst prognosis and the least cases reported. To gain a better understanding and a more decisive precursor for TNBC, two major histone modifications, an activating modification H3K4me3 and a repressive modification H3K27me3, were analyzed using data from normal breast cell lines against TNBC cell lines. The combination of these two histone markers on the gene promoter regions showed a great correlation with gene expression. A list of signature genes was defined as active (highly enriched H3K4me3), including *NOVA1*, *NAT8L*, and *MMP16*, and repressive genes (highly enriched H3K27me3), *IRX2* and *ADRB2*, according to the distribution of these histone modifications on the promoter regions. To further enhance the investigation, potential candidates were also compared with other types of BC to identify signs specific to TNBC. RNA-seq data was implemented to confirm and verify gene regulation governed by the histone modifications. Combinations of the biomarkers based on H3K4me3 and H3K27me3 showed the diagnostic value area under the curve (AUC) 93.28% with P-value of 1.16e-226. The results of this study suggest that histone modification analysis of opposing histone modifications may be valuable toward

developing biomarkers and targets for TNBC and further provide understanding the overall regulation derived by epigenetic modifications.

CHAPTER Ⅲ consists of biomarker study implemented from canine mammary tumors to human BCs. While biomarkers are continuously discovered, specific markers representing the aggressiveness and invasiveness of BC are lacking compared to classification markers. In this study, samples from canine mammary tumors were used in a comparative approach. An extensive 36 fractions of both canine normal and MGC plasma was subjected to high-performance quantitative proteomics analysis. Among the identified proteins, Lecithin-Cholesterol Acyltransferase (LCAT) was discovered to be selectively expressed in mixed tumor samples, which represents an aggressive developed stage of cancer, possibly highly metastatic. With further multiple reaction monitoring (MRM) and western blot validation, we discovered that the LCAT protein is an indicator of aggressive mammary tumor. Interestingly, we also found that LCAT is overexpressed in high grade and lymph node positive BC *in silico* data. We also demonstrated that LCAT is highly expressed in the sera of advanced stage human BCs within the same classification. In conclusion, we identified a possible common plasma protein biomarker, LCAT, that is highly expressed in aggressive human BC and canine mammary tumor.

Keyword: Breast cancer, Comparative medicine, Biomarker, Epigenetics, Proteomics

Student Number: 2016-21753

CONTENTS

LIST OF FIGURES

BACKGROUND

CHAPTER Ⅰ

CHAPTER Ⅱ

CHAPTER Ⅲ

LIST OF TABLES

BACKGROUND

CHAPTER Ⅰ

CHAPTER Ⅱ

CHAPTER Ⅲ

ABBREVIATIONS

AUC	Area under the Curve
BC	Breast Cancer
BP	Biological Process
CC	Cellular Component
CCRC	Canine Cancer Research Center project
CNV	Copy Number Variation
CT	Computed Tomography
DEG	Differentially Expressed Genes
ER	Estrogen
ESI	Electrospray Ionization
FBS	Fetal Bovine Serum
FPKM	Fragments per Kilobase of Exon per Million Fragments Mapped
GO	Gene Ontology
GTF	Gene Transfer Format

HDL	High Density Lipoprotein
HPLC	High Performance Liquid Chromatography
HR	Hazard Ratio
IGV	Integrative Genomic Viewer
KM	Kaplan-Meier
LC	Liquid Chromatogram
lncRNA	Long Noncoding RNA
MALDI	Matrix-Assisted Laser Desorption/Ionization
MEBM	Mammary Epithelial Cell Growth Basal Medium
MF	Molecular Function
MGC	Mammary Gland Carcinoma
miRNA	Micro RNA
MS	Mass Spectrometry
MRM	Multiple Reaction Monitoring
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
ncRNA	Noncoding RNA

PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PR	Progesterone
PROMPT	Promoter Upstream Transcripts
PTM	Post-Translational Modification
RNA-seq	RNA Sequencing
ROC	Receiver Operation Characteristics
rRNA	Ribosomal RNA
SNP	Single Nucleotide Polymorphism
TIN	Transcript Integrity Number
TNBC	Triple-Negative Breast Cancer
TSS	Transcriptional Start Site

BACKGROUND

1. BREAST CANCER

Cancer has been a constant threat in Korea. To make matters worse, cancer incidence and mortality are rapidly growing within the Korean society. Cancer accounts for one in four deaths and more than 200,000 new cancer cases were diagnosed in 2015 (Jung et al., 2018). Statistics measured in Korea reported the incidence rate increased significantly by 3.6% annually from 1999 to 2011 with 229,180 and 78,194 Koreans newly diagnosed and died from cancer in 2016 (Table B-1). Despite the decreased incidence and mortality rate from recent years, cancer will still remain as a major cause of human casualties (Jung et al., 2019).

BC stands as the second most frequent type among cancers and the most common cancer among women that accounted for 24.2% of the cases (Bray et al., 2018). According to the statistics report dating back from 1930 to 2017, cancer originating from female reproductive organs tents to decrease with the exception of BC which is now the leading occurring cancer among women (Fig. B-1). Furthermore, cancer incidence during childhood (ages birth-14 years) is approximately 10% higher in males than in females (18.2 vs 16.4 per 100,000 population), whereas during early adulthood (ages 20-49 years) it is 77% higher in females (203.4 vs 114.9 per 100,000 population), largely because of BC incidence in young women (Siegel et al., 2020)

During the course of cancer study, BC research, therapy, and prediction has been constantly reported by various groups. However, numerous cases of symptoms and subtypes are still far from complete comprehension. Diagnosis is largely depended on computer imaging techniques such as X-rays and computed tomography (CT) scans. Certain types of BC prediction approaches are done with minimal precursors such as estrogen (ER), progesterone (PR), and HER2 expression. The need for additional markers which can pinpoint or predict specific BC types remains.

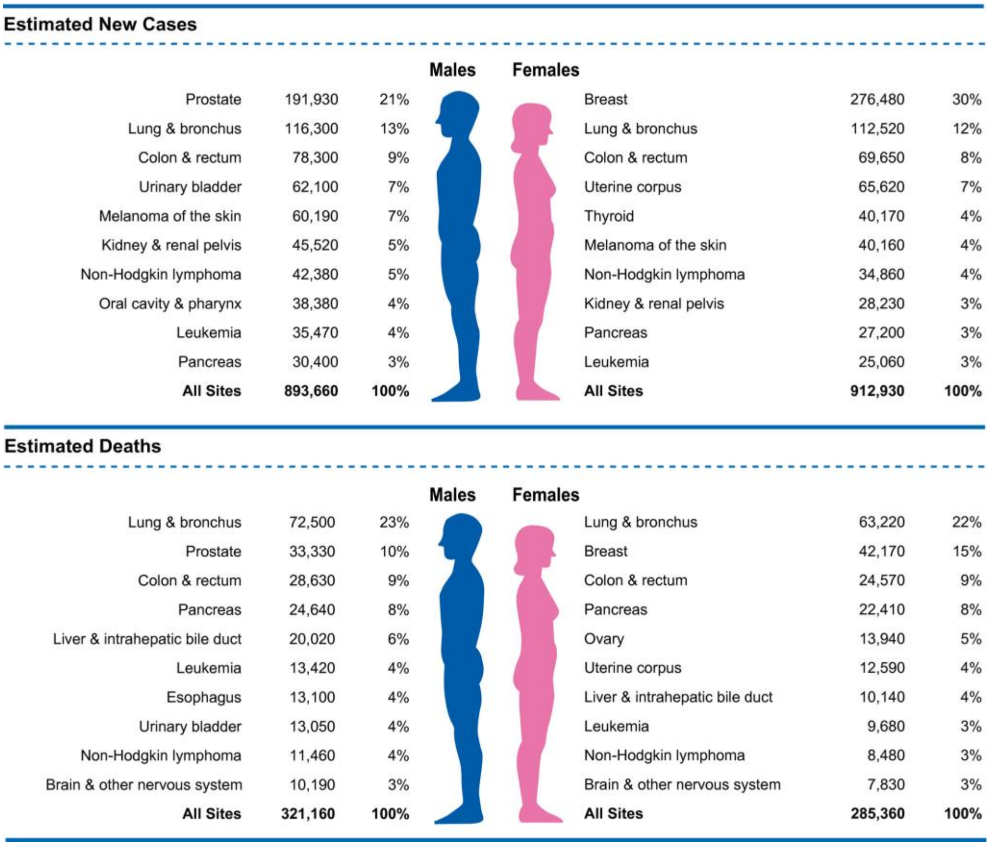
Table B-1. Cancer incidence, deaths, and prevalence by sex in Korea, 2016

SITE/TYPE	**NEW CASES**			**DEATHS**			**PREVALENT CASES**[A)]		
	Both sexes	Men	Women	Both sexes	Men	Women	Both sexes	Men	Women
ALL SITES	229,180	120,068	109,112	78,194	48,208	29,986	1,739,951	764,103	975,848
LIP, ORAL CAVITY, AND PHARYNX	3,543	2,527	1,016	1,203	909	294	23,639	15,847	7,792
ESOPHAGUS	2,499	2,245	254	1,524	1,379	145	9,777	8,780	997
STOMACH	30,504	20,509	9,995	8,264	5,318	2,946	273,701	181,234	92,467
COLON AND RECTUM	28,127	16,672	11,455	8,358	4,659	3,699	236,431	140,852	95,579
LIVER	15,771	11,774	3,997	11,001	8,044	2,957	64,864	48,666	16,198
GALLBLADDER[B)]	6,685	3,490	3,195	4,408	2,248	2,160	21,011	10,776	10,235
PANCREAS	6,655	3,384	3,271	5,614	2,901	2,713	10,595	5,502	5,093
LARYNX	1,167	1,101	66	334	310	24	10,532	9,914	618
LUNG	25,780	17,790	7,990	17,963	13,324	4,639	76,544	47,438	29,106
BREAST	21,839	92	21,747	2,472	16	2,456	198,006	743	197,263
CERVIX UTERI	3,566	-	3,566	897	-	897	52,758	-	52,758
CORPUS UTERI	2,771	-	2,771	313	-	313	23,135	-	23,135

OVARY	2,630	-	2,630	1,204	-	1,204	19,509	-	19,509
PROSTATE	11,800	11,800	-	1,745	1,745	-	77,635	77,635	-
TESTIS	288	288	-	14	14	-	3,204	3,204	-
KIDNEY	5,043	3,410	1,633	1,032	724	308	38,836	26,161	12,675
BLADDER	4,361	3,488	873	1,389	1,029	360	33,543	27,347	6,196
BRAIN AND CNS	2,015	1,104	911	1,327	720	607	11,116	5,734	5,382
THYROID	26,051	5,538	20,513	346	104	242	379,946	65,336	314,610
HODGKIN LYMPHOMA	312	202	110	51	33	18	2,807	1,770	1,037
NON-HODGKIN LYMPHOMA	4,766	2,766	2,000	1,820	1,068	752	30,093	17,130	12,963
MULTIPLE MYELOMA	1,535	837	698	1,010	527	483	5,798	3,050	2,748
LEUKEMIA	3,416	1,991	1,425	1,842	1,025	817	20,751	11,553	9,198
OTHER AND ILL-DEFINED	18,056	9,060	8,996	4,063	2,111	1,952	115,720	55,431	60,289

Source: Cancer statistics, 2016, Statistics Korea.

Adapted from Jung et al., 2019



Estimated New Cases

Males			Females		
Prostate	191,930	21%	Breast	276,480	30%
Lung & bronchus	116,300	13%	Lung & bronchus	112,520	12%
Colon & rectum	78,300	9%	Colon & rectum	69,650	8%
Urinary bladder	62,100	7%	Uterine corpus	65,620	7%
Melanoma of the skin	60,190	7%	Thyroid	40,170	4%
Kidney & renal pelvis	45,520	5%	Melanoma of the skin	40,160	4%
Non-Hodgkin lymphoma	42,380	5%	Non-Hodgkin lymphoma	34,860	4%
Oral cavity & pharynx	38,380	4%	Kidney & renal pelvis	28,230	3%
Leukemia	35,470	4%	Pancreas	27,200	3%
Pancreas	30,400	3%	Leukemia	25,060	3%
All Sites	**893,660**	**100%**	**All Sites**	**912,930**	**100%**

Estimated Deaths

Males			Females		
Lung & bronchus	72,500	23%	Lung & bronchus	63,220	22%
Prostate	33,330	10%	Breast	42,170	15%
Colon & rectum	28,630	9%	Colon & rectum	24,570	9%
Pancreas	24,640	8%	Pancreas	22,410	8%
Liver & intrahepatic bile duct	20,020	6%	Ovary	13,940	5%
Leukemia	13,420	4%	Uterine corpus	12,590	4%
Esophagus	13,100	4%	Liver & intrahepatic bile duct	10,140	4%
Urinary bladder	13,050	4%	Leukemia	9,680	3%
Non-Hodgkin lymphoma	11,460	4%	Non-Hodgkin lymphoma	8,480	3%
Brain & other nervous system	10,190	3%	Brain & other nervous system	7,830	3%
All Sites	**321,160**	**100%**	**All Sites**	**285,360**	**100%**

Figure B-1. Ten Leading Cancer Types for the Estimated New Cancer Cases and Deaths by Sex, United States, 2020.

Estimates are rounded to the nearest 10 and exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder. Ranking is based on modeled projections and may differ from the most recent observed data.

Adapted from Siegel et al., 2020

2. COMPARATIVE MEDICINE

Comparative medicine is an experimental approach that implements animal models of both human and genetically close animal disease in translational and biomedical research (Bradley, 1927). It also substitutes as a means to relate and compare genetic, epigenetic, and other biological characteristics among species to better understand the mechanism and expression profiles of human and animal disease. This method further progress into comparative oncology, which integrates the study of oncology in mammals and implement the biologic, diagnostic and therapeutic knowledge to human cancer for a novel approach (Paoloni and Khanna, 2007). The mouse has been the most frequently used model for genetic studies in human oncology for it had a small size, average lifespan of two years, short gestation period and inexpensiveness in contrast to other mammals but has shown significant limitations and inconsistency when used to study complex human diseases (Gondo et al., 2009, Seok et al., 2013).

Among the animal models, canines have emerged as a strong comparative model due to many advantages as they experience spontaneous disease, genes similar to human, five to seven-fold accelerated ageing, and respond to treatments similarly as humans (Sultan and Ganaie, 2018). Similar to humans, cancer is the leading cause of death in canines of greater than 10 years of age (Gardner et al., 2016). The incidence of mammary tumors in the bitch is approximately three times greater than that in women (Owen, 1979). Canine MGC are biologically heterogeneous neoplasms offering several ways to classify such tumors on the basis of histopathological characteristics or expression of molecular markers (Sleeckx et al., 2011). Despite the appearance of histo-morphological variations between human and canine BC, due to various prognostic indicators, a number of studies have reported that there are significant similarities regarding molecular marker expression, hormone dependency and cancer phenotypes (Ahern et al., 1996, Misdorp, 1964). Recently, in more refined studies employing immunohistochemical approaches and based on the characteristic expression patterns of *ESR1*, *PR* and *EGFR* (*ERBB1*/*HER1*, *ERBB2*/*HER2*, *ERBB3* and *ERBB4*), human-like breast cancer phenotypes for canine MGCs have been developed and classified as luminal A, luminal B, *HER2* positive, and triple-negative (basal-like) (Kabir et al., 2017, Sassi et al., 2010). Such standard classification therefore strongly supports canine MGC as valuable intermediate models for human BC that should be well-placed for developing diagnostic and treatment strategies (Lutful Kabir et al., 2015). Because canine MGCs are considered predictive models for human BC (Vail and MacEwen, 2000), similarities in genetic alterations and cancer predisposition between humans and dogs have raised interest even further. A large number of studies have demonstrated that canine MGCs have many similarities in molecular and clinical features with human BC. Many genetic/epigenetic/tumor biology traits that are most frequently associated with MGC have been identified and comparative gene expression analysis has revealed a significant similarity in the canine and human genes associated with MGC development (Uva et al., 2009). Upon genomic comparison, analysis of deregulated gene sets or cancer signaling pathways showed that a significant proportion of orthologous genes are comparably up- and down-regulated in both human and canine BC. Prominent oncogenic pathways and related genes, such as *PI3K*/*AKT*, *KRAS*, *MAPK*, Wnt, β-catenin, *BRCA2*, *ESR1*, and P-cadherin, are commonly up-regulated while representative tumor suppressive pathways, such as p53, p16/*INK4A*, *PTEN*, and E-cadherin, are down-regulated in human and canine BC (Klopfleisch and Gruber, 2009, Lutful Kabir et al., 2013, Uva et al., 2009). Furthermore, chromosomal studies via molecular and cytogenic mapping of the *INKA*/*ARF* locus depicts high resemblance between human chromosome 9 and canine chromosome 11 (Fig. B-2).

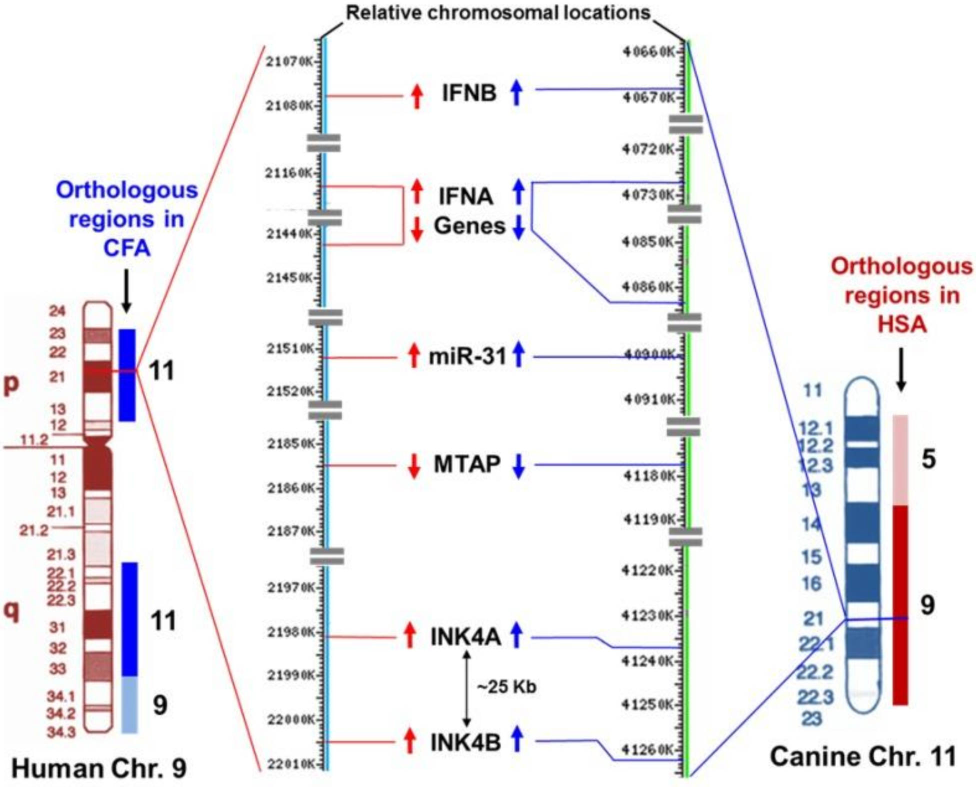


Figure B-2. Frequently deleted regions in human chr. 9p21 and orthologous canine chr. 11

Relative molecular and cytogenetic mapping of the *INKA/ARF* locus and closely related genes with their positions on human and canine chromosome 9 and 11, respectively. The regions at human chromosome 9 and canine chromosome 11 that are frequently deleted in cancers are completely orthologous to each other. The molecular mapping shows the exact chromosomal position of these genes extrapolated from the NCBI map view of each chromosome represented by the current human and canine annotation from releases 106 and 103, respectively. The red and blue arrows indicate the transcriptional orientation of genes in the human and dog chromosomes, respectively. Transcription of genes from the "+ strand" is indicated by down arrows and from the “– strand” by up arrows. (CFA = Canis lupus familiaris; HSA = Homo sapiens; Chr. = Chromosome).

Adapted from Kabir et al., 2016

3. BIOMARKERS

Biomarker, a portmanteau of “biological marker”, is a combined term of measurement that can define the normal and abnormal status of an individual. Biomarkers have been recently emerged as a strong means of diagnostic and therapeutic approach. The discovery of predictive biomarkers will save time and money, and lead to minimal invasion into human organs. Biomarkers include any type of hallmark of physiological states, such as expression profiles, images, genes, or proteins (Dalton and Friend, 2006). An even broader definition takes into account not just incidence and outcome of disease, but also the effects of treatments, interventions, and even unintended environmental exposure, such as to chemicals or nutrients (Strimbu and Tavel, 2010).

Among the variety, genetic biomarkers have proven useful not limited to diagnose and designate appropriate treatments. Combinations of marker expressions further lead to characterization and classification in certain BCs. The most common genomic biomarkers used today are the prognostic markers designed to classify BC in to five distinctive subtypes; Luminal A, Luminal B (*HER2* positive), Luminal B (*HER2* negative), *HER2* positive (non-luminal), and triple-negative breast cancer (TNBC) for an efficient way of diagnosis and therapy (Table B-2). Recent reports studying differential expression patterns of genomic transcriptomes under certain malignant conditions seek to identify more biomarkers that will be more acceptable to understand and characterize BC.

As the human understanding of genomics spread to proteomics and various methods that can identify protein or protein derived modifications develop, protein biomarkers also emerged as a powerful indicator for BC research. The advantages of proteins as a class of biomarkers include their enormous diversity, dynamic turnover and secretion into blood and bodily fluids. There is an estimated number of 20,0300 genes (Legrain et al., 2011), 40,000 unique metabolites (Wishart et al., 2013), ~100,000 mRNA transcripts, and up to 1.8 million of different proteoforms, if posttranslational modifications (PTMs) are considered (Jensen, 2004). Such enormous diversity in proteoforms increases the chances to identify a marker, or a panel of markers, for each disease state. Since protein sequences may also reflect some genomic variations, a single instrumentation platform of mass spectrometry can measure not only changes in protein abundance but also genomic and transcriptomic variations, such as mutant proteins (Drabovich et al., 2015).

Table B-2. Therapy recommendations of the recent St. Gallen Consensus 8. (ET: endocrine therapy; CT: chemotherapy; Anti-HER2: anti-HER2 therapy).

Subtype	Therapy
Luminal A	ET
Luminal B (HER2 negative)	ET ± CT (“after risk assessment”)
Luminal B (HER2 positive)	CT + Anti-HER2 + ET
HER2 positive (non-luminal)	CT + Anti-HER2
Triple negative	CT

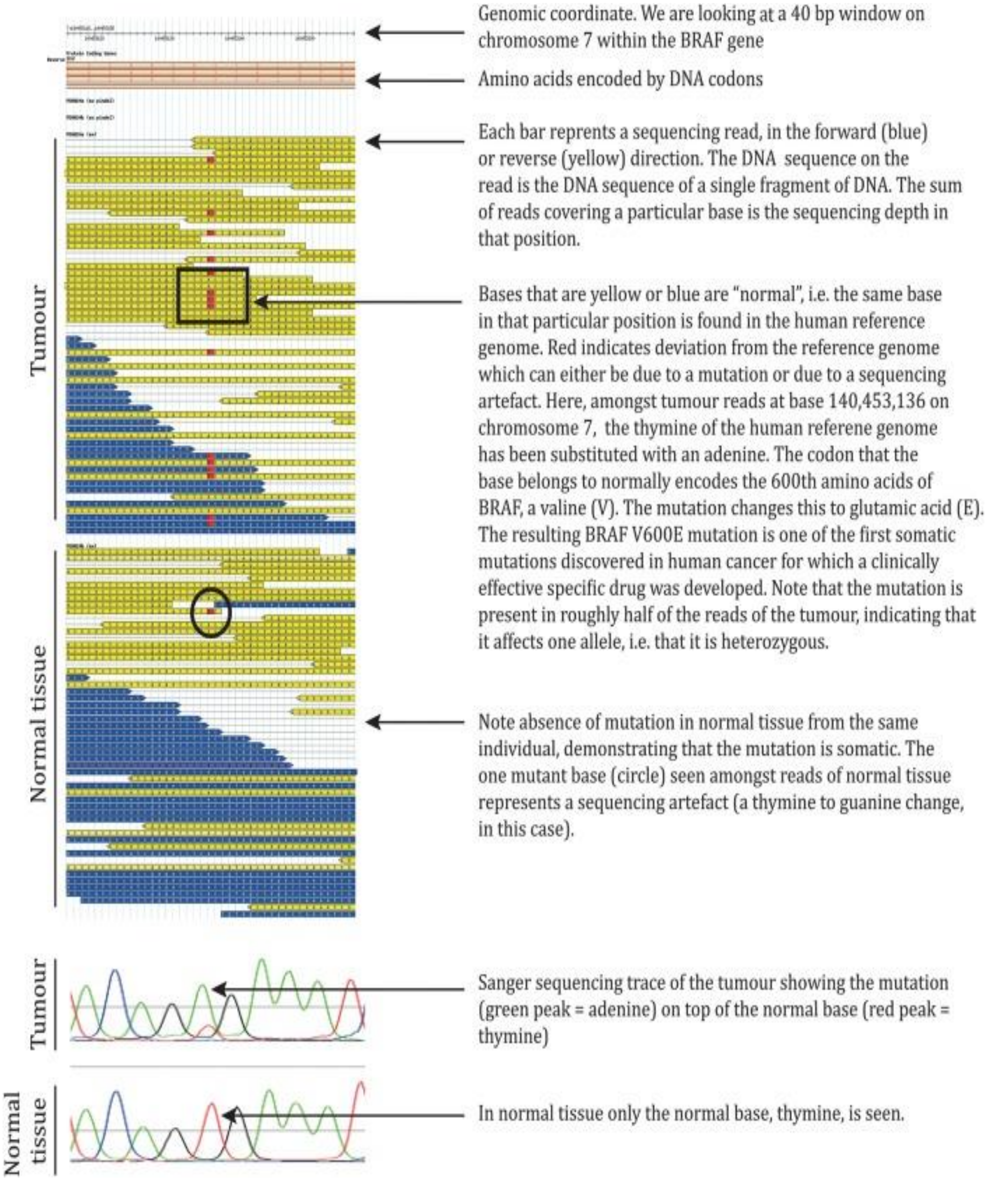
ET: endocrine therapy; CT: chemotherapy; Anti-HER2: anti-HER2 therapy.

Adapted from Schmidt et al.,2012

4. NEXT GENERATION SEQUENCING

Next generation sequencing (NGS) is a term that can be described as a method to determine the nucleic acid sequence of a particular sample in a rapid and cost-efficient way. Early DNA sequences were obtained in the early 1970s by using laborious methods based on two-dimensional chromatography (Padmanabhan et al., 1974). However, as the technology to process high through-put sequencing were enhanced, the total storage and quality of the DNA sequence data were obtained both time and cost efficiently. In current times, there are a number of different NGS platforms using different sequencing technologies. The common mechanism of all NGS platforms is to perform sequencing of millions of small fragments of DNA in parallel. Bioinformatics analyses are used to piece together these fragments by mapping the individual reads to the human reference genome. Each of the three billion bases in the human genome is sequenced multiple times, providing high depth to deliver accurate data and an insight into unexpected DNA variation (Fig. B-3). NGS can be used to sequence entire genomes or constrained to specific areas of interest, including all 22,000 coding genes (a whole exome) or small numbers of individual genes (Behjati and Tarpey, 2013).

Among the methods derived from NGS, RNA sequencing (RNA-seq) is a particular technology-based sequencing technique to reveal the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome (Chu and Corey, 2012). RNA-Seq facilitates the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion, mutations/SNPs and changes in gene expression over time, or differences in gene expression in different groups or treatments (Maher et al., 2009). In addition to mRNA transcripts, RNA-Seq can look at different populations of RNA to include total RNA, small RNA, such as miRNA, tRNA, and ribosomal profiling (Ingolia et al., 2012). RNA-Seq can also be used to determine exon/intron boundaries and verify or amend previously annotated 5' and 3' gene boundaries (Fig. B-4).



Genomic coordinate. We are looking at a 40 bp window on chromosome 7 within the BRAF gene

Amino acids encoded by DNA codons

Each bar reprents a sequencing read, in the forward (blue) or reverse (yellow) direction. The DNA sequence on the read is the DNA sequence of a single fragment of DNA. The sum of reads covering a particular base is the sequencing depth in that position.

Bases that are yellow or blue are "normal", i.e. the same base in that particular position is found in the human reference genome. Red indicates deviation from the reference genome which can either be due to a mutation or due to a sequencing artefact. Here, amongst tumour reads at base 140,453,136 on chromosome 7, the thymine of the human referene genome has been substituted with an adenine. The codon that the base belongs to normally encodes the 600th amino acids of BRAF, a valine (V). The mutation changes this to glutamic acid (E). The resulting BRAF V600E mutation is one of the first somatic mutations discovered in human cancer for which a clinically effective specific drug was developed. Note that the mutation is present in roughly half of the reads of the tumour, indicating that it affects one allele, i.e. that it is heterozygous.

Note absence of mutation in normal tissue from the same individual, demonstrating that the mutation is somatic. The one mutant base (circle) seen amongst reads of normal tissue represents a sequencing artefact (a thymine to guanine change, in this case).

Sanger sequencing trace of the tumour showing the mutation (green peak = adenine) on top of the normal base (red peak = thymine)

In normal tissue only the normal base, thymine, is seen.

Figure B-3. Example of next generation sequencing (NGS) raw data-BRAF V600E mutation in melanoma.

The mutation was found by our group in 2002 as part of several year-long efforts to define somatic mutations in human cancer using Sanger sequencing, prior to the advent of NGS.

Adapted from Behjati et al., 2013

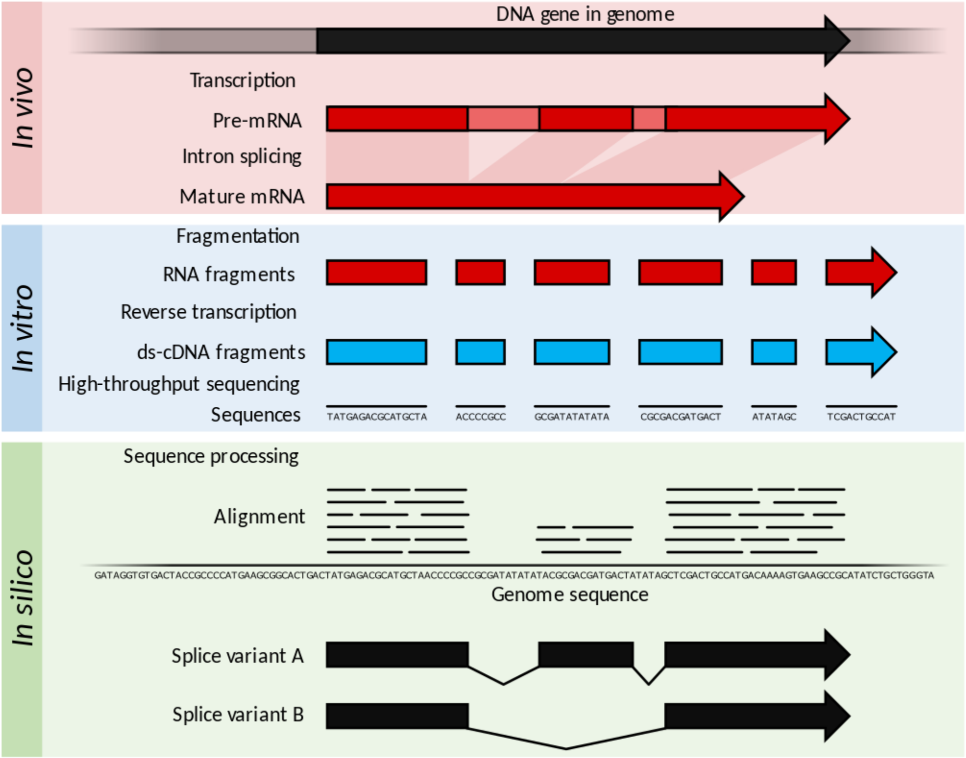


Figure B-4. Summary of RNA-Seq.

Within the organism, genes are transcribed and (in a eukaryotic organism) spliced to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and copied into stable ds-cDNA (blue). The ds-cDNA is sequenced using high-throughput, short-read sequencing methods. These sequences can then be aligned to a reference genome sequence to reconstruct which genome regions were being transcribed. This data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants.

Adopted from Shafee et al., 2017

5. MASS SPECTROMETRY-BASED PROTEOMICS

Edman degradation, which is used to sequence a protein, relies on the identification of amino acids that have been chemically cleaved in a stepwise fashion from the amino terminus of the protein and requires much expertise (Steen and Mann, 2004). In 1996, Mann and colleagues showed that MS could identify gel-separated proteins using a much smaller quantity of the sample than was required by Edman degradation, a method of sequencing amino acids in a peptide, and can fragment the peptides in seconds instead of hours or days (Wilm et al., 1996). Currently, MS-based proteomics has proliferated, and many biologists have access to a service to which they can submit a sample and are handed back a list of proteins that have been identified by MS.

To measure biomolecules, which can be peptides or proteins, by MS, analytes are ionized via electrospray ionization (Uva et al., 2009) or matrix-assisted laser desorption/ionization (MALDI) (Fig. B-5), and their mass is measured by following their specific trajectories in vacuum system. Ionized molecules are recorded as values on the m/z scale, which has units of mass per charge (Steen and Mann, 2004).

Having determined the m/z values and intensities of all the peaks in the spectrum, the mass spectrometer then proceeds to obtain sequence information about these biomolecules. This process is called MS/MS for it couples two stages of MS. In tandem MS, a particular biomolecule ion is isolated, energy is imparted by collisions with an inert gas, and this energy causes the analyte to break apart. A mass spectrum of the resulting fragments is then generated (Fig. B-5).

In general proteomics, the mass spectrometer does not measure proteins, but peptides. First, peptides can be easy to handle and are stable to introduce MS. Second, the sensitivity of MS for peptides is much better than that for proteins, and the protein might be processed and modified such that the combinatorial effect makes determining the masses of the numerous resulting isoforms impossible. Third, the sequence of a peptide is easy to predict, unlike that of a mature protein is not. Finally, MS is most efficient at obtaining sequence information from peptides that are up to ~20 residues long, rather than from whole proteins peptides (Steen and Mann, 2004). The most highly sequence-specific proteases are used to convert proteins to peptides, such as trypsin.

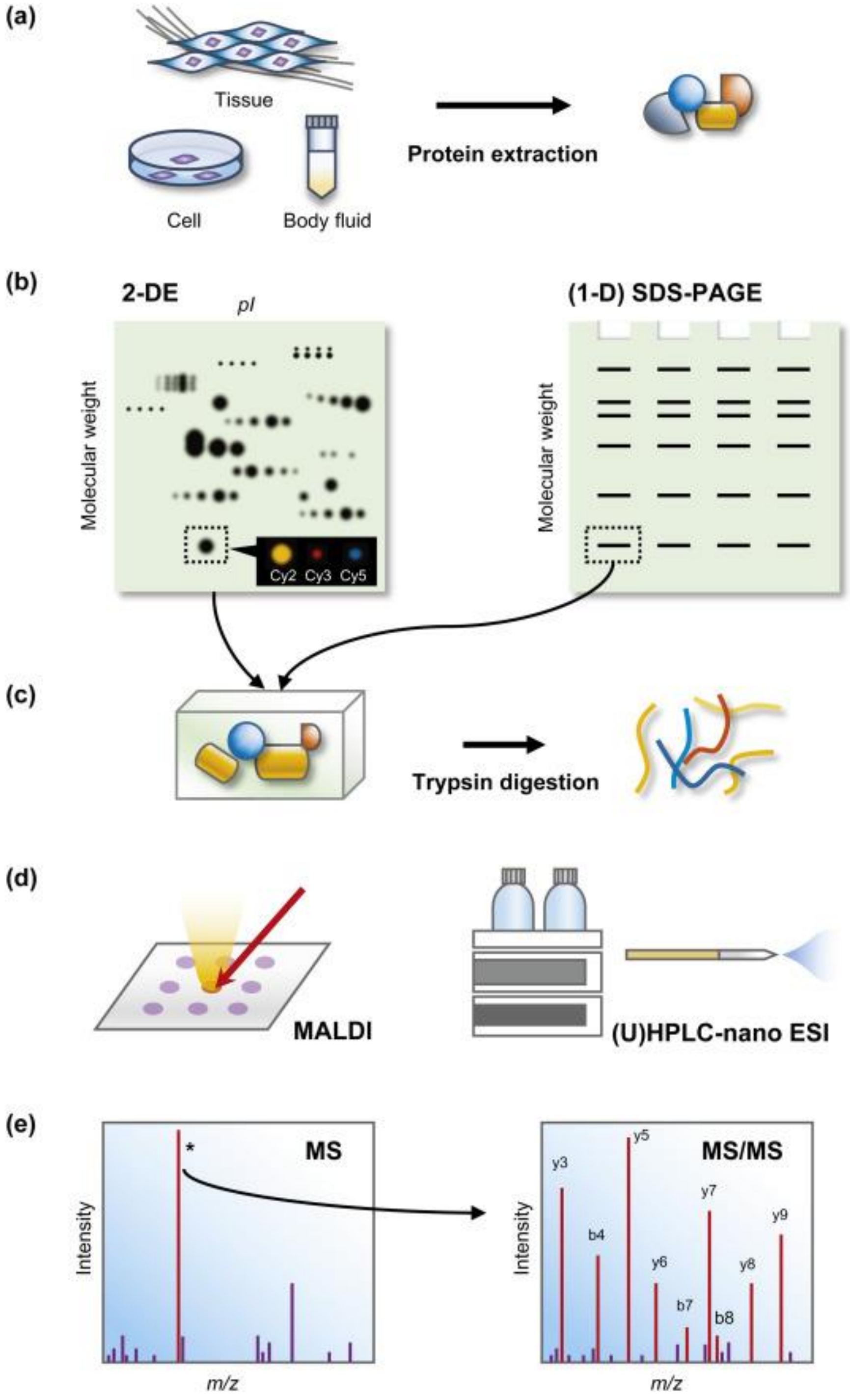


Figure B-5. General workflow of gel-based proteomics.

(a) Proteins are extracted from bio specimen. (b) Extracted protein mixture from biosamples separated by 2-DE or SDS-PAGE. In most case, proteins are quantified on a gel. Using the quantitative difference from DIGE, target spots can be selected from 2-DE. (c) Excised gel pieces are trypsinized and resulting peptides are collected. (d) Peptides are ionized via MALDI or nano ESI and are inducted to MS. (e) Peptide is measured in MS spectrum, followed by selected and isolated, subsequently fragmented to get the sequence information from MS/MS spectrum.

Adopted from Kim et al., 2019

CHAPTER Ⅰ

Transcriptome Signatures of Canine Mammary Gland Carcinomas and Its Comparison to Human Breast Cancers

INTRODUCTION

Human breast cancer (BC) is one of the most common cancers in women and is a leading cause of death worldwide, accounting for 8.8 million deaths in 2015 ("Who Fact Sheet")(WHO Fact Sheet, 2018). Approximately 80% of diagnosed BCs are invasive and heterogeneous, consisting of up to 21 distinct histological subtypes (Sgroi, 2010). Current biological markers used for evaluating molecular subtypes of BC include hormone receptors for estrogen or progesterone, and *HER2*+/−, indicating levels of human epidermal growth factor receptor 2 (*HER2*) (Onitilo et al., 2009). Although large-scale cohort studies using gene expression profiling techniques, such as next-generation sequencing (NGS), have provided better understanding of the molecular regulation of BC, a limited number of studies have been performed in rare and aggressive subtypes of human BC, such as invasive ductal carcinoma, myoepithelial complex type BC, and inflammatory BC (Koczkowska et al., 2016, Li et al., 2017, Ratajska et al., 2015).

Canine mammary gland carcinoma (MGC) is a well-known animal model for human BC, as there are a number of benefits to studying human BC using dogs (Salas et al., 2015). Existing similarities between these species have been reported with respect to genetic, biological, anatomic, and clinical features (Gurda et al., 2017, Liu et al., 2014). Additionally, dogs hold a unique status in human BC studies with respect to epigenetic aberrations since both dogs and humans, especially companion dogs and owners, share neighborhood environments and might be exposed to the same carcinogens (Romagnolo et al., 2016). Moreover, in contrast to human BC, complex/mixed MGC consisting of epithelial masses containing regions of myoepithelial components comprises the majority of MGC in dogs (Im et al., 2014). Thus, since the dog reference genome was unveiled in 2005, a number of comparative analyses using transcriptome data in independent studies have been performed (Klopfleisch et al., 2011, Król et al., 2009, Lindblad-Toh et al., 2005a). However, the results of these studies have been relatively inconsistent and only few biomarkers have been identified for canine MGC as well as human BC (Campos et al., 2012, Vinothini et al., 2009).

In the last few decades, high-throughput sequencing technology in medical oncology has generated a large number of databases including genetic mutations, gene expression profiles, and epigenetic aberrations associated with diverse cancer types (Kamps et al., 2017, Khotskaya et al., 2017). Many gene expression profiling studies on human BC carcinogenesis have also been performed with large BC patient cohorts and have reported many differentially expressed genes (DEGs) and their related cancer pathways (Guo et al., 2017, Li et al., 2016a).

Noncoding RNAs (ncRNAs) have become one of the most highlighted transcriptomic features in diverse organisms, increasing our understanding of the complexity of transcriptomic regulation. More than several tens of thousands of ncRNAs have been identified and have been functionally grouped within human and model organisms, such as yeast and mouse (Ferrero et al., 2018, Liang et al., 2018, Schwarzer et al., 2017). Particularly in human BC, a list of microRNAs (miRNAs) are considered to have crucial roles in cancer development and metastasis, and other studies have shown that miRNA expression profiles of each BC subtype are different (Haakensen et al., 2016, Huo et al., 2016). Moreover, a cluster of oncogenic long ncRNAs (lncRNA) are up-regulated in human BC and seem to be involved in regulating immune system activation (Xu et al., 2017). Additional interesting ncRNAs, including those recently determined and confirmed in existence, are known as promoter upstream transcripts (PROMPTs) (Preker et al., 2011). Interestingly, the presence of PROMPTs may be positively correlated with gene activity. Although PROMPTs are not widespread regulators of gene expression, their existence is tightly regulated by exosome activity, and the analysis of PROMPTs as a part of regulatory mechanisms of transcription in cancer might be important to better understand MGC.

In this study, we sequenced total RNAs from ten pairs of canine MGC and matching adjacent normal tissues to identify canine MGC-associated transcriptomic signatures. We further tested whether these signatures can distinguish canine MGCs from normal tissue using principal component analysis

(PCA) and clustering. To better understand both canine MGCs and human BC, we subsequently extracted a group of canine MGC-associated KEGG pathways and gene ontology (GO) terms. PROMPTs were then suggested as a part of transcriptional regulation mechanisms in cancer. This study will provide new insights into biomarker and target development for human BC as well as canine MGC.

MATERIALS AND METHODS

Specimens

This study was reviewed and approved by the Seoul National University Institutional Animal Care and Use Committee (IACUC# SNU-170602-1). Ten dogs diagnosed with mammary gland tumor were enrolled in this study. Mammary gland tumors and matching adjacent normal tissues were obtained by excisional surgery. Clinical features of eight dogs analyzed in the study are listed in Table S1. Eight pairs of specimens consisting of two simple-, three ductal-, and three complex-subtypes, from diverse breeds including Maltese, Dachshund, and Cocker Spaniel, were processed further for RNA-seq. For total RNA-seq, all tissue samples were immersed in RNAlater solution (Qiagen, Valencia, CA, USA) overnight at 4 ◦C, and stored at −80 ◦C after removal from solution.

RNA Isolation and Total RNA Sequencing

Total RNA was extracted from mammary gland tumors and matched to normal tissues using the RNeasy Mini plus kit (Qiagen, Valencia, CA, USA). Pulverization for sample homogenization was performed with liquid nitrogen before RNA isolation according to the manufacturer's instructions. The RNA quality was assessed by analysis of 18S and 28S rRNA band integrity on RNA 6000 Nano Kit (part # 5067-1511) using an Agilent Bioanalyzer (Agilent, Santa Clara, CA, USA). After ribosomal RNA (rRNA) depletion from 2 µg of total RNA, libraries were constructed using the TruSeq Stranded Total RNA Sample Preparation Kit (RS-122-9007) (Illumina, San Diego, CA, USA) according to the manufacturer's guideline. The cDNA library quality was evaluated electrophoretically with an Agilent DNA 1000 Kit (part # 5067–1504) (Agilent, Santa Clara, CA, USA). Subsequently, libraries were sequenced using Illumina HiSeq2500 that were set to rapid-run mode. Cluster generation, followed by 2 × 100 cycle sequencing reads, separated by paired-end turnaround, were performed on the instrument using HiSeq Rapid SBS Kit v2 (FC-402-4021) and HiSeq Rapid PE Cluster Kit v2 (PE-402-4002) (Illumina, San Diego, CA, USA). Image analysis was performed using the HiSeq control Software version 2.2.58. The raw data were processed, and base-calling was performed using the standard Illumina pipeline (CASAVA version 1.8.2 and RTA version 1.18.64). A summary of statistics of the RNA-seq data is listed in Table S2.

Primary Analysis of RNA-seq Data (Mapping and Quantification)

Initially, transcript integrity was analyzed and transcript integrity number (TIN) was in Table S3. Reads were aligned with the dog reference genome (CanFam 3.1, 2011) using Hiset2 (ver.2.1.0) with cufflink option. Mapped reads were then assembled and counted using Cuffquant (ver. 2.2.1) and our GTF annotation file pre-built with additional transcripts information obtained from 13 different organs based on the Ensembl database (Canis lupus familiaris 3.1.91 gene set). Defaults were used for all other parameters.

Differentially Expressed Gene (Lindblad-Toh et al.) Analysis

For the differential gene expression analysis, three subtypes of MGC (simple, complex, and ductal) and three breeds, as well as all eight MGCs and matching normal tissues, were grouped and compared using Cuffdiff (ver.2.2.1). Genes with expression differences of 2-fold increases or decreases and $p < 0.01$ were evaluated as DEGs and were further analyzed. Fragments per kilobase of exon per million fragments mapped (FPKM) were extracted for all groups, and Plotly package in R was employed to visualize statistically significant changes among the comparisons. Venn diagrams were created using Venny 2.1 (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

Correlation Analysis, Clustering and Principal Component Analysis (PCA)

FPKM values were extracted from a list of DEGs enriched in three subtypes of MGCs. All the FPKM values were log2 transformed to rank correlations among three subtypes of MGCs. Spearman rank correlation was calculated using Perseus ver.1.5.8.5 and visualized as Multiscatter plots in Maxquant software package (Max Planck Institute of Biochemistry, Munich, Germany). Z-scores were calculated from FPKM and further used for gene clustering. Clustering was performed with Kendall clustering method and the heat map was visualized using “pheatmap” in R package. PCA was performed by using ClustVis (https://biit.cs.ut.ee/clustvis/) (Metsalu and Vilo, 2015).

Comparative Gene Expression Analysis among Four Subtypes of Human BC and Three Subtypes of Canine MGC

RNA-seq data for four molecular subtypes (*HER2*+, ER+, ER&*HER2*+, and TNBC) were retrieved from the project (PRJNA305054) in the National Center for Biotechnology Information (NCBI). The expression of orthologous genes, matched with subtype-specific DEGs and summarized in Table S6A, were compared in Spearman correlation and visualized in scatter plot using SPSS program. On the contrary, correlation in gene expression between canine MGC and human BC was computed using the list of genes in PAM50 and Oncotype DX.

Pathway Enrichment Analysis and Gene Ontology (GO) Analysis

To better understand the biological significance of the identified DEGs, we performed GO, gene network analysis, and pathway enrichment analysis. GO was analyzed with overall MGCs-enriched and subtype-enriched DEGs using the web-based functional annotation tool DAVID 6.7 (https://david.ncifcrf.gov) and ClueGo, provide by Cytoscape App Store (apps.cytoscape.org). Three aspects, including biological process (BP), molecular function (MF), and cellular component (CC), were surveyed and the highest enrichment aspect, BP in this study, was documented in detail. GO terms and gene networks were visualized by ClueGo (cytoscape.org) (Lotia et al., 2013). For all GO and KEGG pathway analysis, $p < 0.01$ was considered as significant.

RESULTS

RNA Sequencing in Mammary Gland Carcinomas and Matching Adjacent Normal Dog Tissues

Ten canine MGCs were enrolled in this study as pairs of MGCs and matching adjacent normal tissues which were collected by veterinarians during surgery and pathologically tested. Animal protocols were approved by SNU IACUC (approval#SNU-170602-1, 26 July 2016). Out of ten dogs, two dogs were excluded from this study due to diagnosis of benign adenoma and large differences in the phylogenetic tree of dog breeds. To increase the reliability of the RNA-seq data, each subtype consists of at least two specimens as biological replicates (three specimens in ductal, three in complex, and two in simple type). Ultimately, eight pairs of data of MGC and normal tissues were further analyzed.

Overall, 625.4 and 672.4 million paired-end and strand-specific reads from dog MGC and adjacent normal tissues were sequenced, respectively. The transcript integrity number (TIN) was computed to measure RNA degradation level. Both raw read quality scores (Q30) and median TINs for all the samples were greater than 93.17% and 65%, respectively. Before sequence alignment, gene transfer format (GTF) of Canfam3.1 reference annotation file was updated with our dog transcript library consisting of 10,792 novel transcripts with information obtained from 13 major dog organs. Out of 1.29 billion reads, more than 96.82% reads were mapped onto Canfam3.1, the canine reference genome reinforced by our annotation file. Unique transcripts where the regions had never been annotated in dog were considered "novel". Overall, in a total of eight pairs of transcriptome, the number of transcripts identified with both novel and reference annotations were slightly higher in the adjacent normal tissues (5015 new and 15,602 ref genes) than in the MGC tissues (4683 new and 15,003 ref genes) (Fig. 1-1).

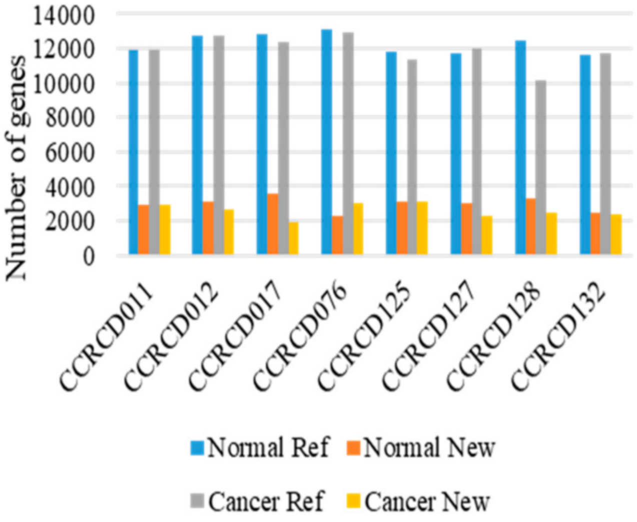


Figure 1-1. Transcript expression found in eight pairs of mammary gland carcinomas (MGCs) and matching adjacent normal tissues.

Ref: Canfam3.1 reference annotation.

Identification of DEGs in Canine MGCs and Their Subtypes

For the differentially expressed gene (Lindblad-Toh et al.) analysis, four comparisons were performed between eight pairs of MGCs and matching adjacent normal tissues and in three subtypes (simple, complex and ductal). DEGs with a p-value < 0.01 and changes greater than 2-fold were determined for each comparison. Cuffdiff analysis identified 350 DEGs, of which 132 and 218 genes were up- and down-regulated, respectively, in a comparison of the eight canine MGCs and matching adjacent normal tissues. Hierarchical clustering with Kendall correlation matrix of the 350 DEGs successfully distinguished MGCs and matching adjacent normal in a heat map analysis (Fig. 1-2A). In total, 454 DEGs (178 up- and 276 down-regulated), 226 DEGs (117 up- and 109 down-regulated) and 171 DEGs (66 up- and 105 down-regulated) were identified as subtype-specific DEGs for complex, ductal, and simple MGCs respectively. Hierarchical clustering with these DEGs successfully separated MGC from normal again (Fig. 1-2B).

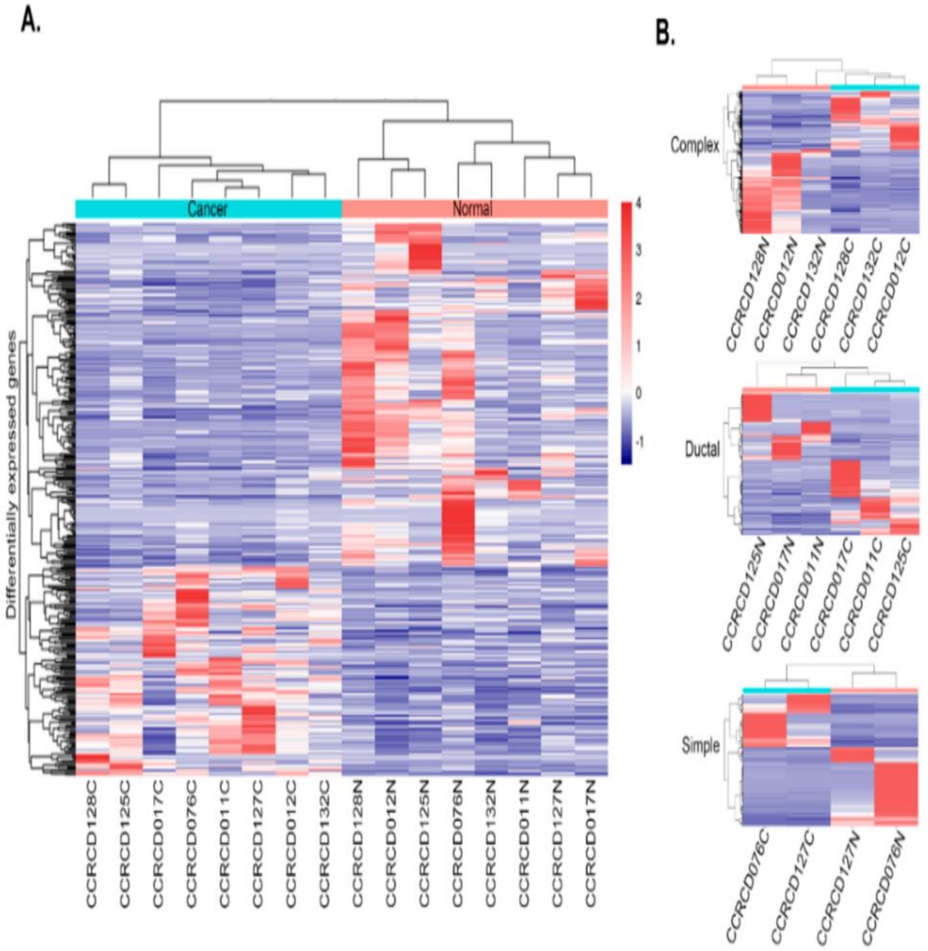


Figure 1-2. Heat map and hierarchical clustering of mammary gland carcinoma (MGCs) and matching adjacent normal tissues.

(A) in eight pairs and (B) in three subtypes of MGCs (complex, simple, and ductal). Eight specimens were labeled with N (normal) and C (Jung et al.). The distance metric used for clustering was Kendall correlation, while the linkage method used was average linkage.

Overall DEGs were summarized and visualized using Venn diagram and Volcano plots (Fig. 1-3). The top five up-/down-DEGs were labeled in Volcano plots (Fig. 1-3) and are listed in Table 1. Out of 851 DEGs, only 16 genes, 1.6% of total DEGs, were shared by all three subtypes, indicating that these three subtypes might have unique RNA expression profiles (Fig. 1-4A). Subsequently, correlations among DEG profiles in these three subtypes were tested and are shown in scatter plots (Fig. 1-4B). All correlation coefficients among subtypes of MGC were between 0.7~0.9, which can be considered highly correlated. There was little difference between the highest correlation (0.849 between ductal and complex subtype) and the lowest correlation (0.784 between simple and complex subtype). Thus, each subtype of MGC had unique transcription signatures, but overall transcriptome profiles might be very similar among MGCs.

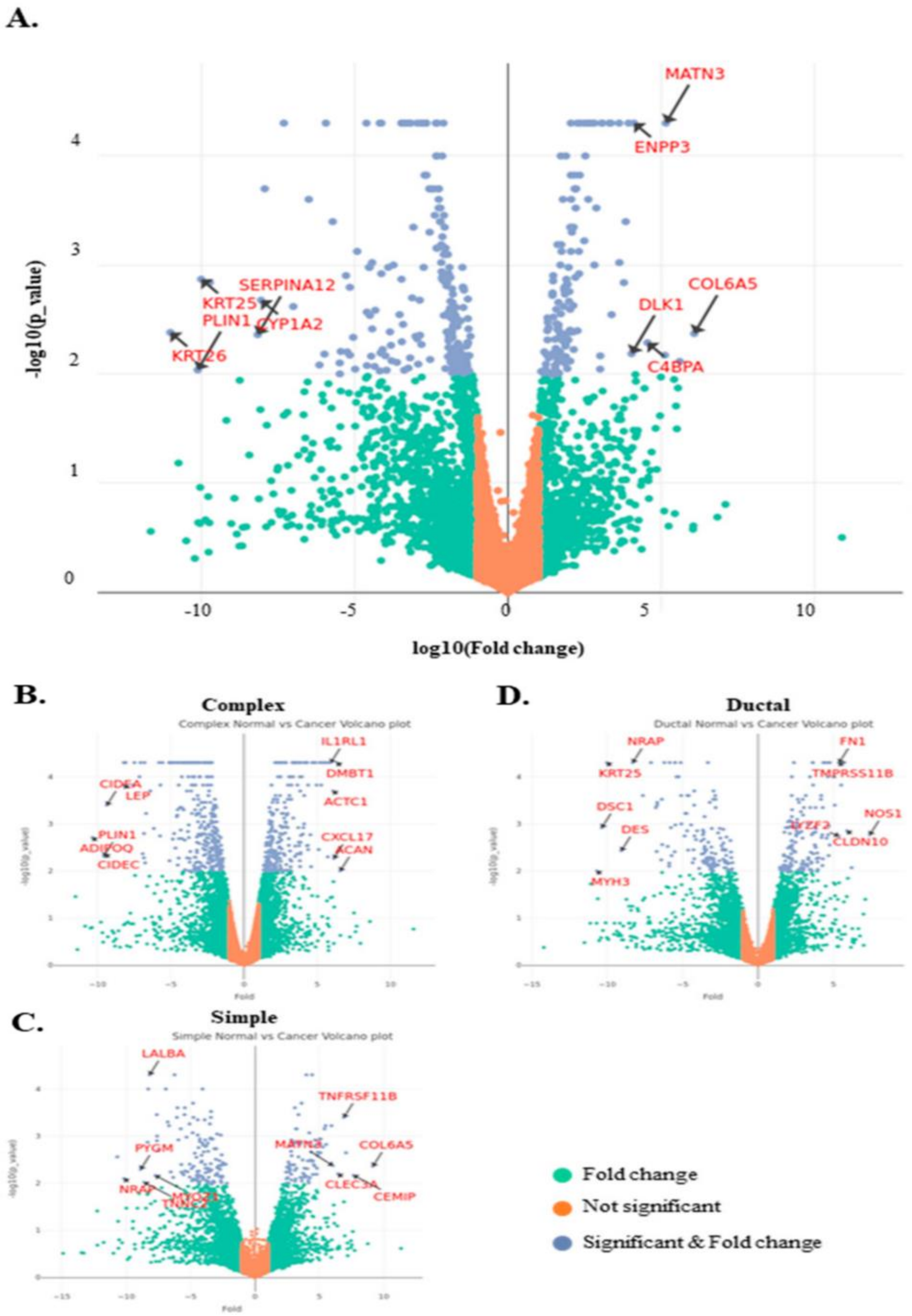


Figure 1-3. Differentially expressed genes (DEGs) in canine MGCs.

(A) Volcano plots of DEG content with larger than two-fold changes (log 2 values) and p-values < 0.001 for total cancer. (B) Complex. (C) Simple. (D) Ductal.

To test whether these DEG signatures represent canine MGCs and/or MGC subtypes, we conducted a principal component analysis (PCA). PCA results indicated that the first principal component (PC1) explains 34.7% of the variability, while PC2 and PC3 explain 13.4% and 10.6% of the variability, respectively, in DEGs of all the canine samples. Three PCs only covered ~58% of total variability. This might represent the complexity of cancer biology in clinical samples. Although these three PCs only covered approximately 58.7% of the total variability in the overall comparison of MGC and the adjacent normal tissue, MGC and the matching normal samples were successfully distinguished from each other in dimensional PCA, illustrated in Fig. 1-4C. Unexpectedly, all eight MGCs were tightly grouped, whereas matching normal tissues were more individually variable (Fig. 1-4C).

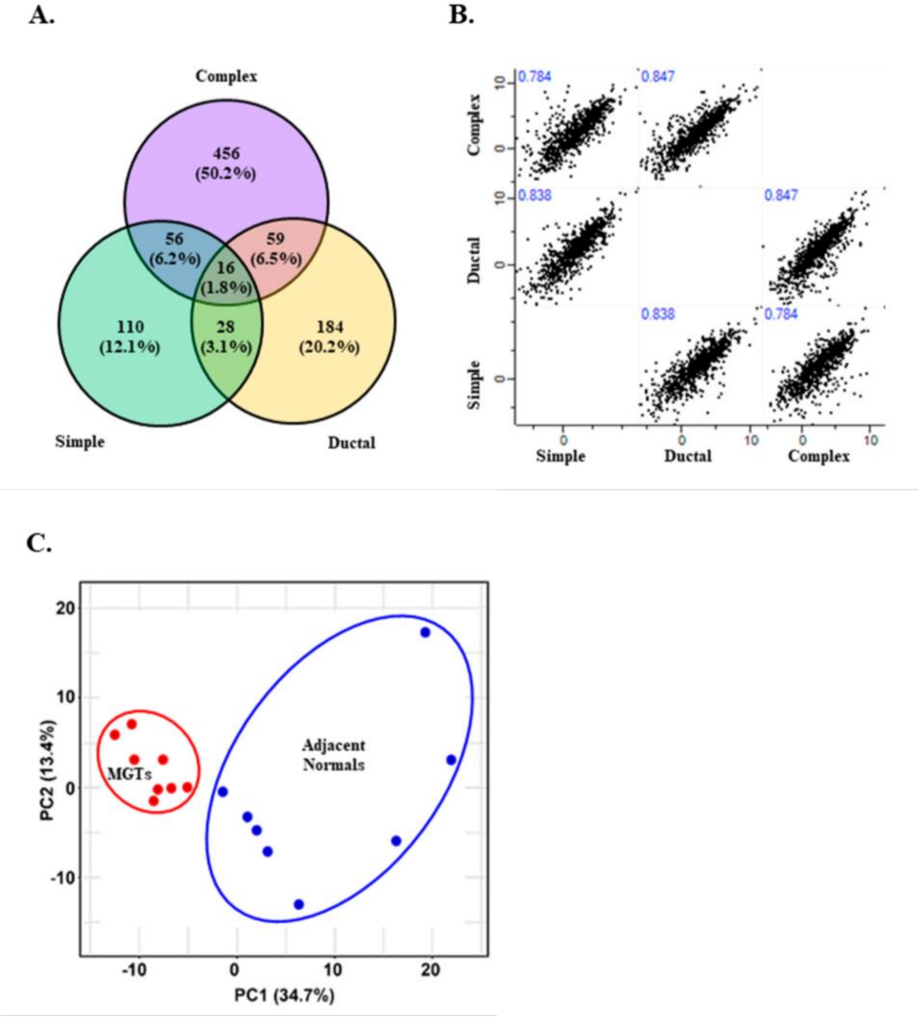


Figure 1-4. Computational analysis of canine differentially expressed genes (DEGs)

(A) Venn diagrams illustrating the number of up- and down-regulated DEGs among three subtypes of MGC. (B) Scatter plots of DEGs among three subtypes of MGC. The Spearman rank correlation based on 555 DEGs was computed by Perseus (ver.1.5.8.5) in Maxquant software. (C) Principal Component Analysis (PCA). The first three principal components explain ~57% of total variations

Table 1-1. Top 5 up-/down-DEGs enriched in overall canine MGC and in three subtypes.

Group	Ensembl ID	Gene	$\log_{10}$(Fold Change)	$-\log_{10}$(*p*-Value)
Overall MGCs	ENSCAFG00000006046	COL6A5	6.06776	2.37161107
	ENSCAFG00000003825	MATN3	5.14522	4.301029996
	ENSCAFG00000024982	C4BPA	4.5425	2.288192771
	ENSCAFG00000000367	ENPP3	4.10668	4.301029996
	ENSCAFG00000017925	DLK1	4.02563	2.187086643
	ENSCAFG00000016014	KRT26	−11.005	2.381951903
	ENSCAFG00000011986	PLIN1	−10.1163	2.038578906
	ENSCAFG00000023806	KRT25	−10.001	2.869666232
	ENSCAFG00000017661	SERPINA12	−8.16046	2.361510743
	ENSCAFG00000017941	CYP1A2	−8.04964	2.677780705
Complex	ENSCAFG00000011534	ACAN	6.56988	2.004364805
	ENSCAFG00000012561	DMBT1	6.33988	4.301029996
	ENSCAFG00000004810	CXCL17	6.13628	2.26760624
	ENSCAFG00000012181	ACTC1	6.08066	3.698970004
	ENSCAFG00000002142	IL1RL1	5.92575	4.301029996
	ENSCAFG00000013694	ADIPOQ	−10.343	2.709965389

	ENSCAFG00000011986	PLIN1	−9.57036	2.314258261
	ENSCAFG00000005266	CIDEC	−9.47559	2.356547324
	ENSCAFG00000018828	CIDEA	−9.36545	3.397940009
	ENSCAFG00000001672	LEP	−8.18516	3.823908741
Ductal	ENSCAFG00000009820	NOS1	7.40628	2.769551079
	ENSCAFG00000005458	CLDN10	5.90219	2.853871964
	ENSCAFG00000014345	FN1	5.3764	4.301029996
	ENSCAFG00000002808	TMPRSS11B	5.3535	4.301029996
	ENSCAFG00000008948	LYZF2	5.3447	2.744727495
	ENSCAFG00000023094	MYH3	−10.6876	2.002176919
	ENSCAFG00000018070	DSC1	−10.3628	2.920818754
	ENSCAFG00000023806	KRT25	−10.0193	4.301029996
	ENSCAFG00000015475	DES	−9.0685	2.431798276
	ENSCAFG00000011103	NRAP	−8.30778	4.301029996

Group	Ensembl ID	Gene	$\log_{10}$(Fold Change)	$-\log_{10}$(p-Value)
Simple	ENSCAFG00000006046	COL6A5	9.15426	2.361510743
	ENSCAFG00000013863	CEMIP	7.65997	2.167491087
	ENSCAFG00000000834	TNFRSF11B	6.89898	3.397940009
	ENSCAFG00000020033	CLEC3A	6.45938	2.200659451
	ENSCAFG00000003825	MATN3	6.0326	2.37675071
	ENSCAFG00000011103	NRAP	−10.1506	2.099632871
	ENSCAFG00000014281	PYGM	−8.87425	2.296708622
	ENSCAFG00000028609	TNNC2	−8.60343	2.019996628
	ENSCAFG00000008950	LALBA	−8.18348	4.301029996
	ENSCAFG00000014842	MYOZ1	−7.72731	2.164309429

Correlation in Gene Expression between Four Molecular Subtypes of Human BC and Three Histological Subtypes of Canine MGC

Eleven RNA-sequencing data for four molecular subtypes (HER2+, ER+, ER&HER2+, and TNBC) were retrieved from the study by Chung W. et al., publicly opened project (PRJNA305054) in the National Center for Biotechnology Information (NCBI) (Chung et al., 2017). DEGs specific to each canine MGC subtype were subjected for correlation analysis. BC with molecular subtype of HER2+ showed significant correlation coefficient (r) with all three canine MGC subtypes (max r = 0.475 with simple subtype, min r = 0.393 with complex subtype, $p < 0.01$) (Fig. 1-5). ER+ and ER+&HER2+ subtypes showed no correlation with 'complex and simple' and ductal subtype, respectively. Only low levels of correlation were found in ER+ with ductal subtype (r = 0.254, $p < 0.05$) and ER+&HER2+ with simple subtype (r = 0.355, $p < 0.05$). Notably, TNBC has strong correlation in both ductal and simple subtypes (r = 0.472 and 0.523, respectively). It is interesting because TNBC is usually defined as basal-like and non-basal-like types in human BC and the most common histological subtype of TNBC is invasive ductal carcinoma. Moreover, the simple subtype showing the highest correlation in TNBC expressed KRT5 and MKI67, which has been known and used as immunohistochemical markers for basal-like breast cancer and proliferation (Jézéquel et al., 2015). Our results indicated that transcriptomic signatures for canine MGC subtypes might represent human BC subtypes and provide new candidates of biomarkers. We then tested the same analysis oppositely using the gene expression profiles listed in PAM50 and Oncotype DX, but no significant correlation was found among subtypes of human BC and canine MGC.

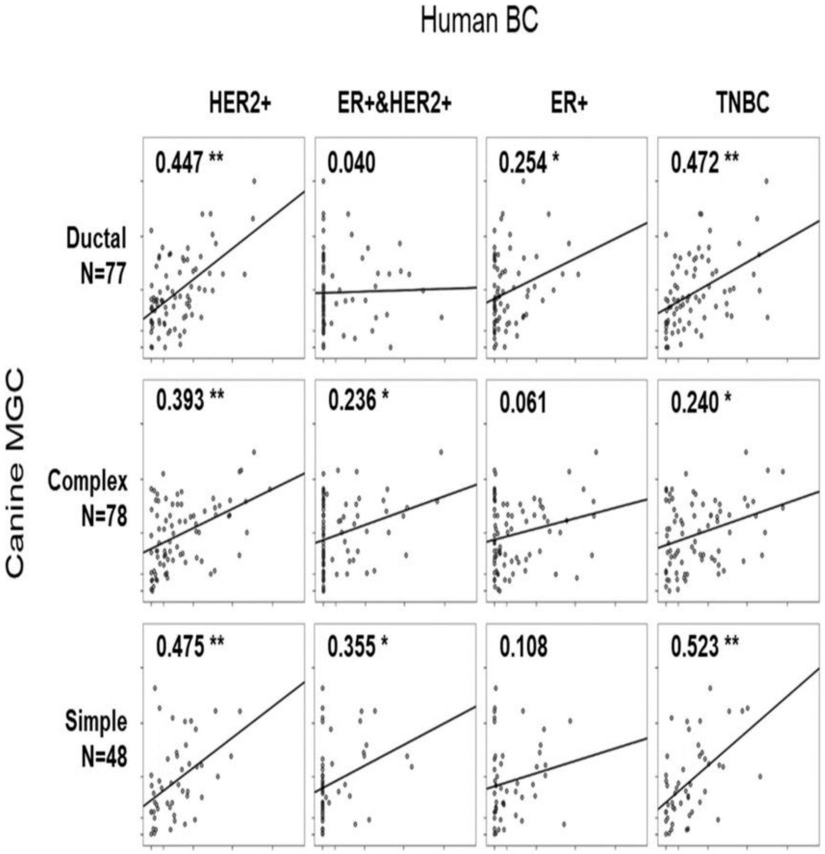


Figure 1-5. Scatter plots showing the correlation between molecular subtypes of human BCs and histological subtypes of canine MGCs. Different numbers of canine MGC subtypes-specific genes were abstracted (Complex: N = 78, Ductal: N = 77, and Simple: N = 48). *, ** indicates $p < 0.05$, $p < 0.01$, respectively.

Gene Ontology (GO) and Network Analysis

To better understand transcriptomic regulation in canine MGCs, we performed GO analysis with DEGs in all MGCs and in each subtype. For GO analysis, only the list of DEGs annotated by Ensembl gene name were subjected to ClueGo software (ver.2.5.0). Three hundred fifteen out of 350 profiled DEGs were assigned to 88 GO terms, including 53 biological processes (BP), 18 cellular components, and 18 molecular function terms. GO terms were mainly categorized into BPs with wide distributions and extensive assignments (53 GO terms). BP assignments in up-regulated DEGs in MGCs were divided into eight groups.

The most prevalent BP group, consisting of eight GO terms, was represented by positive regulation of angiogenesis (GO:0045766). This group also included some important assignments, such as “cell adhesion mediated by integrin (GO:0033627)” and “positive regulation of vasculature development (GO:1904018),” suggesting that the biological processes in MGCs were directionally changed to promote tumor progression with increased vasculature (Niland and Eble, 2011). In contrast, the GO term “release of sequestered calcium ion into cytosol by sarcoplasmic reticulum” (GO:001480) represented BP in down-regulated DEGs. This result is interesting because association between calcium ion homeostasis and cancerization has been reported (Papp et al., 2012). This group consisted of 5 GO terms (GO:0003009, GO:0003009, GO:0055002, GO:0048747 and GO:0055008) covering 33.3% of total GO terms in down-regulated DEGs (Table 1-2).

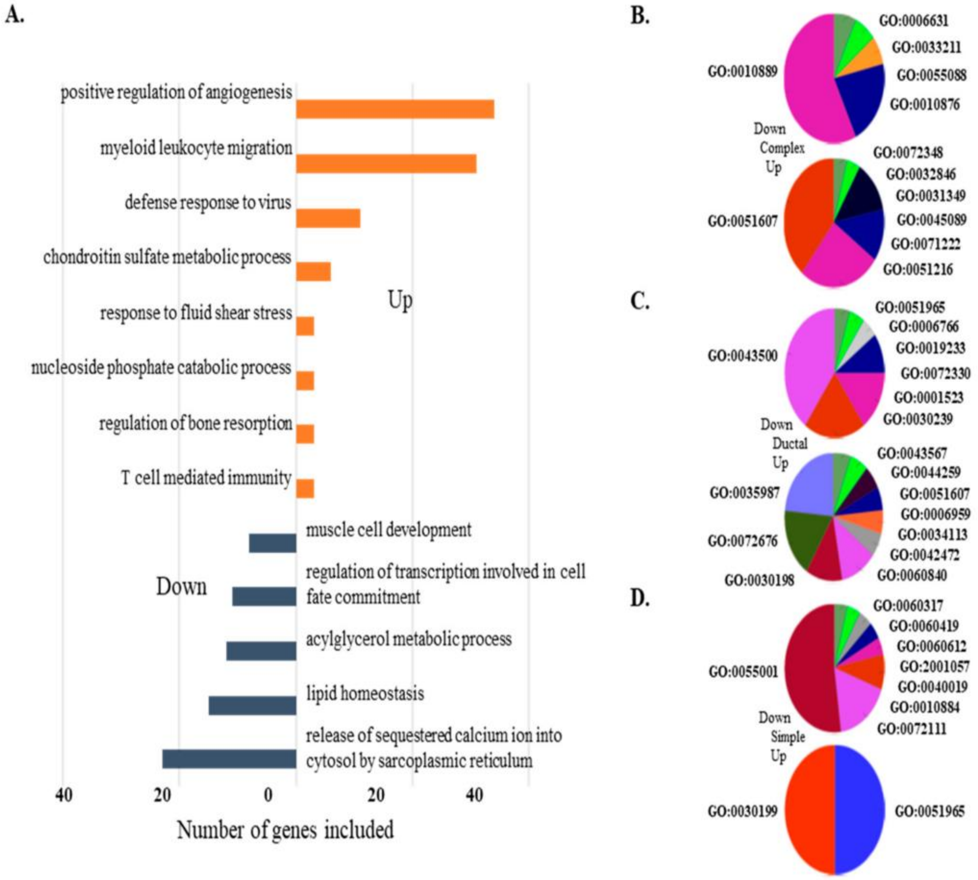


Figure 1-6. Gene ontology (GO) enrichment analysis for DEGs identified in an MGC-specific and subtype-dependent manner.

(A) GO analysis using DEGs from all three subtype comparisons. Orange bar indicates up-regulated GO and dark blue bar represents down-regulated GO. GOID enriched in each comparison of (B) Complex type, (C) Ductal type, and (D) Simple type of MGC.

Table 1-2. Gene ontology (GO) terms biological processes (BP) of up- and down-regulated DEGs in canine MGCs.

Up-Regulated DEGs					
GO groups	**GO ID**	**GO Term**	**% Assoc. Genes**	**No. Genes**	**Associated Genes Found**
0	:1904018	positive regulation of vasculature development	4.58	6	[CHI3L1, CXCL8, FOXC2, SERPINE1, SFRP2, TF]
	:0045766	positive regulation of angiogenesis	4.8	6	[CHI3L1, CXCL8, FOXC2, SERPINE1, SFRP2, TF]
	:0031638	zymogen activation	4.07	5	[PLAU, S100A8, SERPINE1, SERPINE2, TF]
	:0033627	cell adhesion mediated by integrin	5.26	4	[FOXC2, PLAU, SERPINE1, SFRP2]
	:0033628	regulation of cell adhesion mediated by integrin	7.27	4	[FOXC2, PLAU, SERPINE1, SFRP2]
	:1903318	negative regulation of protein maturation	10.34	3	[C4BPA, SERPINE1, SERPINE2]
	:0010955	negative regulation of protein processing	10.34	3	[C4BPA, SERPINE1, SERPINE2]
	:0031639	plasminogen activation	17.65	3	[PLAU, SERPINE1, SERPINE2]
1	:0097529	myeloid leukocyte migration	4.19	7	[CCL8, CMKLR1, CXCL10, CXCL8, S100A8, SERPINE1, SPP1]

	:0097530	granulocyte migration	4.17	5	[CCL8, CMKLR1, CXCL8, S100A8, SPP1]
	:0071222	cellular response to lipopolysaccharide	4.26	6	[CD80, CD86, CXCL10, CXCL8, SERPINE1, TNIP3]
	:0002690	positive regulation of leukocyte chemotaxis	5.26	4	[CMKLR1, CXCL10, CXCL8, SERPINE1]
	:0070098	chemokine-mediated signaling pathway	4.94	4	[CCL8, CMKLR1, CXCL10, CXCL8]
	:0071621	granulocyte chemotaxis	4.39	5	[CCL8, CMKLR1, CXCL8, S100A8, SPP1]
2	:0051607	defense response to virus	4.7	7	[CD86, CXCL10, ITGAX, PTPRC, RSAD2, SAMHD1, TLR7]
	:0002224	toll-like receptor signaling pathway	5.13	4	[CD86, RSAD2, TLR7, TNIP3]
3	:0050654	chondroitin sulfate proteoglycan metabolic process	8.82	3	[BGN, CHST11, NDNF]
	:0030204	chondroitin sulfate metabolic process	11.11	3	[BGN, CHST11, NDNF]
4	:0002456	T cell-mediated immunity	4.05	3	[P2RX7, PTPRC, RSAD2]
5	:0045124	regulation of bone resorption	9.38	3	[P2RX7, TF, TFRC]
6	:1901292	nucleoside phosphate catabolic process	4.11	3	[ENPP3, P2RX7, SAMHD1]
7	:0034405	response to fluid shear stress	9.09	3	[COX-2, P2RX7, SPP1]

Down-Regulated DEGs					
0	:0086036	regulation of cardiac muscle cell membrane potential	27.27	3	[ANK2, FXYD1, TRDN]
	:1903513	endoplasmic reticulum to cytosol transport	9.43	5	[ANK2, DHRS7C, DMD, RYR1, TRDN]
	:1903514	calcium ion transport from endoplasmic reticulum to cytosol	11.11	5	[ANK2, DHRS7C, DMD, RYR1, TRDN]
	:0070296	sarcoplasmic reticulum calcium ion transport	10.64	5	[ANK2, DHRS7C, DMD, RYR1, TRDN]
	:0014808	release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	11.11	5	[ANK2, DHRS7C, DMD, RYR1, TRDN]
1	:0055088	lipid homeostasis	7.37	7	[ANGPTL4, DGAT2, EPHX2, GPAM, LCAT, LPL, RORA]
	:0055090	acylglycerol homeostasis	13.79	4	[ANGPTL4, DGAT2, LPL, RORA]
	:0070328	triglyceride homeostasis	14.81	4	[ANGPTL4, DGAT2, LPL, RORA]
2	:0009755	hormone-mediated signaling pathway	5.56	7	[ACSL1, AR, BMP4, ESR1, PPARG, PRLR, RORA]
	:0060850	regulation of transcription involved in cell fate commitment	17.39	4	[BMP4, PPARG, PROX1, RORA]
3	:0006638	neutral lipid metabolic process	6.32	6	[DGAT2, GPAM, LIPE, LPIN1, SERPINA12, TNXB]
	:0006639	acylglycerol metabolic process	6.45	6	[DGAT2, GPAM, LIPE, LPIN1, SERPINA12, TNXB]
4	:0055001	muscle cell development	4.12	8	[ANK2, BMP4, COL14A1, CSRP3, DMD, PROX1]

Similar analyses were performed for DEGs within the three subtypes. The most prevalent group of BPs in up-regulated genes of the complex subtype is defense response to virus, covering 43.7% of up-regulated DEGs. Furthermore, some important assignments, such as cartilage development (GO:0051216), showed ~28.9%. Interestingly, 14 GO terms obtained from down-regulated DEGs in the complex subtype are grouped into five GO groups associated with lipid-related biological process, such as GO:0010876 that describes lipid localization, GO:0006631 of fatty acid metabolic process, and GO:0033211 of adiponectin-activated signaling. These results indicated the reduction of adipose components in the complex subtype compared to normal tissues. GO terms of defense response to virus (GO:0051607), humoral immune response (GO:0006959), and extracellular matrix organization (GO:0030199) up-regulated in the complex subtype were also shared by GO terms in the ductal subtype (Fig. 1-6B). However, endoderm-related biological processes, such as endodermal cell differentiation (GO:0035987), endoderm formation (GO:0001706), primary germ layer formation (GO:0001704), and endoderm development (GO:0007492), were enriched only in the ductal subtype. Whereas lipid-related BPs were down-regulated in the complex subtype, many GO terms linked to muscles, such as cardiac muscle tissue morphogenesis (GO:0055008), skeletal muscle adaptation (GO:0043501), and muscle adaptation (GO:0043500), were found in down-regulated DEGs in the ductal subtype (Fig. 1-6C). These down-regulated data suggested the dominant origin of ductal epithelium in ductal carcinoma compared to the presence of a certain proportion of myoepithelial cells in normal tissues. Since the sample numbers were relatively small in the simple subtype, only a few GO terms were identified as up-regulated (GO:0030199, GO:0051965). Numbers of GO terms enriched in down-regulated DEGs in the simple subtype were shared by one from the ductal subtype. Various muscle-related biological processes were also down-regulated (GO:0043500. GO:0035994, GO: 0048011, GO:0014888, GO:0055001, and GO:0055008) in simple carcinoma (Fig. 1-6D). Gene networks constructed by DEGs enriched in canine MGCs are shown in Fig. 1-7.

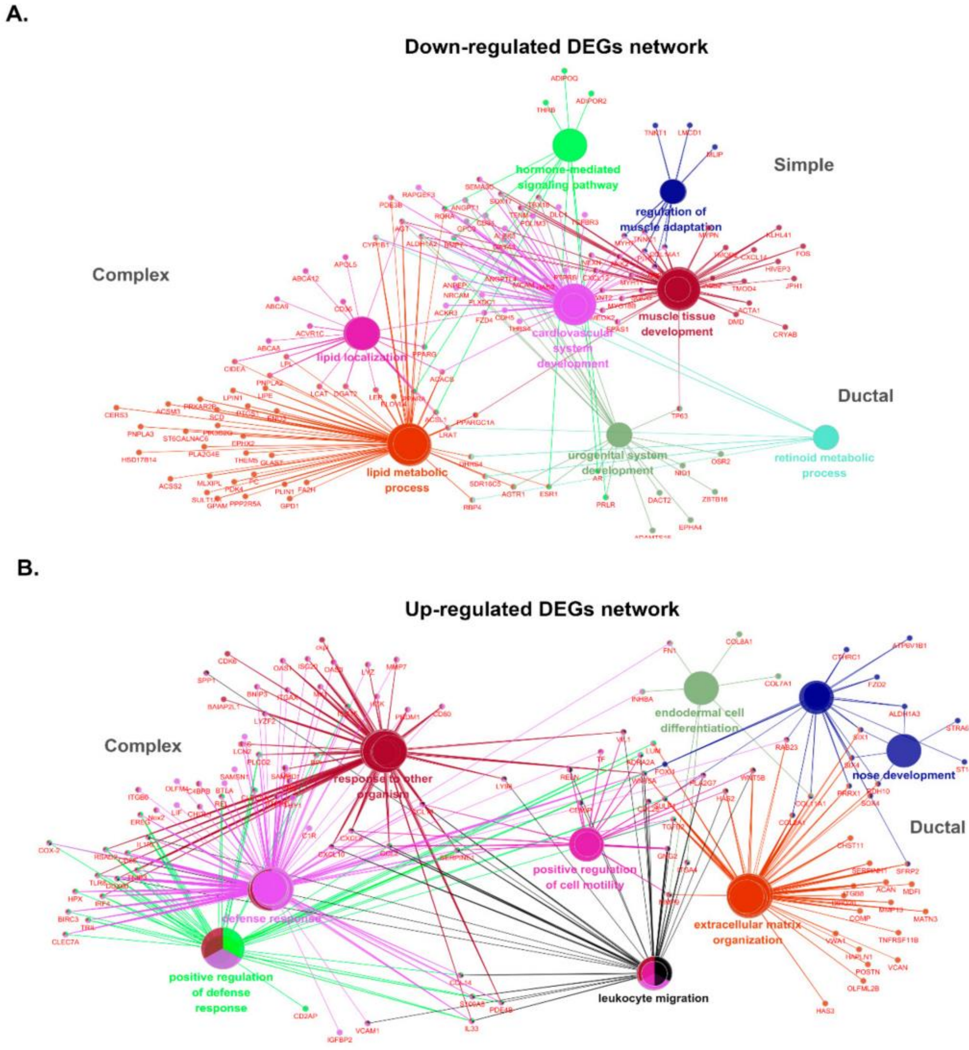


Figure 1-7. Gene network enrichment analysis in three subtypes of MGCs.

(A) Down-regulated DEGs. Lipid metabolism and localization are enriched only in the complex subtype, while muscle-related biological processes are enriched in the ductal subtype. The simple subtype does not construct unique nodes. (B) Up-regulated DEGs. Response to other organisms and defense responses are highlighted in the complex subtype, but cell mobility and extracellular matrix organization are shown in the ductal subtype. No node was found up-regulated in the simple subtype.

Pathways Significantly Enriched in MGC

Many cancer-related pathways including WNT, PI3K/Akt, KRAS, and PTEN pathways have been reported in canines (Campos et al., 2014, Dobbin and Landen, 2013, Terragni et al., 2014). To better understand canine MGC and human BC, we performed KEGG pathway analysis using the web-based DAVID functional annotation tool (https://david.ncifcrf.gov/summary.jsp). For the pathway analysis, we used a list of DEGs summed by the three subtype comparisons because it showed better results than with DEGs from the overall MGC comparison. Out of 727 DEGs, 313 up- and 414 down-regulated DEGs in MGCs were isolated and subjected to KEGG pathway analysis. Three hundred thirteen up-regulated DEGs in MGCs were involved in 24 and 23 KEGG pathways in dog and human databases, respectively. Twenty-one terms from the KEGG pathway analysis, including 'ECM-receptor interaction', 'pathways in cancer', and 'proteoglycan in cancer', were shared by both dog and human databases. However, the terms 'microRNA in cancer', 'salivary secretion', and 'Wnt signaling pathway' were found only in the dog database, while 'dilated cardiomyopathy' and 'Fc gamma R-mediated phagocytosis' were exclusively found only in the human database. The highest assignment of up-regulated DEGs was 'pathways in cancer' which includes WNT and PI3K pathways. Seventeen up-regulated DEGs in canine MGC primarily mapped to ECM-ITGA/B-PI3K signaling and Wnt-Frizzled signaling pathways. ECM signaling is known to be involved in proliferation, migration, invasion, and angiogenesis (Venning et al., 2015). In addition, up-regulation of *COX2, TGFb*, *GLUT1*, *MMP*, and *IL8* genes were involved in angiogenesis, and *BIRC7/2* is known for its function of apoptosis evasion (Bergers and Benjamin, 2003). In contrast, ‘metabolic pathways’ was the highest enriched (45 genes) KEGG pathway among down-regulated DEGs. Interestingly, most DEGs were heavily mapped to glycan biosynthesis and metabolism, and some additionally mapped to lipid metabolism related to glycan biosynthesis and metabolism pathways. These results indicated that aberration of lipid biogenesis and metabolism is associated with canine MGC progression.

Accumulation of Promoter Upstream Transcripts (PROMPTs) and MGC-Associated Gene Transcription

Although some regulatory mechanisms have been suggested, few promoter upstream transcripts (PROMPTs) have been characterized, and many of their functional roles remain unknown (Preker et al., 2011). Here, we measured unknown genome-wide transcripts expressed in the upstream regions of gene promoters. To quantify transcripts upstream of promoter regions, we collected all sequence reads mapped to regions ranging from all genes’ TSS to −1500 upstream. After excluding mapped transcript sequences that are shared with other genes, 28,757 promoter upstream regions consisting of 25,395 Ensembl database genes and 3362 novel transcripts were identified and used for further analysis. These were narrowed down to 41 regions (31 positive and 10 negative correlations) that met the threshold ($p < 0.01$, fold change $\geq$ 2) for genes and (fold change $\geq$ 2)

PROMPTs. Unfortunately, differences in all ten negatively correlated genes and PROMPTs listed in Table S9 were not confirmed by integrative genomic viewer (IGV) due to low expression level of the transcripts. However, the genes and PROMPTs that were positively correlated were confirmed by IGV survey (correlation: 0.71694) (Fig. 1-8. Eleven genes out of 31 were up-regulated in MGCs and positively correlated with PROMPT expression. Some of these promoter regions, such as NOVA1 and GRIA3, have been annotated with antisense RNA and pseudogenes, but most were not. This meant that more comprehensive genome annotations are necessary for the dog genome. Furthermore, it might provide a clue for understanding the regulatory mechanisms of up-regulated gene expression in cancer.

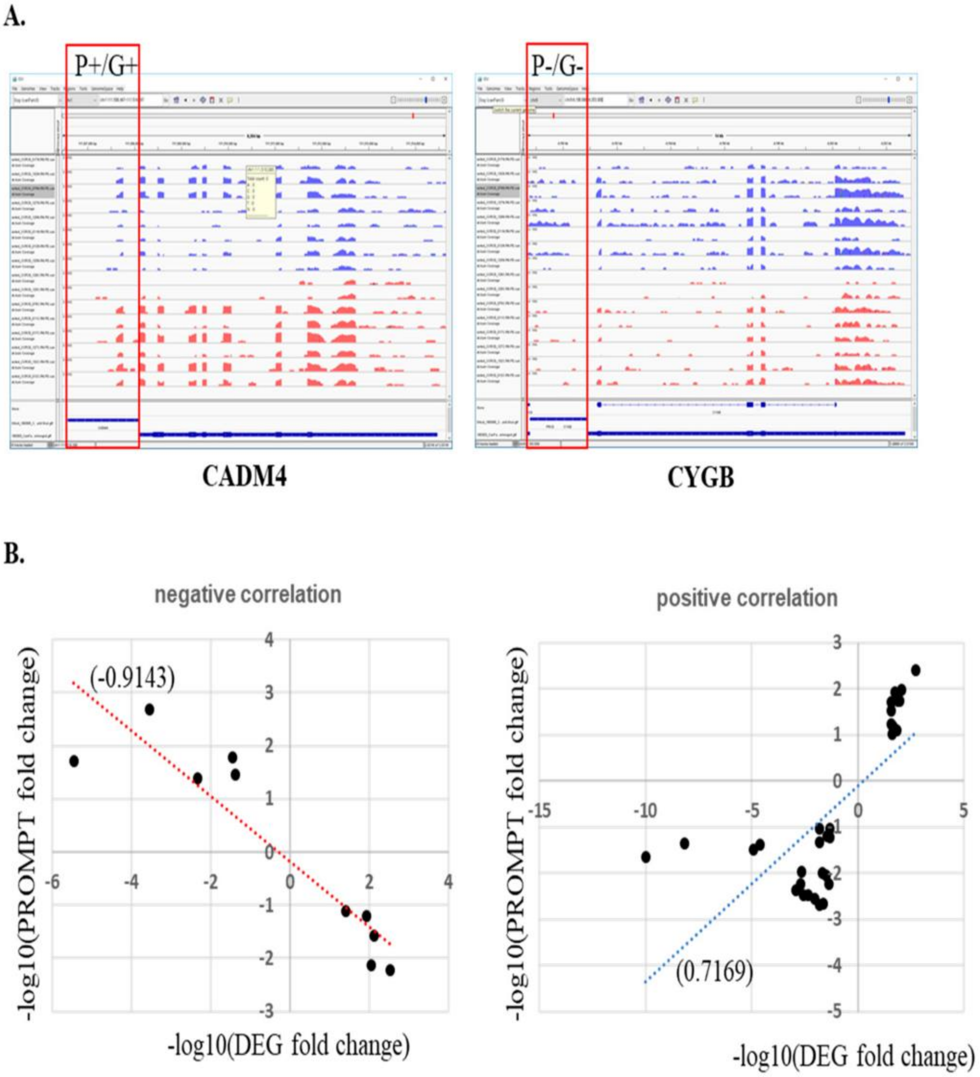


Figure 1-8. Correlation between DEGs and promoter upstream transcripts (PROMPT) expression.

(A) CADM4 and CYGB gene promoter regions as an example of DEGs and PROMPT expression in integrative genomic viewer (IGV). (B) Negative and positive correlation between DEGs and PROMPTs.

Quantitative Real-Time RT-PCR Validation of DEGs in MGCs

To validate our results, quantitative real-time RT-PCR was performed on ten selected DEGs to confirm our RNA-seq data. Out of ten, three well-verified genes were selected for further validation. The three genes belong to divergent functional categories or pathways but are not included in either Mammaprint or OncotypDx. *FN1* (fibronectin 1) is involved in cell adhesion and migration. *BGN* (biglycan) plays a role in collagen fibril assembly in multiple tissues. *SCD* (stearoyl-CoA desaturase) belongs to the fatty acid desaturase family and is involved in fatty acid biosynthesis. Verification was performed in additional pairs of ten MGCs and matching adjacent normal samples using real-time RT-PCR. The relative gene expression to *ATP5B* gene was calculated by the 2−ΔΔCt method and is shown in Fig. 1-9. Up- or down-regulated MGC DEGs in RNA sequencing data were confirmed in most sample pairs. Up-regulated *FN1* and *BGN* were validated in seven out of eight MGCs and matching normal tissues, respectively. In contrast, down-regulated *SCD* was confirmed in six out of eight MGCs (Fig. 1-9A). The Mann–Whitney U test indicated that there was significant difference in gene expression levels between MGCs and adjacent normal tissues (*FN1*; U = 27, p = 0.0083, *BGN*; U = 31, p = 0.0173, *SCD*; U = 34, p = 0.0284). To expand this analysis, we performed a receiver operating characteristics (ROC) analysis for each gene (Fig. 1-9B). A maximum AUC of 0.8125 (95% CI 0.6424–0.9826) was observed in *FN1* gene expression. AUCs of 0.7847 and 0.7639 were observed for *BGN* and *SCD*, respectively.

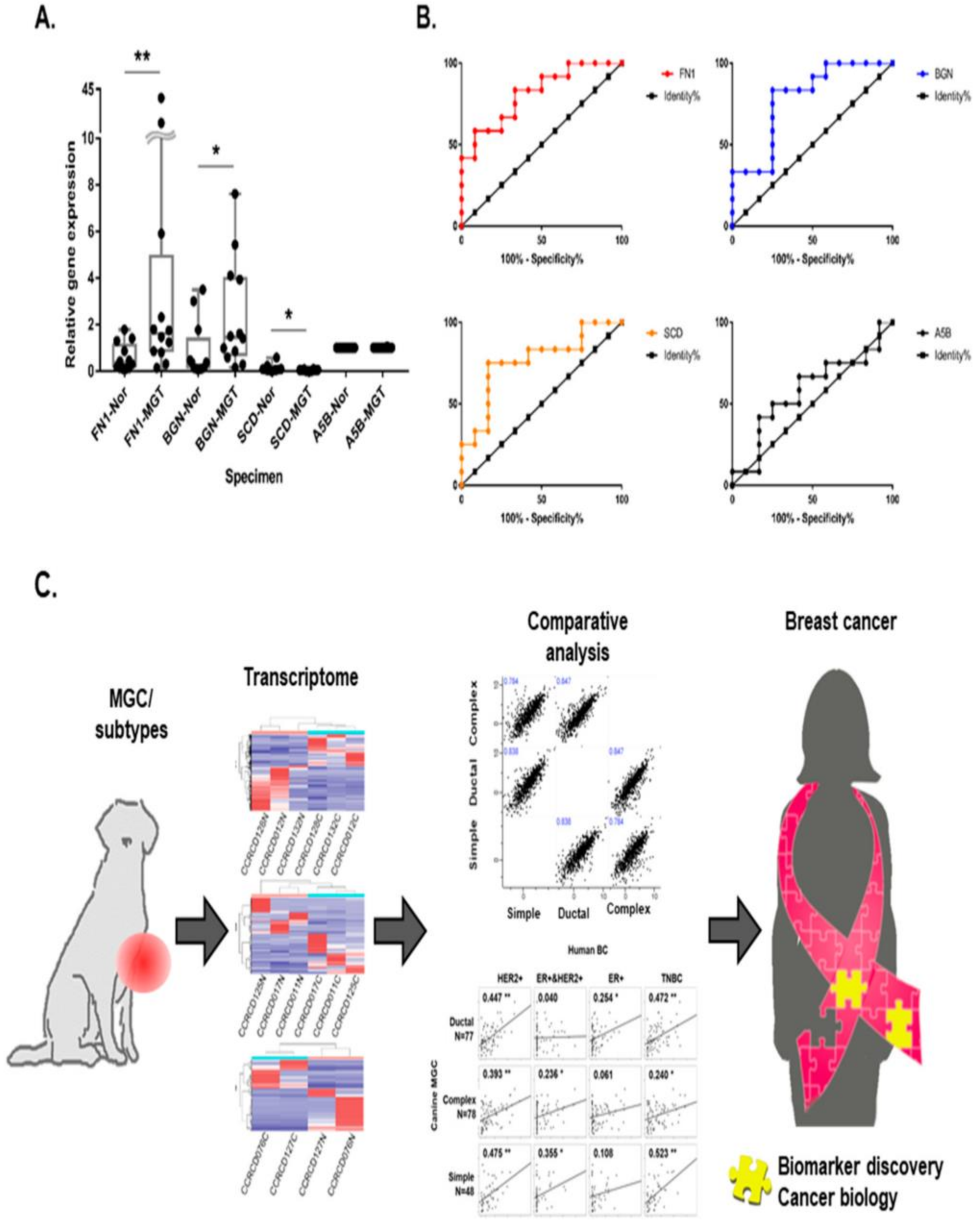


Figure 1-9. Real-time RT-PCR for validation of MGC-enriched RNA expression.

(A) Box-and-whisker plots of relative gene expression levels in MGCs and matching adjacent normal samples. The Mann–Whitney test was performed. Bar graphs representing relative RNA expression of *FN1*, *BGN*, and *SCD* genes in 12 MGCs and adjacent normal tissues. Statistical significance is indicated by asterisks and relative p-value (** $p < 0.01$, * $p < 0.05$) (B) Receiver operating characteristic (ROC) curve for each gene expression level. (C) Conceptional scheme of canine MGC as a model to study human BC and discovery new biomarkers.

DISCUSSION

In this study, we performed genome-wide transcriptome analysis of spontaneous canine MGCs and compared to transcriptome data from four molecular subtypes of human BC. Although the sample size was small, being a pilot study (16 transcriptomes; 8 MGCs with matching adjacent normal tissues), this study could reveal transcriptome signatures enriched in canine MGC and subtypes.

Although there were several reports presenting that canine MGC is a good model for human BC study, subtype levels were still unclear (Abdelmegeed and Mohammed, 2018a, Lutful Kabir et al., 2015). We thus determined whether these two systems are compatible at the transcriptome level. We analyzed correlation in gene expression existing between the subtypes of human BC and canine MGC using the genes differentially expressed in canine MGC. Overall, the level of correlation seemed low between human BC and canine MGC (max r = 0.523, min r = 0.040). However, correlation among the subtypes within canine MGC was not strong either (max r = 0.767). It means that each subtype of canine MGC has a unique gene expression pattern. One of the interesting findings in the correlation analysis was the strongest correlation in human TNBC with canine simple MGC (r

= 0.523) and ductal MGC (r = 0.472). The existence of high transcriptomic correlation between canine MGC subtypes (ductal and simple) and human TNBC might be more important since TNBC has been highlighted in clinical and biomedical research due to its aggressive characteristics with poor prognosis. It has been known that the most common histological subtype of TNBC is invasive ductal carcinoma and their genetic profiles are shared by basal-like BC (Bryan et al., 2006, Plasilova et al., 2016). Thus, our results suggested that transcriptome signature of canine MGC and subtypes is able to represent the origin and characteristics of human BC. On the other hand, ER+-related human BC subtypes (ER+, ER+/HER2+) had few or no significant correlation with any canine MGC subtypes but tend to be shared by ductal and simple subtypes, respectively, in the given groups (Fig. 1-5). However, this result, showing week correlation in ER+-related subtypes with canine MGCs, should be confirmed if it is influenced by spayed dogs.

Since many studies have been performed in human BCs, we reviewed literature regarding human BC and oncogenes to compare our findings to human studies. First, four out of 16 representative DEGs found in all three subtypes of canine MGCs have strong references in human cancer as biomarkers: *CCL23*, *CXCL10*, *SFRP2,* and *FRZB*. These genes are altered in at least four types of human cancers, including BC (Ejaeidi et al., 2015, Ugolini et al., 1999, Veeck et al., 2008). Second, six genes, *CHI3L1*, *CXCL8*, *FOXC2*, *SERPINE1*, *SFRP2*, and *TF*, which are grouped within the highest enrichment GO term, “positive regulation of angiogenesis", have been reported to play roles in various cancer processes, including BC (Kolbl et al., 2016, Libreros et al., 2013, Mazzoccoli et al., 2012, Wang et al., 2018). Moreover, 45 genes enriched in BP GO terms, 'glycan biosynthesis and metabolism' and 'lipid metabolism', may provide strong evidence that cellular metabolism is fundamentally altered in cancer tissue, and lipid metabolism may have crucial roles in cancer progression (Hashmi et al., 2015). This survey confirms that dogs and dog MGCs are good animal models for human breast cancer study at the transcriptome level.

We further investigated the biological roles of MGC subtype-enriched DEGs. KEGG pathway analysis using 211 up- and 306 down-regulated DEGs revealed that cancer signaling in the complex subtype was mainly triggered by Wnt-Frizzled *LRP5/6* and *GPCR* signaling, whereas glycan biosynthesis and metabolism are strongly blocked through down-regulation of *PPAR* signaling, beginning with *CD36-FABP*.

A total of 141 up- and 120 down-regulated DEGs were tested in the ductal subtype. Similar to the complex subtype, both glycan biosynthesis and lipid metabolism were down-regulated, but down-regulated retinol metabolism was found only in the ductal subtype. Although down-regulated biological processes were shared by two different subtypes, there were discrepancies in the list of up-regulated pathways between complex and ductal subtypes. KEGG pathways involved in cancer, such as cell adhesion, PIK3-Akt signaling, and ECM-receptor interaction, are enriched in the ductal subtype. Many ECM molecules have been associated with breast cancer development (Oskarsson, 2013). These discrepancies may partly come from differences in cellular origin, compositions of cell types and the cancer environment.

Since only two pairs of specimens comprised the simple subtype, the number of identified DEGs was small (79 up- and 115 down-regulated). Focal adhesions as well as the Wnt and ECM-ITGB pathways were up-regulated. Interestingly, insulin signaling, including the *FBP1* gene, was the most highly enriched in down-regulated DEGs, but we know that down-regulation of *FBP1* promotes tumor metastasis and indicates poor prognosis in other cancers (Li et al., 2016b). If the results from the canine MGC subtype-enriched transcriptome profiles are validated in a large sample size, it will likely be helpful in developing cancer therapies for human breast cancer counterparts.

As previously stated, only a few aspects of PROMPT, a newly identified class of RNAs produced just upstream of the promoters of active protein-coding genes, have been characterized; due to being rapidly dumped by exosomes, their biological functions remain to be revealed (Liu et al., 2015, Wang et al., 2015). We thus tested whether PROMPT expression can be detected in paired-end stranded total RNA sequencing data. First, we should note that the “PROMPT” we measured in this study might differ from the general term “PROMPT”. We used the term

PROMPT since “promoter upstream transcripts” is exactly what we investigated in this study. However, many transcripts may not satisfy the criteria of the general term PROMPT in size or amount (Preker et al., 2011). Furthermore, our measurements also have a few limitations to calculating accurate levels of transcript expression because small portions of non-coding RNAs including PROMPTs are annotated and characterized with their structures. We then measured all the sequence reads mapped upstream of the promoter region (−1500 bp~TSS) without consideration of RNA structures. It may not represent exact amounts of transcripts if the size is longer than 1500 bp or exon structures vary.

In this study, we selected and showed two gene promoters upstream regions representing each correlation type (Fig. 1-7A). Although negative correlation between genes and PROMPTs were stronger than positive correlations, positive correlations were more reliable because many genes with negative correlations were found as artifacts due to the low number of PROMPTs. Target-enriched high-throughput sequencing for short transcripts may be helpful for this type of analysis. Furthermore, comprehensive annotation with extensive transcriptome analysis in dogs is mandatory for comparative medicine and future study. In addition, diverse small-size non-coding RNAs, including micro RNA, which were not analyzed in this study due to the limitation of RNA isolation method but can be done by miRNA capturing in the future, might have very important roles in canine MGC as well.

Canine MGC has been proposed as a comparative model for spontaneous tumors of human BC due to their genetic, clinical, and biological similarities to human BCs. In addition, closely shared environmental conditions between dog and owner can be beneficial in an approach using epigenetic aberrations. Thus, studies for canine MGCs, counterparts of human BCs, can provide new clues for biomarker screening in human BCs (Fig. 1-9C). We confirmed RNA sequencing data and validated three genes' expression in additional sets of samples using quantitative real-time PCR. *FN1* and *BGN* were targeted here due to their expression pattern being similarly up-regulated in human breast cancers. However, *SCD* was identified as a down-regulated gene in this study but is known to be up-regulated in human BCs. These results might represent similarities and discrepancies that exist between human BC and canine MGCs.

In conclusion, this study reports the comprehensive transcriptome profile of spontaneous canine MGCs and subtypes. Sets of DEGs in canine MGCs were determined from overall canine MGCs for each subtype. Many genes, but not all, listed in this study have been reportedly associated with human cancers including breast cancer. Three canine MGC subtypes then were matched to four human BC subtypes according to their transcriptome profiles. This study may represent the extant similarities between human BCs and canine MGCs. Thus, the current study provides new clues and clinical implications for better understanding of canine MGCs and their application to human BCs. Further validation using large sample numbers will reveal more general features, but our current study provides an important initial understanding of canine MGCs in different canine MGC subtypes.

CHAPTER Ⅱ

Analysis of Opposing Histone Modifications H3K4me3 and H3K27me3 Reveals Candidate Diagnostic Biomarkers for TNBC and Gene Set Prediction Combination

INTRODUCTION

Breast cancer (BC) is one of the most common cancers occurring among females and one of the most dominant causes of cancer related deaths alongside lung cancer (Jemal et al., 2011). Known as highly diverse cancers, BCs are characterized by distinct genetic variations, clinical symptoms, treatments, and prognosis outcomes. In previous studies, breast cancer has been clinically classified by major changes in expression levels (O'Brien et al., 2010) including high expression of the estrogen receptor (ESR), progesterone receptors (PGR), and HER2. Most clinically diagnosed BC types have at least one of these features, but basal-like triple-negative breast cancer (TNBC) presents no expression of the three. TNBC is more aggressive and has poor prognosis, but because of its minor occurrence treatments and therapies, are scarce (Perou, 2011). Some patients suffering from TNBC benefit from chemotherapy, but still need a better method of treatment less toxic and dangerous to the patient. Recent studies of TNBC revealed distinct gene mutation patterns and repressive signal pathways (Carey et al., 2006). Despite the effort of continuing research, the understanding of the governing gene mechanism and systemic regulation of TNBC pathways is lacking compared to other more dominant breast cancer types.

BC molecular identities can be further specified based on epigenetic features. Epigenetic regulation has been a major factor of gene expression control (Jones and Baylin, 2007). Various types of epigenetic control such as DNA methylation and histone modification are crucial for the activation and repression of genes in cancer. Recent studies revealed DNA methylation and histone modification profiles as plausible predictors of well-defined subtypes (Chen et al., 2016). Among the different types of histone modifications, H3K4me3 is a major modification that moderates genes to an active state (Koch et al., 2007). Conversely, histone modification H3K27me3 is a major modification that down-regulates genes when highly enriched (Barski et al., 2007). Continuing efforts to discover various precursors to breast cancer by comparing five or more histone modifications enabled a more thorough understanding and precision of prediction (Xi et al., 2018). However, less is known of the histone modifications specifically contributing to TNBC and the expression differences regulated by the histone regulation.

The purpose of this study was to establish an analytical pipeline for discovering TNBC biomarkers from published histone modification peak data. The combination of the two histone modifications, H3K4me3 and H3K27me3, in TNBC cell lines presented hallmarks of TNBC gene expression against normal breast cell lines. The results providing genes that are epigenetically regulated in

TNBC, were proven successfully by quantitative transcriptional analysis, and suggested biomarker candidates that could specifically diagnose TNBC against normal.

MATERIALS AND METHODS

DATA acquisition and bioinformatics analysis

H3K4me3 and H3K27me3 ChIP-seq data from human BC cell lines, MDA-MB-436, SK-BR-3, ZR-75-1, and human normal breast cell line, HMEC, was obtained from the GEO database GSE62907 (Chaligne et al., 2015). RNA-seq data for all the cancer cell lines was also obtained from the same database. Human normal breast cell line HMEC RNA-seq data was obtained from dataset GSE62820 (Rahman and Mohammed, 2015).

Each ChIP-seq raw dataset was aligned with human reference hg19 using the HISAT2. The peak finding was performed using the 'findPeaks' command of HOMER. Differential peaks of TNBC cell line MDA-MB-436 was analyzed by using HMEC ChIP-seq data as a control group. HOMER software command 'getDifferentialPeaks' was used to identify H3K4me3 enriched peaks, H3K27me3 repressed peaks for activated regions and H3K4me3 repressed, H3K27me3 enriched peaks for down-regulated regions. The fold change cutoff was ≥ 4 for enriched and ≥ 2 for the repressed peak regions. Annotation of all regions was performed using the ‘annotatePeaks.pl’ function of HOMER. Within the annotated list, H3Kme3 regions associated with transcription such as TSS, promoter, and exon regions were selected as potential targets. RNA-seq data was aligned and peak analysis performed using the HOMER transcriptome analysis pipeline. From the ChIP-seq sorted genes, candidates were selected by matching profiles that were high-expressed or low-expressed specifically in the MDA-MB-436 dataset.

Among the histone modifications upregulation of histone H3K4me3 and downregulation of histone H3K27me3 were selected for the peak comparison. After normalization, TNBC H3K4me3 peak data was compared against HMEC H3K4me3 peak data for the differential histone enriched regions. To identify statistically high or low enriched regions, HMEC and MDA-MB-436 ChIP-seq data was used as control groups and experimental groups. As for the potential upregulated regions, only locations in TNBC peaks enriched more than four-fold compared to the HMEC ChIP-seq data and HMEC H3K27me3 locations with a fold enrichment more than two compared to HMEC were sorted (Fig. 2-1A). The opposite method was implemented to sort potential down-regulated regions. HMEC H3K4me3 peaks that were four-fold higher than TNBC and H3K27me3 peaks of TNBC two-fold higher than HMEC were selected. Because the H3K27me3 profile is dispersed across the entire gene structure, highly enriched peaks are difficult to locate. As a result, histone modification H3K27me3 are sorted by a two-fold degree. After differential analysis, sorted regions were annotated with gene names and enriched gene positions. Among the DNA structure, H3K4me3 regions with expression influence were selected as potential histone enriched regions; promoter, TSS, and exon.

Correlation of genes matching the up or down regulating prediction was calculated by comparing each individual ChIP-seq peak log2 fold change and its matching RNA-seq log2 expression fold change. The combined method of sorting candidate genes using H3K4me3 and H3K27me3 histone modification was compared with the methods that sorted the genes using only H3K4me3 or H3K27me3. The accuracy was estimated by a percentage of genes that matched its predicted RNA expression pattern.

Cell culture

MCF-10A normal cell line was cultured in Mammary Epithelial Cell Growth Basal Medium (MEBM) BulletKit (Lonza cat # CC-3150) with an additional 10% fetal bovine serum (FBS, Gibco cat # 16000069) and 1% Antibiotic-Antimycotic product (AA, Gibco cat # 15240062). The cell line SK-BR-3 was cultured using the RPMI media with an additional 10% FBS and 1% AA product. MDA-MB-436 was cultured in the DMEM media with an additional 10% FBS and 1% AA product.

Quantitative RT-qPCR

The RNA isolation was processed using the Rneasy Plus Mini Kit (Qiagen, Hilden, DE). The genomic DNA contamination was eliminated by using the gDNA elimination columns. In addition, 2 μg of the total RNA was used for the cDNA synthesis using the OMMISCRIPT RT KIT (Qiagen, Hilden, DE). The primers for each target gene were designed spanning two different exons. The real-time PCR was performed using the CFX96 Touch Real-Time PCR Detection System (Bio-Rad). The relative gene expression was measured by the ΔΔCTmethod. The data were normalized to the 18S rRNA.

Expression box plot analysis

The gene expression data in the TCGA cancer patients' samples were analyzed with GEPIA (http://gepia2.cancer-pku.cn). The gene expression in the normal data was compared only with the basal-like and TNBC subtypes. The log2 fold change cutoff was set to 1. The p-value cutoff was set to less than 0.01.

Kaplan-Meier plot analysis

The web-based Kaplan-Meier plotter was used to evaluate the effect of candidate genes on survival rates in more than 3,000 BC samples. The hazard ratio (HR) was given with 95% confidence intervals, and log rank P value was calculated and displayed on the web page. The log rank P-values were calculated by auto-selecting the best cutoff option. The affymetrix ID of the top 10 potential up- and down-regulated candidates were listed in Table S2.

ROC analysis

Classification using receiver operating characteristic (ROC) curves was performed using 1,222 normal and breast cancer patients in the TCGA database. The area under the curve (AUC) scores and p-values were calculated using the easyROC web-based tool. The gene set combined logistic regression model was achieved using SPSS statistical analysis software.

RESULTS

Distribution of H3K4me3 and H3K27me3 histone modifications as TNBC-associated epigenomic signatures

ChIP-seq data from HMEC and MDA-MB-436 that represent normal and TNBC cell lines, respectively, were obtained from the public dataset GSE62907. To determine the TNBC-enriched epigenetic alteration, two histone modification signals H3K4me3 and H3K27me3 on the gene promoter regions were compared across the two cell lines. Overall procedures are depicted in Fig. 2-1A. In brief, we normalized all ChIP-signal data to the corresponding inputs. The regions of differentially modified histones were identified from the comparison of HMEC and MDA-MB-436 cell lines. Up- and down-regulated histone modifications in TNBC were selected when regions had a larger than two-fold difference in H3K4me3 and H3K27me3, compared to the normal HMEC (Fig. 2-1A). The promoter region was defined by convention as 2 kb upstream of the TSS of a gene.

Identified as up-regulated genes in TNBC were 1,008 genes with highly enriched H3K4me3 regions and 4,954 genes with depleted H3K27me3 signals in MDA-MB-436 cells. Conversely, 1,608 genes with enriched H3K4me3 and 5,082 genes with low H3K27me3 in HMBC were identified as down-regulated genes in TNBC. As a result, a list of genes exclusively up-regulated (148) and down-regulated (41) in TNBC was determined by combining high H3K4me3 and low H3K27me3 profiles and vice versa (Fig. 2-1B). Each potential candidate was scored by its H3K4me3 peak score. Integrative genomics viewer (IGV) depicted histone modifications on the regions of the NOVA1 and IRX2 genes that were scored in the top (Fig. 2-1C). H3K4me3 signals were enriched and H3K27me3 disappeared on the NOVA1 promoter region in MDA-MB-436, while H3K4me3 signals are very low and H3K27me3 are enriched in HMEC. Oppositely, H3K4me3 signal was highly enriched and the H3K27me3 disappeared on the DUSP6 genes in HMEC, while H3K4me3 signals disappeared and H3K27me3 were enriched in the MDA-MB-436.

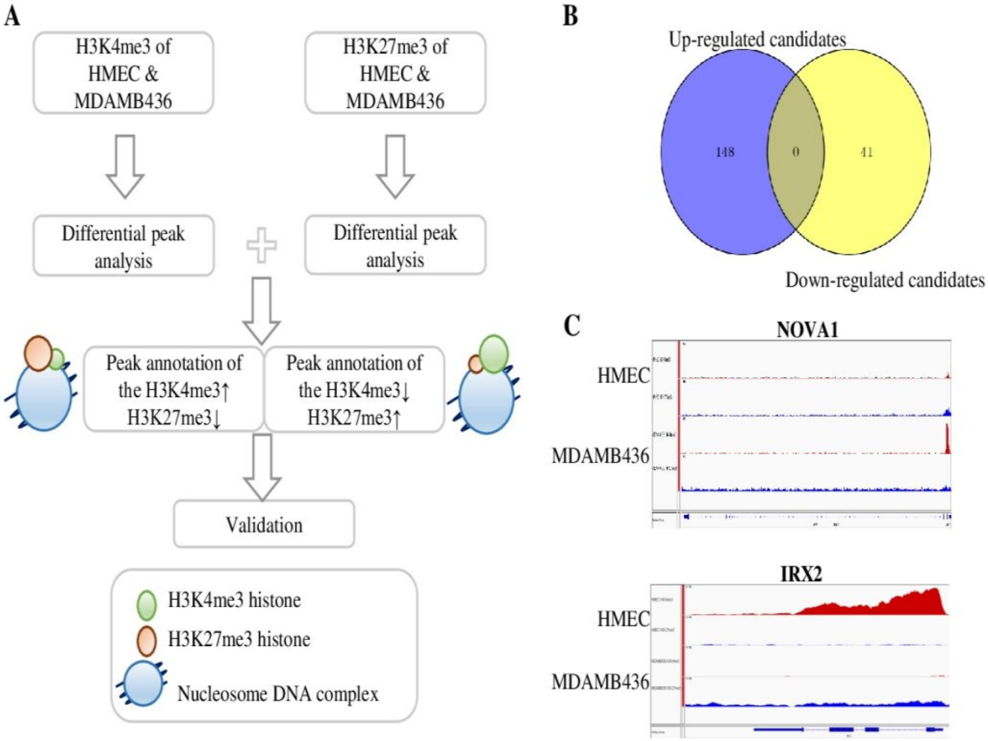

A
H3K4me3 of HMEC & MDAMB436
H3K27me3 of HMEC & MDAMB436
Differential peak analysis
Differential peak analysis
Peak annotation of the H3K4me3↑ H3K27me3↓
Peak annotation of the H3K4me3↓ H3K27me3↑
Validation
H3K4me3 histone
H3K27me3 histone
Nucleosome DNA complex
B
Up-regulated candidates
148
0
41
Down-regulated candidates
C
NOVA1
HMEC
MDAMB436
IRX2
HMEC
MDAMB436

Figure 2-1. Project workflow and ChIP-seq analysis.

(A) Workflow of the ChIP-seq analysis. (B) Venn diagram of 148 potential up-regulated genes and 41 potential down-regulated genes in triple-negative breast cancer. (C) The histone profile H3K4me3 (red) and H3K27me3 (blue) of candidate gene in the HMEC and MDA-MB-436 ChIP-seq datasets.

Integration of transcriptome data revealed TNBC-associated signature genes

Epigenetic profiles such as histone modification represent convincing evidence as biomarker candidates. However, epigenome profiles are not perfectly aligned to their corresponding expression data such as transcriptomic or proteomic data (Gomez-Cabrero et al., 2014). To investigate the effects of epigenomic aberrations on gene expression, matching RNA-seq data of HMEC and MDA-MB-436 were merged with two additional transcriptome datasets obtained from other sub-types of BC cell lines, SK-BR-3 (luminal type) and ZR-75-1 (HER2 expressing) (Chaligne et al., 2015). The influence of histone modification on gene expression was examined by calculating the percentage of RNA-seq expression patterns that match with histone peak fold changes. Overall, H3K4me3 has better correlation than H3K27me3 with gene expression levels in up- and down-regulation. Of note, the combination of histone markers, high H3K4me3 and low H3K27me3 for up-regulated genes and vice-versa for down-regulated genes, presented a remarkable improvement in the correlation with gene expression. The top 10 highest scored genes are indicated by red and blue dots for up- and down-regulated genes. For further analysis, the top 10 scored genes (up- and down-regulated) in ChIP-seq were selected and listed with corresponding RNA expressions in Table 2-1.

Table 2-1. Differentially expressed candidates

Gene name	Score	MDAMB436	SKBR3	ZR751	HMEC
CDH2	1619.5	757.46	428.79	10.67	35.11
DCLK2	857.0	436.45	43.76	1.93	1.28
NOVA1	850.3	120.23	1616.05	27.63	0.53
PLCL2	692.7	121.57	0.00	8.09	0.54
SOX5	826.3	37.11	1.07	6.66	0.05
SALL1	806.2	235.11	0.00	0.08	0.54
SYTL4	780.1	360.80	69.22	112.64	22.07
DNER	744.5	943.92	0.00	15.72	63.47
NAT8L	700.4	284.23	1827.86	320.66	3.46
MMP16	680.3	188.47	99.50	10.53	21.28
DUSP6	2052.3	58.25	9.73	38.26	2706.05
IRX2	1564.5	0.58	0.00	180.80	464.21
ATP2B1	1445.1	212.46	34.82	261.43	1771.94
VSNL1	1296.7	0.02	72.81	0.71	430.11
ADRB2	1248.1	0.01	47.46	0.07	115.33
PLXDC2	1139.4	24.44	0.00	384.42	1078.95
PLD5	1124.9	0.80	196.58	0.79	125.28
TPD52L1	1007.3	59.51	0.00	917.35	332.26
FAM84A	743.7	0.12	4.15	25.43	414.37
SNX19	648.5	553.93	717.26	853.20	1388.00

Candidate selection. Each gene is sorted by a combined data of ChIP-seq data and RNA-seq data. Scores represent ChIP-seq H3K4me3 scores calculated by HOMER. The four FPKM data represent expression profiles retrieved from for cell lines HMEC, SK-BR-3, ZR-75-1 and MDA-MB-436. Ranked by score, potential up-regulating (left) and down-regulating (right) candidates were achieved.

Quantitative RT-PCR validation of candidates' gene expression

The expression of selected candidate genes was confirmed by quantitative real-time RT-PCR in the corresponding breast cancer related cell lines, MCF-10A (normal), MDA-MB-436 (TNBC) and SK-BR-3 (HER2+). For the genes selected by highly-enriched H3K4me3 by depleted H3K27me3, the gene expressions of 10 up-regulated candidates were confirmed by real-time RT-PCR. Except for the *DNER* gene which showed half of the expression in TNBC than in normal cell line, the nine remaining genes expressed highly in TNBC (Fig. 2-2). Interestingly, we found two genes (*MMP16* and *NAT8L*), almost exclusively expressed in TNBC. The largest discrepancy in relative gene expression levels between TNBC and normal was in *MMP16* (-3,000 fold) followed by NAT8L (-2,500 fold). *DCLK2* and *SYTL4* showed higher gene expression levels in cancer cell lines SK-BR-3 (*HER2*+) and MDA-MB-436 (TNBC). *CDH2*, *NOVA1*, *PLCL2*, *SOX5*, and *SALL1* were highly expressed in TNBC, but not in HER2+. Conversely, gene expressions down-regulated in the MDA-MB-436 cells selected from the combination of low H3K4me3 and high H3K27me3 were validated in four of 10 candidates (Fig. 2-3). *DUSP6* and *VSNL1* gene expressions were significantly down-regulated in HER2+ and TNBC breast cancer cell lines. Only *TPD52L1*, and *FAM84A* were most significantly down regulated in TNBC compared to MCF10A and SK-BR-3. The expressions of *ATP2B1*, *ADRB2*, *PLXDC2*, *PLD5*, and *SNX19* grouped in down-regulated genes were not correlated with histone states in TNBC and normal.

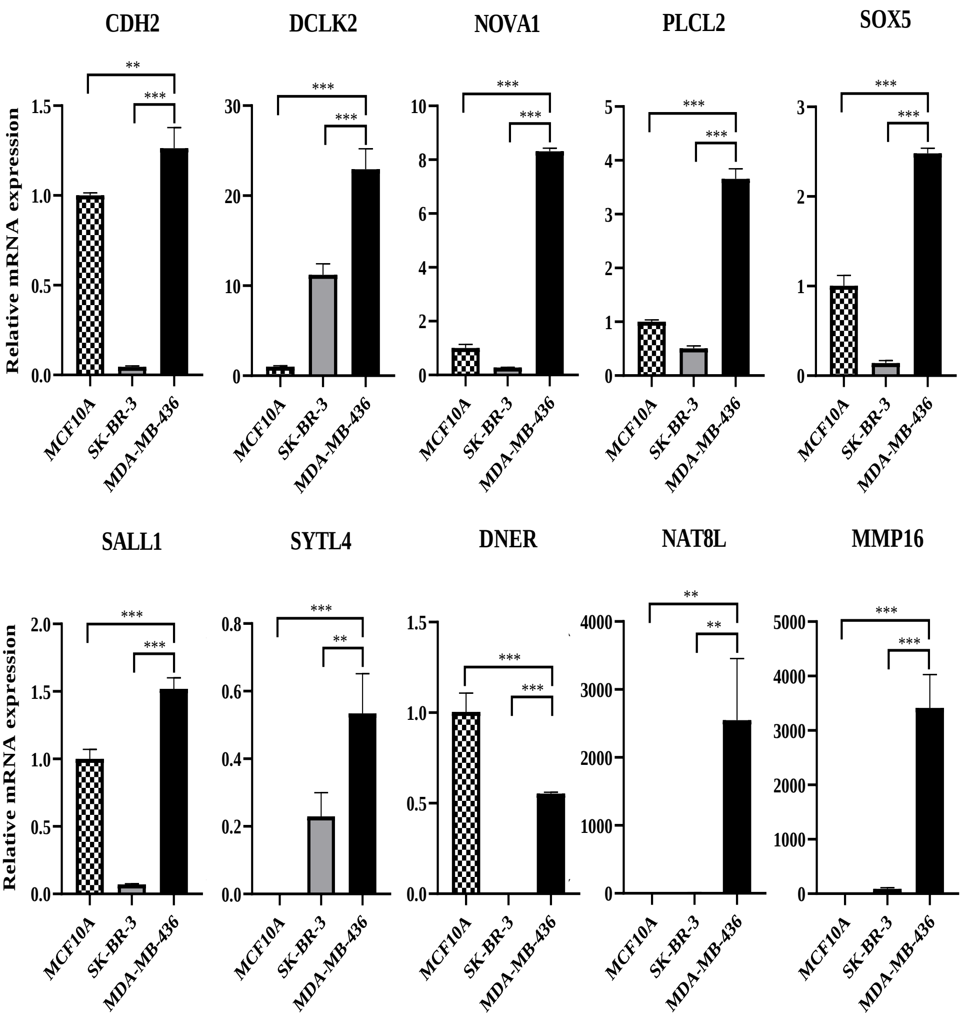


Figure 2-2. Up-regulating biomarker RNA expression validation.

The bar plots of relative RNA expression of 10 genes considered as up-regulating biomarkers for the TNBC in MCF10A (black-stripes), SK-BR-3 (grey), and MDA-MB-436 (black) cell lines.

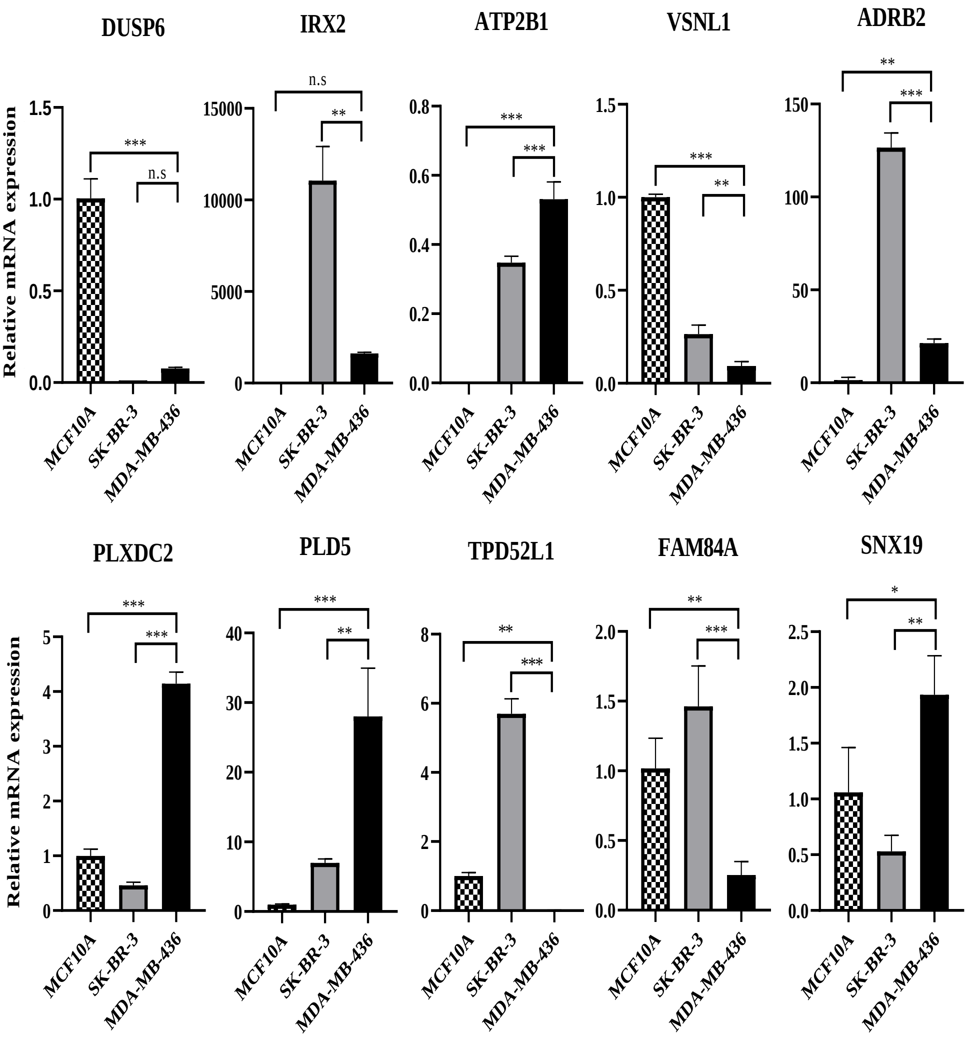


Figure 2-3. Down-regulating biomarker RNA expression validation.

The bar plots of relative RNA expression of 10 genes considered as Down-regulating biomarkers for the TNBC in MCF10A (black-stripes), SK-BR-3 (grey), and MDA-MB-436 (black) cell lines.

Clinical correlation of the genes selected by histone modification data: TCGA data correlation of up- and down-regulated genes and cancer patient survival data

To expand the RT-PCR validated gene set data to cancer patient data in the TCGA public domain, we analyzed the TCGA data for the selected gene. Unfortunately, since TCGA data have not been classified by TNBC, the gene expression pattern in overall BC patients was not nicely correlated with the results in this study targeting TNBC. For example, gene expressions of *NOVA1*, *SOX5*, and *NAT8L* were found higher in the TNBC cell line than in the normal cell line while expressed lower in overall breast cancer than the healthy population (Fig. 2-2). This discrepancy may come from the absence of corresponding classifications in TCGA data, since these three genes whose expressions were upregulated in TNBC were lower in the other cancer cell line (SK-BR-3; *HER2*+) than the normal cell line.

To further correlate the candidate genes with cancer patient data and examine the prognostic value of the candidate genes in BC patient databases, we used the expression box plots (Box-plots: http://gepia2.cancer-pku.cn) of cancer patients and the normal population (Tang et al., 2017) and their Kaplan-Meier plots (KMplot; https://www.kmplot.com). Aberrant gene expression and its influence on overall survival (OS) was presented. Top two representative genes are shown in Fig. 2-4; *CDH2* in up-regulated and *DUSP6* in down-regulated. The *CDH2* found as an up-regulated gene in MDA-MB-436 was highly expressed in basal-like and TNBC patients. Also, patients with higher *CDH2* expression have high mortality when compared to low expressed patients (HR = 1.36, logrank P = 1.3e-07). Conversely, *DUSP6* down-regulated in MDA-MB-436 presented significantly low expression levels in basal-like and TNBC patients when compared to healthy controls (Fig. 2-4A). Survival curves associated with *DUPS6* gene expression indicated that lower *DUPS6* expression in BC patients has an association with worsening OS (Fig. 2-4B). Moreover, we implemented a classification model based on the expression of *CDH2* and *DUSP6* using 1,222 normal and breast cancer patients from the TCGA database. ROC curves from the individual genes had a high AUC with 79% in *CDH2* and 92% in *DUSP6*. When these two genes were combined using the binary logistic regression method, AUC of sensitivity/1-specificity was up to 93%.

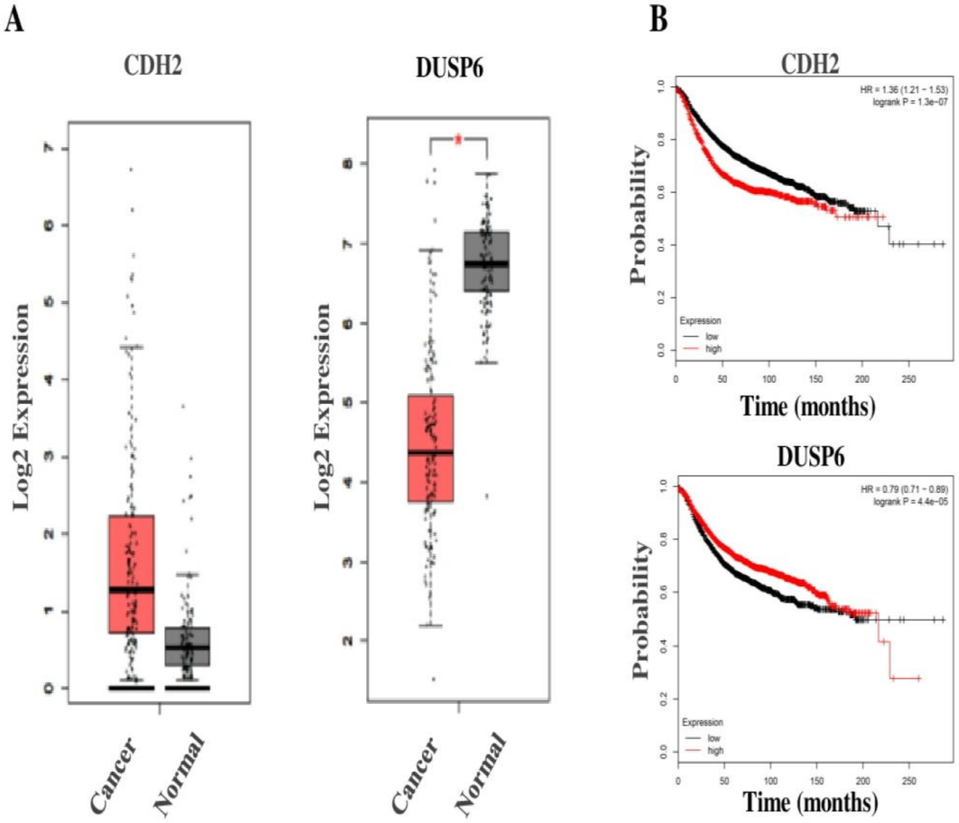


Figure 2-4. Diagnostic and prognostic values of candidates.

The highest ranked genes in up- and down-regulated were subjected to the TCGA data. (A) The gene expression box plot in the TNBC and basal-like BC (red) and normal (grey). (B) The overall survival of breast cancer patients expressing *CDH2* and *DUSP6* (high: red, low: black) in the KM plot.

DISCUSSION

Recently, large amounts of omics data have been produced globally and made publicly available. Most are genomic, such as single nucleotide polymorphism (SNP), copy number variation (CNV), and transcriptomic data. However, recently epigenomic data such as histone modification, methylation status, and chromatin profiles have seen a continuing increase in various cancer studies (Nebbioso et al., 2018), since epigenetic mechanisms are recognized as critical risk factors in the development of cancers.

In this study, we developed a strategy to investigate triple-negative breast cancer (TNBC) biomarkers based on the epigenome dataset of histone modifications (Chaligne et al., 2015). The combination of two different histone modification markers, high H3K4me3 and low H3K27me3 and vice versa, made significant improvement in the correlation with transcriptomic data compared to each marker only. This strategy is similar to the concept of bivalent chromatin that has a role in developmental regulation in pluripotent cells and is defined wherein the region of DNA is bound to histone proteins with repressing and activating epigenetic regulators. However, there were some discrepancies such as the range of region modification that occurred and the combination of epigenetic markers.

We predicted most likely highly up- or down-regulated genes in terms of transcriptome expression based on two histone marks of H3K4me3 and H3K27me3 and listed up top 10 candidates up- and down-regulated in TNBC. Then, we tested the level of gene expression using RNA-seq data and quantitative Real-Time PCR in three BC related cell lines (MCF-10A, MDA-MB-436, and SK-BR-3). Because of availability, the MCF-10A cell line was used as normal breast cell line in our study instead of HMEC used in RNA-seq and ChIP-seq data. This may present unexpected high expression levels of *CDH2*, *SALL1* and *DNER*, and low expression of *IRX2*, *PLXDC2* and *SNX19* in the normal cell line. This result should be confirmed by extended numbers and types of cell lines to exclude cell line specific features.

The in vitro cell line analysis of histone modifications and gene expression was applied to the public clinical data to retrieve the prognostic significance of individual candidates. The top scored genes, *CDH2* and *DUSP6*, up- and down-regulated respectively in TNBC, successfully represented the aggressive pathological phenotype of TNBC, which may directly link to general BC patient's overall survival in TCGA expression plots and KM-plotter (Fig. 2-4).

Many of the candidate genes we selected have been studied regarding their BC-related molecular functions. The remarkable increase of aspartate N-acetyltransferase (*NAD8L*) is reported to develop cancer growth in overall cancer types and is a valuable target for cancer treatment (Zand et al., 2016). Up-regulation of matrix metallopeptidase 16 (*MMP16*) from miR-155 is reported to enhance proliferation and migration in TNBCs. *CDH2*, commonly known as N-cadherin contributes significantly towards transitioning from the epithelial state to the mesenchymal state (EMT) and enacting abnormal cells to invade and metastasize to nearby as well as distant tissues. Sex determining region Y-box protein 5 (*SOX5*) expression is reported to increase *EZH2* expression inducing breast cancer cell proliferation and invasion (Sun et al., 2019). Controversially, *SALL1* is a tumor suppressor in luminal BC types, as well as in TNBCs (Ma et al., 2018, Wolf et al., 2014). Notably, we newly identified four novel candidate genes never been reported in BC (Nova Alternative Splicing Regulator 1 (*NOVA1*), Phospholipase C Like 2 (*PLCL2*), Synaptotagmin Like 4 (*SYTL4*), and Delta/Notch Like EGF Repeat Containing (Gardner et al., 2016)). Since *NOVA1* (52.37%) and *DNER* (34.45%) have been studied in various other cancers, but not in BC.

BCs are continuously separated by different measures for more precise classification. We used the top-ranked genes in up- and down-regulated markers to observe if it could contribute to enhancing BC classification. Each gene showed high differentiation, but the combination of differentially expressed candidate genes, predicted by H3K4me3 and H3K27me3 histone marks analysis, using the logistic regression models further improved the accuracy of BC diagnosis.

In conclusion, we suggested a bioinformatical strategy to reveal TNBC biomarkers using histone modifications of H3K4me3 and H3K27me3 and combining transcriptomic datasets. The functional study of the candidate genes found in this study in BC, especially in TNBC, is necessary in more extensive datasets and cancer types for better understanding and discovering novel biomarkers and therapeutic targets.

CHAPTER Ⅲ

Common Plasma Protein Marker LCAT in Aggressive Human Breast Cancer and Canine Mammary Gland Carcinoma

INTRODUCTION

Among all the malignant tumors, breast cancer (BC) is known to be one of the most frequently diagnosed cancers. In fact, it is the most-studied malignancy in the world (Woolston, 2015). Despite the efforts of various researchers, the struggle to understand and cure breast cancer continues on many fronts. A large number of genes have been selected as biomarkers to further understand BC: whether it is invasive or non-invasive (Hoag, 2015), whether it is classified to a certain category, etc. BC biomarkers can be organized into three major categories: prognostic, predictive, and pharmacodynamic markers (Ulaner et al., 2016). The most frequently used biomarkers are the prognostic and therapy-decision biomarkers, consisting of tissue-based biomarkers such as estrogen receptor (ER), progesterone receptor (PgR), and human epidermal growth factor receptor 2 (*HER2*) (Harris et al., 2007). Additional protein biomarkers were able to be identified by the improvement of mass spectrometry (MS)-based proteomics technologies, which enabled blood analysis of solid tumors (Geyer et al., 2017). Although sequencing technology have been increased and enhanced, biomarkers that depict advanced stage malignancies or cancers that undergo metastasis are considerably scarce compared to early-stage prognostic biomarkers. The most well-known markers would be carcinoma antigen 15-3 (CA-15-3), CA-27/29, and carcinoembryonic antigen (CEA) (Banin Hirata et al., 2014), which respectively indicate relevant data related to breast cancer, but still more indicators are needed .

Canine MGC are frequently studied alongside human BC. Not only is it studied due to dogs' close relations to humans, but it is also a well-known animal model for alternative human BC investigation (Salas et al., 2015). However, when it comes to cancer indicating markers, canine prognostic biomarkers are rare. Most are inferred indicators, such as CA-15-3, that derived from data in human samples. To further understand and diagnose MGC, it is certainly a necessity to find suitable biomarkers that depict stage-wise and aggressive MGC.

This study focused on discovering aggressiveness biomarkers of canine MGC using canine normal and cancer plasma samples. After an extensive search consisting of 36 fractions of each sample run in mass spectrometry (MS), potential targets were further filtered through MRM data and validated by Western blot. Once a suitable biomarker was discovered, *in silico* data of human breast cancer was implemented to investigate its possibility as a human aggressiveness-indicating biomarker and further validated in human plasma samples and cell lines. This study will provide a novel aggressiveness biomarker that can be applied to both human and canine malignant cancer patients.

MATERIALS AND METHODS

Plasma sampling

Canine normal and cancer plasma were obtained from the Canine Cancer Research Center project (CCRC). No live animals were directly involved in this study. For each sample, 50ul of plasma was used. The depletion process was done by using the multi affinity removal spin cartridge top 2 depletion kit (Agilent, location, Cat # 5188-8825). Every product was concentrated by using a speed-vac and 200ul of HPLC water was added to dissolve for further processing. Digestion was done via the filter-aided sample preparation protocol (Wisniewski et al., 2009). Desalting was done by using SDB-RPS resin. The initial 6 canine normal and cancer samples were pooled into 3 samples. 6-plex tandem mass tagging was implemented for the normal and cancer samples. Thirty-six fractions were made by using Waters' HPLC columns. The column length is 25 cm, consisting of C18 with a pore size of 5um attached to an HPLC separation unit. Twenty-four canine plasma samples underwent an identical depletion and digestion process. Fractionation and desalting was done by using the SDP-RPS 3 fraction method. Human normal and cancer plasma were obtained from a local hospital which was involved in the CCRC project. No live patients were directly involved in this study. Each sample was identically processed as the canine samples, with a resulting fraction of 3 using the SDP-RPS method.

Mass spectrometry and peptide analysis

Proteomics analysis was done as previously reported by our group (Kim et al., 2016). Each fraction was identified by a Q-Exactive Orbitrap mass spectrometer located at the Korea Brain Research Institute (KBRI). Additional samples used for validation was identified using the Orbitrap fusion mass spectrometer located at the Ulsan National Institute of Science and Technology (UNIST). Raw data was collected for initial peptide research. Protein identification was done with the Maxquant protein search engine (https://www.maxquant.org/). Major search options were assembled with 6 minimal peptides, 1 unique and razor peptide. Additional modifications included methyl oxidation and N-term acetylation. The data was analyzed by the Perseus protein analysis tool attached to the Maxquant software. Differential analysis was done by sorting cancer proteins expressed more or less than 1.2 fold compared to normal samples. P-value cutoff was set to 0.05.

MRM measurements

Multiple reaction monitoring (MRM) analysis was done as previously reported by our group (Kim and Cho, 2019). Briefly, identical plasma samples used in the initial protein search was picked for MRM validation. To each desalted peptide product 20ul of 0.1% TFA in HPLC water was added. Each sample was subjected to a 60 min length liquid chromatogram (LC). Peptide intensity was identified by the triple quad LC/MS 6490. Identified peptides were compared and validated using the skyline software (https://skyline.ms/project/home/begin.view?).

Kaplan-Meier (KM) plot analysis

Survival analysis was done using a web-based Kaplan-Meier (KM) plotter to evaluate the candidate gene and survival rates in more than 3,000 breast cancer samples. Grades and subtypes were sorted by the options provided within the KM plotter tool. The hazard ratio (HR) was given with 95% confidence intervals, and log rank P value was calculated by auto selecting the best cutoff option.

Western blot

Western blot analysis was done as previously reported in our laboratory (Cho et al., 2017). Depleted proteins were dried by the speed vacuum centrifuge method and prepared in HPLC grade water. SDS-PAGE was performed using a 10% polyacrylamide gel. LCAT (Abcam, Cambridge, UK) antibodies were used at a 1:1000 dilution.

RESULTS

To maintain MS quality among samples, a basic three step process was followed for every canine normal and cancer plasma sample: protein depletion, digestion, and fractionation. A total of 12 normal and cancer samples were used for basic profiling and primary targeting. Plasma protein is not easily acquired in normal protein preprocessing, so an extensive 36 fractions with 6-plex TMT labels were used for both sample types (Fig. 3-1). Each fraction contained proteins located in various timelines, which allowed for a higher yield for comparative analysis. Among the samples, cancer subtypes that were diagnosed as aggressive or highly metastatic were handpicked. Histologically, canine mixed tumors are characterized by the presence of myoepithelial cancer cells habited with bone/cartilage mesenchymal cells (Dantas Cassali et al., 2012), which can be categorized as highly developed metastatic cancer. After analysis of LC-MS/MS results, a total of 292 proteins were identified, with 54 proteins elevated in cancer compared to normal plasma (Fig. 3-2). Elevated proteins included SERPING1 and SERPINA6, which are known to be increased in canine MGC and are currently recognized markers in human BC patients (de Ronde et al., 2013).

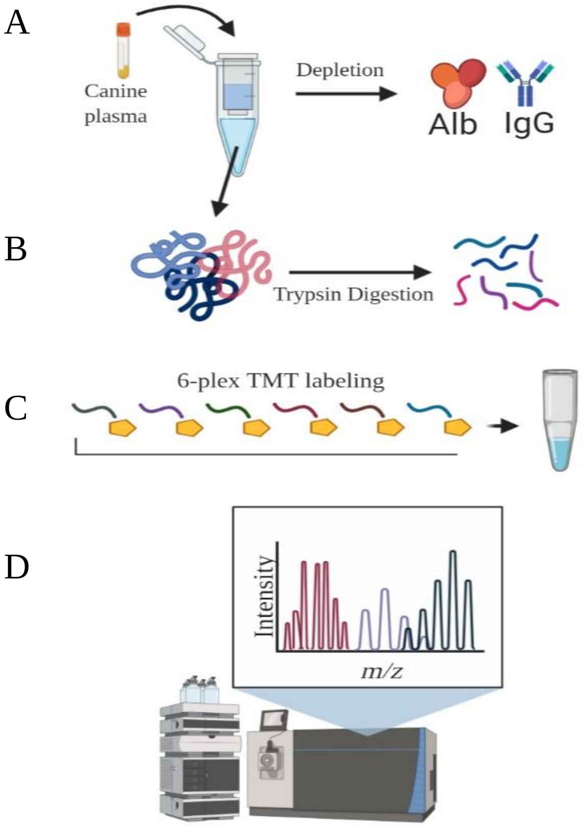


Figure 3-1. Proteome profiling of canine normal & cancer plasma samples.

Basic schematics of the proteomic procedure. (A) Each sample is depleted of plasma abundant proteins Alb and IgG. (B) After depletion, samples were treated with trypsin enzymes for protein digestion. (C) Peptides were then labeled with 6-plex tandem mass tag (TMT) systems for quantification. (D) Finally, peptides were analyzed by MS for protein identification and relative expression analysis.

Among the increased proteins, we focused on LCAT, or lecithin-cholesterol acyltransferase, for several reasons. First, LCAT was the fourth highest elevated protein with a significant p-value of 0.03 (Table 3-1). The other proteins that had higher fold change did not suffice the statistical significance, since LC-MS/MS intensities can vary due to the sample's natural characteristics such as age, dog breed etc. Second, LCAT is a highly abundant protein that converts free cholesterol into a more hydrophobic form, which eventually synthesizes into high density lipoproteins (HDL) that gain mobility to move unidirectionally (Dobiasova and Frohlich, 1999). Highly abundant proteins are much more viable biomarker candidates since they can be detected with ease. Third and most interestingly, human LCAT is well known to be decreased in overall BC tissues (Subbaiah et al., 1997). Because the plasma samples analyzed in our study were highly developed or metastatic carcinomas, this reason alone raised a possibility that LCAT expression patterns can be altered when mammary tumors become more invasive and aggressive. To further address LCAT as a protein highly expressed in mixed tumors, additional MRM analysis was done as a validation. The result indicated that LCAT levels in mixed tumor cancer were more than ten times higher than in normal plasma (Fig. 3-2). While LCAT was easily detected in mixed tumors, only a small portion was identified in normal samples. Our data showed that the LCAT protein is elevated in the plasma of mixed type MGC.

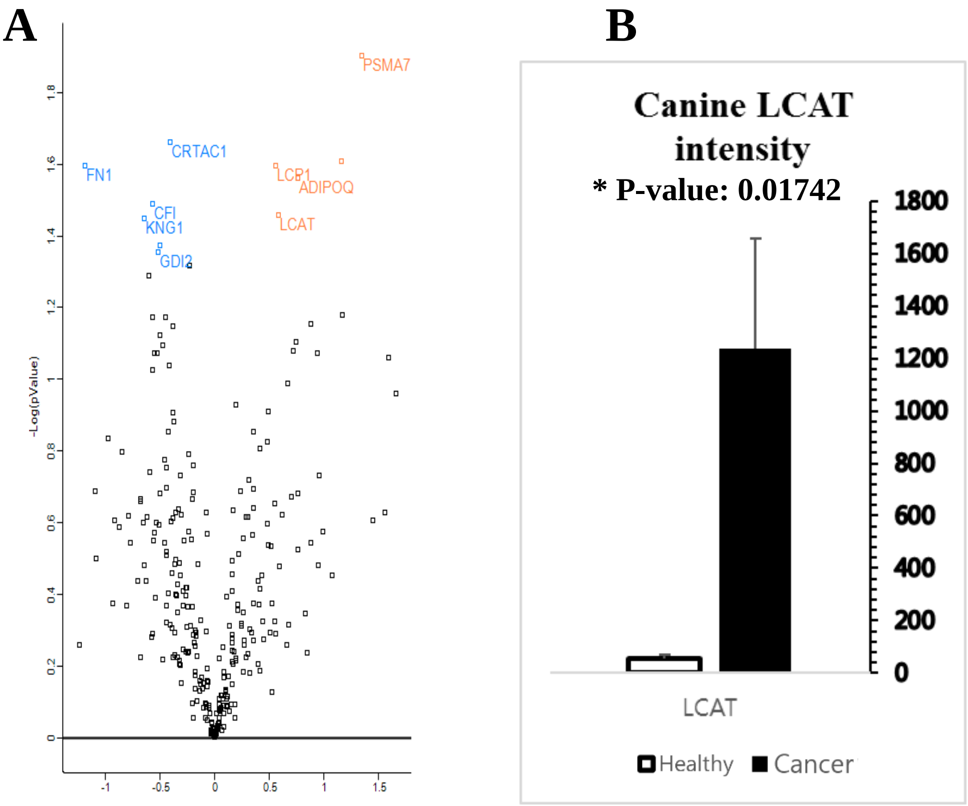


Figure 3-2. Proteome expression analysis

(A) Proteins profiled in each plasma group. Volcano plot indicates top 5 up-regulated genes (Red) and down-regulated genes (Blue) in cancer state. (B) Protein LCAT intensity identified by MRM.

Table 3-1. Differential analysis of protein expression in canine plasma

Gene Name	Fold change	Enriched sample	P-value
PSMA7;PSMA8	2.545446	Cancer	0.012536934
Ig heavy chain V region MOO	2.239375	Cancer	0.024700876
ADIPOQ	1.701391	Cancer	0.025296237
LCAT	1.49674	Cancer	0.027445789
LCP1;PLS3	1.477428	Cancer	0.034921527

Canine MGC can be classified into various types when categorized by histological diagnosis. We further focused on whether LCAT expression is elevated only in mixed tumors or in other cancer subtypes as well. A total of 23 canine plasma samples consisting of normal and distinct cancer subtypes were processed via the four basic steps explained in Fig. 3-1. Due to excessive labor, three fractions were performed for each sample. Every sample contained 110~230 proteins with visible LCAT expressions. When concisely compared between normal and cancer specimens, the normal LCAT level was observed to be slightly higher (Fig. 3-3A), which correlates with recent human studies (Subbaiah et al., 1997). However, when cancer samples were classified into cancer subtypes, mixed tumor samples presented the highest intensity compared to simple or complex tumor samples (Fig. 3-3B). Simple tumor is comprised of tubular or papillary adenocarcinomas, usually consisting of individual cancer cells derived from their respective tissue origin. Complex tumors are microscopically diagnosed by a formation of epithelial and myoepithelial cells. Mixed type tumors are histologically and microscopically more advanced and metastatic compared to the other tumor types. Our data further emphasizes that LCAT expression is elevated in a highly developed cancer environment. To further prove the expressive traits of LCAT, Western blotting was performed in each group. Undergoing blind selection of normal as well as each cancer subtype, we confirmed that mixed type tumors tended to have higher levels of LCAT compared to normal and other distinct cancer subtypes (Fig. 3-4). Our results showed that the LCAT protein is elevated in the plasma of highly developed, invasive and metastatic mammary tumors such as mixed tumors.

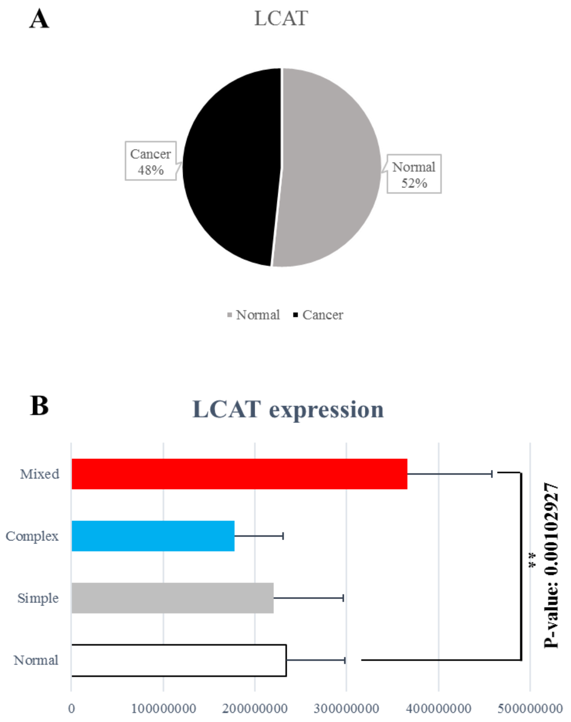


Figure 3-3. LCAT expression of canine plasma samples.

(A) overall expression comparison of LCAT in 25 normal and cancer plasma. (B) LCAT expression difference among normal and cancer subtypes.

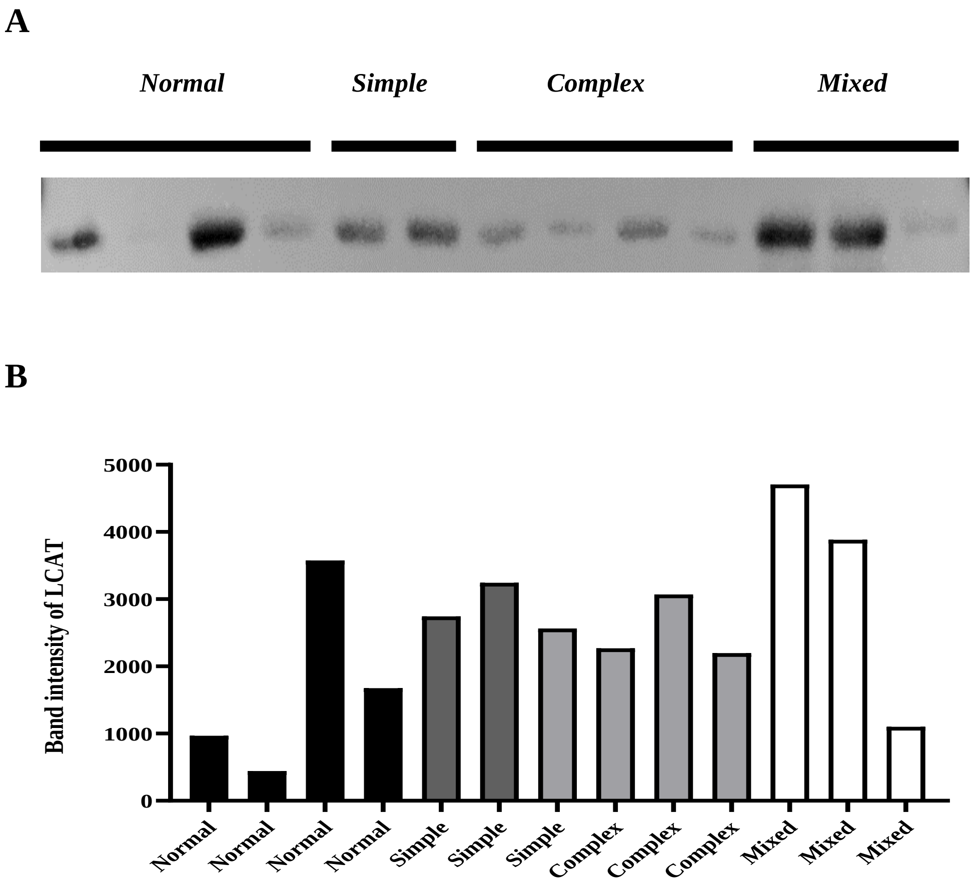


Figure 3-4. Wester blotting of LCAT expression in canine plasma

(A) Cropped gel indicating LCAT expression of selected canine samples from normal and each cancer subtype. (B) Numeric intensity of LCAT expression derived from western blot. Samples range from Normal (Black) to Simple (Dark grey), Complex (light grey), and Mixed type tumors (white).

Human LCAT activity is reported to be constrained in BC. However, detailed expression profiles of distinct cancer subtypes have not yet been addressed. We analyzed LCAT expression *in silico* data of human BC tissues provided by the TCGA database. Because this data lacks specific stage information, human BC was separated into grades, from lowest normal-like grade 1 to highly aggressive and invasive grade 3 (Rakha et al., 2018) (Fig. 3-5A). Survival analysis of each grade presented dissimilar outcomes. Expression levels of LCAT in grades 1 and 2 did not seem to influence the mortality. On the other hand, high expression of LCAT decreased the survival rate of patients with grade 3 BC. This provided evidence that LCAT might have a role in aggressive types of BC. We further sorted grade 3 patients into lymph node positive and negative types. Lymph nodes can be viable metastasis indicators, as they are one of the most common organs involved in aggressive BC metastasis (Rahman and Mohammed, 2015). Interestingly, patients with lymph node metastasis indicated a drastic decline in survival rates compared to non-metastatic BC (Fig 3-5B). Therefore, it was clear that LCAT has a negative correlation with the survival rate of patients with BC that has substantially developed and undergone metastasis.

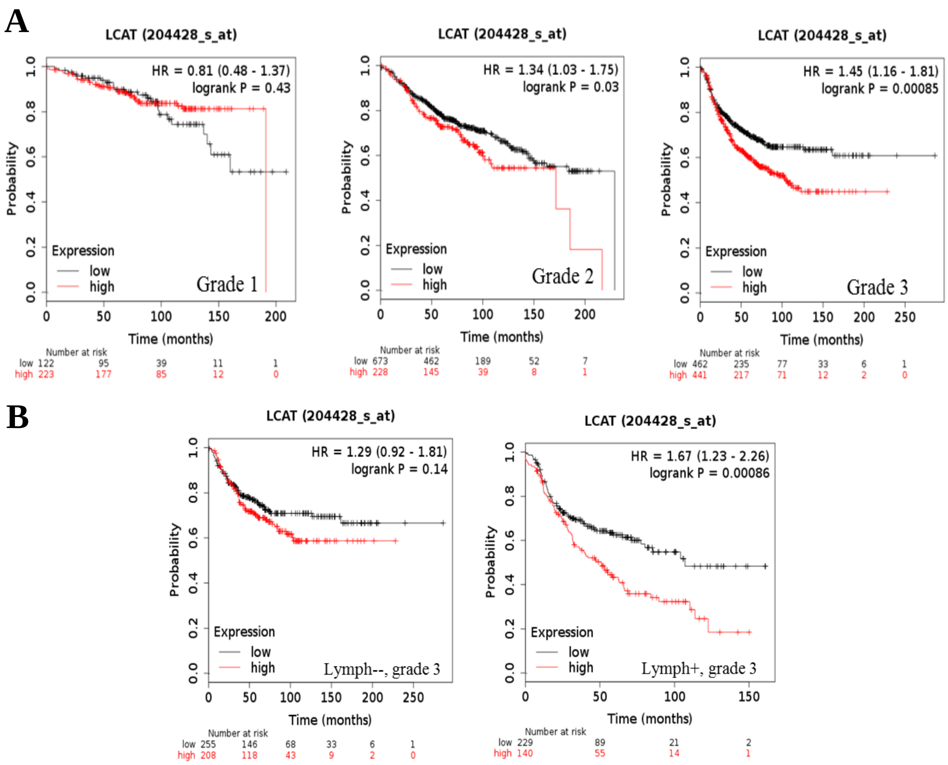


Figure 3-5. Survival analysis of LCAT expression.

Kaplan-Meier (KM) plots representing patient mortality of increased and decreased LCAT expression**. (A)** survival rate of BC patients with grades starting from 1 to 3. **(B)** LCAT expression survival rate of BC patients with or without lymph node metastasis

Survival analysis of BC patients' data alone does not fully explain whether the change of mortality is caused by differential expression or functional alteration. To validate human LCAT expression among BC plasma, various subtypes and stages suitable for stage-wise analysis were selected for Western blotting from 24 samples. Surprisingly, among luminal A plasma, LCAT expression was correlated with an increased stage of cancer development (Fig. 3-6A). Stages of BC consisted from the *in situ* stage 0 to the very invasive and developed stage 3. Stage 0 was mainly intact within the ductal parts of the breast, which expressed similar LCAT patterns to normal patients. However, as the cancer's characteristics became more aggressive with continuous development, LCAT expression increased. The second stage of BC could be further separated by the possibility of lymph node metastasis. Stage 2B was considered to be more invasive into the lymph nodes than 2A, which was reflected in the further elevated expression of LCAT (Fig. 3-6B). This data, although tested in a limited number of samples and needing to be studied with more samples, indicated again that the LCAT plasma level correlates with the aggressiveness of BC.

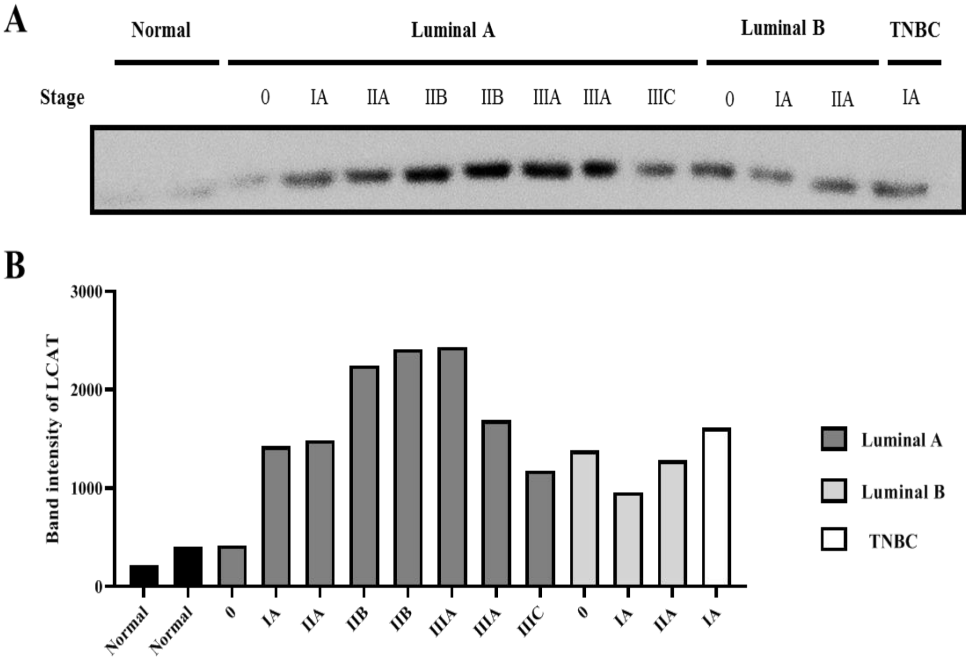


Figure 3-6. Western blot of LCAT expression among human BC plasma.

(A) Cropped gels of Normal, Luminal A, Luminal B, and TNBC subtype BC patients with different stages indicating LCAT expression. Full-length gels are presented in Supplementary Fig. 2. Stages range from the least cancer-like 0 stage to the most aggressive and metastatic IIIC. (B) Intensity of each sample listed in Fig. 3-6A

DISCUSSION

This study was able to identify a potentially strong protein biomarker as an aggressive BC indicator through a comparative approach via canine MGC plasma. Not only are dogs environmentally related to humans, but they are also genetically highly conserved compared to other universal experimental animals such as mice or rats (Lindblad-Toh et al., 2005b). This aspect gives access to human BC from a different perspective. Previous studies that implemented canine MGC as a suitable model for human BC research has been widely used to understand certain BC subtypes that seldom occur but have high mortality rates, such as TNBC and other myoepithelial carcinomas (Abdelmegeed and Mohammed, 2018b) Our group also reported resemblances of canine MGC to human BC in terms of transcriptomic analyses (Lee et al., 2018). Though canine MGC do not have criteria related to cancer development stages, tumor subtypes are diagnosed by grades to indicate severity. Complex type MGC are mainly graded 1~2, which indicates relatively stable cancer cells with a good prognosis. Mixed type tumors are far more metastatic due to the multiple cancer cells related to breast and bones (Tavasoly et al., 2013). By using developed, metastatic canine mixed tumor plasma, we were able to discover a novel biomarker that could be used as a precursor of aggressiveness and metastasis of both human BC and canine MGC.

MGC cells secrete certain proteins which can be used to identify their nature. Furthermore, compared to the direct tissue approach, plasma proteins are considered a much more viable biomarker able to be implemented in both basic and clinical research (Surinova et al., 2011). Despite the provided advantages, the majority of reported plasma protein biomarkers are not easily accessed due to their low abundance and co-habitation with abundant proteins such as albumin and various immunoglobulins. Initially through extensive plasma sampling using 6-plex TMT labeling and a wide selection of fractions, the protein, LCAT, was identified in all normal and cancer samples. LCAT is a well-known enzyme that participates in transporting cholesterol (Kosek et al., 1999). To validate the protein's practical role as a biomarker compared to other low abundant proteins, an additional 24 plasma samples of canine normal and various MGC was inspected with a very compact and reproducible method. As a result, we were able to identify LCAT as a selective biomarker highly elevated in MGC that have undergone a series of developments and metastases.

Before investigating LCAT in human BC, we were well aware of the previous studies that in general mentioned that LCAT activity is decreased in BC patients. The main LCAT products, lipoproteins LDL and HDL, are considered as potential new causes in BC development (Cedo et al., 2019). Decreased LCAT activity leading to low level high-density lipoproteins (HDL) was reported in BC patients before radiotherapy (Ozmen and Askin, 2013). However, more recent reports described LCAT and HDL levels rising in breast cancer subjects. A genome-wide study using 164 discrete variants associated with HDL, LDL and cholesterol among 101,424 BC cases and 80,253 controls provided strong evidence that increased HDL may be related to BC occurrence (Beeghly-Fadiel et al., 2019). The controversial debate of lipoproteins in BC may be due to the lack of understanding of how LCAT activity influences high- and low-density lipoproteins. The dispute also infers the possibility that LCAT expression is not correlated to BC as a whole, but rather only to certain types. Through *in silico* research and Western blot validation, we were able to demonstrate that human LCAT resembles canine LCAT expression level patterns, as it was increased in highly progressed breast cancers within the same classified subtypes. These results not only indicate a discovery of a novel protein biomarker in breast cancer, but could also provide further understanding of the lipoprotein pathway that is involved in aggressive breast cancer development.

In conclusion, this study reveals the plasma protein LCAT as a biomarker for indicating advanced breast cancer as well as mammary tumor undergoing metastasis using a comparative analysis approach from canine to human cases. Further extending the comparative analysis using more than 150 samples of canine and human plasma revealed proteins which altered expression when MGC is developed in both canine and humans. Among the commonly regulated proteins, protein LCP1 was identified to have a significant increase in human cohorts with MGC dogs as companions. The identified biomarkers will provide further evidence in diagnosing clinical samples of dogs and humans. LCP1 will not only serve as a biomarker for MGC diagnosis, but also might serve as a guiding protein which can warn humans by investigating their canine partners.

GENERAL DISCUSSION

Many researches regarding canine MGC as an appropriate model for researching human BC has been mainly demonstrated by antibody based immunohistochemical staining or targeted proteins directly expressed from the malignant tissue. To further address the advantages of canine MGC as appropriate BC comparative models, I used transcriptomic data to compare canine MGC and its adjacent normal tissue for identifying transcriptomes specifically expressing in overall MGCs and each respective subtypes simple, ductal and complex MGCs. As a result, genes differentially expressed in all subtypes such as *CCL23*, *CXCL10*, *SFRP2*, and *FRZB* have been known to be altered in human BC. Not just only individual gene expression shared common ground. Once performing GO term analysis within the DEGs, term "positive regulation of angiogenesis" consisting of genes *CHI3L1*, *CXCL8*, *FOXC2*, *SERPINE1*, *SFRP2*, and *TF* have also been reported to play roles in various malignancies including BC. Furthermore, by implementing public RNA-seq data of human BC subtypes and comparing its transcriptomic expression profiles with canine MGC histological subtypes, I discovered that while ER+ and ER+&HER2+ subtypes showed no correlation with 'complex and simple' and ductal subtypes, TNBC had a strong correlation in both simple and ductal subtypes.

Other reports describing invasive ductal carcinoma showing histological resemblance with TNBC and similar transcriptomic profiles from simple subtypes such as *KRT5* and *MKI67* present a possibility of which the transcriptomic signatures for canine MGC might indicate certain human BC subtypes. This further leads to finding biomarkers which were previously not considered as subtype specific target.

Even though BC research through *in silico* data have been gradually beneficial, a large portion of the targets did not correlate with real experiments. Meanwhile, various methods of omics technology were implemented to analyze the genetic and epigenetic characteristics of human BC. As more data was processed, correlating the genetic and epigenetic omics datasets made possible to yield outcomes which tend to be more accurate than those discovered with only one type of expression data. By combining the epigenetic histone modification profiles of H3K4me3 and H3K27me3, which tends to exist within the same genomic region, sorted out cancer specifically enriched genes that highly matched with the transcriptomic expression. The results not only represented a possible marker for TNBC, but also provide information on how the gene is regulated in an epigenetic matter. Further validating the top ten up- and down-regulating candidates in TNBC cell lines and other breast cancer cell lines resulted in a high correlation. Among the candidates, genes such as *NAD8L*, *MMP16*, *CDH2* and *SOX5* have been studied regarding their BC-related molecular functions, with *MMP16* reporting a specific role in TNBC. While some targets were matched with previous researches, I have found novel biomarkers which expressions are specific to TNBC. *NOVA1*, *PLCL2*, *SYTL4* and *DNER* did not show any information considering human BC. The discovered markers may prove as a viable target for understanding TNBC.

Proteomic expression in cancer also serves as a strong indicator of the patients' anomaly. As I addressed canine MGC models as an appropriate comparative medicinal approach, I further investigated on comparing canine and human malignancies in a proteomic level. During the process, I deliberately selected canine MGC subtypes which were histologically diagnosed as late-stage, or highly metastatic, to see if MGC of a more metastatic and aggressive stage would alter the proteome expression compared to not only normal, but also other MGC subtypes. By processing proteomic data using canine plasma samples, I reported plasma protein LCAT as a biomarker to highly advanced stage, metastatic mixed MGC subtype.

GENERAL CONCLUSION

The continuous efforts to fully overcome BC has led to various methods of analysis branching from producing sequencing data which supervises the overall gene and protein expression of the human construct to implementing comparative models of relatively close species for discovering aspects which were neglected or overlooked in the human approach. Recent advances in sequencing technology made possible to manufacture various omics data regarding the genomic, epigenetic and proteomic region such as transcriptomic expression, histone modification profiles and protein expression. Along with the datasets, personalized medicine has been gaining importance, with predictive biomarkers indicating disease progression and target approaches for therapy and monitoring.

In these studies, I applied various datasets of different omics technology to discover potential biomarkers which express distinct profiles from normal states in both canine MGC and human BC. The analysis resulted in the identification of DEGs and DEPs of each representative species and led to commonly expressed targets both previously reported and novel in breast malignancies. The verified biomarkers will suit as a potential target for not just overall MGCs, but also in certain advanced stage specific cancer of high metastatic features.

Considering the series of omics data comparison of two genetically and environmentally close species, this research might provide further insight to establish an appropriate understanding of comparative medicinal BC biomarker development which encompass canines and humans alike.

REFERENCES

1. Jung, K. W., Y. J. Won, H. J. Kong, E. S. Lee, and Registries Community of Population-Based Regional Cancer. "Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2015." *Cancer Res Treat* 50, no. 2 (2018): 303-16.

2. Jung, K. W., Y. J. Won, H. J. Kong, and E. S. Lee. "Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2016." *Cancer Res Treat* 51, no. 2 (2019): 417-30.

3. Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. "Global Cancer Statistics 2018: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA Cancer J Clin* 68, no. 6 (2018): 394-424.

4. Siegel, E., J. Tseng, A. Giuliano, F. Amersi, and R. F. Alban. "Treatment at Academic Centers Increases Likelihood of Reconstruction after Mastectomy for Breast Cancer Patients." *J Surg Res* 247 (2020): 156-62.

5. Bradley, O. C. "What Is Comparative Medicine?" *Proc R Soc Med* 21, no. 1 (1927): 129-34.

6. Paoloni, M. C., and C. Khanna. "Comparative Oncology Today." *Vet Clin North Am Small Anim Pract* 37, no. 6 (2007): 1023-32; v.

7. Gondo, Y., R. Fukumura, T. Murata, and S. Makino. "Next-Generation Gene Targeting in the Mouse for Functional Genomics." *BMB Rep* 42, no. 6 (2009): 315-23.

8. Seok, J., H. S. Warren, A. G. Cuenca, M. N. Mindrinos, H. V. Baker, W. Xu, D. R. Richards, G. P. McDonald-Smith, H. Gao, L. Hennessy, C. C. Finnerty, C. M. Lopez, S. Honari, E. E. Moore, J. P. Minei, J. Cuschieri, P. E. Bankey, J. L. Johnson, J. Sperry, A. B. Nathens, T. R. Billiar, M. A. West, M. G. Jeschke, M. B. Klein, R. L. Gamelli, N. S. Gibran, B. H. Brownstein, C. Miller-Graziano, S. E. Calvano, P. H. Mason, J. P. Cobb, L. G. Rahme, S.

F. Lowry, R. V. Maier, L. L. Moldawer, D. N. Herndon, R. W. Davis, W. Xiao, R. G. Tompkins, Inflammation, and Large Scale Collaborative Research Program Host Response to Injury. "Genomic Responses in Mouse Models Poorly Mimic Human Inflammatory Diseases." *Proc Natl Acad Sci U S A* 110, no. 9 (2013): 3507-12.

9. Sultan, F., and B. A. Ganaie. "Comparative Oncology: Integrating Human and Veterinary Medicine." *Open Vet J* 8, no. 1 (2018): 25-34.

10. Gardner, H. L., J. M. Fenger, and C. A. London. "Dogs as a Model for Cancer." *Annu Rev Anim Biosci* 4 (2016): 199-222.

11. Owen, L. N. "A Comparative Study of Canine and Human Breast Cancer." *Invest Cell Pathol* 2, no. 4 (1979): 257-75.

12. Sleeckx, N., H. de Rooster, E. J. Veldhuis Kroeze, C. Van Ginneken, and L. Van Brantegem. "Canine Mammary Tumours, an Overview." *Reprod Domest Anim* 46, no. 6 (2011): 1112-31.

13. Ahern, T. E., R. C. Bird, A. E. Bird, and L. G. Wolfe. "Expression of the Oncogene C-Erbb-2 in Canine Mammary Cancers and Tumor-Derived Cell Lines." *Am J Vet Res* 57, no. 5 (1996): 693-6.

14. Misdorp, W. "[Tumors in Humans and Domestic Animals]." *Jaarb Kankeronderz Kankerbestrijd Ned* 14 (1964): 261-74.

15. Kabir, F. M. L., P. DeInnocentes, P. Agarwal, C. P. Mill, D. J. Riese Nd, and R. C. Bird. "Estrogen Receptor-Alpha, Progesterone Receptor, and C-Erbb/Her-Family Receptor Mrna Detection and Phenotype Analysis in Spontaneous Canine Models of Breast Cancer." *J Vet Sci* 18, no. 2 (2017): 149-58.

16. Sassi, F., C. Benazzi, G. Castellani, and G. Sarli. "Molecular-Based Tumour Subtypes of Canine Mammary Carcinomas Assessed by Immunohistochemistry." *BMC Vet Res* 6 (2010): 5.

17. Lutful Kabir, F. M., C. E. Alvarez, and R. C. Bird. "Canine Mammary Carcinomas: A Comparative Analysis of Altered Gene Expression." *Vet Sci* 3, no. 1 (2015).

18. Vail, D. M., and E. G. MacEwen. "Spontaneously Occurring Tumors of Companion Animals as Models for Human Cancer." *Cancer Invest* 18, no. 8 (2000): 781-92.

19. Uva, P., L. Aurisicchio, J. Watters, A. Loboda, A. Kulkarni, J. Castle, F.

Palombo, V. Viti, G. Mesiti, V. Zappulli, L. Marconato, F. Abramo, G. Ciliberto, A. Lahm, N. La Monica, and E. de Rinaldis. "Comparative Expression Pathway Analysis of Human and Canine Mammary Tumors." *BMC Genomics* 10 (2009): 135.

20. Klopfleisch, R., and A. D. Gruber. "Differential Expression of Cell Cycle Regulators P21, P27 and P53 in Metastasizing Canine Mammary Adenocarcinomas Versus Normal Mammary Glands." *Res Vet Sci* 87, no. 1 (2009): 91-6.

21. Lutful Kabir, F. M., P. Agarwal, P. Deinnocentes, J. Zaman, A. C. Bird, and R. C. Bird. "Novel Frameshift Mutation in the P16/Ink4a Tumor Suppressor Gene in Canine Breast Cancer Alters Expression from the P16/Ink4a/P14arf Locus." *J Cell Biochem* 114, no. 1 (2013): 56-66.

22. Dalton, W. S., and S. H. Friend. "Cancer Biomarkers--an Invitation to the Table." *Science* 312, no. 5777 (2006): 1165-8.

23. Strimbu, K., and J. A. Tavel. "What Are Biomarkers?" *Curr Opin HIV AIDS* 5, no. 6 (2010): 463-6.

24. Legrain, P., R. Aebersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C. Borchers, G. L. Corthals, C. E. Costello, E. W. Deutsch, B. Domon, W. Hancock, F. He, D. Hochstrasser, G. Marko-Varga, G. H. Salekdeh, S. Sechi, M. Snyder, S. Srivastava, M. Uhlen, C. H. Hu, T. Yamamoto, Y. K. Paik, and G. S. Omenn. "The Human Proteome Project: Current State and Future Direction." *Mol Cell Proteomics* (2011).

25. Wishart, D. S., T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. "Hmdb 3.0--the Human Metabolome Database in 2013." *Nucleic Acids Res* 41, no. Database issue (2013): D801-7.

26. Jensen, O. N. "Modification-Specific Proteomics: Characterization of Post-Translational Modifications by Mass Spectrometry." *Curr Opin Chem Biol* 8, no. 1 (2004): 33-41.

27. Drabovich, A. P., E. Martinez-Morillo, and E. P. Diamandis. "Toward an Integrated Pipeline for Protein Biomarker Development." *Biochim Biophys Acta* 1854, no. 6 (2015): 677-86.

28. Padmanabhan, R., E. Jay, and R. Wu. "Chemical Synthesis of a Primer and

Its Use in the Sequence Analysis of the Lysozyme Gene of Bacteriophage T4." *Proc Natl Acad Sci U S A* 71, no. 6 (1974): 2510-4.

29. Behjati, S., and P. S. Tarpey. "What Is Next Generation Sequencing?" *Arch Dis Child Educ Pract Ed* 98, no. 6 (2013): 236-8.

30. Chu, Y., and D. R. Corey. "Rna Sequencing: Platform Selection, Experimental Design, and Data Interpretation." *Nucleic Acid Ther* 22, no. 4 (2012): 271-4.

31. Maher, C. A., C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan. "Transcriptome Sequencing to Detect Gene Fusions in Cancer." *Nature* 458, no. 7234 (2009): 97-101.

32. Ingolia, N. T., G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman. "The Ribosome Profiling Strategy for Monitoring Translation in Vivo by Deep Sequencing of Ribosome-Protected Mrna Fragments." *Nat Protoc* 7, no. 8 (2012): 1534-50.

33. Steen, H., and M. Mann. "The Abc's (and Xyz's) of Peptide Sequencing." *Nat Rev Mol Cell Biol* 5, no. 9 (2004): 699-711.

34. Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, and M. Mann. "Femtomole Sequencing of Proteins from Polyacrylamide Gels by Nano-Electrospray Mass Spectrometry." *Nature* 379, no. 6564 (1996): 466-9.

35. "Who Fact Sheet." http://www.who.int/news-room/fact-sheets/detail/cancer (

36. Sgroi, Dennis C. "Preinvasive Breast Cancer." *Annual Review of Pathology: Mechanisms of Disease* 5 (2010): 193-221.

37. Onitilo, Adedayo A, Jessica M Engel, Robert T Greenlee, and Bickol N Mukesh. "Breast Cancer Subtypes Based on Er/Pr and Her2 Expression: Comparison of Clinicopathologic Features and Survival." *Clinical medicine & research* 7, no. 1-2 (2009): 4-13.

38. Koczkowska, Magdalena, Monika Zuk, Adam Gorczynski, Magdalena Ratajska, Marzena Lewandowska, Wojciech Biernat, Janusz Limon, and Bartosz Wasag. "Detection of Somatic Brca 1/2 Mutations in Ovarian Cancer–Next-Generation Sequencing Analysis of 100 Cases." *Cancer medicine* 5, no. 7 (2016): 1640-46.

39. Li, Guoli, Xinwu Guo, Lili Tang, Ming Chen, Xipeng Luo, Limin Peng, Xunxun Xu, Shouman Wang, Zhi Xiao, and Wenjun Yi. "Analysis of Brca1/2 Mutation Spectrum and Prevalence in Unselected Chinese Breast Cancer Patients by Next-Generation Sequencing." *Journal of cancer research and clinical oncology* 143, no. 10 (2017): 2011-24.

40. Ratajska, Magdalena, Magdalena Krygier, Maciej Stukan, Alina Kuźniacka, Magdalena Koczkowska, Mirosław Dudziak, Marcin Śniadecki, Jarosław Dębniak, Dariusz Wydra, and Izabela Brozek. "Mutational Analysis of Brca1/2 in a Group of 134 Consecutive Ovarian Cancer Patients. Novel and Recurrent Brca1/2 Alterations Detected by Next Generation Sequencing." *Journal of Applied Genetics* 56, no. 2 (2015): 193-98.

41. Salas, Y., A. Marquez, D. Diaz, and L. Romero. "Epidemiological Study of Mammary Tumors in Female Dogs Diagnosed During the Period 2002-2012: A Growing Animal Health Problem." *PloS one* 10, no. 5 (2015): e0127381.

42. Gurda, Brittney L, Allison M Bradbury, and Charles H Vite. "Focus: Comparative Medicine: Canine and Feline Models of Human Genetic Diseases and Their Contributions to Advancing Clinical Therapies." *The Yale journal of biology and medicine* 90, no. 3 (2017): 417.

43. Liu, Deli, Huan Xiong, Angela E Ellis, Nicole C Northrup, Carlos O Rodriguez, Ruth M O'Regan, Stephen Dalton, and Shaying Zhao. "Molecular Homology and Difference between Spontaneous Canine Mammary Cancer and Human Breast Cancer." *Cancer research* 74, no. 18 (2014): 5045-56.

44. Romagnolo, Donato F, Kevin D Daniels, Jonathan T Grunwald, Stephan A Ramos, Catherine R Propper, and Ornella I Selmin. "Epigenetics of Breast Cancer: Modifying Role of Environmental and Bioactive Food Compounds." *Molecular nutrition & food research* 60, no. 6 (2016): 1310-29.

45. Im, KeumSoon, NH Kim, HY Lim, HW Kim, JI Shin, and JH Sur. "Analysis of a New Histological and Molecular-Based Classification of Canine Mammary Neoplasia." *Veterinary Pathology* 51, no. 3 (2014): 549-59.

46. Klopfleisch, R, D Lenze, M Hummel, and AD Gruber. "The Metastatic Cascade Is Reflected in the Transcriptome of Metastatic Canine Mammary Carcinomas." *The Veterinary Journal* 190, no. 2 (2011): 236-43.

47. Król, M, J Skierski, NA Rao, E Hellmen, JA Mol, and T Motyl. "Transcriptomic Profile of Two Canine Mammary Cancer Cell Lines with Different Proliferative and Anti-Apoptotic Potential." *Journal of Physiology and Pharmacology* 60, no. Suppl. 1 (2009): 95-106.

48. Lindblad-Toh, Kerstin, Claire M Wade, Tarjei S Mikkelsen, Elinor K Karlsson, David B Jaffe, Michael Kamal, Michele Clamp, Jean L Chang, Edward J Kulbokas, and Michael C Zody. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." *Nature* 438, no. 7069 (2005a): 803-19.

49. Campos, LC, GE Lavalle, A Estrela-Lima, JC Melgaco de Faria, JE Guimarães, AP Dutra, E Ferreira, LP de Sousa, É ML Rabelo, and AFD Vieira da Costa. "Ca 15.3, Cea and Ldh in Dogs with Malignant Mammary Tumors." *Journal of veterinary internal medicine* 26, no. 6 (2012): 1383-88.

50. Vinothini, G, C Balachandran, and S Nagini. "Evaluation of Molecular Markers in Canine Mammary Tumors: Correlation with Histological Grading." *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* 18, no. 5-6 (2009): 193-201.

51. Kamps, Rick, Rita D Brandão, Bianca J Bosch, Aimee DC Paulussen, Sofia Xanthoulea, Marinus J Blok, and Andrea Romano. "Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification." *International journal of molecular sciences* 18, no. 2 (2017): 308.

52. Khotskaya, Yekaterina B, Gordon B Mills, and Kenna R Mills Shaw. "Next-Generation Sequencing and Result Interpretation in Clinical Oncology: Challenges of Personalized Cancer Therapy." *Annual review of medicine* 68 (2017): 113-25.

53. Guo, X, H Xiao, S Guo, L Dong, and J Chen. "Identification of Breast Cancer Mechanism Based on Weighted Gene Coexpression Network Analysis." *Cancer gene therapy* 24, no. 8 (2017): 333-41.

54. Li, J., Y. Wang, Q. G. Li, J. J. Xue, Z. Wang, X. Yuan, J. D. Tong, and L. C. Xu. "Downregulation of Fbp1 Promotes Tumor Metastasis and Indicates Poor Prognosis in Gastric Cancer Via Regulating Epithelial-Mesenchymal Transition." *PloS one* 11, no. 12 (2016a): e0167857.

55. Ferrero, Giulio, Francesca Cordero, Sonia Tarallo, Maddalena Arigoni, Federica Riccardo, Gaetano Gallo, Guglielmo Ronco, Marco Allasia, Neha

Kulkarni, and Giuseppe Matullo. "Small Non-Coding Rna Profiling in Human Biofluids and Surrogate Tissues from Healthy Individuals: Description of the Diverse and Most Represented Species." *Oncotarget* 9, no. 3 (2018): 3097.

56. Liang, Chaojie, Zining Qi, Hua Ge, Chaowei Liang, Yu Zhang, Zhimin Wang, Ruihuan Li, and Jiansheng Guo. "Long Non-Coding Rna Pcat-1 in Human Cancers: A Meta-Analysis." *Clinica Chimica Acta* 480 (2018): 47-55.

57. Schwarzer, Adrian, Stephan Emmrich, Franziska Schmidt, Dominik Beck, Michelle Ng, Christina Reimer, Felix Ferdinand Adams, Sarah Grasedieck, Damian Witte, and Sebastian Käbler. "The Non-Coding Rna Landscape of Human Hematopoiesis and Leukemia." *Nature communications* 8, no. 1 (2017): 1-17.

58. Haakensen, Vilde D, Vegard Nygaard, Liliana Greger, Miriam R Aure, Bastian Fromm, Ida RK Bukholm, Torben Lüders, Suet-Feung Chin, Anna Git, and Carlos Caldas. "Subtype-Specific Micro-Rna Expression Signatures in Breast Cancer Progression." *International journal of cancer* 139, no. 5 (2016): 1117-28.

59. Huo, Lei, Yan Wang, Yun Gong, Savitri Krishnamurthy, Jing Wang, Lixia Diao, Chang-Gong Liu, Xiuping Liu, Feng Lin, and William F Symmans. "Microrna Expression Profiling Identifies Decreased Expression of Mir-205 in Inflammatory Breast Cancer." *Modern Pathology* 29, no. 4 (2016): 330-46.

60. Xu, Shouping, Dejia Kong, Qianlin Chen, Yanyan Ping, and Da Pang. "Oncogenic Long Noncoding Rna Landscape in Breast Cancer." *Molecular cancer* 16, no. 1 (2017): 129.

61. Preker, Pascal, Kristina Almvig, Marianne S Christensen, Eivind Valen, Christophe K Mapendano, Albin Sandelin, and Torben Heick Jensen. "Promoter Upstream Transcripts Share Characteristics with Mrnas and Are Produced Upstream of All Three Major Types of Mammalian Promoters." *Nucleic acids research* 39, no. 16 (2011): 7179-93.

62. Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, 3rd, M. C. Zody, E. Mauceli, X. Xie, M. Breen, R. K. Wayne, E. A. Ostrander, C. P. Ponting, F. Galibert, D. R. Smith, P. J. DeJong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C. W. Chin, A. Cook, J. Cuff, M. J. Daly, D. DeCaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L.

Goodstadt, A. Heger, C. Hitte, L. Kim, K. P. Koepfli, H. G. Parker, J. P. Pollinger, S. M. Searle, N. B. Sutter, R. Thomas, C. Webber, J. Baldwin, A. Abebe, A. Abouelleil, L. Aftuck, M. Ait-Zahra, T. Aldredge, N. Allen, P. An, S. Anderson, C. Antoine, H. Arachchi, A. Aslam, L. Ayotte, P. Bachantsang, A. Barry, T. Bayul, M. Benamara, A. Berlin, D. Bessette, B. Blitshteyn, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, A. Brown, P. Cahill, N. Calixte, J. Camarata, Y. Cheshatsang, J. Chu, M. Citroen, A. Collymore, P. Cooke, T. Dawoe, R. Daza, K. Decktor, S. DeGray, N. Dhargay, K. Dooley, K. Dooley, P. Dorje, K. Dorjee, L. Dorris, N. Duffey, A. Dupes, O. Egbiremolen, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, P. Ferreira, S. Fisher, M. FitzGerald, K. Foley, C. Foley, A. Franke, D. Friedrich, D. Gage, M. Garber, G. Gearin, G. Giannoukos, T. Goode, A. Goyette, J. Graham, E. Grandbois, K. Gyaltsen, N. Hafez, D. Hagopian, B. Hagos, J. Hall, C. Healy, R. Hegarty, T. Honan, A. Horn, N. Houde, L. Hughes, L. Hunnicutt, M. Husby, B. Jester, C. Jones, A. Kamat, B. Kanga, C. Kells, D. Khazanovich, A. C. Kieu, P. Kisner, M. Kumar, K. Lance, T. Landers, M. Lara, W. Lee, J. P. Leger, N. Lennon, L. Leuper, S. LeVine, J. Liu, X. Liu, Y. Lokyitsang, T. Lokyitsang, A. Lui, J. Macdonald, J. Major, R. Marabella, K. Maru, C. Matthews, S. McDonough, T. Mehta, J. Meldrim, A. Melnikov, L. Meneus, A. Mihalev, T. Mihova, K. Miller, R. Mittelman, V. Mlenga, L. Mulrain, G. Munson, A. Navidi, J. Naylor, T. Nguyen, N. Nguyen, C. Nguyen, T. Nguyen, R. Nicol, N. Norbu, C. Norbu, N. Novod, T. Nyima, P. Olandt, B. O'Neill, K. O'Neill, S. Osman, L. Oyono, C. Patti, D. Perrin, P. Phunkhang, F. Pierre, M. Priest, A. Rachupka, S. Raghuraman, R. Rameau, V. Ray, C. Raymond, F. Rege, C. Rise, J. Rogers, P. Rogov, J. Sahalie, S. Settipalli, T. Sharpe, T. Shea, M. Sheehan, N. Sherpa, J. Shi, D. Shih, J. Sloan, C. Smith, T. Sparrow, J. Stalker, N. Stange-Thomann, S. Stavropoulos, C. Stone, S. Stone, S. Sykes, P. Tchuinga, P. Tenzing, S. Tesfaye, D. Thoulutsang, Y. Thoulutsang, K. Topham, I. Topping, T. Tsamla, H. Vassiliev, V. Venkataraman, A. Vo, T. Wangchuk, T. Wangdi, M. Weiand, J. Wilkinson, A. Wilson, S. Yadav, S. Yang, X. Yang, G. Young, Q. Yu, J. Zainoun, L. Zembek, A. Zimmer, and E. S. Lander. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." *Nature* 438, no. 7069 (2005b): 803-19.

63. Metsalu, T., and J. Vilo. "Clustvis: A Web Tool for Visualizing Clustering of Multivariate Data Using Principal Component Analysis and Heatmap." *Nucleic Acids Res* 43, no. W1 (2015): W566-70.

64. Lotia, S., J. Montojo, Y. Dong, G. D. Bader, and A. R. Pico. "Cytoscape App Store." *Bioinformatics* 29, no. 10 (2013): 1350-1.

65. Chung, Woosung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-

Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, and Yeon Hee Park. "Single-Cell Rna-Seq Enables Comprehensive Tumour and Immune Cell Profiling in Primary Breast Cancer." *Nature communications* 8, no. 1 (2017): 1-12.

66. Jézéquel, Pascal, Delphine Loussouarn, Catherine Guérin-Charbonnel, Loïc Campion, Antoine Vanier, Wilfried Gouraud, Hamza Lasla, Catherine Guette, Isabelle Valo, and Véronique Verrièle. "Gene-Expression Molecular Subtyping of Triple-Negative Breast Cancer Tumours: Importance of Immune Response." *Breast Cancer Research* 17, no. 1 (2015): 43.

67. Niland, Stephan, and Johannes A Eble. "Integrin-Mediated Cell-Matrix Interaction in Physiological and Pathological Blood Vessel Formation." *Journal of oncology* 2012 (2011).

68. Papp, Béla, Jean-Philippe Brouland, Atousa Arbabian, Pascal Gélébart, Tünde Kovács, Régis Bobe, Jocelyne Enouf, Nadine Varin-Blank, and Ágota Apáti. "Endoplasmic Reticulum Calcium Pumps and Cancer Cell Differentiation." *Biomolecules* 2, no. 1 (2012): 165-86.

69. Campos, Miguel, MMJ Kool, Sylvie Daminet, Richard Ducatelle, G Rutteman, HS Kooistra, S Galac, and JA Mol. "Upregulation of the Pi3k/Akt Pathway in the Tumorigenesis of Canine Thyroid Carcinoma." *Journal of veterinary internal medicine* 28, no. 6 (2014): 1814-23.

70. Dobbin, Zachary C, and Charles N Landen. "The Importance of the Pi3k/Akt/Mtor Pathway in the Progression of Ovarian Cancer." *International journal of molecular sciences* 14, no. 4 (2013): 8213-27.

71. Terragni, Rossella, Andrea Casadei Gardini, Silvia Sabattini, Giuliano Bettini, Dino Amadori, Chiara Talamonti, Massimo Vignoli, Laura Capelli, Jimmy H Saunders, and Marianna Ricci. "Egfr, Her-2 and Kras in Canine Gastric Epithelial Tumors: A Potential Human Model?" *PloS one* 9, no. 1 (2014): e85388.

72. Venning, Freja A, Lena Wullkopf, and Janine T Erler. "Targeting Ecm Disrupts Cancer Progression." *Frontiers in oncology* 5 (2015): 224.

73. Bergers, Gabriele, and Laura E Benjamin. "Tumorigenesis and the Angiogenic Switch." *Nature reviews cancer* 3, no. 6 (2003): 401-10.

74. Abdelmegeed, Somaia M, and Sulma Mohammed. "Canine Mammary Tumors as a Model for Human Disease." *Oncology letters* 15, no. 6

(2018a): 8195-205.

75. Bryan, B. B., S. J. Schnitt, and L. C. Collins. "Ductal Carcinoma in Situ with Basal-Like Phenotype: A Possible Precursor to Invasive Basal-Like Breast Cancer." *Mod Pathol* 19, no. 5 (2006): 617-21.

76. Plasilova, M. L., B. Hayse, B. K. Killelea, N. R. Horowitz, A. B. Chagpar, and D. R. Lannin. "Features of Triple-Negative Breast Cancer: Analysis of 38,813 Cases from the National Cancer Database." *Medicine (Baltimore)* 95, no. 35 (2016): e4614.

77. Ejaeidi, A. A., B. S. Craft, L. V. Puneky, R. E. Lewis, and J. M. Cruse. "Hormone Receptor-Independent Cxcl10 Production Is Associated with the Regulation of Cellular Factors Linked to Breast Cancer Progression and Metastasis." *Exp Mol Pathol* 99, no. 1 (2015): 163-72.

78. Ugolini, F., J. Adelaide, E. Charafe-Jauffret, C. Nguyen, J. Jacquemier, B. Jordan, D. Birnbaum, and M. J. Pebusque. "Differential Expression Assay of Chromosome Arm 8p Genes Identifies Frizzled-Related (Frp1/Frzb) and Fibroblast Growth Factor Receptor 1 (Fgfr1) as Candidate Breast Cancer Genes." *Oncogene* 18, no. 10 (1999): 1903-10.

79. Veeck, J., E. Noetzel, N. Bektas, E. Jost, A. Hartmann, R. Knuchel, and E. Dahl. "Promoter Hypermethylation of the Sfrp2 Gene Is a High-Frequent Alteration and Tumor-Specific Epigenetic Marker in Human Breast Cancer." *Mol Cancer* 7 (2008): 83.

80. Kolbl, A. C., U. Jeschke, K. Friese, and U. Andergassen. "The Role of Tf- and Tn-Antigens in Breast Cancer Metastasis." *Histol Histopathol* 31, no. 6 (2016): 613-21.

81. Libreros, S., R. Garcia-Areas, and V. Iragavarapu-Charyulu. "Chi3l1 Plays a Role in Cancer through Enhanced Production of Pro-Inflammatory/Pro-Tumorigenic and Angiogenic Factors." *Immunol Res* 57, no. 1-3 (2013): 99-105.

82. Mazzoccoli, G., V. Pazienza, A. Panza, M. R. Valvano, G. Benegiamo, M. Vinciguerra, A. Andriulli, and A. Piepoli. "Arntl2 and Serpine1: Potential Biomarkers for Tumor Aggressiveness in Colorectal Cancer." *J Cancer Res Clin Oncol* 138, no. 3 (2012): 501-11.

83. Wang, T., L. Zheng, Q. Wang, and Y. W. Hu. "Emerging Roles and Mechanisms of Foxc2 in Cancer." *Clin Chim Acta* 479 (2018): 84-93.

84. Hashmi, S., Y. Wang, D. S. Suman, R. S. Parhar, K. Collison, W. Conca, F. Al-Mohanna, and R. Gaugler. "Human Cancer: Is It Linked to Dysfunctional Lipid Metabolism?" *Biochim Biophys Acta* 1850, no. 2 (2015): 352-64.

85. Oskarsson, T. "Extracellular Matrix Components in Breast Cancer Progression and Metastasis." *Breast* 22 Suppl 2 (2013): S66-72.

86. Li, Jian-Rong, Chuan-Hu Sun, Wenyuan Li, Rou-Fang Chao, Chieh-Chen Huang, Xianghong Jasmine Zhou, and Chun-Chi Liu. "Cancer Rna-Seq Nexus: A Database of Phenotype-Specific Transcriptome Profiling in Cancer Cells." *Nucleic acids research* 44, no. D1 (2016b): D944-D51.

87. Liu, J., L. Shen, J. Yao, Y. Li, Y. Wang, H. Chen, and P. Geng. "Forkhead Box C1 Promoter Upstream Transcript, a Novel Long Non-Coding Rna, Regulates Proliferation and Migration in Basal-Like Breast Cancer." *Mol Med Rep* 11, no. 4 (2015): 3155-9.

88. Wang, Y., J. Yao, H. Meng, Z. Yu, Z. Wang, X. Yuan, H. Chen, and A. Wang. "A Novel Long Non-Coding Rna, Hypoxia-Inducible Factor-2alpha Promoter Upstream Transcript, Functions as an Inhibitor of Osteosarcoma Stem Cells in Vitro." *Mol Med Rep* 11, no. 4 (2015): 2534-40.

89. Jemal, A., F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman. "Global Cancer Statistics." *CA Cancer J Clin* 61, no. 2 (2011): 69-90.

90. O'Brien, K. M., S. R. Cole, C. K. Tse, C. M. Perou, L. A. Carey, W. D. Foulkes, L. G. Dressler, J. Geradts, and R. C. Millikan. "Intrinsic Breast Tumor Subtypes, Race, and Long-Term Survival in the Carolina Breast Cancer Study." *Clin Cancer Res* 16, no. 24 (2010): 6100-10.

91. Perou, C. M. "Molecular Stratification of Triple-Negative Breast Cancers." *Oncologist* 16 Suppl 1 (2011): 61-70.

92. Carey, L. A., C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston, S. L. Deming, J. Geradts, M. C. Cheang, T. O. Nielsen, P. G. Moorman, H. S. Earp, and R. C. Millikan. "Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study." *JAMA* 295, no. 21 (2006): 2492-502.

93. Jones, P. A., and S. B. Baylin. "The Epigenomics of Cancer." *Cell* 128, no. 4 (2007): 683-92.

94. Chen, X., H. Hu, L. He, X. Yu, X. Liu, R. Zhong, and M. Shu. "A Novel

Subtype Classification and Risk of Breast Cancer by Histone Modification Profiling." *Breast Cancer Res Treat* 157, no. 2 (2016): 267-79.

95. Koch, C. M., R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaoz, G. K. Clelland, S. Wilcox, D. M. Beare, J. C. Fowler, P. Couttet, K. D. James, G. C. Lefebvre, A. W. Bruce, O. M. Dovey, P. D. Ellis, P. Dhami, C. F. Langford, Z. Weng, E. Birney, N. P. Carter, D. Vetrie, and I. Dunham. "The Landscape of Histone Modifications across 1% of the Human Genome in Five Human Cell Lines." *Genome Res* 17, no. 6 (2007): 691-707.

96. Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. "High-Resolution Profiling of Histone Methylations in the Human Genome." *Cell* 129, no. 4 (2007): 823-37.

97. Xi, Y., J. Shi, W. Li, K. Tanaka, K. L. Allton, D. Richardson, J. Li, H. L. Franco, A. Nagari, V. S. Malladi, L. D. Coletta, M. S. Simper, K. Keyomarsi, J. Shen, M. T. Bedford, X. Shi, M. C. Barton, W. L. Kraus, W. Li, and S. Y. R. Dent. "Histone Modification Profiling in Breast Cancer Cell Lines Highlights Commonalities and Differences among Subtypes." *BMC Genomics* 19, no. 1 (2018): 150.

98. Chaligne, R., T. Popova, M. A. Mendoza-Parra, M. A. Saleem, D. Gentien, K. Ban, T. Piolot, O. Leroy, O. Mariani, H. Gronemeyer, A. Vincent-Salomon, M. H. Stern, and E. Heard. "The Inactive X Chromosome Is Epigenetically Unstable and Transcriptionally Labile in Breast Cancer." *Genome Res* 25, no. 4 (2015): 488-503.

99. Rahman, M., and S. Mohammed. "Breast Cancer Metastasis and the Lymphatic System." *Oncol Lett* 10, no. 3 (2015): 1233-39.

100. Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner. "Data Integration in the Era of Omics: Current and Future Challenges." *BMC Syst Biol* 8 Suppl 2 (2014): I1.

101. Tang, Z., C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang. "Gepia: A Web Server for Cancer and Normal Gene Expression Profiling and Interactive Analyses." *Nucleic Acids Res* 45, no. W1 (2017): W98-W102.

102. Nebbioso, A., F. P. Tambaro, C. Dell'Aversana, and L. Altucci. "Cancer Epigenetics: Moving Forward." *PLoS Genet* 14, no. 6 (2018): e1007362.

103. Zand, B., R. A. Previs, N. M. Zacharias, R. Rupaimoole, T. Mitamura, A. S. Nagaraja, M. Guindani, H. J. Dalton, L. Yang, J. Baddour, A. Achreja, W.

Hu, C. V. Pecot, C. Ivan, S. Y. Wu, C. R. McCullough, K. M. Gharpure, E. Shoshan, S. Pradeep, L. S. Mangala, C. Rodriguez-Aguayo, Y. Wang, A. M. Nick, M. A. Davies, G. Armaiz-Pena, J. Liu, S. K. Lutgendorf, K. A. Baggerly, M. B. Eli, G. Lopez-Berestein, D. Nagrath, P. K. Bhattacharya, and A. K. Sood. "Role of Increased N-Acetylaspartate Levels in Cancer." *J Natl Cancer Inst* 108, no. 6 (2016): djv426.

104. Sun, C., Y. Ban, K. Wang, Y. Sun, and Z. Zhao. "Sox5 Promotes Breast Cancer Proliferation and Invasion by Transactivation of Ezh2." *Oncol Lett* 17, no. 3 (2019): 2754-62.

105. Ma, C., F. Wang, B. Han, X. Zhong, F. Si, J. Ye, E. C. Hsueh, L. Robbins, S. M. Kiefer, Y. Zhang, P. Hunborg, M. A. Varvares, M. Rauchman, and G. Peng. "Sall1 Functions as a Tumor Suppressor in Breast Cancer by Regulating Cancer Cell Senescence and Metastasis through the Nurd Complex." *Mol Cancer* 17, no. 1 (2018): 78.

106. Wolf, J., K. Muller-Decker, C. Flechtenmacher, F. Zhang, M. Shahmoradgoli, G. B. Mills, J. D. Hoheisel, and M. Boettcher. "An in Vivo Rnai Screen Identifies Sall1 as a Tumor Suppressor in Human Breast Cancer with a Role in Cdh1 Regulation." *Oncogene* 33, no. 33 (2014): 4273-8.

107. Woolston, C. "Breast Cancer." *Nature* 527, no. 7578 (2015): S101.

108. Hoag, H. "Molecular Biology: Marked Progress." *Nature* 527, no. 7578 (2015): S114-5.

109. Ulaner, G. A., C. C. Riedl, M. N. Dickler, K. Jhaveri, N. Pandit-Taskar, and W. Weber. "Molecular Imaging of Biomarkers in Breast Cancer." *J Nucl Med* 57 Suppl 1 (2016): 53S-9S.

110. Harris, Lyndsay, Herbert Fritsche, Robert Mennel, Larry Norton, Peter Ravdin, Sheila Taube, Mark R Somerfield, Daniel F Hayes, and Robert C Bast Jr. "American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer." *Journal of clinical oncology* 25, no. 33 (2007): 5287-312.

111. Geyer, P. E., L. M. Holdt, D. Teupser, and M. Mann. "Revisiting Biomarker Discovery by Plasma Proteomics." *Mol Syst Biol* 13, no. 9 (2017): 942.

112. Banin Hirata, B. K., J. M. Oda, R. Losi Guembarovski, C. B. Ariza, C. E. de Oliveira, and M. A. Watanabe. "Molecular Markers for Breast Cancer:

Prediction on Tumor Behavior." *Dis Markers* 2014 (2014): 513158.

113. Wisniewski, J. R., A. Zougman, N. Nagaraj, and M. Mann. "Universal Sample Preparation Method for Proteome Analysis." *Nat Methods* 6, no. 5 (2009): 359-62.

114. Kim, Y. I., J. M. Ahn, H. J. Sung, S. S. Na, J. Hwang, Y. Kim, and J. Y. Cho. "Meta-Markers for the Differential Diagnosis of Lung Cancer and Lung Disease." *J Proteomics* 148 (2016): 36-43.

115. Kim, Y. I., and J. Y. Cho. "Gel-Based Proteomics in Disease Research: Is It Still Valuable?" *Biochim Biophys Acta Proteins Proteom* 1867, no. 1 (2019): 9-16.

116. Cho, H. M., P. H. Kim, H. K. Chang, Y. M. Shen, K. Bonsra, B. J. Kang, S. Y. Yum, J. H. Kim, S. Y. Lee, M. C. Choi, H. H. Kim, G. Jang, and J. Y. Cho. "Targeted Genome Engineering to Control Vegf Expression in Human Umbilical Cord Blood-Derived Mesenchymal Stem Cells: Potential Implications for the Treatment of Myocardial Infarction." *Stem Cells Transl Med* 6, no. 3 (2017): 1040-51.

117. Dantas Cassali, G., A. Cavalheiro Bertagnolli, E. Ferreira, K. Araujo Damasceno, C. de Oliveira Gamba, and C. Bonolo de Campos. "Canine Mammary Mixed Tumours: A Review." *Vet Med Int* 2012 (2012): 274608.

118. de Ronde, J. J., E. H. Lips, L. Mulder, A. D. Vincent, J. Wesseling, M. Nieuwland, R. Kerkhoven, M. J. Vrancken Peeters, G. S. Sonke, S. Rodenhuis, and L. F. Wessels. "Serpina6, Bex1, Agtr1, Slc26a3, and Laptm4b Are Markers of Resistance to Neoadjuvant Chemotherapy in Her2-Negative Breast Cancer." *Breast Cancer Res Treat* 137, no. 1 (2013): 213-23.

119. Dobiasova, M., and J. J. Frohlich. "Advances in Understanding of the Role of Lecithin Cholesterol Acyltransferase (Lcat) in Cholesterol Transport." *Clin Chim Acta* 286, no. 1-2 (1999): 257-71.

120. Subbaiah, P. V., M. Liu, and T. R. Witt. "Impaired Cholesterol Esterification in the Plasma in Patients with Breast Cancer." *Lipids* 32, no. 2 (1997): 157-62.

121. Rakha, E. A., M. A. Aleskandarany, M. S. Toss, N. P. Mongan, M. E. ElSayed, A. R. Green, I. O. Ellis, and L. W. Dalton. "Impact of Breast Cancer Grade Discordance on Prediction of Outcome." *Histopathology* 73, no. 6 (2018): 904-15.

122. Abdelmegeed, S. M., and S. Mohammed. "Canine Mammary Tumors as a Model for Human Disease." *Oncol Lett* 15, no. 6 (2018b): 8195-205.

123. Lee, K. H., H. M. Park, K. H. Son, T. J. Shin, and J. Y. Cho. "Transcriptome Signatures of Canine Mammary Gland Tumors and Its Comparison to Human Breast Cancers." *Cancers (Basel)* 10, no. 9 (2018).

124. Tavasoly, A., H. Golshahi, A. Rezaie, and M. Farhadi. "Classification and Grading of Canine Malignant Mammary Tumors." *Vet Res Forum* 4, no. 1 (2013): 25-30.

125. Surinova, S., R. Schiess, R. Huttenhain, F. Cerciello, B. Wollscheid, and R. Aebersold. "On the Development of Plasma Protein Biomarkers." *J Proteome Res* 10, no. 1 (2011): 5-16.

126. Kosek, A. B., D. Durbin, and A. Jonas. "Binding Affinity and Reactivity of Lecithin Cholesterol Acyltransferase with Native Lipoproteins." *Biochem Biophys Res Commun* 258, no. 3 (1999): 548-51.

127. Cedo, L., S. T. Reddy, E. Mato, F. Blanco-Vaca, and J. C. Escola-Gil. "Hdl and Ldl: Potential New Players in Breast Cancer Development." *J Clin Med* 8, no. 6 (2019).

128. Ozmen, H. K., and S. Askin. "Lecithin: Cholesterol Acyltransferase and Na(+)-K(+)-Atpase Activity in Patients with Breast Cancer." *J Breast Cancer* 16, no. 2 (2013): 159-63.

129. Beeghly-Fadiel, A., N. K. Khankari, R. J. Delahanty, X. O. Shu, Y. Lu, M. K. Schmidt, M. K. Bolla, K. Michailidou, Q. Wang, J. Dennis, D. Yannoukakos, A. M. Dunning, P. D. P. Pharoah, G. Chenevix-Trench, R. L. Milne, D. J. Hunter, H. Per, P. Kraft, J. Simard, D. F. Easton, and W. Zheng. "A Mendelian Randomization Analysis of Circulating Lipid Traits and Breast Cancer Risk." *Int J Epidemiol* (2019).

국문초록

오믹스 데이터를 이용한 개와 사람의 바이오마커 비교연구

박형민

서울대학교 대학원
수의학과 수의생명과학 전공

지도교수 조제열

유방암은 여성과 암캐에서 가장 빈번하게 진단되는 악성종양 중 하나이다. 이러한 암과 관련된 이상현상을 완전히 이해하고 극복하려는 수많은 노력에도 불구하고, 유방 조직의 특정 부위에서 발생하는 여러 유형들은 드물지만 위협적인 악성 종양으로 발달한다. 비교 의학적 접근법은 인간의 유방암 연구에 기존과는 다른 관점으로 접근하는 효과적인 방법으로 등장했다. 다양한 오믹스 기술의 등장과 함께 유방암 치료의 전반적인 방향이 대규모 데이터를 이용하여 특정 유방암을 지칭하는 바이오마커

발굴로 기울었다. 차세대 염기서열 분석(NGS)을 이용한 후생유전체 데이터부터 질량분석기(MS)에서 생산하는 단백체 정보까지, 이러한 오믹스 데이터를 통합분석하는 것이 악성 유방암 진단과 약물 표적 발견을 위한 해결책이다. 본 연구는 총 3장으로 구성된다.

제1장에서는 10쌍의 개 유선암 및 인접 정상 조직에서 추출한 RNA-seq 데이터로 개 유선암과 연관된 신호를 식별하는 방법을 설명한다. 유방암(BC)/유선암(MGC)은 가장 빈번한 암중 하나이며 암과 관련된 사망률에서 선두를 차지하고 있다. 개 유선암과 사람 유방암 특이적 유전자를 이해하기 위해, 우리는 개의 8쌍의 발암과 인접한 정상 조직에서 얻은 RNA의 염기서열을 분석했다. 전사체 분석을 통해 개 전체 유선암에서 351개의 특이적 발현유전자를 확인했다. 비교분석 결과, 개 유선암의 세 가지 조직학적 유형(단순형, 관상형, 복합형)과 인간 유방암의 네 가지 분자 유형(HER2+, ER+, ER&HER2+, TNBC) 사이에 존재하는 상관관계를 밝혔다. 세 종류의 개 유선암을 모두 공유하는 8개의 DEG는 이전에 인간 연구에서 암과 관련된 유전자로 보고됐다. 확인된 DEG를 이용한 유전자 온톨로지 및 발현 경로 분석 결과, 세포 증식, 접착, 염증 반응 과정이 유선암 DEG에서 나타났다. 이와는 대조적으로, 세포사멸과 관련된 전사체 조절 및 지방산 항상성에 연관된 유선암 DEG들은 하향 조절되었다. 더욱이, 상류 프로모터 전사체(PROMPT)와 DEG 사이에 상관관계가 있음을 밝혔다. 개 유선암 및 조직학적 유형 특이적 발현 유전자를 통해 우리는 인간의 유방암과 개 유선암을 더 잘 이해할

수 있게 되었으며, 두 질병의 바이오마커의 진단과 개발에 대한 새로운 통찰력을 얻을 수 있을 것이다.

제2장은 특정 유방암 유형의 판별과 치료에 초점을 맞추고있다. 여러 유방암 유형 중 삼중음성유방암(TNBC)은 예후가 가장 나쁘며 보고된 사례가 가장 적다. TNBC에 대한 보다 나은 이해와 효과적인 전구체을 얻기 위해 TNBC 세포와 정상 유방 세포의 데이터를 사용하여 두 가지 주요 히스톤 변형인 활성화 변형체 H3K4me3와 억압 변형체 H3K27me3를 분석하였다. 프로모터 유전자에 두 히스톤 변형체의 조합을 통해 유전자 발현과 높은 상관관계가 있음을 확인했다. 유전자의 목록은 *NOVA1*, *NAT8L*, *MMP16*을 포함한 활성화된 유전자(H3K4me3이 많이 포진된)와 *IRX2*, *ADRB2*와 같은 억제된 유전자(H3K27me3이 많이 포진된)로 정의됐다. 추가적인 조사를 위해, 후보 유전자들은 TNBC에 특이적인 발현함을 식별하기 위해 다른 종류의 유방암과 비교했다. RNA-seq 데이터는 히스톤 변형에 의해 지배되는 유전자 조절을 확인하고 검증하기 위해 구현됐다. H3K4me3와 H3K27me3를 통합하여 분석한 바이오마커 조합은 P-값이 1.16e-226인 AUC 93.28%를 보였다. 이 연구 결과는 프로모터 지역에 위치한 서로 반대되는 히스톤 변형 분석이 TNBC의 바이오마커의 진단 및 개발에 활용될 수 있음을 시사하며 발현의 과정이 후성유전체에 의한 조절 기작과 관련되어 있기에 이러한 유전자 발현에 대한 연구방향을 제시해 줄 수 있을 것이다.

제3장은 개 유선암에서 시작하여 인간 유방암까지 적용될 수 있는 바이오마커 연구로 구성된다. 바이오마커는 지속적으로 발견되지만, 유방암의 공격성과 지속성을 대표해주는 바이오마커는 유방암의 유형을 분류시키는 바이오마커에 비해 부족하다. 이 연구는 비교의학적 접근법을 통한 개 유선 종양 샘플을 사용했다. 개암 정상 혈장과 유선암 혈장 모두 36분할을 통한 광범위한 정량적 단백체 분석을 진행했다. 확인된 단백질 중 LCAT는 전이 가능성이 높은 공격적인 암 발병 단계를 나타내는 혼합형 종양 검체에서 특이적으로 발현되는 것으로 밝혀졌다. 추가적인 질량분석과 Western Blot 검증을 통해 우리는 LCAT 단백질이 전이성이 높은 유선종양의 지표단백질이 될 수 있음을 발견했다. 흥미롭게도, 사람의 림프절 양성 유방암에서 과발현된 LCAT이 환자의 수명을 유의미하게 줄이며 유방암 중 2기 이상 진행되었을 때에도 개 유선암과 동일하게 높게 발현되는 것을 확인하였다. 이것으로 단백질 LCAT은 사람과 개에서 공격적인 형태의 유방암 및 유선암을 지칭하는 지표단백질로서의 가능성을 밝혔다.

주요어: 유방암, 비교의학, 바이오마커, 후성유전학, 단백체학
학번: 2016-21753