A comparative genomics approach to identify novel inherited cancer risk variants in dogs and humans

by

Anna Laureen Watkins Huskey

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 7, 2021

Key Words: comparative genomics, comparative oncology, whole genome sequencing, inherited cancer risk, CEACAM, dog

Approved by

Dr. Nancy Merner, Professor Pathobiology
Dr. Randall Clark Professor Drug Discovery and Development
Dr. Laurie Stevison Professor Biology
Dr. Raj Amin Professor Drug Discovery and Development
Dr. Muralikrishnan Dhanasekaran Professor Drug Discovery and Development

Abstract

Dogs provide a very special and unique opportunity for novel discovery in inherited disease studies. Through breeding practices to aid in the development of specific breeds, dog breeds are a very homogenous population, which has resulted in an increase of inherited diseases in purebred dog populations. Early genetic studies of the dog often employed linkage maps within familial studies, which were largely helpful due to the high linkage disequilibrium that exists in the dog population from breeding. However, as sequencing technology has developed, more dog genetic studies are being carried out, and many of them utilizing next generation sequencing (NGS) technology. From this technology, whole genome sequencing (WGS) is a sequencing option that provides an unbiased survey of the entire genome. This complete genome of information has become more crucial in genetic studies as it is estimated that the vast majority of disease influencing mutations within dogs will be outside of the coding portion of the genome.

Furthermore, there are many diseases that have genetic similarities in both dogs and humans, allowing the dog to benefit from previous human disease studies and also to serve as a model for human diseases. Dogs have been successful models for very heterogeneous human diseases. WGS has been an effective method for identifying mutations associated with inherited diseases through multiple different analyses methods, and identifying disease influencing risk genes in dogs can be easier due to the high homogeneity within breeds. This can then be translated to human disease studies, potentially as candidate gene approaches. This approach also translates well to cancer studies, as cancer is a genetic disease, and WGS can aid in identifying mutations in both species.

Due to similar presentations and previously known similar genetic links between breast cancer and canine mammary tumors (CMT), a cohort of purebred CMT-affected dogs were investigated through pedigree analysis and WGS to identify risk variants within the cohort. This involved an initial analysis of mutations in orthologs of human breast cancer risk genes. Variants within both *BRCA2* and *STK11* were associated with CMT risk; breed-specific associations were identified. This initial analysis highlighted the effectiveness of WGS and elucidating CMT risk in small breed-specific cohorts.

In search for novel risk variants, the WGS data of five Golden Retrievers were subsequently analyzed. Upon identifying and validating mutations shared amongst all five Golden Retrievers, the results were compared to human breast cancer cases to elucidate risk.

Rare protein truncating variants (PTVs, nonsense, frameshifting and splice-site affecting mutations) were investigated in the Golden Retrievers WGS data and then genotyped in the remaining Golden Retriever cohort. From this a frameshifting mutation in *CEACAM24* was identified in the CMT-affected Golden Retriever cohort, which translated to a significant association of rare PTVs in the *CEACAM* gene family in human breast cancer cases. This was the first time inherited mutations the *CEACAM* gene family were associated with inherited breast cancer risk.

The *CEACAM* gene family has long been tied to colorectal cancer (CRC) development and progression; however, there is limited to no information on this gene family and inherited CRC risk. An analysis to investigate an association with *CEACAM* genes and inherited CRC risk was carried out. Rare PTVs and missense mutations were both investigated for influence, and no gene-based or –family associations were identified. However, certain individual mutations were associated, highlighting the need for further exploration. Ultimately, this work represents one of the first investigations of the *CEACAM* gene family and inherited CRC risk. This dissertation highlights the power of WGS of dogs and how such studies can benefit human health through comparative oncology.

Acknowledgments

I would like to thank my graduate co-advisors, Dr. Nancy Merner and Dr. Randall Clark for guiding me through my graduate career. I would also like to express my gratitude to my current and past laboratory team members for their assistance in these research endeavors. Also, to my husband, thank you for your continuous support during my graduate career. I would also like to thank my Graduate Advisory Committee for their help and guidance, along with the faculty members of the Department of Drug Discovery and Development in the Harrison School of Pharmacy and the Department of Pathobiology in the College of Veterinary Medicine for their constant support and available resources. Additionally, I would like to thank the Office of Information Technology for compute time and technical support with all high-performance computer needs.

These research efforts would not have been possible without the Orthopedic Foundation for Animals CHIC DNA Repository, who provided the dog DNA samples for the canine mammary tumor affected cohort and the breed specific control DNA. Furthermore, I would like to thank the Auburn University Research Initiative in Cancer, both for a seed grant allowing for the whole genome sequencing and for my graduate fellowship to support my research endeavors. I would also like to thank Auburn University's College of Science and Mathematics for the Auburn University Cellular and Molecular Biosciences (CMB) Peaks of Excellence Research Fellowship (to A.L.W.H.; NSF-EPS-1158862, grant G00006750) to support my research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abstract

NGS	next generation sequencing
WGS	whole genome sequencing
CMT	canine mammary tumor
PTV	protein truncating variant
CRC	colorectal cancer

Chapter 1

LD	linkage disequilibrium
SNP	single nucleotide polymorphism
GWAS	genome-wide association study
bp	base pairs
NGS	next generation sequencing
WES	whole exome sequencing
WGS	whole genome sequencing
Mb	megabase pair
DCM	dilated cardiomyopathy
DMD	Duchenne muscular dystrophy
NCL	neuronal ceroid lipofuscinosis
PxD	paroxysmal dyskinesias
cPxD	paroxysmal dyskinesias
BHD	*Birt-Hogg-Dubé*
CMT	canine mammary tumor

Chapter 2

RCND	renal cystadenocarcinoma and nodular dermatofibrosis
BHD	*Birt-Hogg-Dubé*
CMT	canine mammary tumor
WGS	whole genome sequencing

ESS	English Springer Spaniels
AKC	American Kennel Club
CHIC	Canine Health Information Center
OFA	Orthopedic Foundation for Animals
GATK	Genome Analysis Toolkit
GVCF	genomic variant calling format
EVA	European Variation Archive
SNV	single nucleotide variant
WES	whole exome sequencing
VUS	variants of unknown significance

Chapter 3

CMT	canine mammary tumor
PTV	protein truncating variant
CHIC	Canine Health Information Center
WGS	whole genome sequencing
PCR	polymerase chain reaction
ELM	Eukaryotic Linear Motif
TCGA	The Cancer Genome Atlas
SNV	single nucleotide variant
BAM	binary sequence alignment mapping
GDC	Genomic Data Commons
GATK	Genome Analysis Toolkit
gVCF	genome variant calling format
NHLBI	National Heart, Lung, and Blood Institute
HBOC	hereditary breast and ovarian cancer

Chapter 4

CRC	colorectal cancer
TCGA	The Cancer Genome Atlas
HBOC	hereditary breast and ovarian cancer

BAM	binary sequence alignment mapping
GDC	Genomic Data Commons
GATK	Genome Analysis Toolkit
VCF	variant calling format
PTV	protein truncating variant
MAF	minor allele frequency
ELM	Eukaryotic Linear Motif

Chapter 5

LD	linkage disequilibrium
WGS	whole genome sequencing
CMT	canine mammary tumors
TCGA	The Cancer Genome Atlas
CRC	colorectal cancer

Chapter 1: Introduction

1.1 History of Dog Breed Structure

Dogs provide a very special and somewhat unique opportunity in inherited disease studies; based on the specific breeding patterns of domestic dogs, they are groups of very homogenous populations.[1] The first step towards developing these homogeneous populations resulted from the domestication of wolves around 15,000 years ago. Since this time, increasing numbers of dogs have shared their environment with humans, and both species have benefited from the shared living spaces. Dogs have helped humans with hunting, protection, herding, and more recently, companionship. Through this time, dogs have been enriched for behaviors that mimic humans and support human needs. Due to the increased desire for specific dog traits and characteristics, humans have very specific breeding patterns to result in particular breeds. These dogs can vary from their ancestral wolf populations by more than 40-fold in overall size, and often have a variety of skills and traits that aid them in modern society.[2] There are currently around 400 modern dog breeds, most of these generating from a common stock of dogs for each breed.[1; 3] These breeds represent a greater species diversity in physiological differences including skeletal size than any other mammalian species.[4] These differences across breeds have developed as people desired specific aspects from their dogs to aid as working dogs, such as herding breeds, all the way to purely companion animals. This has resulted in a vast genetic variation between dog breeds, 27.5%, as compared to human populations with 5.4%. These breeding patterns have also resulted in the preservation of disease influencing mutations.[3; 5] This has resulted in both a general increase of inherited diseases among purebred dogs, along with increased numbers of specific diseases in certain breeds.[1; 6]

1.2 Linkage Disequilibrium & Early Dog Genetics Disease Studies

While purebred dogs are at increased risk of inherited diseases, the breeding patterns also result in a very high linkage disequilibrium (LD) that can result in increased ability to detect the genetic variants influencing disease risk.[7] Linkage disequilibrium is defined as "the nonrandom association of alleles at two or more loci in a general population".[8] Within a dog breed genetic homology is estimated to be around 95%.[2] It is estimated that LD is approximately 100 times higher in purebred dogs than in humans, which can aid in canine genetic studies and has resulted in fewer markers needed to map genes in the dog.[7] While dog breeds are typically considered to be closed breeding populations, some breeds do have higher degrees of LD, which has been influenced by breed popularities over time.[2] Golden Retrievers, who have been known as an exceptionally popular breed, have an LD at half maximum value at around 500kb.[2; 9] In Bernese Mountain dogs and Pekingese, at least 3 Mega bases (Mb) was the determined length at half maximum LD value.[9] These breeds with greater LD have had more limited popularity and, therefore, a more restricted gene pool, decreasing the ability for genetic diversity.[2] This restricted gene pool within dog breeds has resulted in an estimated need of only 10,000 single nucleotide polymorphisms (SNPs) for genome-wide association studies (GWASs) as compared to 500,000 SNPs needed in humans. In general, a GWAS utilizes a case-control analysis to identify common SNPs or regions that could be linked to a specific phenotype.[10-12] The need to use a smaller number of SNPs in dog studies can aid in efficiency and feasibility of identifying genetic factors influencing phenotypes, including diseases and disease risk.

Over 2 decades have passed since the first dog genetic disease studies were carried out.[13; 14] Traditionally, genetic linkage aided in discovery, which fostered the identification of a marker linked to copper toxicosis susceptibility in terriers. These initial studies were greatly aided by the increased LD that exists within the dog genome, and GWASs in dogs have greatly benefited from the high LD in dogs as well. An early GWAS on atopic dermatitis in dogs utilized a SNP chip with SNPs from both the poodle and boxer references[15] and was able to determine two regions of interest that were independently association with the disease in multiple dog breeds.

1.3 Sequencing the dog genome and next generation sequencing (NGS) development

Early genetic studies of the dog utilized radiation hybrid and meiotic linkage maps.[16] These maps provided the basis of many analyses since their availability in the late 1990s. Early dog studies also benefited from comparisons to the more well-developed human and mouse genomes.[2] However, later studies have benefitted from the development of the dog reference genome in the early 2000s. A poodle was shotgun sequenced with approximately 1.5X depth coverage with up to 80% genome coverage in 2003 and provided the first option for a good reference genome for genetic studies in domestic dogs.[2; 17] Shortly after a 7.5X sequencing of a boxer was generated and provided a stronger option for a reference genome for dog studies with greater genome coverage as well.[1; 2] This reference genome determined a decrease amount of repeat insertion as compared to the human and mouse genomes, resulting in an estimation of the dog genome being approximately 18% smaller than the human genome.[1] The dog reference genome underwent further improvement in 2014 to aid in closing sequence gaps in the previous boxer dog reference genome.[18] This dog reference genome has 99.8% of the base pairs (bp) covered at least once and is a widely used canine reference genome canFam3.1, which has continued to empower canine genetic disease studies.

Through advancements in sequencing technology and the development and improvement of the dog reference genome, greater numbers of dog genetic studies are being carried out. Most of these are utilizing the newest sequencing technology, next generation sequencing (NGS). This is a massively parallel sequencing method that sequences millions of small DNA fragments and utilizes bioinformatics strategies to align the fragments.[19] NGS allows for a broader spectrum of genetic alterations to be captured from large-scale inversions to single base changes. Three common platforms are used for sequencing: Ion Torrent, Complete Genomics (BGI/MGI), and Illumina.[20] Illumina sequencing platforms have an average read length of a 150 bp per read. Ion Torrent technology utilizes longer read lengths, typically from 200-400 bp per read. BGI/MGI sequencing has ranges from 50-400 bp reads. Sequencing with shorter read lengths may cause some issues with aligning to the reference genome.[21] In particularly repetitive regions, it can be more problematic for alignment software to correctly align the shorter reads and can make identifying copy number variation more challenging. The alignment difficulties can be aided however, by the use of paired-end sequencing.[10] Often Illumina sequencing is performed through paired-end sequencing, Ion Torrent with single end sequencing, and BGI/MGI sequences with most options as paired end sequencing.[20]

Targeted sequencing, whole exome sequencing (WES) and whole genome sequencing (WGS) represent three of the more common sequencing regions within NGS and can be carried out with the three different sequencing platforms.[19; 20] Illumina is the most commonly used platform in NGS sequencing studies.[20] Targeted sequencing has benefitted from the efficiency of NGS and many gene panels have been designed to look at disease specific candidate genes.[22; 23] This has proven to be a cost-efficient approach for many human research and clinical based genetic analyses[22]. This is not overly common in dog studies; however, a human panel was used to identify mutations in a canine cohort with lymphoma.[24] Targeted sequencing also often provides the highest depth of coverage with well over 500X sequencing depths on average within the targeted locations,[22] and some studies achieving average coverage of at least 1650X.[25] This method allows for a very high sequencing sensitivity;[22] however, targeted sequencing limits the analysis to those already suggested regions in genetic analysis and does not allow for a more exploratory approach if less is known about the phenotype or disease and can introduce biases into the results by only examining certain known regions.[26]

WES provides information on the entire coding portion of the genome and does require that the coding regions already be known.[27] This is a similar approach to targeted sequencing, but targets a much larger span of regions, and therefore tends to have lower coverage depths, 50X-80X average,[28] than targeted sequencing, which could decrease the sensitivity of the mutational analysis.[27] However, there is substantially more genetic information provided by this approach as compared to targeted sequencing. A dog exome panel was developed for the dog reference genome canFam3.1 and published in 2014.[29] It had around 204,000 regions with a total size of approximately 53 Mb. Additional developments in the dog reference genome and coding regions have resulted in an updated exome panel with multiple options in 2015.[30] The exome-plus option with 152Mb covered includes several known non-coding RNA regions and an exome-CDS option that only includes the mRNA regions for the protein coding regions and excludes 3’ and 5’ UTR with 71Mb covered.

A final option of NGS is whole genome sequencing (WGS). This method provides an unbiased survey of the entire genome (coding and non-coding), gives an extensive amount of data, and does not contain some of the sequencing biases that other methods contain.[31-33] Furthermore, there are some studies indicating that WGS provides a more powerful approach to identify coding variants compared to WES.[34] Using Illumina sequencing technology, WGS commonly offers average coverage options of 15X or 30X, and, as the cost of WGS continues to decrease and data management strategies continue to evolve, this method is becoming more commonly used as an approach in canine genetic disease studies.[35-38]

1.4 WGS as an approach to identify inherited mutations in dogs

WGS is an effective method for identifying mutations associated with inherited diseases through multiple different analyses methods. In combination, WGS and GWAS have been a successful approach, allowing for the identification of disease regions through GWAS and then mutations in those regions through WGS.[39-42] These two analysis methods can greatly aid each other; especially when first analyzing a large case/control cohort during the GWAS, followed by WGS of a single dog to identify the disease associated/causing mutation, which is a very cost-effective approach. This approach has been successful in identifying a causative mutation in dogs with Dandy-Walker-Like Malformation which presented with an autosomal recessive mode of inheritance.[39] Another GWAS-WGS based study identified a specific causative mutation associated with neurodegeneration.[40] An additional study successfully utilizing this approach analyzed cobalamin malabsorption in Border Collies and identified an associated frameshift mutation.[41]

WGS a family with an affected offspring and healthy parents also allows for the opportunity discover the causal mutation, through a trio study. Through WGS the three, the entire genome can be explored for the variants with differing presentation between parents and offspring.[35; 43] This can be used to identify *de novo* mutations, "mutations that appear in an individual despite not being seen in their parents"[44], that are dominantly influencing disease.[45] This approach can be very helpful for a dog who has a novel presentation for a disease and doesn't have the previously identified mutation influencing the disease. This such case occurred with a dog who had ichthyosis, but not a mutation previously associated with the disease and a missense mutation in *ASPRV1* was identified.[43] Trio studies can also be used to identify causal mutations with a recessive mode of inheritance.[45-47] A WGS trio study with hereditary footpad hyperkeratosis identified a homozygous single missense variant in *FAM83G* was isolated as the disease influencing variant.[35] WGS can also be used in combination with candidate gene studies. One such example, sequenced the genome of a single dog and identified a mutation of interest in a candidate gene that the lab had previously connected to cobalamin malabsorption in Border Collies,[48] and associated a frameshift mutation in this gene to the same disease in Beagles.

Simple filtering strategies and statistical analyses of WGS data alone can also identify disease-causing variants. Complex diseases can greatly utilize the power that WGS provide; a mutation influencing spinocerebellar ataxia with myokymia and/or seizures in Jack Russell Terrier and related breeds, also known as Russel group terriers (RGT), was isolated through WGS combined with a case-control analysis.[49] A study on Doberman pinschers with dilated cardiomyopathy (DCM) was able to identify a missense mutation in TTN after WGS five dogs.[50] This disease is characterized by genetic heterogeneity even within a breed, and the disease presentation in Doberman pinschers shares many similarities to human DCM including the genetic heterogeneity, along with the importance of *TTN* in influence of DCM; this study highlights how helpful dogs can be as a model of even largely heterogeneous human diseases.

1.5 Comparative Genomics – Dogs and Humans

Comparative genomics is a field that focuses on the conserved genomic regions among species and their likelihood to perform similar functions.[51] This field has been largely helpful in evolutionary biology, but has growing utility in disease studies allowing research into diseases that affect both species to benefit the other.[14; 51-57] There are many diseases that occur in similar manners and presentations in both dogs and humans.[5] There are an estimated 360 canine hereditary diseases that have human equivalents, which is the vast majority of the approximately 450 known genetic dog diseases, which is a higher amount of shared genetic diseases among humans than other domesticated animals.[55] Due to the increased homogeneity and phenotypic expressions within a dog breed, they serve as a more controlled population for both genetic and potentially environmental influences. Furthermore, dogs often share human environments, so the environmental impacts are fairly controlled for between the species as compared to other species for comparative analyses.[58] This can aid in more easily identifying mutations or regions that could be influencing both dog and human diseases. Dogs are also helpful models for modeling disease progression and monitoring therapeutic approaches.[5] This allows the dogs with similar diseases to be used as models of the human disease.

Studying these diseases and their causes often helps shed light on information that benefits both species.[59] Duchenne muscular dystrophy (DMD) is a genetic disease that is found in both species and is characterized by similar sex-linked inheritance pattern.[55] This disease is present in many different dog breeds and genetic insights on both human and dog populations have aided in greater understanding of the disease in both species. Furthermore, WGS has become more crucial to dog disease studies as humans and canines share a similarity in genetic risk for many diseases[60] and greater than 80 percent of variants associated with disease in humans are outside of the coding portion of the genome according to more recent association studies.[61] Additional candidate gene analysis from WGS data can be aided by discoveries from human disease studies.[37; 52; 62-64] These studies have been successful by investigating orthologs of known human disease genes. This has been greatly beneficial in several dog populations to identify associated mutations. Furthermore, this approach takes advantage of the high degree of human research that has already occurred for some genetic diseases shared between the species. One disease in dogs, where a candidate gene approach from human studies has been very helpful in, is neuronal ceroid lipofuscinosis (NCL). This disease has a very similar presentation in dogs and humans, so orthologs of genes associated with the human NCL are often used as a candidate gene approach to determine causal variants in dogs of multiple different breeds.[54; 64-66] These studies carried out WGS a single affected dog of a specific breed and then further investigated mutations of interest in controls and other breed-specific affected dogs. Due to the similarities of human and canine paroxysmal dyskinesias (PxD) research into canine PxDs (cPxDs) has already been heavily influenced by what is already known about human cases.[53] In a dog study investigating cPxDs in Soft Coated Wheaton Terriers a missense mutation was identified in two WGS dogs in *PIGN* that is likely the causal mutation. In previous human studies, this gene has been investigated, and they determined that it was possible that mutations in human PIGN could possibly cause PxD or other similar phenotypes in humans, highlighting the usefulness dogs could be as a model for this disease.

Dogs have served as model for many genetic diseases and have led to findings in humans already.[48; 59; 67; 68] With the increased data and power of WGS, genome sequencing dogs provide excellent opportunities to be models for human diseases. A GWAS in conjunction with WGS on dogs and humans with cleft lips or palates was carried out.[69] A frameshift mutation in *ADAMTS20* was determined to be the likely mutation, and found to be likely breed specific in Nova Scotia Duck Trolling Retrievers, and in Guatemalan humans, the gene was also associated with cleft lips or palates. The findings from this study highlight the usefulness of canine approaches to aid in identifying causal mutations or influential genes in disease that affect both species. Due to the increased homogeneity in dogs, it can be easier to identify these risk genes in them; then, those findings can translate to human studies, potentially as candidate gene approaches in human disease studies.

As stated before, dogs present an excellent model for many inherited diseases that includes inherited cancer.[70] Comparative oncology studies companion animals with naturally-occurring cancers to elucidate information on cancer biology and therapy to provide information benefitting both species.[71] As dogs share a similar cancer incidence to humans, with approximately 30% of both species developing the disease, dogs can provide a model of many types of cancer.[58] One genetic comparison of human and dog cancers resulted in the discovery of a mutation in the *BHD* gene in the dog influencing risk for multifocal renal cystadenocarcinoma.[72] In humans, mutations in this gene are linked to Birt-Hogg-Dubé syndrome. This is a syndrome which is associated with many types of non-cancerous tumors along with cancerous ones such as renal tumors.[73] This finding was identified in dogs through an analysis of a Germen Shepherd Dog.[72] While the sequencing used in the analysis was not WGS, it did allow for a low coverage survey of the genome that was generated through mini-libraries from BAC clones. This discovery showed the benefits of a genome survey when identifying mutations and was able to show a shared link with human and dog renal cancers, and the potential for dogs as models of inherited cancers. Furthermore, WGS has been exceptionally useful identifying mutations in inherited diseases, and this includes its usefulness in identifying both somatic and inherited cancer causes. As a genetic disease, cancer is often investigated at a genetic level. WGS in dogs only increases the likelihood of finding causal mutations for cancers syndromes in dogs, like those found in human cancer syndromes.

The following chapters describe my approach in identifying mutations influencing inherited cancer risk in both dogs and humans with cancer. Canine mammary tumors (CMTs) and human breast cancers have many tumor similarities, including genetic risk factors[74-79]. Therefore, CMTs can serve as a model of hereditary breast cancer susceptibility, especially considering similar genetics and familial clustering.[79; 80] Utilizing WGS data for 14 dogs across four different breeds, an initial analysis of orthologs of human breast cancer susceptibility genes was carried out within a canine mammary tumor (CMT) cohort. Stemming from this, whole genome analysis was also carried out in whole genome sequenced CMT-affected Golden Retrievers within the cohort to search for novel risk factors and the findings were translated to human cancers. The work within this dissertation highlights how dogs can serve as a model for breast cancer and provide insights that can benefit both species, and also how these results can lead to other analysis for human diseases.

Chapter 2: Whole genome sequencing for the investigation of canine mammary tumor inheritance - an initial assessment of high risk breast cancer genes reveal *BRCA2* and *STK11* variants potentially associated with risk in purebred dogs.

This information was published in Canine Medicine and Genetics in 2020

Anna LW Huskey[*†], Katie Goebel[†], Carlos Lloveras-Fuentes[†], Isaac McNeely[†], Nancy D Merner[*†]

2.1 Abstract

Background: Although, in general, cancer is considered a multifactorial disease, clustering of particular cancers in pedigrees suggests a genetic predisposition and could explain why some breeds appear to have an increased risk of certain cancers. To our knowledge, there have been no published reports of whole genome sequencing to investigate inherited canine cancer risk, and with little known about canine mammary tumor genetic susceptibility, we carried out whole genome sequencing on 14 purebred dogs diagnosed with mammary tumors from four breed-specific pedigrees. Following sequencing, each dog's data was processed through a bioinformatics pipeline. This initial report highlights variants in orthologs of human breast cancer susceptibility genes.

Results: The overall whole genome and exome coverage averages were 26.0X and 25.6X, respectively, with 96.1% of the genome and 96.7% of the exome covered at least 10X. Of the average 7.9 million variants per dog, initial analyses involved surveying variants in orthologs of human breast cancer susceptibility genes, *BRCA1*, *BRCA2*, *CDH1*, *PTEN*, *STK11,* and *TP53*. Nineteen unique coding variants were identified and validated through PCR and Sanger sequencing. Potential CMT-associated variants were identified in *BRCA2* and *STK11*, and breed-specific analyses revealed the breeds at the highest risk. Several additional *BRCA2* variants showed trends toward significance, but have conflicting interpretations of pathogenicity, and correspond to variants of unknown significance in humans, which require further investigation. Variants in other genes were noted but did not appear to be associated with disease.

Conclusions: Whole genome sequencing proves to be an effective method to elucidate risk of CMT. Risk variants in orthologs of human breast cancer susceptibility genes have been identified. Ultimately, these whole genome sequencing efforts have provided a plethora of data that can also be assessed for novel discovery and have the potential to lead to breakthroughs in canine and human research through comparative analyses.

Key Words: Whole Genome Sequencing (WGS), Canine Mammary Tumors (CMT), inherited risk, germline mutation, purebred dogs

2.2 Plain English Summary

Despite the advances in sequencing technology and the success of previous canine whole genome sequencing research, we know of no other publications that report using whole genome sequencing to investigate a genetic risk (aka. a risk that can be passed down through generations) for canine mammary tumors in purebred dogs. This canine cancer type is comparable to human breast cancer, and as a result, genes that are known to influence inherited risk for breast cancer were investigated to determine if those same genes played a role in risk for dogs. We whole genome sequenced 14 purebred dogs from four different breeds; each of the dogs within a breed had been tied back to the same family tree (pedigree). From this study, we have identified mutations in genes *BRCA2* and *STK11* that could increase risk for those dogs with the mutations. These mutations seem to be present in some breeds more than other, thus affecting risk differently. Furthermore, the large dataset from this research allows for further exploration to find additional mutations in other dogs that influence their risk for canine mammary tumors.

2.3 Background

The practice of breeding dogs for specific characteristics and traits has resulted in over 190 phenotypically diverse breeds, according to the American Kennel Club.[81] Defined as selective breeding, this practice has cultivated breed-specific gene pools that not only contribute towards each breed's defining features but also disease susceptibility. To date, over 450 canine genetic diseases have been reported, many of which are monogenic and limited to a specific dog breed(s).[55; 82; 83] Efforts to understand the genetic causes of such diseases began in the 1980s with the first canine genetic mutation identified in 1989 for hemophilia B, an X-linked disorder.[84] Since then, investigating hereditary diseases that segregate in purebred lines/pedigrees have fostered numerous genetic discoveries; over 130 canine hereditary diseases are now genetically explained.[55; 82; 83] Through these discoveries, it has been determined that there is much genetic overlap between canine and human disease. Importantly, the elucidation of certain hereditary canine diseases has even led to breakthroughs in human medicine, with disease such as sleep disorders, *Birt-Hogg-Dubé syndrome* and more.[68; 83; 85-87]

Interestingly, despite the fact that some dog breeds appear to have an increased risk of certain cancer types, little is known about the etiology. Although, in general, cancer is considered a multifactorial disease, clustering of particular cancers in pedigrees suggests a genetic predisposition.[88] In humans, the study of cancer families has revealed genetic mutations that severely increase lifetime risk of developing particular cancers; for instance, high-risk mutations in *BRCA1*, *BRCA2*, *CDH1*, *PTEN*, *STK11,* and *TP53* all result in hereditary cancer syndromes (such as hereditary breast cancer syndrome, Li Fraumeni syndrome and Cowden Syndrome) associated with an increased risk of breast cancer as well as other cancer types.[89] Therefore, breed or kennel/pedigree-based studies should be a beneficial approach to determine cancer genetic risk in dogs. This approach was successful in identifying the susceptibility locus for multifocal renal cystadenocarcinoma and nodular dermatofibrosis (RCND) in a German Shepherd pedigree.[87] RCND, an inherited cancer syndrome, is an autosomal dominantly inherited trait that is caused by a mutation in the *Birt-Hogg-Dubé* (*BHD*) gene, which is named after the equivalent human cancer syndrome.[72; 90; 91] Similar to how the *BHD* mutation in German Shepherds predisposes them to RCND, there are likely yet-to-be-discovered mutations that explain particular cancer incidences in other breeds.

With little known about canine mammary tumor (CMT) genetic susceptibility,[80] we decided to carry out whole genome sequencing (WGS) on 14 purebred dogs diagnosed with CMT from four different breeds (Golden Retriever, Siberian Husky, Dalmatian, and Standard Schnauzer). The CMT-affected dogs from each breed were linked back to a common ancestor through pedigree analysis. Even though it is highly debated as to which dog breeds have the greatest CMT susceptibility or prevalence, we hypothesized that a cluster of CMT in these pedigrees is indicative of a genetic predisposition. Previous attempts to study CMT genetics either focused on small cohorts of multiple breeds or English Springer Spaniels (ESS).[80] Multiple studies have indicated that the ESS from Sweden is a high-risk breed; however, it is worth noting that dogs in Sweden are rarely spayed - a procedure known to greatly reduce the risk of CMT.[74; 92; 93] Nevertheless, studying ESSs has revealed apparent CMT-associated SNVs, including ones in *BRCA1* and *BRCA2*, but the causative alleles have yet to be identified.[94-96] To our knowledge, there have been no published reports of WGS to investigate CMT inherited-genetic risk. Furthermore, outside of ESS, there have been no breed-specific analyses. Considering that different WGS efforts in dogs have recently proven to be advantageous in elucidating genetic susceptibility to disease,[35; 49; 50; 52; 66; 97] differences in body types,[98] as well as adaptions against parasites,[99] we have compiled and processed WGS data to begin the exploration of breed-specific CMT-risk alleles and, in this initial report, specifically reveal the coding variants detected in orthologs of the high-risk human breast cancer susceptibility genes.

2.4 Materials and Methods

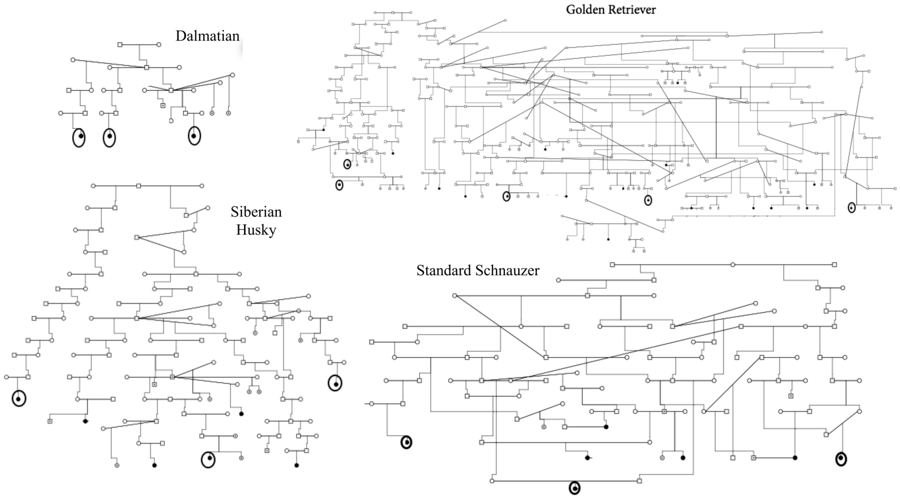
2.4.1 CMT sample acquisition: DNA from 85 purebred CMT-affected dogs, representing 32 different American Kennel Club (AKC) recognized breeds (Table 2.1), was obtained from the Canine Health Information Center (CHIC) DNA Repository, which is a part of the Orthopedic Foundation for Animals (OFA; https://www.ofa.org/about/dna-repository). Briefly, this repository stores canine DNA samples and corresponding genealogic and phenotypic information to facilitate genetics research. Dog owners submit either blood or buccal samples to the repository along with their pets' health history. Ultimately, researchers request access to samples pertaining to a disease of interest along with any additional information submitted. An unfortunate limitation of this resource is the lack of collected data. Being reliant on the owner's knowledge and willingness to share, along with a generic survey used for all collected samples/phenotypes, information such as CMT pathology/histology, age of onset, and spay/neuter status were not provided to the research team.

Table 2.1: The total number of DNA samples from CMT-affected dogs obtained from the CHIC repository

Breed	**Dogs per Breed**	**Dogs Connected to a Common Ancestor**
Akita	1	--
Alaskan Malamute	2	--
Australian Cattle Dog	2	2
Beauceron	1	--
Bichon Frise	2	--
Border Terrier	1	--
Bouvier des Flandres	1	--
Boxer	1	--
Bullmastiff	1	--
Chesapeake Bay Retriever	1	--
Collie	1	--
Dalmatian	3	3
Doberman Pinscher	3	3
French Bulldog	1	--
Golden Retriever	18	18
Gordon Setter	4	4
Great Pyrenees	1	--
Irish Setter	2	2
Keeshond	2	2
Kerry Blue Terrier	1	--
Kuvasz	1	--
Leonberger	1	--
Mastiff	1	--
Newfoundland	4	4
Parson Russell Terrier	1	--
Pembroke Welsh Corgi	2	2
Petit Basset Griffon des Vendeen	2	--
Schipperke	1	--
Siberian Husky	8	7
Standard Schnauzer	7	7
Welsh Springer Spaniel	5	5
Total dogs	**82**	**59**
Total breeds	**32**	**12**

Of the 85 acquired samples, both blood-extracted DNA and buccal swabs were obtained. DNA was purified from the provided buccal swabs using the QIAamp DNA Mini Kit (Cat No./ID: 51304). Of the 32 represented breeds, 15 had multiple samples per breed (Table 2.1); thus, pedigree analyses were performed to identify breed-specific common ancestors and determine the level of relationship. Specifically, a dog's registration and breeding information were entered into online (and mainly breed-specific) databases to build pedigrees. From this, 12 different pedigrees were generated.

Figure 2.1 Purebred dog pedigrees and selected samples for WGS.



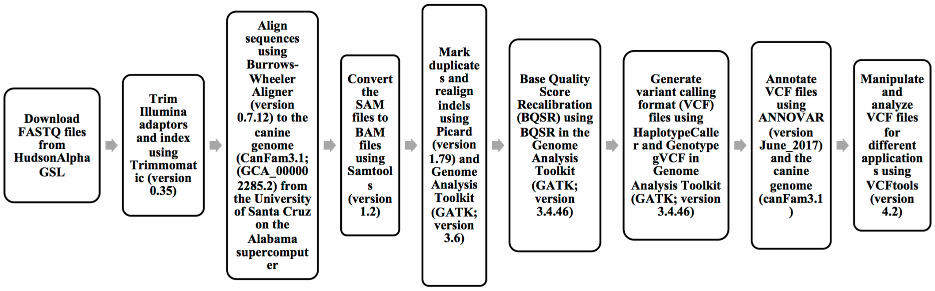
Pedigree symbols: square = male; circle = female; filled black symbol = CMT-affected individual with DNA sample in the Merner laboratory; dot in symbol = DNA is available for this individual in the CHIC repository; larger circle around an individual = CMT-affected samples sent for WGS

Offspring of WGS samples are not depicted here, See Additional File 1 for offspring information.

2.4.2 Sequencing and bioinformatics: Fourteen DNA samples from four pedigrees were chosen for WGS. This included five Golden Retriever samples (three female, two male), three Siberian Husky females, three Standard Schnauzer females, and three Dalmatian females (Figure 2.1). The selected dogs from each breed were AKC-registered and located within the same pedigree. Also, utilizing the CHIC database, offspring information of each dog was recorded to attempt to determine intact status (spay or neuter status) as hormone exposure can affect the likelihood of development of CMT (Additional File 2.1). Samples were prepared for Illumina platform WGS at HudsonAlpha Institute for Biotechnology's Genome Sequencing Laboratory and the sequencing was carried out on Illumina HiSeq X. Paired FASTQ files were obtained from

HudsonAlpha with sequencing data for each sample; the quality of the raw FASTQ files was determined using FASTQC. After assuring quality files, this sequencing data was carried through an in-house bioinformatics canine pipeline that was adapted from the Genome Analysis Toolkit (GATK) best practices bioinformatics pipeline (Figure 2.2).[100] In brief, each sample file had Illumina adapters trimmed using the program Trimmomatic.[101] Samples were then aligned to the canine genome CanFam3.1[18] using BWA mem[102]. Duplicate reads were marked using a Picard tool from version 1.79 (http://broadinstitute.github.io/picard/.); then indels were realigned and base quality scores were recalibrated referencing the CanFam3.1 dbsnp data using Base Quality Score Recalibrator (BQSR) as part of the GATK v.3.4.46.[103] Additionally, using GATK, coverage was calculated using the Depth of Coverage tool and genomic variant calling format (GVCF) files were generated using Haplotype Caller and then merged through genotyping GVCF files. ANNOVAR[104] was used to annotate the VCF files using gene prediction from Ensembl build version 75. Variants were filtered by a Quality by Depth threshold of at least 12.

Figure 2.2 Bioinformatics pipeline for canine WGS data: This pipeline has been adapted from GATK's Best Practice Pipeline for use on canine WGS data.


Download FASTQ files from HudsonAlpha GSL
Trim Illumina adaptors and index using Trimmomatic (version 0.35)
Align sequences using Burrows-Wheeler Aligner (version 0.7.12) to the canine genome (CanFam3.1; (GCA_000002285.2) from the University of Santa Cruz on the Alabama supercomputer
Convert the SAM files to BAM files using Samtools (version 1.2)
Mark duplicates and realign indels using Picard (version 1.79) and Genome Analysis Toolkit (GATK; version 3.6)
Base Quality Score Recalibration (BQSR) using BQSR in the Genome Analysis Toolkit (GATK; version 3.4.46)
Generate variant calling format (VCF) files using HaplotypeCaller and Genotype gVCF in Genome Analysis Toolkit (GATK; version 3.4.46)
Annotate VCF files using ANNOVAR (version June_2017) and the canine genome (canFam3.1)
Manipulate and analyze VCF files for different applications using VCFtools (version 4.2)

Coding variants within orthologs of human breast cancer susceptibility genes were isolated using the following coordinates: *BRCA1(*ENSCAFT00000023190.4*)*:chr9:19960910-20024390, *BRCA2(*ENSCAFT00000010309.3):chr25:7734450-7797815, *CDH1*(ENSCAFT00000032333.3): chr5:80759112-80834940, *PTEN*(ENSCAFT00000024821.3): chr26:37853135-37913097, *STK11*(ENSCAFT00000031055.3): chr20:57556289-57625288, and *TP53*(ENSCAFT00000026465.3): chr5:32560598-32574109. All coding variants identified through WGS were validated through PCR and Sanger sequencing. Once the variant list was finalized, protein sequences for the orthologous human genes (*BRCA1* (NP_009231*)*, *BRCA2* *(*NP_000050), *CDH1* (NP_004351), *PTEN* (NP_000305), *STK11* (NP_000446), and *TP53* (NP_000537)) were compared to the canine protein sequences (that corresponded to the above canine gene accession numbers) through EMBOSS Water alignment (https://www.ebi.ac.uk/Tools/psa/emboss_water/). These alignments were used to determine the corresponding human amino acid of each coding variant. The ClinVar database was then checked to see if a human mutation was identified in that position.[105]

2.4.3 Controls: Control data was obtained through Ensembl by accessing each canine gene's variant table,[106] which reports population genetic information from the European Variation Archive (EVA; https://www.ebi.ac.uk/eva/?eva-study=PRJEB24066). EVA provides data from the "High quality variant calls from multiple dog genome project – Run 1" representing WGS data of over 200 dogs from multiple breeds. Variants were filtered based on GATK's best practices filtering guidelines, and the resulting variants and corresponding frequencies are accessible on the web through Ensembl's database. Exact breed and sex information of these 200 dogs was unknown. This EVA control dataset is similar to the use of publically available databases that present general population control data for human disease genetic studies.[107-111]

2.4.4 Statistical Analyses: For all the *BRCA1, BRCA2, CDH1, PTEN, STK11,* and *TP53* coding variants validated in the 14 CMT cases, allele frequencies were calculated in both cases and controls. Major and minor alleles were defined based on EVA control data. Subsequently, the Fisher Exact test was carried out to determine any statistically significant allele frequency differences between the EVA controls and the overall CMT cohort, as well as each specific breed. The Fisher Exact test, a test of contingency tables that calculates statistical significance based on a probability scale, is typically used as a statistical test for allele frequency.[107; 112] This statistical analysis method has been considered a solution for analysis with small cell counts, which is why this analysis method was chosen for our analyses.[113] P-values were calculated using Fisher Exact test in R (v 3.5.1), which were not adjusted for multiple testing.

2.5 Results

2.5.1 Sequencing and Annotation: WGS of the 14 dogs yielded an average sequencing depth of 26.0X (Table 2). On average, 99.13% of the reads aligned to the reference, resulting in 99.7%, 99.1%, 96.1% and 75.6% of the genome being covered at least 1X, 5X, 10X, and 20X, respectively (Table 2.2). Altogether, the total number of unique variant calls was 17,867,633, comprised of 12,071,269 single nucleotide variants (SNVs) and 4,081,564 indels. An average of 7,909,896 variants were called for each dog, the majority of which were non-coding, with an average of 40,965 coding variants per dog. The overall average sequencing depth of the exome, according to Ensembl build version 75, was 25.6X; 99.8%, 99.4%, 96.7%, and 76.0% of the exome was covered at least 1X, 5X, 10X, and 20X, respectively (Additional File 2.2).

Table 2.2: Whole genome coverage summary.

Sample	**Number of Mapped Reads to canFam3.1**	**% of Reads Mapped to canFam3.1**	**Average Sequencing Depth**	**% of bases covered greater than or equal to:**						
				1X	**10X**	**20X**	**25X**	**50X**	**75X**	**100X**
Dal 1	432,798,423	99.0	29	99.7	98.9	92.3	73.6	1.2	0.5	0.3
Dal 2	479,265,395	99.1	24.6	99.7	98.7	79.0	45.9	0.8	0.4	0.2
Dal 3	517,919,216	99.1	27.4	99.7	98.8	89.1	64.7	1.0	0.4	0.3
GoldenR 1	514,850,463	99.2	29	99.7	98.9	92.4	73.3	1.2	0.5	0.3
GoldenR 2	521,394,202	99.3	29.2	99.7	98.9	92.7	74.5	1.2	0.5	0.3
GoldenR 3	469,958,383	99.1	26.9	99.7	98.1	85.4	62.1	1.0	0.4	0.3
GoldenR 4	420,815,898	99.0	23.4	99.6	97.2	72.4	39.3	0.7	0.3	0.2
GoldenR 5	435,936,648	99.2	24.3	99.7	98.5	77.1	43.9	0.8	0.4	0.2
SibH 1	439,440,441	99.2	25.1	99.7	98.6	81.3	49.3	0.9	0.4	0.2
SibH 2	676,505,498	99.2	35.1	99.7	99.0	97.3	92.3	2.9	0.7	0.4
SibH 3	306,005,622	99.2	16	99.6	91.7	18.0	3.6	0.4	0.2	0.1
StandSch 1	233,378,490	99.3	12.1	99.6	71.0	3.9	1.0	0.2	0.1	0.1
StandSch 2	716,837,887	98.7	36.6	99.7	99.2	97.6	93.5	4.2	0.8	0.5
StandSch 3	444,186,080	99.1	25	99.7	98.6	80.2	48.3	0.9	0.4	0.2
Average	472,092,332	99.1	26	99.7	96.1	75.6	54.7	1.2	0.4	0.3

2.5.2 Variant analyses: A total of 19 coding variants, 13 nonsynonymous and six synonymous, were detected in *BRCA1*, *BRCA2*, *CDH1*, *PTEN*, *STK11,* and *TP53* (Table 2.3; Additional File 2.3). The nonsynonymous variants included ten missense variants (only one of which was considered possibly damaging based on Polyphen analysis), two non-frameshifting deletions, one non-frameshifting indel (Table 2.3; Additional File 2.3). Of the 19 total variants, 11 had been previously reported in CMT canine cohorts (Table 2.3). Three *STK11* missense variants were identified (Table 2.3), one of which was detected in a single breed (Additional File 2.3). These three *STK11* variants have yet to be reported, not only in CMT studies, but also in the EVA control dataset (Table 2.3). Consequently, they appear to be associated with an increased risk of CMT and each variant may affect breeds differently (Table 2.3 and 2.4). Additionally, significant P-values were generated for *BRCA2* variants (Table 2.3 and 2.4). Variants in other genes were noted but did not appear to be associated with disease.

Table 2.3: Summary of canine coding variants found within orthologs of human breast cancer susceptibility genes.

Gene	RS ID Number	Variant Name	Protein Name	Variant Type	Polyphen Score	MAF in EVA Control Cohort (%)	MAF in CMT cases Cohort (%)	P-values (Total CMT Cases versus EVA Controls)	Initially Reported - CMT Heritability Study (Reference #)
BRCA1: ENSCAFT00000043953.1	rs397509570**	c.G3075A**	p.S1025S**	synonymous	NA	49.3	46.4	0.8465	*Borge et al.* 2011 *(46)*
BRCA2: ENSCAFT00000010309.3	rs23250374	c.A428G	p.H143R	missense	BENIGN	25.7	42.9	0.0749	*Yoshikawa et al.* 2008 *(47)*
	rs850935038**	c.T1158G**	p.C386W**	missense	BENIGN	20.6	42.9	**0.0095**	*Yoshikawa et al.* 2008 *(47)*
	rs851104585**	c.C2144A**	p.P715Q**	missense	BENIGN	0	0	1	-
	rs852009320**	c.C2154A**	p.S718S**	synonymous	NA	0	0	1	-
	rs851813778**	c.C2183T**	p.A728V**	missense	BENIGN	0	0	1	-
	rs851048998**	c.A2222G**	p.N741S**	missense	BENIGN	0	0	1	-
	rs23244160	c.A2401C	p.K801Q	missense	POSSIBLY DAMAGING	31.2	14.3	0.0868	*Borge et al.* 2011 (46)
	rs8676219	c.A4304G	p.K1435R	missense	BENIGN	25.9	42.9	0.0758	*Yoshikawa et al.* 2008 *(47)*
	rs397511123	c.6918_6920delGTT	p.L2307del	In frame deletion	NA	31.2	14.3	0.0868	*Borge et al.* 2011 *(46)*

	rs23255542	c.C6930T	p.F2310F	synonymous	NA	28.9	42.9	0.1359	*Yoshikawa et al.* 2008 *(47)*
	rs853007536**	c.9995_9996insAAA**	p.M3332delinsIK**	indel	NA	20.9	42.9	**0.0162**	*Yoshikawa et al.* 2005 *(48)*
CDH1: ENSCAFT00000032333.3	rs852509306	c.387_389delCCA	p.129delH	In frame deletion	NA	18.9	17.9	1	*Borge et al.* 2011 *(46)*
	rs397512866	c.C945T	p.S315S	synonymous	NA	12.3	14.3	0.7659	*Borge et al.* 2011 *(46)*
	rs851557759	c.A2448G	p.E816E	synonymous	NA	8.6	3.6	0.7187	-
PTEN: ENSCAFT00000024821.3	rs397513087	c.C909T	p.L303L	synonymous	NA	3.7	7.1	0.2970	*Borge et al.* 2011 *(46)*
STK11: ENSCAFT00000031055.3	-	c.C109T^	p.P37S^	missense	UNKNOWN	0	3.6	0.0654	-
	-	c.A286G^	p.M96V^	missense	BENIGN	0	10.7	**0.0003**	-
	-	c.T293C^	p.F98S^	missense	BENIGN	0	10.7	**0.0003**	-
TP53: ENSCAFT00000026465.3	no mutations were found								

Table 2.4: Significant breed-specific P-values for nonsynonymous variants

Gene	RS ID Number	Variant Name	Protein Name	Polyphen Score	MAF in EVA Control Cohort (%)	CMT Cohort									
						Total Cohort		Breed Specific							
								Dalmatian		Golden Retriever		Siberian Husky		Standard Schnauzer	
						MAF (%)	P-value	MAF (%)	P-value	MAF (%)	P-value	MAF (%)	P-value	MAF (%)	P-value
BRCA2: ENSCAFT00000010309.3	rs23250374	c.A428G	p.H143R	BENIGN	25.7	42.9	0.0749	33.3	0.6506	70	**0.0048**	0	0.3447	50	0.1847
	rs850935038**	c.T1158G**	p.C386W**	BENIGN	20.6	42.9	**0.0095**	50	0.1107	10	0.6948	83.3	**0.0096**	50	0.1107
	rs853007536**	c.9995_9996insAAA**	p.M3332delinsIK**	NA	20.9	42.9	**0.01621**	50	0.1136	10	0.6950	83.3	**0.0006**	50	0.1136
STK11: ENSCAFT00000031055.3	-	c.C109T^	p.P37S^	UNKNOWN	0	3.6	0.0654	16.7	**0.0148**	0	1	0	1	0	1
	-	c.A286G^	p.M96V^	BENIGN	0	10.7	**0.0003**	0	1	0	1	33.3	**0.0002**	16.7	**0.0148**
	-	c.T293C^	p.F98S^	BENIGN	0	10.7	**0.0003**	0	1	0	1	33.3	**0.0002**	16.7	**0.0148**

** Major allele corresponds to the alternate allele, not the reference allele (based on EVA control data)

^ P-values for these variants were generated following the assumption that 200 of the control dogs were successfully sequenced in this location, and no mutations were identified

2.6 Discussion

In an effort to study CMT heritability, our group acquired germline DNA from CMT-affected purebred dogs whose samples were submitted to the CHIC repository by the owner. Based on the hypothesis that dogs from the same breed/lineage share ancestral CMT-genetic risk factors, WGS was carried out on 14 samples from four generated pedigrees, including Golden Retriever, Siberian Husky, Standard Schnauzer, and Dalmatian. However, it is important to note that even if our hypothesis holds true in future studies that validate our findings or through novel CMT-gene discovery efforts, some cases within each pedigree could be phenocopies, representing sporadic cases not due to a familial genetic variant. This has to be kept in mind since ages of onset were not available through CHIC and early ages of onset are associated with hereditary risk.

Our CMT-affected cohort represents dogs from the United States and did not include any ESS, which is the only breed to date that has had breed-specific CMT-genetic analyses.[94-96] To our knowledge, there have been no published reports of WGS to investigate the inherited risk of CMT. However, a compilation of next-generation sequencing efforts was used to compare human breast tumors to CMTs and somatic mutations were identified.[74] Additionally, a limited number of studies have investigated germline CMT risk, and only a few risk variants have been identified with significance.[80] On our initial quest to find inherited breed-specific CMT-risk alleles, it is important to note that all CMT-affected dogs chosen for WGS were female except two closely related Golden Retrievers males. In addition to family history, male breast cancer is a hallmark of hereditary breast cancer in humans;[89] in fact, genetic predisposition significantly elevates the risk of male breast cancer, which is otherwise rare in the general population.[114] Therefore, assuming CMT genetic risk is similar to human genetic susceptibility, these two CMT-affected males suggest genetic factors are playing a role and were selected to enhance the prospects of discovery.

Unlike human disease gene discovery efforts, which have capitalized on whole-exome sequencing (WES) to facilitate discovery upon the introduction of next-generation sequencing,[115] WGS has been the methodology of choice for identifying the genetic factors associated with inherited canine diseases. WGS and WES involve the re-sequencing of a genome or exome, respectively, which was made possible for canines once the first reference genome was published in 2005.[1] In 2013, the first WGS[41; 116; 117] and WES[118] studies identified mutations associated with inherited canine disorders. Since that time, despite improvements to canine exome designs,[30] the use of WES lagged behind. A possible reason for this is the cost. From our experience, when determining which of the two sequencing approaches to take for this study, the cost of WES was surprisingly expensive. WES baits alone were ~$1000 per sample, which was the total cost per sample for WGS (to yield an average sequencing depth of at least 15X). Additional benefits to WGS include, (a) avoiding technical enrichment biases associated with exome sequencing capture, (b) more uniformity regarding sequencing-quality parameters, (c) the ability to explore both coding and non-coding regions, (d) the ability to better detect variants in coding regions (including regions targeted by a WES kit), and (e) the continued usefulness of the data as the annotation of the canine reference genome improves and gaps are filled.[18; 34; 35; 119; 120]

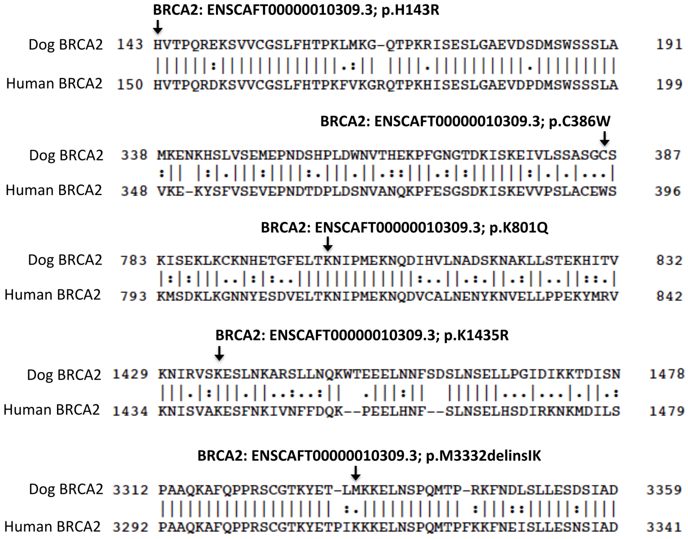
Upon WGS of the 14 CMT-affected dogs, individual average sequencing depths ranged from 12.1 to 36.6X and overall averaged 26.0X. Aiming to achieve an average sequencing depth of, at least, 15X, all but one dog yielded such results (Table 2.2). Ultimately, the overall average was comparable to other canine WGS studies using Illumina technology.[49; 52; 66; 97] On average, 99.7% of the genome was covered at least 1X, which is comparable to the Illumina-generated data in Gilliam *et al*.[49] Noteworthy, it was higher than Sayyab *et al.* who used Ion Proton technology and reported an average sequencing depth of 9.2X and that 96% of the genome was covered at least 1X.[35] Viluma *et al*., who carried out another Ion Proton study, determined that 80% of the genome was covered at least 4X;[38] this is vastly different from our data, which covered 99.1% of the genome at 5X or greater. Similar to the two Ion Proton studies, our group also sought to determine the coverage of the canine exome through our WGS efforts. Not only did our study produce greater coverage for the canine genome, we additionally determined higher coverage results for the canine exome. Previously, Sayyab *et al.* reported that 91% of the exome was covered at least 1X, and Viluma *et al*. reported 77% of the exome was covered at least 4X. Contrarily, we obtained 99.8% and 99.4% of the exome at 1X and 5X, respectively (Additional File 2.2). In fact, these results far surpass the 5X coverage noted by Broeckx *et al.* regarding their improved canine exome design; they stated that just over 90% of the targeted base pairs were covered at least 5X.[30] Furthermore, Broeckx *et al.* had an average sequencing depth of 68.3X, which emphasizes the issue of lack of uniformity regarding targeted captures.

On average, each of the 14 dogs had 7.9 million variants called. Overall, this is comparable to the number of variants reported in the WGS studies that had similar sequencing depths.[14; 49; 66; 97] The majority of the variants were non-coding, which, in the future, provides data for exploration. However, for this study, we focused on coding variants, specifically in orthologs of high-risk human breast cancer susceptibility genes, *BRCA1, BRCA2, CDH1, PTEN, STK11, and TP53,* [89] as an initial gene exclusion approach, acknowledging that this dataset will be subsequently analyzed to investigate risk in other coding and non-coding regions of the genome. Through our initial analysis, 19 different coding variants were identified through WGS and confirmed through PCR and Sanger sequencing (Table 2.3). Interestingly, this list of variants gave insight regarding the complications of next-generation sequencing in dogs. Using a reference sequence derived from a Boxer for the alignment and, similarly, gene transcripts derived from the latest assembly for the annotation, we noted instances when the data could have easily been misconstrued. Firstly, four *BRCA2* variants were homozygous in all 14 CMT-affected dogs. This observation hinted that each alternate allele could in fact be the true wild-type (major) allele for the species since the four reference alleles appear to be unique to the Boxer. This was confirmed when we determined that all EVA control dogs were also homozygous for the four alternate alleles, as well as when we compared the Boxer reference protein sequence to the BRCA2 protein sequence for the Basenji (Basenji-breed-1.1) and the dingo dog (ASM325472v1). The reference genome is of an unaffected female Boxer, but that is the difficulty when studying a disease with age-related risk. These four *BRCA2* variants, with alleles that appear to be extremely rare in the species according to the control data, need to be further investigated to determine if they contribute toward disease risk in the Boxer. Unfortunately, we did not sequence any Boxers in this study, but their assessment would require a careful analysis of controls to properly interpret the data, which stresses that analyzing controls from multiple breeds can have extreme benefits.

Similar to the example above, there were other instances where the alternate allele in the Boxer was in fact the major allele in controls. This was the case for two *BRCA2* variants that appear to be associated with CMT risk, particularly in the Siberian Huskies. According to the Boxer reference sequence and annotation using transcript ID ENSCAFT00000010309.3, these two variants were named *BRCA2*:c.T1158G (p.C386W) and *BRCA2*:c.9995_9996insAAA (p.M3332delinsIK), which were previously reported in CMT heritability studies.[121-124] Thus, the Boxer had a cysteine at amino acid 386 and a methionine at 3332. However, interestingly, the major allele in the EVA control dogs translated to most dogs having a tryptophan at amino acid

386, and isoleucine-lysine at position 3332, which also resembles that of the references for Basenji dog breed and the dingo dog and, most interestingly, corresponds to the conserved human residues. Comparing allele frequencies between the CMT cases and EVA controls revealed that cysteine at amino acid 386 and a methionine at 3332 were associated with an increased risk of CMT. In fact, these alleles appear to be most strongly associated with CMT risk in Siberian Huskies (Table 2.4). These associations will need to be validated by studying larger cohorts. Boxers should also be studied to determine the true allele frequencies in that breed. If a cysteine at position 386 and a methionine at 3332 are actually more common in Boxers, they could be at an elevated disease risk. Noteworthy, the human BRCA2 residue W395 corresponds to W386 in these dogs (Figure 2.3), and while a cysteine mutation at W395 has not been found in human hereditary breast cancer cases, two pathogenic truncation mutations have been reported at that position (ClinVar Variation IDs: 266612 and 265053), along with the missense variant, W395G, which is considered a variant of unknown significance (VUS; ClinVar Variation ID: 51078).[105] Similarly, human BRCA2 residues I3312 and K3313 correspond to the conserved isoleucine-lysine in dogs at 3332 (Figure 3), and BRCA2 p.I3312M has been reported as another VUS (ClinVar Variation ID: 52921).[105] VUS are defined as genetic variants for which there is no clear association with disease risk, and it has been reported that as many as 15% of people who undergo *BRCA1* and *BRCA2* genetic screening are informed of a detected VUS.[125]

Figure 2.3 BRCA2 dog and human protein alignment for non-synonymous variants previously reported in CMT heritability studies.


```
                 BRCA2: ENSCAFT00000010309.3; p.H143R
                 ↓
Dog BRCA2    143 HVTPQREKSVVCGSLFHTPKLMKG-QTPKRISESLGAEVDSDMSWSSSLA   191
                 ||||||:|||||||||||||.:|| ||||.||||||||||.||||||||
Human BRCA2  150 HVTPQRDKSVVCGSLFHTPKFVKGRQTPKHISESLGAEVDPDMSWSSSLA   199

                          BRCA2: ENSCAFT00000010309.3; p.C386W
                                                                ↓
Dog BRCA2    338 MKENKHSLVSEMEPNDSHPLDWNVTHEKPFGNGTDKISKEIVLSSASGCS   387
                 :|| |:|.|||:||||:.|||.||.::|||.:|:||||||:|.|.|...|
Human BRCA2  348 VKE-KYSFVSEVEPNDTDPLDSNVANQKPFESGSDKISKEVVPSLACEWS   396

                     BRCA2: ENSCAFT00000010309.3; p.K801Q
                                   ↓
Dog BRCA2    783 KISEKLKCKNHETGFELTKNIPMEKNQDIHVLNADSKNAKLLSTEKHITV   832
                 |:|:|||..|:|:..||||||||||||:..||.:.||.:||..||::.|
Human BRCA2  793 KMSDKLKGNNYESDVELTKNIPMEKNQDVCALNENYKNVELLPPEKYMRV   842

                   BRCA2: ENSCAFT00000010309.3; p.K1435R
                      ↓
Dog BRCA2   1429 KNIRVSKESLNKARSLLNQKWTEEELNNFSDSLNSELLPGIDIKKTDISN  1478
                 |||.|:|||.||..:..:||  .|||:||  ||||||...|...|.||.:
Human BRCA2 1434 KNISVAKESFNKIVNFFDQK--PEELHNF--SLNSELHSDIRKNKMDILS  1479

                 BRCA2: ENSCAFT00000010309.3; p.M3332delinsIK
                                     ↓
Dog BRCA2   3312 PAAQKAFQPPRSCGTKYET-LMKKELNSPQMTP-RKFNDLSLLESDSIAD  3359
                 ||||||||||||||||||| :.||||||||||| :|||::|||||:||||
Human BRCA2 3292 PAAQKAFQPPRSCGTKYETPIKKKELNSPQMTPFKKFNEISLLESNSIAD  3341
```

In addition to *BRCA2* c.T1158G (p.C386W), we identified three other *BRCA2* missense variants that had been previously reported in CMT studies assessing inherited risk; this included *BRCA2*:c.A428G (p.H143R), *BRCA2*:c.A2401C (p.K801Q), and *BRCA2*:c.A4304G (p.K1435R; Table 3).[80; 122; 124] Even though neither of these variants generated a significant P-value when investigating the overall CMT cohort, those P-values appeared to be trending towards significance. Nonetheless, breed-specific analyses suggested that BRCA2 p.H143R is associated with CMT-risk in Golden Retrievers (Table 2.4). This variant was previously described as possibly damaging by Borge *et al.*,[124; 126] but PolyPhen2 analysis predicts it to be benign.[127] Similarly, contradictory pathogenicity predictions were noted for BRCA2 p.K801Q. It was predicted to be possibly damaging using PolyPhen2 but was initially reported by Borge *et al.* in 2011 and predicted to be benign.[124; 126] Moreover, the Polyphen2-suggested benign variant, p.K1435R, was reported by Yoshikawa *et al.* in 2008 as possibly deleterious upon blood and CMT analyses, including loss-of-heterozygosity studies.[122] Altogether, knowing that current computational prediction methods misclassify a significant percentage of clinically valid missense variants,[128] and that the P-values generated for those variants were, at least, trending towards significance, larger genotyping and functional studies will be required for true classification. Additionally, all three missense variants are conserved in humans (Figure 2.3), and, most interestingly, the equivalent mutations of canine p.H143R and p.K1435R have been identified in humans as p.H150R and p. K1440R, respectively (ClinVar Variation IDs: 51657 and 51632).[105] These variants are classified as VUS, similar to the other *BRCA2* VUS mentioned above. Overall, VUS include missense variants as well as in-frame insertions and deletions, both of which were detected in this study; this overlap with human and dogs offers another avenue for exploration since the reclassification of VUS is a current hot topic.[129; 130]

Regarding the other assessed genes, *STK11* displayed the most interesting results. Three missense variants were identified, *STK11* c.C109T (p.P37S), *STK11* c.A286G (p.M96V), and *STK11* c.T293C (p.F98S), all of which appear to play a role in CMT risk. Our findings suggest that *STK11* is a CMT susceptibility gene, corroborating a similar claim in a recent publication by Canadas *et al.*[131] Canadas and colleagues suggested that the minor allele (T) of rs22928814, which lies within an intron of *STK11,* was associated with an increased risk of CMT. Interestingly, this allele, which the authors reported to have a frequency of 25.7% and 14.9% in cases and controls, respectively,[131] has a frequency of 26.6% in EVA controls according to

Ensembl,[106] which is more similar to the frequency reported in the CMT cases and stresses the need for validation studies. Of note, this variant was not detected in any of the CMT-affected dogs sequenced in this study. However, the three missense variants identified in this study appear to be extremely rare alleles since they were not reported in EVA controls. Regarding STK11 p.M96V and p.F98S, breed-specific P-values of 1.824E-04 and 0.01478 were generated for the Siberian Huskies and Standard Schnauzers, respectively (Table 2.4). Additionally, STK11 p.P37S was only detected in one Dalmatian and breed-specific analyses suggests that this variant possibly increases risk of CMT in that breed. Overall, these findings mimic the phenomena in humans that rare *STK11* variants increase risk of disease.[89] However, it is worth noting that these variants are not in a conserved region with human STK11 protein sequence. How these *STK11* variants, along with the identified *BRCA2* variants, specifically contribute towards risk needs to be further studied. Firstly, variants in both *STK11* and *BRCA2* appear to be tightly linked, thus determining the true risk alleles in both *BRCA2* and *STK11* is important. Also, polygenic risk assessment in humans is another hot topic,[132] and demonstrating the same concept in dogs would further validate their usefulness as a model of hereditary breast cancer.[80]

2.7 Conclusions

To our knowledge, we carried out the first study to assess inherited CMT risk through WGS data analysis, and we investigated risk through both multiple breed and breed-specific analyses. This manuscript specifically reports the variants detected in six orthologs of high-risk human breast cancer susceptibility genes as an initial gene exclusion approach, acknowledging that this WGS dataset will be subsequently analyzed to investigate risk in other coding and non-coding regions of the genome. Through our initial efforts, we identified variants in *BRCA2* and *STK11* that appear to be associated with CMT risk. These variants could alter risk in many breeds but appear to be more prevalent in some breeds compared to others. Additionally, we identified several *BRCA2* variants that correspond to VUS in humans. Indeed, these results need to be validated; the identified variants now require further investigation to determine the role they play in risk in both humans and dogs, which we plan to promptly address. For instance, noting the limitation of using a control dataset of multiple unknown breeds, we plan to acquire control samples to determine breed-specific allele frequencies. Furthermore, functional studies are pertinent to determine pathogenicity. Ultimately, in addition to this initial gene exclusion effort, this dataset provides the opportunity for novel discovery and has the potential to lead to further breakthroughs in canine and human breast cancer research through comparative analyses. Overall, in the era of personalized medicine, identifying risk variants not only provides better risk assessment and opportunities to selectively breed out a pathogenic mutation, it also can provide insight towards disease mechanism and aid in the development of targeted therapies.[88; 133]

2.8 Additional Files

Additional File 2.1:

Table summary of offspring information from CHIC repository for the 14 WGS samples. (XLSX 10kb)

Breed	Sample	Sex	Total # of Litters	Age at First Reported Litter	Age at Last Reported Litter
Siberian Husky	SibH 1	F	None reported in CHIC	-	-
	SibH 2	F	1	4	4
	SibH 3	F	None reported in CHIC	-	-
Dalmatian	Dal 1	F	6	2	7
	Dal 2	F	2	3	6
	Dal 3	F	1	3	3
Golden Retreiver	GoldR 1	F	1	5	5
	GoldR 2	F	1	3	3
	GoldR 3	M	Male (1 litter)	2	2
	GoldR 4	M	Male (2 litter)	3	4
	GoldR 5	F	None reported in CHIC	-	-
Standard Schnauzer	StandSch 1	F	1	6	6
	StandSch 2	F	3	2	4
	StandSch 3	F	1	4	4

Additional File 2.2:

Table of exome Coverage Summary for the 14 canines sequenced. Exome regions according to Ensembl build version 75 for CanFam3.1. (XLSX 12kb)

Sample	**Average Sequencing Depth**	**% of bases covered greater than or equal to:**								
		1X	**5X**	**10X**	**15X**	**20X**	**25X**	**50X**	**75X**	**100X**
Dal 1	28.7	99.8	99.6	99.4	98.4	93.1	74.0	0.6	0.2	0.1
Dal 2	24.2	99.8	99.6	99.1	95.9	79.3	45.6	0.3	0.1	0
Dal 3	27.0	99.8	99.6	99.3	97.9	89.7	64.8	0.5	0.1	0.1
Golden R 1	28.6	99.8	99.7	99.7	98.4	93.1	73.6	0.6	0.2	0.1
Golden R 2	28.8	99.8	99.6	99.4	98.4	93.3	74.6	0.6	0.2	0.1
Golden R 3	26.6	99.8	99.6	98.7	95.4	86.4	62.5	0.5	0.1	0.1
Golden R 4	23.1	99.8	99.5	97.9	92.0	73.1	39.2	0.3	0.1	0
Golden R 5	23.9	99.8	99.6	99.0	95.2	77.2	43.3	0.3	0.1	0
Sib H 1	24.7	99.8	99.6	99.6	96.4	81.7	49.1	0.4	0.1	0.1
Sib H 2	34.6	99.8	99.6	99.5	99.1	98.0	92.9	2.1	0.3	0.1
Sib H 3	15.6	99.8	99.3	92.1	57.0	17.2	2.9	0.1	0	0
Stand Sch 1	11.9	99.7	97.6	70.8	22.8	3.2	0.5	0.1	0	0
Stand Sch 2	35.9	99.8	99.7	99.5	99.2	98.1	93.9	3.4	0.3	0.1
Stand Sch 3	24.5	99.8	99.6	99.1	95.9	80.3	47.7	0.4	0.1	0
Average	25.6	99.8	99.4	96.7	88.7	76.0	54.6	0.7	0.1	0.1

Additional File 2.3:

Table details of canine coding variants found within orthologs of human breast cancer susceptibility genes. (XLSX 20kb)

Gene	**RS ID Number**	**Variant Name**	**Protein Name**	**Variant Type**	**Polyphen Score**	**Mutation Carriers per breed**																**Cases - Allele and Genotype Frequencies in total CMT Cohort**					**Controls - Allele and Genotype Frequencies in EVA Study (PRJEB24066)**				
						Dalmation			**Golden Retriever**					**Siberian Husky**			**Standard Schnauzer**			**Allele Frequency (Allele: Frequency (Count))**		**Genotype Frequency (Genotype: Frequency (Count))**			**Allele Frequency (Allele: Frequency (Count))**		**Genotype Frequency (Genotype: Frequency (Count))**		
						Dal 1	**Dal 2**	**Dal 3**	**Gold R 1**	**Gold R 2**	**Gold R 3***	**Gold R 4***	**Gold R 5**	**Sib H 1**	**Sib H 2**	**Sib H 3**	**StandSch 1**	**StandSch 2**	**StandSch 3**	**Minor**	**Major**	**Minor Homozygous**	**Heterozygous**	**Major Homozygous**	**Minor**	**Major**	**Minor Homozygous**	**Heterozygous**	**Major Homozygous**
BRCA1: ENSCAFT00000023190.4	rs397509570**	c.G3075A**	p.S1319S**	synonymous	NA	HOM	HET	HOM	-	HET	HET	-	HET	HOM	HOM	HOM	-	-	HET	**G**: 0.464 (13)	**A**: 0.536 (15)	**G|G**: 0.286 (4)	**A|G**: 0.357 (5)	**A|A**: 0.357 (5)	**G**: 0.493 (213)	**A**: 0.507 (219)	**G|G**: 0.352 (76)	**A|G**: 0.282 (61)	**A|A**: 0.366 (79)
BRCA2: ENSCAFT00000010309.3	rs23250374	c.A428G	p.H143R	missense	BENIGN	HET	-	HET	HET	HET	HET	HOM	HOM	-	-	-	HET	HET	HET	**C**: 0.429 (12)	**T**: 0.571 (16)	**C|C**: 0.143 (2)	**C|T**: 0.571 (8)	**T|T**: 0.286 (4)	**C**: 0.257 (111)	**T**: 0.743 (321)	**C|C**: 0.125 (27)	**C|T**: 0.264 (57)	**T|T**: 0.611 (132)
	rs850935038**	c.T1158G**	p.C386W**	missense	BENIGN	HET	HET	HET	HOM	HET	HOM	HOM	HOM	-	HET	-	HET	HET	HET	**A**: 0.429 (12)	**C**: 0.571 (16)	**A|A**: 0.143 (2)	**A|C**: 0.571 (8)	**C|C**: 0.285 (4)	**A**: 0.206 (90)	**C**: 0.794 (346)	**A|A**: 0.060 (13)	**A|C**: 0.294 (64)	**C|C**: 0.647 (141)
	rs851104585**	c.C2144A**	p.P715Q**	missense	BENIGN	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	0	**T**: 1.000 (28)	0	0	**T|T**: 1.000 (14)	0	**T**: 1.000 (436)	0	0	**T|T**: 1.000 (218)
	rs852009320**	c.C2154A**	p.S718S**	synonymous	NA	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	0	**T**: 1.000 (28)	0	0	**T|T**: 1.000 (14)	0	**T**: 1.000 (436)	0	0	**T|T**: 1.000 (218)
	rs851813778**	c.C2183T**	p.A728V**	missense	BENIGN	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	0	**A**: 1.000 (28)	0	0	**A|A**: 1.000 (14)	0	**A**: 1.000 (436)	0	0	**A|A**: 1.000 (218)
	rs851048998**	c.A2222G**	p.N741S**	missense	BENIGN	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	0	**C**: 1.000 (28)	0	0	**C|C**: 1.000 (14)	0	**C**: 1.000 (436)	0	0	**C|C**: 1.000 (218)
	rs23244160	c.A2401C	p.K801Q	missense	POSSIBLY DAMAGING	-	HET	-	HET	-	HET	-	-	-	HET	-	-	-	-	**G**: 0.143 (4)	**T**: 0.857 (24)	**G|G**: 0 (0)	**G|T**: 0.286 (4)	**T|T**: 0.714 (10)	**G**: 0.312 (136)	**T**: 0.688 (300)	**G|G**: 0.133 (29)	**G|T**: 0.358 (78)	**T|T**: 0.509 (111)
	rs8676219	c.A4304G	p.K1435R	missense	BENIGN	HET	-	HET	HET	HET	HET	HOM	HOM	-	-	-	HET	HET	HET	**C**: 0.429 (12)	**T**: 0.571 (16)	**C|C**: 0.143 (2)	**C|T**: 0.571 (8)	**T|T**: 0.286 (4)	**C**: 0.259 (113)	**T**: 0.741 (323)	C|C: 0.119 (26)	**C|T**: 0.280 (61)	**T|T**: 0.601 (131)

	rs397511123	c.6918_6920delGTT	p.L2307del	In frame deletion	NA	-	HET	-	HET	-	HET	-	-	-	HET	-	-	-	-	- : 0.143 (4)	**AAC**: 0.857 (24)	-|-: 0 (0)	-|**AAC**: 0.286 (4)	**AAC|AAC**: 0.714 (10)	-: 0.312 (136)	**AAC**: 0.688 (300)	-|-: 0.133 (29)	-|**AAC**: 0.358 (78)	**AAC|AAC**: 0.509 (111)
	rs23255542	c.C6930T	p.F2310F	synonymous	NA	HET	-	HET	HET	HET	HET	HOM	HOM	-	-	-	HET	HET	HET	**A**: 0.429 (12)	**G**: 0.571 (16)	**A|A**: 0.143 (2)	**A|G**: 0.571 (8)	**G|G**: 0.286 (4)	**A**: 0.289 (114)	**G**: 0.711 (280)	**A|A**: 0.127 (25)	**A|G**: 0.325 (64)	**G|G**: 0.548 (108)
	rs853007536**	c.9995_9996insAAA**	p.M3332delinsIK**	indel	NA	HET	HET	HET	HOM	HET	HOM	HOM	HOM	-	HET	-	HET	HET	HET	-: 0.429 (12)	**TTT**: 0.571 (16)	-|-: 0.143 (2)	-|**TTT**: 0.571 (8)	**TTT|TTT**: 0.286 (4)	-: 0.209 (91)	**TTT**: 0.791 (345)	-|-: 0.064 (14)	-|**TTT**: 0.289 (63)	**TTT|TTT**: 0.647 (141)
CDH1: ENSCAFT00000032333.3	rs852509306	c.387_389delCCA	p.129delH	In frame deletion	NA	-	-	-	HET	-	HET	-	-	-	HET	-	-	HOM	-	-: 0.179 (5)	**TGG**: 0.821 (23)	-|-: 0.071 (1)	-|**TGG**: 0.214 (3)	**TGG|TGG**: 0.714 (10)	-: 0.189 (81)	**TGG**: 0.804 (345)	-|-: 0.079 (17)	-|**TGG**: 0.219 (47)	**TGG|TGG**: 0.693 (149)
	rs397512866	c.C945T	p.S315S	synonymous	NA	-	-	-	HET	-	HET	-	-	-	-	-	-	HOM	-	**A**: 0.143 (4)	**G**: 0.857 (24)	**A|A**: 0.071 (1)	**A|G**: 0.143 (2)	**G|G**: 0.786 (11)	**A**: 0.123 (53)	**G**: 0.877 (379)	**A|A**: 0.051 (11)	**A|G**: 0.144 (31)	**G|G**: 0.806 (174)
	rs851557759	c.A2448G	p.E816E	synonymous	NA	-	-	-	-	-	-	-	-	-	HET	-	-	-	-	**C**: 0.036 (1)	**T**: 0.964 (27)	**C|C**: 0 (0)	**C|T**: 0.071 (1)	**T|T**: 0.929 (13)	**C**: 0.086 (37)	**T**: 0.914 (395)	**C|C**: 0.032 (7)	**C|T**: 0.106 (23)	**T|T**: 0.861 (186)
PTEN: ENSCAFT00000024821.3	rs397513087	c.C909T	p.L303L	synonymous	NA	-	-	-	HET	-	-	-	-	-	-	-	-	-	HET	**T**: 0.071 (2)	**C**: 0.929 (26)	**T|T**: 0 (0)	**C|T**: 0.071 (1)	**C|C**: 0.929 (13)	**T**: 0.037 (16)	**C**: 0.963 (420)	**T|T**: 0.018 (4)	**C|T**: 0.037 (8)	**C|C**: 0.945 (206)
STK11: ENSCAFT00000031055.3	-	c.C109T^	p.P37S^	missense	UNKNOWN	-	HET	-	-	-	-	-	-	-	-	-	-	-	-	**A**: 0.036 (1)	**G**: 0.964 (27)	**A|A**: 0 (0)	**A|G**: 0.071 (1)	**G|G**: 0.929 (13)	0	0	0	0	0
	-	c.A286G^	p.M96V^	missense	BENIGN	-	-	-	-	-	-	-	-	-	HET	HET	-	-	HET	**C**: 0.107 (3)	**T**: 0.893 (25)	**C|C**: 0 (0)	**C|T**: 0.214 (3)	**T|T**: 0.786 (11)	0	0	0	0	0
	-	c.T293C^	p.F98S^	missense	BENIGN	-	-	-	-	-	-	-	-	-	HET	HET	-	-	HET	**G**: 0.107 (3)	**A**: 0.893 (25)	**G|G**: 0 (0)	**A|G**: 0.214 (3)	**A|A**: 0.786 (11)	0	0	0	0	0
TP53: ENSCAFT00000026465.3	no mutations were found					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Chapter 3: *CEACAM* gene family mutations associated with inherited breast cancer risk - a comparative oncology approach to discovery

This work is submitted for publication in Frontiers in Genetics in 2021: Anna L.W. Huskey, Isaac McNeely, Nancy D. Merner*

3.1 Abstract

Introduction: Recent studies comparing canine mammary tumors (CMTs) and human breast cancers have revealed remarkable tumor similarities, identifying shared expression profiles and acquired mutations. CMTs can also provide a model of inherited breast cancer susceptibility in humans; thus, we investigated breed-specific whole genome sequencing (WGS) data in search for novel CMT risk factors that could subsequently explain inherited breast cancer risk.
Methods: WGS was carried out on five CMT-affected Gold Retrievers. Protein truncating variants (PTVs) within human orthologs detected in all five samples were validated and genotyped in the 13 remaining CMT-affected Golden Retrievers. Allele frequencies were compared to canine controls. Subsequently, human blood-derived exomes from The Cancer Genome Atlas breast cancer cases were analyzed and allele frequencies were compared to Exome Variant Server ethnic-matched controls.
Results: *Carcinoembryonic Antigen-related Cell Adhesion Molecule 24* (*CEACAM24*) c.247dupG;p.(Val83Glyfs*48) was the only validated variant and had a frequency of 66.7% amongst the 18 Golden Retrievers with CMT. This was significant compared to the European Variation Archive (*p-value* 1.52x10^{-8}) and non-Golden Retriever American Kennel Club breeds (*p-value* 2.48x10^{-5}). With no direct ortholog of *CEACAM24* in humans but high homology to all CEACAM gene family proteins, all human *CEACAM* genes were investigated for PTVs. A total of six and sixteen rare PTVs were identified in African and European American breast cancer cases, respectively. Single variant assessment revealed five PTVs associated with breast cancer risk. Gene-based aggregation analyses revealed that rare PTVs in *CEACAM6*, *CEACAM7*, and *CEACAM8* are associated with European American breast cancer risk, and rare PTVs in *CEACAM7* are associated with breast cancer risk in African Americans. Ultimately, rare PTVs in the entire *CEACAM* gene family are associated with breast cancer risk in both European and African Americans with respective *p-values* of 1.75x10^{-13} and 1.87x10^{-04}.

Conclusion: This study reports the first association of inherited *CEACAM* mutations and breast cancer risk, and potentially implicates the whole gene family in genetic risk. Precisely how these mutations contribute to breast cancer needs to be determined; especially considering our current knowledge on the role that the *CEACAM* gene family plays in tumor development, progression, and metastasis.

Keywords: breast cancer, canine mammary tumor (CMT), *CEACAM*, whole genome sequencing (WGS), comparative oncology, inherited risk, rare protein truncating variants (PTVs), splice mutations

3.2 Introduction

Breast cancer is a serious health concern. Amongst both sexes, it globally ranks as the second most commonly diagnosed type of cancer and the second leading cause of cancer-related deaths, accounting for ~2.1 million new diagnoses and 626,679 deaths in 2018.[134] Worldwide, it is also the most common cancer diagnosed in women and the overall leading cause of cancer-related female deaths.[134] In the United States, 2020 estimates predicted breast cancer to be the leading site of new cancer diagnoses in women and the second leading cause of cancer-related deaths, resulting in 276,480 new diagnoses and 42,170 deaths.[135] Advances in breast cancer research have translated to better disease screening, diagnosis, and treatment, but new research questions continuously arise as time and medical needs progress.[136]

Comparative oncology, which is the study of cancer biology and therapy in spontaneous, naturally-occurring cancers in companion animals, provides valuable models of human cancer that have and will continue to make research advances.[71] Recent studies comparing canine mammary tumors (CMTs) and human breast cancers have revealed notable tumor similarities, identifying shared expression profiles and acquired mutations.[74-79] CMTs can also provide a model of hereditary breast cancer susceptibility genes in humans, especially considering similar genetics and familial clustering.[79; 80] While most CMT studies investigating inherited risk have focused on identifying genetic variants in orthologs of known human breast cancer risk genes,[80; 137] in this study, we investigate breed-specific whole genome sequencing (WGS) data in search for novel CMT risk factors. WGS studies have been used to make numerous disease gene discoveries in dogs, many of which clearly translated to human health.[35; 49; 50; 53; 66; 97] Taking a similar approach, we identified a *Carcinoembryonic Antigen-related Cell Adhesion Molecule 24*

(*CEACAM24)* protein-truncating variant (PTV) in a Golden Retriever CMT pedigree, which ultimately revealed that rare PTVs in the *CEACAM* gene family are associated with breast cancer risk in humans.

3.3 Materials and Methods

3.3.1 Golden Retriever pedigree and WGS:

As previously described by Huskey *et al*., blood- or buccal-derived DNA samples were obtained from 18 CMT-affected Golden Retrievers from the Canine Health Information Center (CHIC) DNA repository, and a pedigree was constructed linking all 18 dogs in one large pedigree [137]. Five of those Golden Retrievers (three female, two male) were selected for WGS and the data was processed through a bioinformatics pipeline [137]. Upon alignment to the CanFam3.1 reference genome and annotation using gene predictions from Ensembl build version 75, a script was written to isolate PTVs found in all five Golden Retriever samples. PTVs were defined as single nucleotide variants (SNVs) that resulted in a premature stop codon or abrogated a splice site, and small insertions or deletions (indels) that changed a transcript's reading frame. Upon filtering, the genes with PTVs were classified into two different groups, orthologs of human genes or non-orthologs. Polymerase chain reaction (PCR) and Sanger sequencing were carried out to validate the PTVs in human orthologs. *CEACAM24* c.247dupG;p.(Val83Glyfs*48) was the only validated variant. Following validation, the 13 remaining CMT-affected Golden Retrievers underwent PCR and Sanger sequencing to determine their mutation status.

3.3.2 Canine controls:

As a convenient, publically available, online canine genetic variant repository, the European Variation Archive (https://www.ebi.ac.uk/eva/?eva-study=PRJEB24066) was initially used to note the allele frequency of *CEACAM24* c.247dupG;p.(Val83Glyfs*48). The European Variation Archive provides high quality WGS variant calls of over 200 dogs from multiple breeds (breed and sex information was unknown). The data was obtained through Ensembl by accessing the canine gene's 'Variant table' under 'Genetic Variation'; for a particular variant, 'Population genetics' information was given, including European Variation Archive allele frequencies [106]. Furthermore, additional splicing, frame-shifting, and stop gain mutations within the other dog *CEACAM* genes were investigated through Ensembl transcripts (*CEACAM16*:

ENSCAFT00000044174; *CEACAM18*: ENSCAFT00000004587; *CEACAM20*: ENSCAFT00000047731; *CEACAM24*: ENSCAFT00000047960; *CEACAM28*: ENSCAFT00000022623). *CEACAM1*, *CEACAM23*, and *CEACAM30* did not have variant information available in Ensembl for European Variation Archive data.

Through the CHIC repository, blood or buccal-swab derived DNA from purebred, American Kennel Club registered dogs were randomly selected and obtained to determine the frequency of *CEACAM24* c.247dupG;p.(Val83Glyfs*48). This included DNA from Golden Retrievers (n=87), as well as 13 other breeds, including Petit Basset des Griffon (n=10), Gordon Setter (n=8), Australian Cattle Dog (n=10), Siberian Husky (n=10), Dalmatian (n=10), Irish Setter (n=9), Welsh Pembroke Corgi (n=10), Standard Schnauzer (n=10), Newfoundland (n=10), Keeshond (n=10), Great Dane (n=8), Doberman Pinscher (n=10), and Boxer (n=10). PCR and Sanger sequencing were carried out to determine *CEACAM24* c.247dupG;p.(Val83Glyfs*48) genotypes of each dog.

3.3.3 Canine statistical analyses:

Upon determining *CEACAM24* c.247dupG;p.(Val83Glyfs*48) allele frequencies, p-values were generated using the Fisher's Exact Test in R (v 3.5.1), comparing allele differences in Golden Retriever to control dogs, including both European Variation Archive and CHIC DNA samples.

3.3.4 Dog and human CEACAM protein analyses:

EMBOSS water alignment[138] was carried out to determine the level of homogeneity between the dog CEACAM24 protein and other dog and human CEACAM proteins. Additionally, InterPro[139] and the Eukaryotic Linear Motif (ELM) resource[140] were used to identify CEACAM domains and binding motifs, respectively.

3.3.5 *Human CEACAM gene analysis – The Cancer Genome Atlas*

Due to the homogeneity of the CEACAM gene family and no direct ortholog of dog *CEACAM24* in humans, all human *CEACAM* family genes were investigated for rare PTVs in The Cancer Genome Atlas (TCGA) breast cancer cohort. Investigating inherited risk, only blood-derived exomes of breast cancer cases were analyzed. Overall, whole-exome binary sequence alignment mapping (BAM) files were downloaded using the Genomic Data Commons (GDC) Data Portal Repository through approved research project #10805. To acquire the samples, the specific filters under the 'Cases' category included: Project (TCGA-BRCA), Samples Sample Type (Blood Derived Normal), and Race ('Black or African American' and 'White'). The samples were further filtered under the 'Files' category, including Experimental Strategy (WXS) and Data Format (BAM). A total of 170 sample files were obtained for African Americans and 650 for European Americans. These files were downloaded using the GDC Data Transfer Tool (version 1.2.0). Only individuals with known ages of breast cancer onset were used in this study; as a result, one European American and two African American BAM files were removed from further bioinformatics processing and statistical analysis.

The downloaded BAM files, which had previously been aligned to the hg38 human reference genome, were processed using the remaining steps of a pipeline adapted from the Genome Analysis Toolkit's (GATK's) best practices pipeline.[100] Base quality scores were recalibrated using BaseRecalibrator and then HaplotypeCaller was used to generate genome variant calling format (gVCF) files (GATK version 3.6). GenotypeGVCFs was used to merge the individual gVCF files based on ethnicity (GATK version 3.6). The European American files were merged in batches of approximately 200 using GATK's (version 3.6) CombineGVCFs prior to merging into a single VCF file with GenotypeGVCFs. The two ethnic specific VCF files were then processed through a variant quality score recalibration using VariantRecalibrator (GATK version 3.6), and, as recommended, SNVs were filtered using a pass filter of 99.5%, and indels were filtered using a slightly lower pass filter of 99.0%.[100] Variants in *CEACAM1* (NM_001184815; chr19:42507306-42528481), *CEACAM3* (NM_001815 at chr19:41796587-41811554), *CEACAM4* (NM_001817; chr19:41618971-41627074), *CEACAM5* (NM_004363; chr19:41708626-41730421), *CEACAM6* (NM_002483; chr19:41755530-41772210), *CEACAM7* (NM_006890; chr19:41673303-41688270), *CEACAM8* (NM_001816 at chr19:42580243-42594924), *CEACAM16* (NM_001039213; chr19:44699151-44710718), *CEACAM18* (NM_001278392; chr19:51478643-51490605), *CEACAM19* (NM_020219; chr19:44671452-44684355), *CEACAM20* (NM_001102597; chr19:44506159-44529675), and *CEACAM21* (NM_001098506; chr19:41576166-41586844) were then extracted from the ethnic specific VCF files and annotated using ANNOVAR (version June2017). Variants were filtered to include rare

PTVs with ethnic-specific minor allele frequencies of <1% in Exome Variant Server (EVS; National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project).[108]

3.3.6 Human statistical analyses:

Using the Fisher's exact test[141; 142] in R (v 3.5.1), individual PTVs were assessed to compare allele frequency differences between ethnic-specific TCGA breast cancer cases and EVS controls. The Fisher's method was used for gene-based and gene family-based aggregation analyses.[143; 144] The R tool 'sumlog' (in the 'metap' package) was used to combine *p-values* for each aggregation test. To accommodate for the one-sided nature of the Fisher exact test p-values, compliments of *p-values* in the opposite direction were used in the calculations for the Fisher's method aggregation analyses.

3.3.7 Human mutation analysis:

Mutalyzer was used to determine the effect of frame-shifting and nonsense variants on the coded protein.[145] Human splicing mutations that affected non-protein-coding exons of the mRNA, specifically in the 3' untranslated region (UTR), were analyzed using the miRDB tool to identify microRNA binding sites potentially lost due to a splicing mutation.[146] For each gene harboring a splice mutation affecting non-protein-coding exons, microRNA binding sites within the 3'UTR with a target score of ≥80 were noted. The top five ranked microRNA targets were investigated for previous cancer (specifically, hereditary breast and ovarian cancer (HBOC) syndrome) associations.

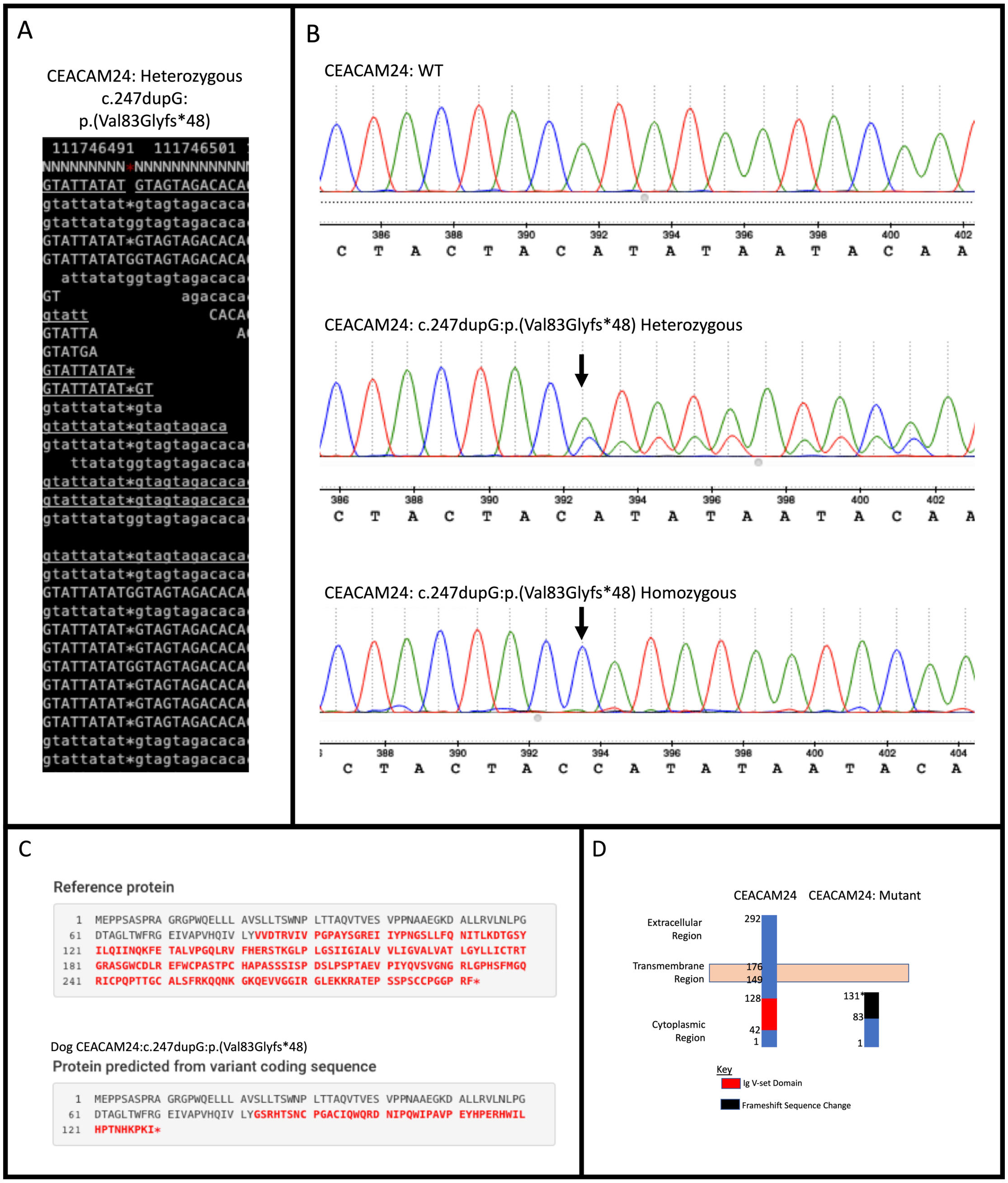
3.4 Results

Upon filtering the WGS data, 12 different PTVs were detected in all five Golden Retrievers, four of which were within human orthologs. Only one PTV, a frame-shifting mutation in *CEACAM24* (c.247dupG;p.(Val83Glyfs*48)) was validated (Figure 3.1). This mutation had a frequency of 66.7% amongst the 18 Golden Retrievers with CMT in this study (Table 3.1).

Table 3.1: *CEACAM24* c.247dupG; p.(Val83Glyfs*48) genotypes and allele frequencies

Data Set / Cohort	**Dog Breed**	**# of Dogs**	**# of HOM**	**# of HET**	**Minor Allele Frequency**	**P-value for Comparison to CMT affected Golden Cohort**
CMT Affected	Golden Retriever	18	6	9	66.7	**-**
CHIC USA Breed Specific Controls	Golden Retriever	87	42	34	67.8	0.3334
CHIC USA Non-Golden Retriever Controls	Petit Basset Griffon Vendeen	10	7	2	80.0	**2.48x10^{-5}**
	Gordon Setter	8	5	2	75.0	
	Australian Cattle Dog	10	4	2	50.0	
	Siberian Husky	10	4	1	45.0	
	Dalmatian	10	3	2	40.0	
	Irish Setter	9	0	1	5.6	
	Welsh Pembroke Corgi	10	0	0	0.0	
	Standard Schnauzer	10	0	0	0.0	
	Newfoundland	10	0	0	0.0	
	Keeshond	10	0	0	0.0	
	Great Dane	8	0	0	0.0	
	Doberman Pinscher	10	0	0	0.0	
	Boxer	10	0	0	0.0	
	Totals & Avg MAF of CHIC Non-Golden Retriever Controls	125	23	10	22.4	
European Variation Archive Controls	European General Dog Population	196	12	44	17.3	**1.52x10^{-8}**

Figure 3.1: *CEACAM24* (c.247dupG; p.(Val83Glyfs*48)) mutation summary; (**A**) samtools tview image capture of the mutation in a WGS CMT-affected Golden Retriever; (**B**) Sanger sequencing results of validation in CMT-affected Golden Retriever cohort depicting wildtype (WT), heterozygous, and homozygous sequences at the mutation location; (**C**) Mutalyzer prediction of the change in protein squence with frameshifting mutation; (**D**) Depiction of the WT and mutated protein and lost regions and domains of the dog CEACAM24 protein with the frameshift mutation.



```
 111746491  111746501
NNNNNNNNN*NNNNNNNNNNNNN
GTATTATAT GTAGTAGACACA
gtattatat*gtagtagacaca
gtattatatggtagtagacaca
GTATTATAT*GTAGTAGACACA
GTATTATATGGTAGTAGACACA
  attatatggtagtagacaca
GT              agacaca
gtatt              CACA
GTATTA                A
GTATGA
GTATTATAT*
GTATTATAT*GT
gtattatat*gta
gtattatat*gtagtagaca
gtattatat*gtagtagacaca
   ttatatggtagtagacaca
gtattatat*gtagtagacaca
gtattatat*gtagtagacaca
gtattatatggtagtagacaca

gtattatat*gtagtagacaca
gtattatat*gtagtagacaca
GTATTATATGGTAGTAGACACA
gtattatat*gtagtagacaca
GTATTATAT*GTAGTAGACACA
GTATTATAT*GTAGTAGACACA
GTATTATATGGTAGTAGACACA
GTATTATAT*GTAGTAGACACA
GTATTATAT*GTAGTAGACACA
GTATTATAT*GTAGTAGACACA
gtattatat*gtagtagacaca
gtattatat*gtagtagacaca
```

Reference protein

```
  1  MEPPSASPRA GRGPWQELLL AVSLLTSWNP LTTAQVTVES VPPNAAEGKD ALLRVLNLPG
 61  DTAGLTWFRG EIVAPVHQIV LYVVDTRVIV PGPAYSGREI IYPNGSLLFQ NITLKDTGSY
121  ILQIINQKFE TALVPGQLRV FHERSTKGLP LGSIIGIALV VLIGVALVAT LGYLLICTRT
181  GRASGWCDLR EFWCPASTPC HAPASSSISP DSLPSPTAEV PIYQVSVGNG RLGPHSFMGQ
241  RICPQPTTGC ALSFRKQQNK GKQEVVGGIR GLEKKRATEP SSPSCCPGGP RF*
```

Dog CEACAM24:c.247dupG:p.(Val83Glyfs*48)
Protein predicted from variant coding sequence

```
  1  MEPPSASPRA GRGPWQELLL AVSLLTSWNP LTTAQVTVES VPPNAAEGKD ALLRVLNLPG
 61  DTAGLTWFRG EIVAPVHQIV LYGSRHTSNC PGACIQWQRD NIPQWIPAVP EYHPERHWIL
121  HPTNHKPKI*
```

Upon comparing that frequency to the 17.3% allele frequency in the European Variation Archive, a *p-value* of 1.52x10^{-8} was generated. Representing dogs from another continent and not knowing the breeds of the European Variation Archive, the frequency of *CEACAM24* c.247dupG;p.(Val83Glyfs*48) was subsequently determined in different American Kennel Club breeds (Table 3.1). There was no statistically significant difference between Golden Retriever CMT cases and controls. However, there was a significant difference between Golden Retrievers cases and other American Kennel Club breeds (2.48x10^{-5}; Table 3.1). The *CEACAM24* c.247dupG;p.(Val83Glyfs*48) allele frequency ranged from 0-80% in the assessed breeds (Table 1). *CEACAM24* c.247dupG;p.(Val83Glyfs*48) abolishes the extracellular region, the transmembrane domain, and part of the cytoplasmic region, including the Ig V-set domain (Figure 3.1C & D).

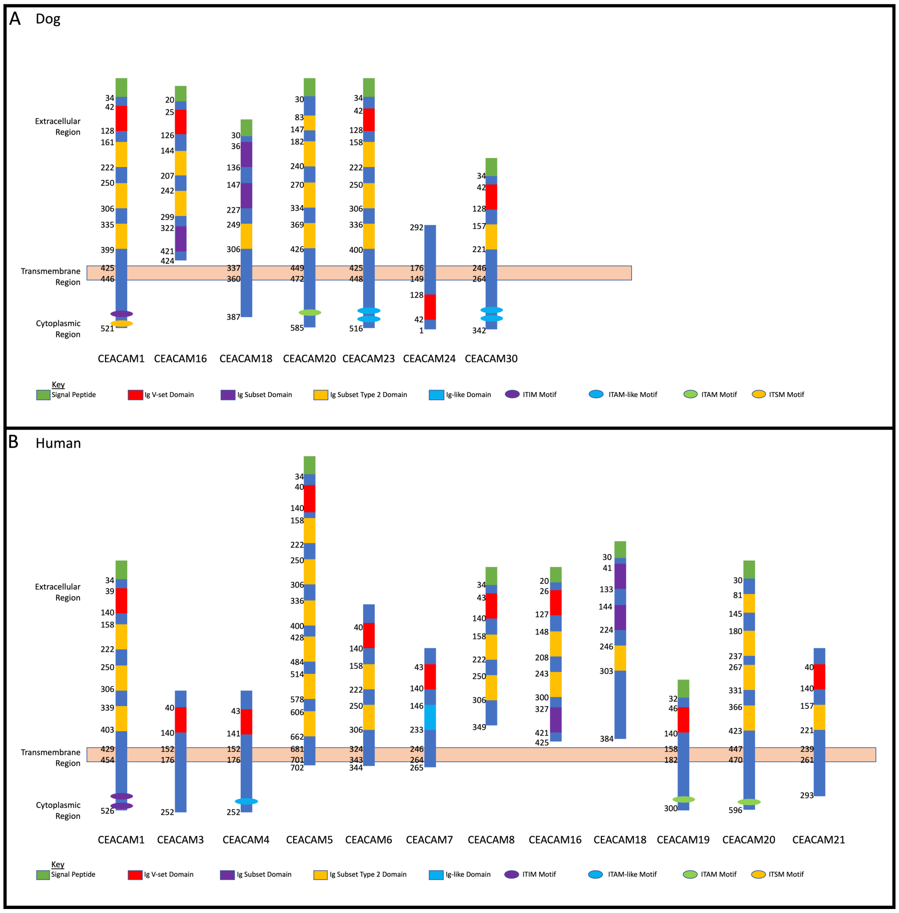
Homology analysis revealed that the dog CEACAM proteins were, on average, 43.7% similar to the dog CEACAM24 protein (Table 3.2 and Figure 3.2A). Similarly, there were many related functional domains and high homology between the dog CEACAM24 protein and the human CEACAM proteins, averaging 51.9% similarity (Table 3.2 and Figure 3.2). This homology, along with the fact that there is no direct human ortholog of dog *CEACAM24*, prompted all human *CEACAM* genes (Figure 2B) to be investigated for rare PTVs in the TCGA breast cancer cohort.

Table 3.2: Homology of Dog and Human CEACAM proteins to Dog CEACAM24 Protein

Species	Gene Name	Protein Accession	% Identity	% Similarity
Dog	CEACAM1	NP_00101026	52.2	58.4
	CEACAM16	ENSCAFP00000039084	22.5	37.7
	CEACAM18	ENSCAFP00000058450	19.3	32.5
	CEACAM20	ENSCAFP00000036293	21.2	31.9
	CEACAM23	NP_001091021	38.4	40.8
	CEACAM24	NP_001091023	100	100
	CEACAM28	NP_001091015	42.2	46.3
	CEACAM30	NP_001091022	53.6	58.3
Average of all Dog CEACAM proteins compared to Dog CEACAM24 (excluding CEACAM24 from analysis)			35.6	43.7
Human	CEACAM1	NP_001171744	53.1	60.8
	CEACAM3	NP_001806	47	58.2
	CEACAM4	NP_001808	50.4	63.4
	CEACAM5	NP_004354	53.2	61

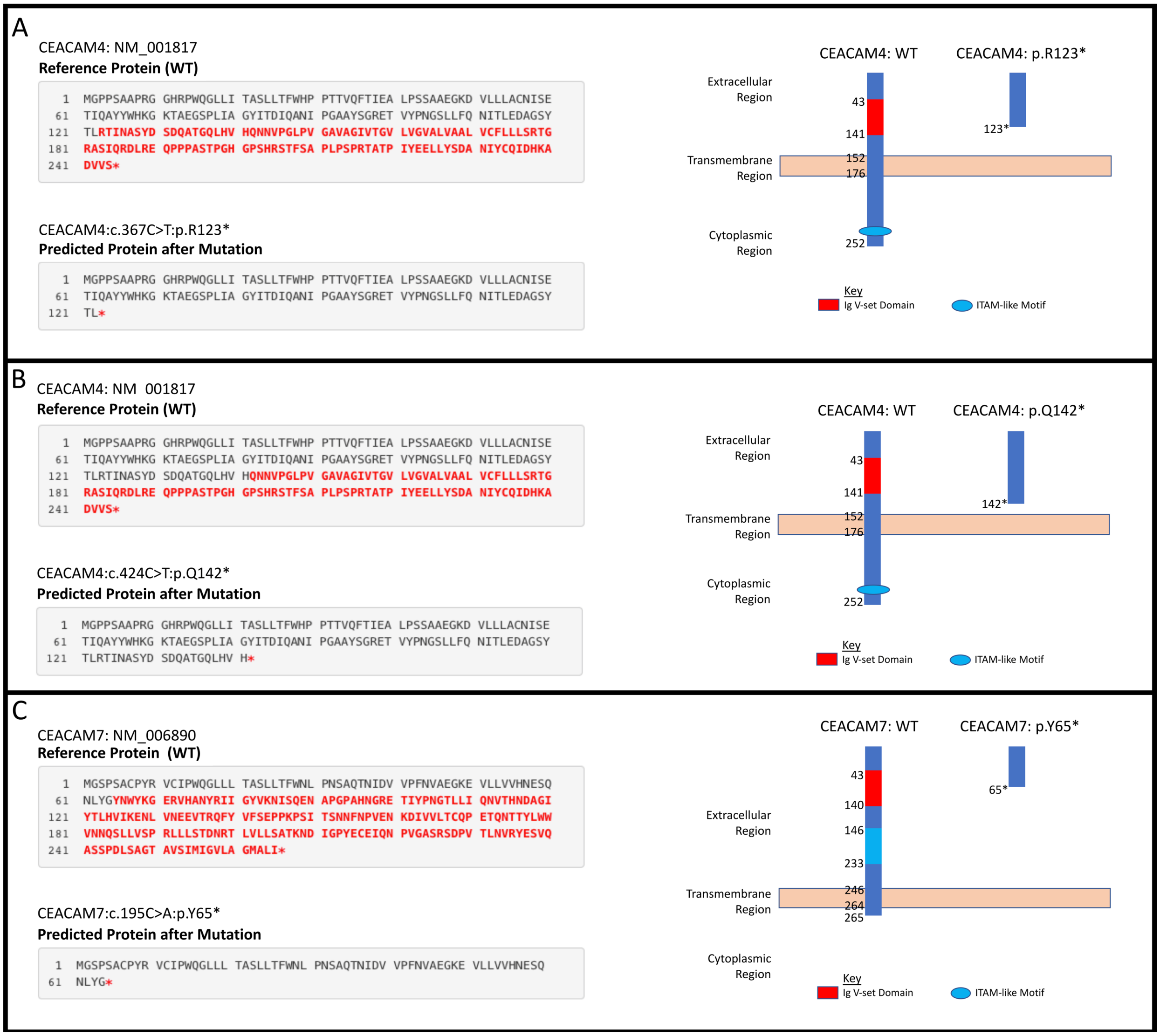
	CEACAM6	NP_002474	37.8	48
	CEACAM7	NP_008821	45.1	58.3
	CEACAM8	NP_001807	53.8	63.6
	CEACAM16	NP_001034302	28	43.5
	CEACAM18	NP_001265321	26.9	46.2
	CEACAM19	NP_064604	23.7	38.1
	CEACAM20	NP_001096067	25.7	39.9
	CEACAM21	NP_001091976	34.1	42.3
Average of all Human CEACAM proteins compared to Dog CEACAM24			39.9	51.9

Figure 3.2: Dog and human *CEACAM* gene family protein domain analysis; (**A**) Dog CEACAM protein domain and binding site depictions with membrane regions; (**B**) Human CEACAM protein domain and binding site depictions with membrane regions.



A total of six rare PTVs were identified in African Americans and sixteen in European Americans breast cancer cases (Supplementary Tables 3.1 and 3.2). Single variant assessment revealed five variants associated with breast cancer risk, three of which were associated each with European and African American breast cancer (Table 3.3, Figures 3.3 and 3.4). One variant, *CEACAM7* c.195C>A;p.(Y65X), was associated with breast cancer risk in both ethnicities (Table 3 and Figure 3). Two stop gain mutations in *CEACAM4* were associated with African American breast cancer (Table 3.3 and Figure 3.3**)**, and two splicing mutations were associated with European American breast cancer, one in *CEACAM6* and another within *CEACAM8* (Table 3 and Figure 4).

Figure 3.3: Individual significant stop gain mutations; (**A**) *CEACAM4* c.367C>T;p.(Arg123*); (**B**) *CEACAM4* c.424C>T;p.(Gln142*); (**C**) *CEACAM7* c.195C>A;p.(Tyr65*).



A

CEACAM4: NM_001817
Reference Protein (WT)

```
  1  MGPPSAAPRG GHRPWQGLLI TASLLTFWHP PTTVQFTIEA LPSSAAEGKD VLLLACNISE
 61  TIQAYYWHKG KTAEGSPLIA GYITDIQANI PGAAYSGRET VYPNGSLLFQ NITLEDAGSY
121  TLRTINASYD SDQATGQLHV HQNNVPGLPV GAVAGIVTGV LVGVALVAAL VCFLLLSRTG
181  RASIQRDLRE QPPPASTPGH GPSHRSTFSA PLPSPRTATP IYEELLYSDA NIYCQIDHKA
241  DVVS*
```

CEACAM4:c.367C>T:p.R123*
Predicted Protein after Mutation

```
  1  MGPPSAAPRG GHRPWQGLLI TASLLTFWHP PTTVQFTIEA LPSSAAEGKD VLLLACNISE
 61  TIQAYYWHKG KTAEGSPLIA GYITDIQANI PGAAYSGRET VYPNGSLLFQ NITLEDAGSY
121  TL*
```

B

CEACAM4: NM 001817
Reference Protein (WT)

```
  1  MGPPSAAPRG GHRPWQGLLI TASLLTFWHP PTTVQFTIEA LPSSAAEGKD VLLLACNISE
 61  TIQAYYWHKG KTAEGSPLIA GYITDIQANI PGAAYSGRET VYPNGSLLFQ NITLEDAGSY
121  TLRTINASYD SDQATGQLHV HQNNVPGLPV GAVAGIVTGV LVGVALVAAL VCFLLLSRTG
181  RASIQRDLRE QPPPASTPGH GPSHRSTFSA PLPSPRTATP IYEELLYSDA NIYCQIDHKA
241  DVVS*
```

CEACAM4:c.424C>T:p.Q142*
Predicted Protein after Mutation

```
  1  MGPPSAAPRG GHRPWQGLLI TASLLTFWHP PTTVQFTIEA LPSSAAEGKD VLLLACNISE
 61  TIQAYYWHKG KTAEGSPLIA GYITDIQANI PGAAYSGRET VYPNGSLLFQ NITLEDAGSY
121  TLRTINASYD SDQATGQLHV H*
```

C

CEACAM7: NM_006890
Reference Protein (WT)

```
  1  MGSPSACPYR VCIPWQGLLL TASLLTFWNL PNSAQTNIDV VPFNVAEGKE VLLVVHNESQ
 61  NLYGYNWYKG ERVHANYRII GYVKNISQEN APGPAHNGRE TIYPNGTLLI QNVTHNDAGI
121  YTLHVIKENL VNEEVTRQFY VFSEPPKPSI TSNNFNPVEN KDIVVLTCQP ETQNTTYLWW
181  VNNQSLLVSP RLLLSTDNRT LVLLSATKND IGPYECEIQN PVGASRSDPV TLNVRYESVQ
241  ASSPDLSAGT AVSIMIGVLA GMALI*
```

CEACAM7:c.195C>A:p.Y65*
Predicted Protein after Mutation

```
 1  MGSPSACPYR VCIPWQGLLL TASLLTFWNL PNSAQTNIDV VPFNVAEGKE VLLVVHNESQ
61  NLYG*
```

Figure 3.4: *CEACAM6* and *CEACAM8* significant splicing mutations; (**A**) Depiction of the change in genomic sequence with splice site mutation; (**B**) Depiction of the top five miRNA binding sites for *CEACAM6* and *CEACAM8* within the mature mRNA. Blue is coding and red is non-coding.

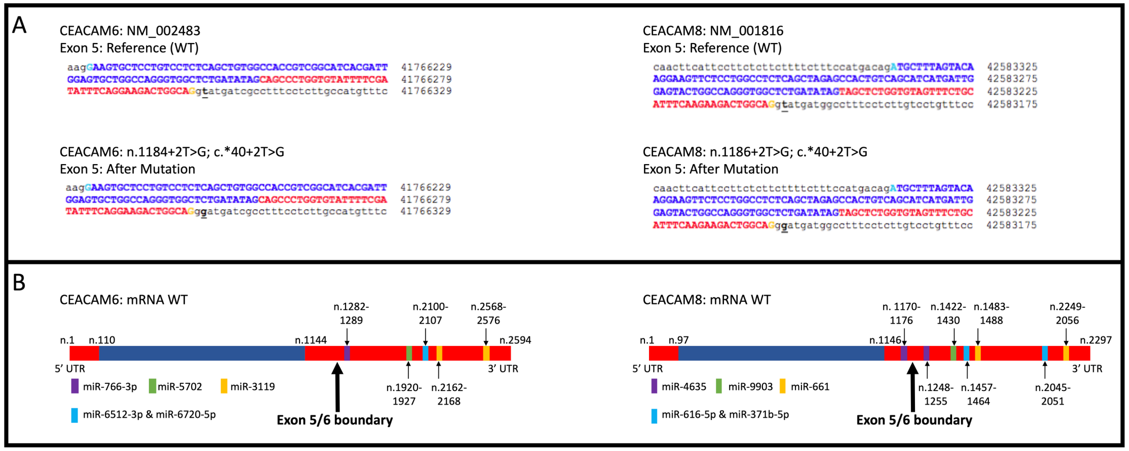

A
CEACAM6: NM_002483
Exon 5: Reference (WT)
aagGAAGTGCTCCTGTCCTCTCAGCTGTGGCCACCGTCGGCATCACGATT 41766229
GGAGTGCTGGCCAGGGTGGCTCTGATATAGCAGCCCTGGTGTATTTTCGA 41766279
TATTTCAGGAAGACTGGCAGgtatgatcgcctttcctcttgccatgtttc 41766329
CEACAM6: n.1184+2T>G; c.*40+2T>G
Exon 5: After Mutation
aagGAAGTGCTCCTGTCCTCTCAGCTGTGGCCACCGTCGGCATCACGATT 41766229
GGAGTGCTGGCCAGGGTGGCTCTGATATAGCAGCCCTGGTGTATTTTCGA 41766279
TATTTCAGGAAGACTGGCAGggatgatcgcctttcctcttgccatgtttc 41766329
CEACAM8: NM_001816
Exon 5: Reference (WT)
caacttcattccttctcttcttttctttccatgacagATGCTTTAGTACA 42583325
AGGAAGTTCTCCTGGCCTCTCAGCTAGAGCCACTGTCAGCATCATGATTG 42583275
GAGTACTGGCCAGGGTGGCTCTGATATAGTAGCTCTGGTGTAGTTTCTGC 42583225
ATTTCAAGAAGACTGGCAGgtatgatggcctttcctcttgtcctgtttcc 42583175
CEACAM8: n.1186+2T>G; c.*40+2T>G
Exon 5: After Mutation
caacttcattccttctcttcttttctttccatgacagATGCTTTAGTACA 42583325
AGGAAGTTCTCCTGGCCTCTCAGCTAGAGCCACTGTCAGCATCATGATTG 42583275
GAGTACTGGCCAGGGTGGCTCTGATATAGTAGCTCTGGTGTAGTTTCTGC 42583225
ATTTCAAGAAGACTGGCAGggatgatggcctttcctcttgtcctgtttcc 42583175
B
CEACAM6: mRNA WT
n.1282-1289
n.2100-2107
n.2568-2576
n.1
n.110
n.1144
n.2594
5' UTR
3' UTR
n.1920-1927
n.2162-2168
miR-766-3p
miR-5702
miR-3119
miR-6512-3p & miR-6720-5p
Exon 5/6 boundary
CEACAM8: mRNA WT
n. 1170-1176
n.1422-1430
n.1483-1488
n.2249-2056
n.1
n.97
n.1146
n.2297
5' UTR
3' UTR
n.1248-1255
n.1457-1464
n.2045-2051
miR-4635
miR-9903
miR-661
miR-616-5p & miR-371b-5p
Exon 5/6 boundary

Table 3.3: Significant mutations in CEACAM gene family. Individual mutation p-values were calculated using Fisher's Exact test.

Gene Name	Variant Type	Genomic Position on Chr 19	mRNA Variant Name	Protein Variant Name	rs ID	EA			AA		
						MAF (%)		Mutation Specific P-values	MAF (%)		Mutation Specific P-values
						EVS EA	TCGA EA	TCGA EA	EVS AA	TCGA AA	TCGA AA
CEACAM4: NM_001817	stopgain	41625658	c.367C>T	p.R123X	rs147663846	-	-	-	0.20	0.89	**0.04803**
	stopgain	41625601	c.424C>T	p.Q142X	rs199937487	-	-	-	0.02	0.60	**0.01431**
CEACAM6: NM_002483	splicing	41766301	c.*40+2T>G	-	rs782698255	0.00	0.46	**7.40E-06**	0.00	0.30	0.07636
CEACAM7: NM_006890	stopgain	41687091	c.195C>A	p.Y65X	rs782316651	0.00	10.79	**2.20E-16**	0.00	4.46	**2.20E-16**
CEACAM8: NM_001816	splicing	42583204	c.*40+2T>G	-	rs748512513	0.00	1.62	**2.20E-16**	-	-	-

Both of those splicing mutations affect non-protein-coding exons in the 3' UTR, which, instead of truncating the protein, potentially disrupt key microRNA binding sites previously associated with cancer (Table 3.4 and Figure 3.4). Overall, gene-based aggregation analyses revealed that rare PTVs in *CEACAM6*, *CEACAM7*, and *CEACAM8* are associated with European American breast cancer risk, and rare PTVs in *CEACAM7* are associated with breast cancer risk in African Americans (Table 3.5). Ultimately, rare PTVs in the entire *CEACAM* gene family are associated with breast cancer risk in both European and African Americans with respective *p-values* of 1.75×10^{-13} and 1.87×10^{-04} (Table 3.5).

Table 3.4: Top five miRNA binding sites for both *CEACAM6* and *CEACAM8* and previous cancer associations

Gene Target Name	**miRNA Name**	**Previous Cancer Association**	**Previous HBOC Association**
CEACAM6	miR-3119	**Yes** 147	No
	miR-766-3p	**Yes** 148-154	**Yes** 152-154
	miR-6512-3p	**Yes** 155	**Yes** 155
	miR-6720-5p	**Yes** 155-157	**Yes** 155
	miR-5702	**Yes** 158; 159	**Yes** 159
CEACAM8	miR-661	**Yes** 160-165	**Yes** 162-165
	miR-9903	**Yes** 166	**Yes** 166
	miR-616-5p	**Yes** 167-170	**Yes** 169; 170
	miR-371b-5p	**Yes** 171; 172	No
	miR-4635	**Yes** 173-177	**Yes** 176

Table 3.5: Aggregation analysis for rare (<1% MAF) PTVs in the *CEACAM* gene family

Gene Name	**Gene Specific p-values**	
	AA	**EA**
CEACAM1: NM_001184815	1	0.8784262
CEACAM3: NM_001815	1	0.3978745
CEACAM4: NM_001817	0.148726	0.7479721
CEACAM5: NM_004363	1	0.8516203
CEACAM6: NM_002483	0.07636	**1.4423E-05**
CEACAM7: NM_006890	**1.8694E-12**	**1.2241E-11**
CEACAM8: NM_001816	0.2727805	**6.4189E-12**
CEACAM16: NM_001039213	0.923479	0.9930833
CEACAM18: NM_001278392	1	1
CEACAM19: NM_020219	1	1
CEACAM20: NM_001102597	1	0.9190567
CEACAM21: NM_001098506	0.9604724	0.7104384
***CEACAM* gene family**	**1.87E-04**	**1.75E-13**

3.5 Discussion

Utilizing a comparative oncology approach, our team identified *CEACAM24* c.247dupG;p.(Val83Glyfs*48) in Golden Retrivers with CMT and subsequently determined that rare PTVs in the entire *CEACAM* gene family were associated with inherited breast cancer risk in humans. We previously described a large Golden Retriever pedigree with segregating CMT, carried out WGS on five selected Golden Retriever cases, and highlighted variants in orthologs of human breast cancer susceptibility genes.[137] In this current study, we used the same WGS dataset to identify novel variants that could be influencing Golden Retriever CMT susceptibility. We isolated PTVs found in all five sequenced Golden Retriever samples, and, upon validation, determined the mutation status in the 13 remaining CMT-affected Golden Retrievers within the pedigree. *CEACAM24* c.247dupG;p.(Val83Glyfs*48) was the only validated variant and had an allele frequency of 66.7% amongst the 18 CMT-affected dogs. Despite not being recognized as a breed highly affected by CMT, Golden Retrievers have a higher prevalence of cancer compared to many dog breeds with 65% of Golden Retrievers in the United States succumbing to the disease.[88; 178; 179] The Golden Retriever *CEACAM24* c.247dupG;p.(Val83Glyfs*48) allele frequency and cancer mortality rate are very similar.

The CMT-affected Golden Retrievers within this study can all be linked back to a sire in the USA from the 1950s, which was shortly after the registration of the breed with the American Kennel Club. Since importation to and registration in the United States, Golden Retrievers in Europe and the United States are considered two distinct populations, as breeding between the two continents is rare and unique gene pools have been established due to strict breeding standards and the popular-sire effect.[180] Cancer mortality in European-bred Golden Retrievers has been reported to be 38.8%, which is much lower than Golden Retrievers in the United States (65%).[88; 178] These differences could be explained by distinct genetic risk factors. The allele frequency of *CEACAM24* c.247dupG;p.(Val83Glyfs*48) in the European Variant Archive was 17.3%, which corresponded to a *p-value* of $1.52\text{x}10^{-8}$ when compared to our CMT-affected Golden Retrievers from the United States. However, in addition to not knowing breed-specific information in the European Variant Archive, genetic bottlenecks upon importation to the United States need to be acknowledged. Thus, comparing allele frequencies to a United States dog population with known breed status was important, which can be determined through American Kennel Club registration. Overall, *CEACAM24* c.247dupG;p.(Val83Glyfs*48) appears to be common in Golden Retrievers in the United States with an allele frequency of 67.8%, which is not significantly different from the CMT-affected Golden Retriever cases. However, that allele frequency was determined by screening 87 Golden Retrievers from the CHIC repository with unknown disease diagnoses and age at sample submission, not ideal for canine cancer studies.[181; 182]

Regarding the assessment of other American Kennel Club breeds, an overall *CEACAM24* c.247dupG;p.(Val83Glyfs*48) allele frequency of 22.4% was revealed, which was significantly different from CMT-affected Golden Retriever cases. Noting the small sample sizes of each breed, over half of the assessed breeds showed no presence of the variant. However, some breeds contained the variant at higher levels; most notably, Petit Basset Griffon Vendeen, Gordon Setter, Australian Cattle Dog, Siberian Husky, and Dalmatian. Petit Basset Griffon Vendeen, which had the highest allele frequency, has a cancer mortality rate of 33%.[88] In a United Kingdom study, Dalmatians, Gordon Setters, and Siberian Huskies were found to have cancer mortality rates ranging from 19.1 – 31.8%,[88] and Australian Cattle Dogs have a rate of 27%.[183]

CEACAM24 c.247dupG;p.(Val83Glyfs*48) abolishes the extracellular region, the transmembrane domain, and part of the cytoplasmic region, including the Ig V-set domain, a key domain that makes it a part of the Ig superfamily.[184; 185] Thus, it is presumed to be a loss-of-function mutation. *CEACAM24* is a part of the dog *CEACAM* gene family and, according to Ensembl, no other stop gain or frame-shifting variants have been identified in dog *CEACAM* genes. However, one splicing mutation in *CEACAM28* (c.1415-2A>G) was identified, which had a 34% allele frequency within the European Variation Archive. *CEACAM* genes have diverse functions in cell-cell adhesion, cell signaling, immunity/inflammation, angiogenesis, and tumor development, progression and metastasis.[185-187] The *CEACAM* gene family is present in many mammalian species but has evolved in a highly species-specific manner, heavily influenced by pathogen/host coevolution.[188-190] Despite phylogenetic discordance of dog and human *CEACAM* genes,[190] our analyses revealed there is high homology between the dog CEACAM24 protein and the human CEACAM proteins, averaging 51.9% similarity. This homology, along with the fact that there is no direct human ortholog of the *CEACAM24* gene, prompted all human *CEACAM* genes to be investigated for rare PTVs in the TCGA breast cancer cohort.

There are 12 human *CEACAM* genes, all of which are clustered on chromosome 19q13.2-19q13.4. Over the years, genetic markers in that region have been associated with many different types of cancer susceptibility, including breast cancer.[191-197] Nonetheless, inherited mutations in *CEACAM* genes have yet to be associated with inherited risk of cancer.[198-200] Aberrant expression of many *CEACAM* genes have been associated with tumorigenesis, and *CEACAM* gene products are recognized as clinically-relevant tumor markers.[185-187] Regarding breast cancer, *CEACAM1* has been shown to be down-regulated compared to normal breast tissue,[201] similar to its expression in prostate,[202; 203] endometrial,[204] gastric,[205] and colon cancer,[206; 207] identifying it as a tumor suppressor. It has also been demonstrated that *CEACAM5* [208]*, CEACAM6,*[209-211] and *CEACAM19*[212; 213] are overexpressed in breast cancer and are associated with enhanced tumor invasiveness and metastasis. Conversely, *CEACAM6* and *CEACAM8* co-expression inhibits proliferation and invasiveness of breast cancer cells.[214] Additionally, *CEACAM* gene splice variants have been suggested to play a role in breast cancer tumorigenesis.[215; 216] Lastly, through exome sequencing, Li *et al.* observed loss of heterozygosity of *CEACAM1*, *CEACAM3*, *CEACAM5*, *CEACAM6*, *CEACAM7* and *CEACAM8* in breast cancer tumors that were associated with metastasis, suggesting that this closely-linked gene family regulates tumorigenesis and metastasis synergistically.[217] Corroborating those preliminary findings, we have now determined that rare inherited PTVs in the entire *CEACAM* gene family are associated with breast cancer risk in both European and African Americans with respective p-values of $1.75\text{x}10^{-13}$ and $1.87\text{x}10^{-04}$. The p-value generated for African American breast cancer risk was likely influenced by the small sample size in TCGA.

We analyzed blood-derived exomes of European and African American breast cancer cases in TCGA to identify inherited PTVs in all human *CEACAM* genes, and detected sixteen and six rare PTVs in each ethnicity, respectively. Gene-based analyses determined that rare PTVs in *CEACAM6*, *CEACAM7*, and *CEACAM8* are associated with European American breast cancer risk, and rare PTVs in *CEACAM7* are associated with breast cancer risk in African Americans. *CEACAM7*, which was associated with breast cancer risk in both ethnicities, has no current link to breast cancer. However, down-regulation of *CEACAM7* in hyperplastic polyps and early adenomas represent some of the earliest observable molecular events leading to colorectal tumors.[218] Though expression was thought to be restricted to the epithelial cells of the colon and pancreas, according to the Human Protein Atlas, grandular cells of the breast have moderate CEACAM7 protein expression.[219; 220] How *CEACAM7* plays a role in breast cancer is currently unknown, but the link could even be indirect and due to expression in non-breast tissue[221]. *CEACAM7* c.195C>A;p.(Y65X), which was detected in 10.8% and 4.5% of European and African American cases, respectively, was absent in all EVS controls. It severely truncates the 265 amino acid proteins and results in a loss of the cytoplasmic region, as well as a large portion of the extracellular region, including disruption of the Ig-like and Ig V-set domains. It is likely a loss-of-function mutation (Figure 3.3).

Rare PTVs in *CEACAM6* and *CEACAM8* appear to only be associated with European American breast cancer risk. Considering that *CEACAM6/8* co-expression inhibits proliferation and invasiveness of breast cancer cells,[214] having a rare PTV in one of those two genes may be sufficient to override their synergistic tumor-suppressing relationship. While a number of PTVs were detected in these genes, two splicing mutations, *CEACAM6* c.*40+2T>G and *CEACAM8* c.*40+2T>G, were individually determined to be associated with European American breast cancer, both of which affect non-coding exons in the 3' UTR. Both mutations affect the donor site immediately following exon 5 of their respective genes, which contains both coding and non-coding DNA. The mutated donor sites likely affect the downstream sequence of the mature mRNA product, either retaining (all or a part of) intron 5 or removing exon 6, the last non-coding exon, where many microRNA binding sites are located (Figure 4). Based on miRDB rankings, the top five microRNAs that bind to the 3' UTRs of *CEACAM6* and *CEACAM8* have previous links to cancer (Table 3.4); thus, disrupted microRNA binding likely leads to aberrant *CEACAM6* and *CEACAM8* expression.

Two stop gain mutations in *CEACAM4* (c.367C>T;p.R123X and c.424C>T;p.Q142X) were associated with African American breast cancer. These mutations were not detected in European American cases or controls, and are very rare in the general African American population. They were detected in significantly more African American breast cancer cases compared to ethnic-matched controls, suggesting their involvement in African American breast cancer risk. However, gene-based aggregation analyses did not support *CEACAM4* as a breast cancer risk gene. Larger African American breast cancer cohorts will need to be studied to validate these findings. Interestingly, in a study of parous women with and without breast cancer, *CEACAM4* has been reported to be up-regulated in normal breast compared to breast tumor samples.[222] Though race/ethnicity was not revealed in that study, the results suggest that *CEACAM4* could be a breast cancer tumor suppressor.

It has long been reported that minimal genetic changes can have radical effects on the function of *CEACAM* genes.[223] Residues in CEACAM6 and CEACAM8 have been identified that are critical for CEACAM6 homodimerization as well as the formation of *CEACAM6* and *CEACAM8* heterodimers, which is important in preventing breast cancer cell proliferation.[214; 224] There have also been residues reported in *CEACAM1* that are crucial for determining the risk of infection by receptor-binding pathogens[225] and preventing the killing activity of NK cells.[226] Furthermore, somatic missense mutations in colorectal cancers have been detected in *CEACAM1*[207] and *CEACAM5,*[227] the latter of which has been shown to increase proliferation by inhibiting TGFB signaling and altering the intestinal microbiome. The microbiome has been reported as a new breast cancer risk factor.[228; 229] In fact, differences have been reported in the microbiome of normal and cancerous breast tissue, as well as the gut microbiota of breast cancer cases versus controls.[229] Disrupted *CEACAM* genes could be the underlying mechanism through altered TGFB signaling, bacteria docking, and/or estrogen metabolism.[225; 227; 229; 230] This study reports the first association of inherited *CEACAM* mutations and breast cancer risk, and potentially implicates the whole gene family in genetic risk. Precisely how these mutations contribute to breast cancer needs to be determined, especially considering our current knowledge on the role that the *CEACAM* gene family plays in tumor development, progression, and metastasis.

3.6 Supplementary Material

Supplementary Table 3.1: Summary of all rare PTVs found in African American TCGA Cohort and EVS control Cohort. Individual Mutation P-values were calculated using Fisher's Exact Test. Gene specific and full gene family aggregate P-values were generated using Fisher's Method for combining p-values.

** The complement was generated for all p-values not equaling one for variants that were more common in controls than cases to correct for directionality.

^^ The mutation was named according to hg38 and rsID reported in dbSNP instead of hg19 as reported in EVS

Gene Name	**Variant Type**	**Variant Name**	**Variant Name**	**Genomic Position on Chr 19**	**rs ID**	**MAF (%)**		**Mutation Specific P-values**	**Gene Specific P-values**
						EVS AA	**TCGA AA ALL**	**TCGA AA ALL**	**TCGA AA ALL**
CEACAM21: NM_001098506	stop-gained	c.44G>A	p.(W15*)	41576318	rs371372590	0.0227	0.0000	1	0.9604724
	frameshift	c.91del1	p.(T32Pfs*47)	41577225	rs535449616	0.6009	0.2976	0.2827**	
	splicing	c.424+1G>A	.	41577560	rs370750766	0.0228	0.0000	1	
	frameshift	c.471_472del2	p.(K159Gfs*11)	41579398	.	0.1265	0.0000	1	
CEACAM4: NM_001817	splicing	c.670-2A>T	.	41619397	rs372504368	0.0227	0.0000	1	0.148726
	stopgain	c.424C>T	p.Q142X	41625601	rs199937487	0.0227	0.5952	**0.01431**	
	stopgain	c.367C>T	p.R123X	41625658	rs147663846	0.2043	0.8929	**0.04803**	
	frameshift	c.13_14insT	p.(S5Ffs*35)	41626950	.	0.0235	0.0000	1	
	frameshift	c.12_13insC	p.(S5Lfs*35)	41626951	.	0.0938	0.0000	1	

CEACAM7: NM_006890	stop-gained	c.295C>T	p.(R99*)	41686991	rs150543831	0.0227	0.0000	1	**1.8694E-12**
	frameshift	c.269del1	p.(N90Mfs*20)	41687016	.	0.0235	0.0000	1	
	stopgain	c.195C>A	p.Y65X	41687091	rs782316651	0.0000	4.4643	**2.20E-16**	
	splicing	c.64+1G>T	.	41688101	rs141024482	0.0227	0.0000	1	
CEACAM5: NM_004363	splicing	c.425-1G>A	.	41714970	rs201377769	0.0454	0.0000	1	1
	frameshift	c.1010del1	p.(D337Vfs*5)	41717505	.	0.0235	0.0000	1	
CEACAM6: NM_002483	splicing	c.*40+2T>G	.	41766301	rs782698255	0.0000	0.2976	0.07636	0.07636
CEACAM3: NM_001815	stop-gained	c.44G>A	p.(W15*)	41796721	rs377467224	0.0227	0.0000	1	1
CEACAM1: NM_001184815	frameshift	c.1379delA	p.K460fs	42509116	rs781044252	0.0469	0.0000	1	1
CEACAM8: NM_001816	frameshift	c.743delA:	p.Y248fs	42588999	.	0.0000	0.2976	0.07104	0.2727805
	frameshift	c.550_551del2	p.(L185Pfs*24)	42589608	.	0.2111	0.0000	1	
CEACAM20: NM_001102597	frameshift	c.1622delC	p.P541fs	44511145	rs150406547	0.0539	0.0000	1	1
	stop-gained	c.1537C>A	p.(C512*)	44512056	rs150222142	0.2753	0.0000	1	
	frameshift	c.1448dup1^^	p.(D484Gfs*5)^^	44,512,933	rs5828200	0.0530	0	1	

CEACAM19: NM_020219	stop-gained	c.289C>T	p.(R97*)	44672829	rs367774891	0.0227	0.0000	1	1
CEACAM16: NM_001039213	frameshift	c.276del1	p.(L93Wfs*125)	44703586	.	0.3792	0.0000	0.3756**	0.923479
	frameshift	c.287_290del4	p.(Q96Pfs*121)	44703597	.	0.1775	0.0000	1	
	frameshift	c.857del1	p.(Q287Rfs*34)	44705784	.	0.2675	0.0000	1	
CEACAM18: NM_001278392	stop-gained	c.387G>A	p.(W129*)	51480667	rs369762254	0.0255	0	1	1
Aggregate P-value using Fishers Method									**0.000187025**

Supplemental Table 3.2: Summary of all rare PTVs found in European American TCGA Cohort and EVS control Cohort. Individual Mutation P-values were calculated using Fisher's Exact Test. Gene specific and full gene family aggregate P-values were generated using Fisher's Method for combining p-values.

** The complement was generated for all p-values not equaling one for variants that were more common in controls than cases to correct for directionality.

Gene Name	**Variant Type**	**mRNA Variant Name**	**Protein Variant Name**	**Genomic Position on Chr 19**	**rs ID**	**MAF (%)**		**Mutation Specific P-values**	**Gene Specific P-values**
						EVS EA	**TCGA EA ALL**	**TCGA EA ALL**	**TCGA EA ALL**
CEACAM21: NM_001098506	frameshift	c.91del1	p.(T32Pfs*47)	41577225	rs535449616	0.0247	0.0000	1	0.7104384
	stopgain	c.139G>T	p.E47X	41577274	rs200535080	0.0000	0.0770	0.1327	
	stopgain	c.292C>T	p.R98X	41577427	rs369283885	0.0116	0.0770	0.2454	
	frameshift	c.471_472del2	p.(K159Gfs*11)	41579398	.	0.0619	0.0000	1	
	splicing	c.882+1G>A	.	41585872	rs62119455	0.6809	0.6163	1	
CEACAM4: NM_001817	splicing	c.64+1G>C	.	41626899	rs115582444	0.0116	0.0000	1	0.7479721
	frameshift	c.12_13insC	p.(S5Lfs*35)	41626951	.	0.1090	0.0000	0.3803**	
CEACAM7: NM_006890	frameshift	c.727_728insGGGGAAA	p.S243fs	41677482	.	0.0000	0.0770	0.1341	**1.22411E-11**
	stop-gained	c.397G>T	p.(E133*)	41686889	rs150439369	0.0116	0.0000	1	
	stop-gained	c.295C>T	p.(R99*)	41686991	rs150543831	0.0116	0.0000	1	

	frameshift	c.269del1	p.(N90Mfs*20)	41687016	.	0.0485	0.0000	1	
	stopgain	c.195C>A	p.Y65X	41687091	rs782316651	0.0000	10.7858	**2.20E-16**	
	splicing	c.64+1G>T	.	41688101	rs141024482	0.4186	0.5393	0.4978	
CEACAM5: NM_004363	stop-gained	c.83G>A	p.(W28*)	41709698	rs369263590	0.0116	0.0000	1	0.8516203
	splicing	c.424+1G>A	.	41710040	rs377398784	0.0116	0.0000	1	
	stopgain	c.1880G>A	p.W627X	41721030	rs74946020	0.0000	0.0770	0.1313	
CEACAM6: NM_002483	splicing	c.424+1G>A	.	41756960	rs782773255	0.0000	0.0770	0.1313	**1.4423E-05**
	splicing	c.*40+2T>G	.	41766301	rs782698255	0.0000	0.4622	**7.40E-06**	
CEACAM3: NM_001815	splicing	c.424+1G>A	.	41797949	rs781898698	0.0000	0.0770	0.1314	0.3978745
	splicing	c.542+1G>T	.	41808931	rs368228701	0.0116	0.0000	1	
CEACAM1: NM_001184815	frameshift	c.1379delA	p.K460fs	42509116	rs781044252	0.4240	0.0770	0.91819**	0.8784262
	splicing	c.1182-1G>A	.	42511629	rs376067131	0.0116	0.0000	1	
	frameshift	c.791_792insG	p.(N264Kfs*25)	42521433	.	0.0121	0.0000	1	
	frameshift	c.553_554insAGGC	p.(L185Qfs*26)	42522073	.	0.1333	0.0000	0.62**	
	frameshift	c.464delA	p.N155fs	42522163	rs773369383	0.0000	0.0770	0.1314	
CEACAM8: NM_001816	splicing	c.*40+2T>G	.	42583204	rs748512513	0.0000	1.6179	**2.20E-16**	**6.41894E-12**
	frameshift	c.981_982insT	p.(P328Sfs*6)	42583314	.	0.0363	0.0000	1	

	stop-gained	c.970C>T	p.(Q324*)	42583326	rs139611602	0.0116	0.0000	1	
	frameshift	c.550_551del2	p.(L185Pfs*24)	42589608	.	0.1333	0.0000	0.62**	
	frameshift	c.364delC	p.L122fs	42593601	rs572450516	0.0242	0.0770	0.3548	
	stopgain	c.27C>A	p.C9X	42594802	.	0.0000	0.0770	0.1452	
CEACAM20: NM_001102597	frameshift	c.1622delC	p.P541fs	44511145	rs150406547	0.0253	0.0770	0.3666	0.9190567
	splicing	c.1310-1G>C	.	44513290	rs201465799	0.0360	0.0000	1	
	stop-gained	c.742C>T	p.(R248*)	44522643	rs368941407	0.0120	0.0000	1	
CEACAM19: NM_020219.3	frameshift	c.384_387del4	p.(E129Gfs*3)	44672923	.	0.0242	0.0000	1	1
	splicing	c.792+1G>C	.	44681313	rs369842569	0.0116	0.0000	1	
CEACAM16: NM_001039213.2	frameshift	c.276del1	p.(L93Wfs*125)	44703586	.	0.2386	0.0000	0.90332**	0.9930833
	frameshift	c.287_290del4	p.(Q96Pfs*121)	44703597	.	0.1755	0.0000	0.7583**	
	frameshift	c.857del1	p.(Q287Rfs*34)	44705784	.	0.5792	0.0000	0.997948**	
CEACAM18: NM_001278392.1	stop-gained	c.182G>A	p.(W61*)	51480462	rs370424604	0.0119	0.0000	1	1
Aggregate P-value using Fishers Method									**1.75193E-13**

Chapter 4: An investigation into the role of inherited CEACAM gene family variants and colorectal cancer (CRC) risk.

4.1 Abstract

Colorectal cancer (CRC) is the fourth most common cancer diagnosis in the US, and this risk can increase with a family history. Of inherited cases only 30% are explained by mutations in known CRC risk genes associated with inherited CRC syndromes. Of these risk syndromes two are shared by breast cancer and CRC, along with many other similar factors. In a previous analysis the CEACAM gene family was associated with inherited breast cancer risk. This work represents an investigation of the CEACAM gene family into inherited CRC risk. Utilizing The Cancer Genome Atlas (TCGA) CRC cohort, rare protein truncating variants and missense variants were investigated in a gene aggregation analysis along with individually. There was no overall association of either class of mutation or together with CRC risk; however, 9 individual missense mutations were associated CRC risk, and small changes in CEACAM genes has been known to influence gene functions. Three of these mutations occurred within the Ig V-set domains of CEACAM1, -3 and -4. The Ig V-set domains are crucial for dimer formation and this is likely how these mutations are influencing CRC risk. Additionally, two mutations in between functional domains of CEACAM8 were also associated. Two mutations in CEACAM18 were associated but occur after the functional domains. A single missense mutation in both CEACAM 19 and -20 were also associated outside of functional domains. The exact impact of the many of these mutations in unknown, highlighting the need for further studies investigating the CEACAM genes in CRC cases for risk influences.

4.2 Introduction

Colorectal cancer (CRC) is the fourth most common cancer diagnosis in the US for both men and women, and has a rising trend of diagnosis in younger adults.[231] The lifetime risk of CRC development is between 4.0% and 5% for both men and women.[231; 232] However, this risk can increase with a multitude of factors, including a family history of CRC.[231] Approximately 30% of CRC cases are familial.[232; 233] Of the inherited cases with a known genetic cause, the majority are a result of Lynch syndrome.[234] Additional syndromes linked to CRC risk are polyposis syndromes (including classic familial adenomatous polyposis (FAP), attenuated FAP, MUTYH-associated polyposis, Peutz-Jeghers syndrome, juvenile polyposis syndrome, hyperplastic polyposis and serrated polyposis syndrome),[231; 235] Lynch-like syndrome,[236] familial colorectal cancer type X (FCCX),[236] and hereditary breast and ovarian cancer syndrome (HBOC), resulting from *BRCA1/2* mutations.[231] However, up to 30% of the inherited cases are estimated to still be genetically unsolved.[235]

Interestingly, CRC and breast cancer share many risk factors.[231; 237; 238] In addition to increased risk of both cancers in certain hereditary cancer syndromes (i.e., Lynch syndrome and *BRCA1/2* mutations[231; 237]), women diagnosed with CRC have a higher risk of developing breast cancer as a secondary cancer diagnosis.[239] Previously, rare protein-truncating variants (PTVs) in the *CEACAM* gene family have been associated with inherited breast cancer risk (Chapter 3). This gene family is composed of 12 genes clustered on chromosome 19q13.2-19q13.4. They are a part of the Ig superfamily and have diverse functions, including cell adhesion and signaling, and play roles in immunity, angiogenesis, and cancer.[185-187] Aberrant expression of *CEACAM* genes have long been associated with tumorigenesis and *CEACAM* gene products are recognized as tumor markers for many different cancers,[185-187] including breast[240] and CRC.[241; 242] However, their impact on inherited cancer risk is vastly understudied.

Atypical *CEACAM* gene expression has been heavily linked to CRC development and progression.[185; 186] In 1965, CEA (more currently known as CEACAM5) was first identified as a tumor marker for CRC.[241; 242] In addition to *CEACAM5*, *CEACAM6* is overexpressed in CRC and has been determined to increase invasiveness.[243] Contrarily, *CEACAM1*[206; 207] and *CEACAM7*[244] have decreased expression in CRC, and *CEACAM7* expression has been shown to be maintained through different stages and can serve as a predictor of reoccurrence. Furthermore, *CEACAM1*[207] and *CEACAM5*[227] somatic missense mutations have been detected in CRC tumors. Considering the above and the fact that both CRC and breast cancer share many risk factors, including genetics,[231; 237] herein, the *CEACAM* gene family was investigated to determine if harboring mutations are associated with CRC inherited risk.

4.3 Methods and Materials

Blood-derived exomes of CRC cases in the TCGA were analyzed to investigate if CEACAM mutations play a role in inherited risk. Through approved research project #10805, whole-exome binary sequence alignment mapping (BAM) files were downloaded through the

Genomic Data Commons (GDC) Data Portal Repository. Samples were acquired by setting specific filters; under the 'Cases' category: Project (TCGA-COAD), Samples Sample Type (Blood Derived Normal), and Race ('Black or African American' and 'White'). The samples were further filtered under the 'Files' category, including Experimental Strategy (WXS), and Data Format (BAM). A total of 48 sample files were obtained for African Americans and 199 for European Americans. These files were downloaded using the GDC Data Transfer Tool (version 1.2.0).

The downloaded BAM files, which had previously been aligned to the hg38 human reference genome, were processed using the remaining portions of a pipeline adapted from the Genome Analysis Toolkit's (GATK's) best practices pipeline.[100] Base quality scores were recalibrated using BaseRecalibrator and then HaplotypeCaller was used to generate genome variant calling format (gVCF) files (GATK version 4.1.9). GenomicsDBImportant was used to generate CEACAM gene family ethnic-specific datasets, and was carried out through a *CEACAM* gene family specific set of intervals (*CEACAM1* (NM_001184815; chr19:42507306-42528481), *CEACAM3* (NM_001815 at chr19:41796587-41811554), *CEACAM4* (NM_001817; chr19:41618971-41627074), *CEACAM5* (NM_004363; chr19:41708626-41730421), *CEACAM6* (NM_002483; chr19:41755530-41772210), *CEACAM7* (NM_006890; chr19:41673303-41688270), *CEACAM8* (NM_001816 at chr19:42580243-42594924), *CEACAM16* (NM_001039213; chr19:44699151-44710718), *CEACAM18* (NM_001278392; chr19:51478643-51490605), *CEACAM19* (NM_020219; chr19:44671452-44684355), *CEACAM20* (NM_001102597; chr19:44506159-44529675), and *CEACAM21* (NM_001098506; chr19:41576166-41586844)). This was followed by GenotypeGVCFs function to generate ethnic specific variant calling format (VCF) files (GATK version 4.1.9). The two ethnic specific VCF files were then annotated using ANNOVAR (version June2020). Variants were filtered to include rare, protein truncating variants (PTVs; nonsense mutations, frameshifting mutations or splice-site affecting mutations) and missense variants with ethnic-specific minor allele frequencies (MAFs) of <1% in Exome Variant Server (EVS; National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project).[108] Each PTV and missense variant was individually investigated using the Fisher's exact test[49; 141] in R (v 3.5.1), to generate p-values comparing MAFs of ethnic-specific TCGA CRC cases and EVS controls. Subsequently, PTVs and missense variants were investigated together and as individual groups in a gene-based and gene family-based aggregation analyses using the Fisher method through the 'sumlog' command as part of the 'metap' packages within R.[143; 144] P-values were not corrected for multiple testing. Lastly, missense pathogenicity was predicted using Polyphen2.[127] For all significant mutations, protein analysis using InterPro[139] and the Eukaryotic Linear Motif (ELM) resource[140] was carried out to identify CEACAM domains and binding motifs, respectively.

4.4 Results

After filtering for rare PTVs and missense variants in the entire *CEACAM* gene family within the TCGA-COAD cohort, a total of 14 different variants were identified in African American cases (one frameshift and 13 missense; Supplementary Table 4.1) and 34 different variants were identified in European American cases (one frameshift, two splice, and 31 missense; Supplementary Table 4.2). In African American cases, 5 of the 13 missense variants were classified as probably damaging; however, none of those mutations were found to be associated with CRC risk. Only two variants were determined to be individually associated with African American CRC risk *CEACAM3*, c.283T>A; p.(Y95N) and *CEACAM8* c.739A>G: p.(T247A); however, both of those variants are likely benign (Table 4.1). In European American cases, 10 of the missense variants were determined to be probably damaging but only two of those variants were found to be associated with CRC risk, *CEACAM1* c.203A>G; p.(Y68C) and *CEACAM18* c.1069T>G; p.(C357G). A total of seven variants were determined to be individually associated in CRC in European Americans, all of which were missense variants. This included the two aforementioned probably damaging missense variants and five that were predicted to be benign (Table 4.2). Gene family and gene-specific aggregation analyses did not yield any significant results, including a combined assessment of PTVs and missense variants, as well as group analyses of PTVs, missense mutations, and probably damaging missense mutations.

Table 4.1: Summary of statistically significant variants in African American TCGA CRC cohort

Gene	Chr 19 Position	Mutation Type	Functional Prediction - Polphen	cDNA Change	Protein Change	TCGA AA Colon MAF (%)	EVS AA MAF (%)	AA Indivdual P-values
CEACAM3: NM_001815	41797807	missense	benign:0.159	c.283T>A	p.(Y95N)	5.208	0.894	0.002
CEACAM8: NM_001816	42589003	missense	benign:0.001	c.739A>G	p.(T247A)	4.167	0.931	0.015

Table 4.2: Summary of statistically significant variants in European American TCGA CRC cohort

Gene	Chr 19 Position	Mutation Type	Functional Prediction - Polyphen	cDNA Change	Protein Change	TCGA EA Colon MAF (%)	EVS EA MAF (%)	EA Indivdual P-values
CEACAM1: NM_001184815	42527262	missense	probably-damaging:1.0	c.203A>G	p.(Y68C)	0.503	0.070	0.046
CEACAM4: NM_001817	41625657	missense	benign:0.325	c.368G>A	p.(R123E)	0.503	0.000	0.002
CEACAM8: NM_001816	42589735	missense	benign:0.005	c.425C>T	p.(P142L)	0.503	0.012	0.006
CEACAM18: NM_001080405	51483229	missense	probably-damaging:1.0	c.1069T>G	p.(C357G)	0.503	0.059	0.036
	51483284	missense	benign:0.013	c.1124A>G	p.(Q375R)	0.503	0.059	0.036
CEACAM19: NM_020219	44681293	missense	benign:0.01	c.773G>C	p.(R258T)	1.005	0.093	0.001
CEACAM20: NM_001102597	44512936	missense	benigns:0.062	c.1445C>T	p.(T482I)	0.503	0.000	0.002

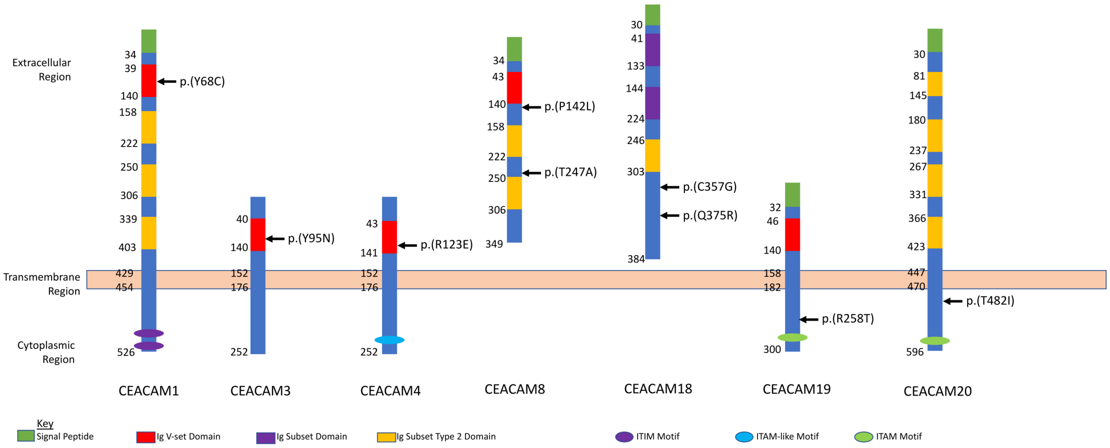
4.5 Discussion

Upon surveying the *CEACAM* gene family for rare PTVs and missense variants in CRC cases from TCGA and controls from the EVS, no gene-based or gene family-based associations with inherited risk of CRC were revealed. This was unexpected due to the previous association of rare PTVs in the *CEACAM* gene family with inherited breast cancer risk (Chapter 3), and the known similarities between breast cancer and CRC risk.[231; 237; 238] The results were also surprising because somatic *CEACAM* mutations had previously been detected in CRC tumors,[207; 227] and abnormal *CEACAM* expression has been linked to CRC development and progression.[186; 187] Furthermore, it has been demonstrated that *CEACAM* gene function can be affected by even minor genetic changes,[223] and specific residues within CEACAM proteins are known to be crucial for normal function.[206; 224; 245]

Despite the lack of association from aggregation analyses, individual variants were found to be associated with CRC inherited risk (Table 4.1 and 4.2). All associations involved individual missense variants; none involved PTVs, unlike the association of *CEACAM* PTVs with breast cancer risk (Chapter 3). In fact, only four different PTVs were detected amongst all CRC cases, none of which overlapped between ethnicities. In European American CRC cases, one splice variant was detected in *CEACAM7* (c.64+1G>T) and *CEACAM21* (c.882+1G>A), and a frameshift mutation was detected in *CEACAM20,* c.1623del1; p.(F542Sfs*56). Additionally, a frameshift mutation in *CEACAM21*, c.91del1; p.(T32Pfs*47), was detected in an African American CRC case.

Of the nine total missense variants that were associated with either African or European American CRC risk, three were within the Ig V-set (variable) domain (Figure 4.1). This included *CEACAM1*, c.203A>G; p.(Y68C) and *CEACAM4* c.368G>A; p.(R123E), which were associated with European American CRC risk, and *CEACAM3*, c.283T>A; p.(Y95N), which was associated with African American risk (Figure 4.1). Despite the fact that only *CEACAM1*, c.203A>G; p.(Y68C) was predicted to be pathogenic through PolyPhen2, the Ig V-set (variable) domain is crucial for the dimerization of many CEACAM proteins and their ability to function within normal ranges.[245; 246] It has even been demonstrated that mutating particular residues within the Ig V-set domain of CEACAM1 can affect the monomer-homodimer exchange and result in the protein staying in a monomeric state,[245] and CEACAM1's ability to dimerize with itself and other CEACAM proteins is required for proper function.[247-250] Knowing that dimerization is crucial and CEACAM1's current role in CRC,[206; 207] *CEACAM1*, c.203A>G; p.(Y68C) is a probable CRC inherited risk factor. *CEACAM3* c.283T>A; p.(Y95N), has been considered benign based on PolyPhen2 prediction and has been reported as benign in ClinVar; however, limited information was provided for that clinical classification.[105] Considering that *CEACAM3* has previously been shown to serve as a possible biomarker for CRC, with potentially greater use than the historically used *CEACAM5,*[251; 252] validating the association of *CEACAM3* c.283T>A; p.(Y95N) with African American CRC inherited risk is crucial in identifying possibly risk factors for this understudied population. Lastly, *CEACAM4* has been previously associated with thyroid cancer[253] but its role in CRC is unknown. Overall, missense variants within the Ig V-set domain identified in this study could result in repressed dimerization; this plausible disease mechanism requires further investigation.

Figure 4.1: CEACAM protein analysis for significant mutations in CRC cohort.



Two statistically significant missense variants were identified in both *CEACAM8* and *CEACAM18*. The two variants in *CEACAM8*, c.425C>T; p.(P142L) and c.739A>G; p.(T247A), were associated with CRC risk in European and African American cases, respectively. *CEACAM8* p.(P142L) is located between the IgV-set domain and the first Ig subset type 2 domain and p.(T247A) is in between the two extracellular Ig subset type 2 domains (Figure 4.1). Even though the role of these variants is unclear, they could influence dimerization, and since CEACAM8 forms dimers with CEACAM6 and CEACAM1,[246; 249] both of which have previous associations with CRC,[206; 207; 243] this could be a possible route of influencing CRC risk. *CEACAM18* c.1069T>G; p.(C357G) and c.1124A>G; p.(Q375R), were significantly associated in European American CRC, and p.(C357G) was predicted to be pathogenic through

PolyPhen2.[127] These mutations occur after known functional domains for CEACAM18 (Figure 4.1), but could potentially influence how the protein interacts with the cell membrane as exactly how the C-terminus interacts with the membrane is unknown.[186; 189; 230] Beyond these two *CEACAM18* variant associations, there is no known link between CEACAM18 and CRC.

A single missense mutation in both *CEACAM19*, c.773G>C; p.(R258T), and *CEACAM20*, c.1445C>T; p.(T482I), was associated with European American CRC. Both of these mutations occur within the cytoplasmic region of the protein, but not in the ITAM binding motifs (Figure 4.1). Again, the possible impacts of these mutations are unclear; however, *CEACAM19* and *-20* have previous cancer links.[212; 213; 216; 254; 255] *CEACAM19* has been determined to be over expressed in breast cancer, with the potential for use as a biomarker;[213] thus, detecting *CEACAM19*, c.773G>C; p.(R258T) in this CRC cohort could be further establishing similar risk factors for both cancers. Recently, CEACAM20 has been determined to play a role in gut microbiome regulation, and it's expression can also be influenced by gut bacteria.[256; 257] The microbiome is a known factor influencing CRC risk and progression,[231] and this could be a method by which mutations in *CEACAM20* influence CRC risk.

Overall, this study aimed to determine if inherited *CEACAM* gene variants play a role in CRC risk. No gene- or gene family-based associations were identified, but individual missense variants in seven different *CEACAM* genes appear to be associated with inherited CRC risk. It is important to note that the TCGA CRC cohort is not a hereditary/familial CRC cohort. The cases simply represent individuals diagnosed with colorectal adenocarcinomas. Though *CEACAM* variants do not appear to play a significant role in this cohort, studying hereditary/familial CRC cohorts could reveal different findings, especially considering that a large percentage of inherited CRC is suspected to be influenced by lower penetrant variants compounded with environmental factors.[231; 235] Furthermore, the TCGA CRC cohort was subdivided by ethnicity, and European Americans cases were represented ~4X more than African American cases, which likely affected the number of variants detected in each ethnic group. This is a concerning limitation, as African Americans have the highest CRC incidence and mortality rates of any ethnicity in the United States[258]. Both TCGA CRC ethnic groups have a limited number of cases, and with the prevalence of previous research linking the *CEACAM* genes to spontaneous CRC,[185; 186; 206; 207; 210; 218; 227; 243; 244; 251; 252; 259] more genetic and functional investigations of the *CEACAM* gene family should be carried out.

4.6 Supplementary Material

Supplemental Table 4.1: Full list of rare (MAF <1%) Stop Gain, Frameshifting, splice-site and missense mutations identified in the "Black or African American" TCGA-COAD cohort and the CEACAM EVS AA cohort

Gene	Position	Function	Functonal Prediction - PH	cDNA	Protein	TCGA AA Colon			EVS AA			AA Indivdual P-values
						Alt Allele Frequency	Ref Allele Frequency	MAF (%)	Alt Allele Frequency	Ref Allele Frequency	MAF (%)	
CEACAM1: NM_001184815	42509115	frameshift	.	c.1379del1	p.(K460Sfs*23)	0	96	0	2	4262	0.0469	1.000
	42509227	missense	possibly-damaging:0.654	c.1268T>G	p.(M423R)	0	96	0	1	4405	0.0227	1.000
	42510897	missense	possibly-damaging:0.738	c.1258C>T	p.(P420S)	0	96	0	6	4400	0.1362	1.000
	42521953	missense	benign:0.095	c.674G>A	p.(R225H)	0	96	0	1	4405	0.0227	1.000
	42522002	missense	probably-damaging:1.0	c.625G>A	p.(D209N)	0	96	0	1	4405	0.0227	1.000
	42522031	missense	probably-damaging:0.972	c.596C>T	p.(T199I)	0	96	0	1	4405	0.0227	1.000
	42522185	missense	possibly-damaging:0.956	c.442T>C	p.(S148P)	0	96	0	1	4405	0.0227	1.000
	42527041	missense	benign:0.307	c.424C>T	p.(P142S)	0	96	0	1	4405	0.0227	1.000
	42528341	missense	benign:0.019	c.34C>T	p.(R12C)	0	96	0	3	4403	0.0681	1.000
CEACAM3: NM_001815	41796721	nonsense	.	c.44G>A	p.(W15*)	0	96	0	1	4405	0.0227	1.000
	41796726	missense	benign:0.211	c.49G>A	p.(G17R)	0	96	0	2	4404	0.0454	1.000
	41797801	missense	benign:0.002	c.277G>C	p.(A93P)	3	93	3.125	41	4317	0.9408	0.068
	41797807	missense	benign:0.159	c.283T>A	p.(Y95N)	5	91	5.20833333	39	4323	0.8941	0.002
	41797858	missense	probably-damaging:0.96	c.334G>T	p.(V112F)	0	96	0	2	4404	0.0454	1.000
	41797922	missense	probably-damaging:1.0	c.398A>G	p.(E133G)	0	96	0	169	4237	3.8357	0.051
	41808845	missense	benign:0.143	c.457G>A	p.(V153I)	0	96	0	1	4405	0.0227	1.000
	41808848	missense	probably-damaging:0.981	c.460G>A	p.(A154T)	0	96	0	1	4405	0.0227	1.000
	41808893	missense	benign:0.414	c.505G>A	p.(A169T)	0	96	0	2	4404	0.0454	1.000

CEACAM4: NM_001817	41619340	missense	possibly-damaging:0.691	c.725T>C	p.(V242A)	0	96	0	1	4405	0.023	1.000
	41619397	splice	.	c.670-2A>T	.	0	96	0	1	4405	0.023	1.000
	41621730	missense	probably-damaging:0.999	c.463G>A	p.(G155S)	0	96	0	1	4405	0.023	1.000
	41621760	missense	benign:0.0	c.433G>A	p.(V145I)	0	96	0	3	4403	0.068	1.000
	41625601	nonsense	.	c.424C>T	p.(Q142*)	0	96	0	1	4405	0.023	1.000
	41625658	nonsense	.	c.367C>T	p.(R123*)	0	96	0	9	4397	0.204	1.000
	41625685	missense	probably-damaging:0.992	c.340C>G	p.(L114V)	0	96	0	5	4401	0.113	1.000
	41625754	missense	benign:0.04	c.271C>A	p.(P91T)	0	96	0	3	4403	0.068	1.000
	41625900	missense	probably-damaging:1.0	c.125C>T	p.(P42L)	0	96	0	3	4403	0.068	1.000
	41625955	missense	probably-damaging:1.0	c.70C>T	p.(L24F)	1	95	1.04166667	41	4365	0.931	0.597
	41626915	missense	probably-damaging:1.0	c.49G>T	p.(G17W)	0	96	0	1	4405	0.023	1.000
	41626938	missense	benign:0.003	c.26G>A	p.(R9H)	0	96	0	3	4401	0.068	1.000
	41626950	frameshift	.	c.13_14insT	p.(S5Ffs*35)	0	96	0	1	4263	0.023	1.000
	41626951	frameshift	.	c.12_13insC	p.(S5Lfs*35)	0	96	0	4	4260	0.094	1.000
CEACAM5: NM_004363	41709737	missense	benign:0.135	c.122C>T	p.(T41M)	0	96	0	1	4405	0.0227	1.000
	41709746	missense	possibly-damaging:0.742	c.131A>C	p.(N44T)	0	96	0	2	4404	0.0454	1.000
	41709845	missense	benign:0.051	c.230G>A	p.(R77H)	0	96	0	4	4402	0.0908	1.000
	41709860	missense	possibly-damaging:0.498	c.245A>G	p.(Y82C)	0	96	0	1	4405	0.0227	1.000
	41709932	missense	probably-damaging:0.996	c.317C>G	p.(S106C)	0	96	0	1	4405	0.0227	1.000
	41709959	missense	possibly-damaging:0.59	c.344A>G	p.(N115S)	0	96	0	1	4405	0.0227	1.000
	41710011	missense	possibly-damaging:0.846	c.396A>T	p.(E132D)	0	96	0	2	4404	0.0454	1.000
	41714970	splice	.	c.425-1G>A		0	96	0	2	4404	0.0454	1.000
	41714998	missense	probably-damaging:0.999	c.452G>A	p.(S151N)	0	96	0	1	4405	0.0227	1.000
	41715033	missense	probably-damaging:1.0	c.487G>A	p.(V163M)	0	96	0	1	4405	0.0227	1.000
	41715103	missense	possibly-damaging:0.894	c.557C>A	p.(P186Q)	2	94	2.08333333	18	4388	0.4085	0.067
	41715194	missense	probably-damaging:1.0	c.648A>T	p.(E216D)	0	96	0	1	4405	0.0227	1.000
	41715202	missense	probably-damaging:1.0	c.656A>G	p.(N219S)	1	95	1.04166667	6	4400	0.1362	0.140

	41715210	missense	possibly-damaging:0.796	c.664A>G	p.(S222G)	0	96	0	2	4404	0.0454	1.000
	41715219	missense	probably-damaging:0.973	c.673C>A	p.(R225S)	0	96	0	2	4404	0.0454	1.000
	41715220	missense	probably-damaging:0.988	c.674G>A	p.(R225H)	0	96	0	1	4405	0.0227	1.000
	41715740	missense	probably-damaging:1.0	c.794C>T	p.(P265L)	0	96	0	2	4404	0.0454	1.000
	41715763	missense	benign:0.024	c.817G>A	p.(V273I)	0	96	0	1	4405	0.0227	1.000
	41717505	frameshift	.	c.1010del1	p.(D337Vfs*5)	0	96	0	1	4263	0.0235	1.000
	41717520	missense	benign:0.044	c.1024G>A	p.(A342T)	0	96	0	2	4404	0.0454	1.000
	41717587	missense	possibly-damaging:0.507	c.1091C>T	p.(P364L)	1	95	1.04166667	3	4403	0.0681	1.000
	41717632	missense	possibly-damaging:0.876	c.1136C>T	p.(T379I)	0	96	0	2	4404	0.0454	1.000
	41717635	missense	probably-damaging:0.998	c.1139T>C	p.(L380P)	0	96	0	1	4405	0.0227	1.000
	41717664	missense	possibly-damaging:0.551	c.1168C>A	p.(P390T)	0	96	0	1	4405	0.0227	1.000
	41717695	missense	probably-damaging:0.999	c.1199G>T	p.(S400I)	0	96	0	2	4404	0.0454	1.000
	41718139	missense	benign:0.072	c.1249G>A	p.(D417N)	0	96	0	1	4405	0.0227	1.000
	41718152	missense	probably-damaging:0.987	c.1262C>T	p.(S421F)	0	96	0	3	4403	0.0681	1.000
	41718307	missense	probably-damaging:1.0	c.1417G>A	p.(G473R)	0	96	0	1	4405	0.0227	1.000
	41719974	missense	benign:0.406	c.1537G>A	p.(V513M)	0	96	0	1	4405	0.0227	1.000
	41720133	missense	benign:0.018	c.1696G>A	p.(A566T)	0	96	0	2	4404	0.0454	1.000
	41720178	missense	possibly-damaging:0.83	c.1741C>T	p.(R581C)	0	96	0	17	4389	0.3858	1.000
	41721033	missense	possibly-damaging:0.669	c.1883G>A	p.(R628H)	0	96	0	4	4402	0.0908	1.000
	41721153	missense	benign:0.0	c.2003T>C	p.(I668T)	0	96	0	2	4404	0.0454	1.000
	41727239	missense	benign:0.414	c.2032G>A	p.(G678R)	1	95	1.04166667	25	4381	0.5674	0.430
CEACAM6: NM_002483	41755660	missense	benign:0.094	c.22C>T	p.(P8S)	0	96	0	1	4405	0.0227	1.000
	41755699	missense	probably-damaging:0.992	c.61A>C	p.(T21P)	0	96	0	1	4405	0.0227	1.000
	41756698	missense	benign:0.322	c.163G>A	p.(A55T)	0	96	0	10	4396	0.2270	1.000
	41756720	missense	probably-damaging:0.971	c.185G>A	p.(R62H)	0	96	0	1	4405	0.0227	1.000
	41756834	missense	benign:0.011	c.299C>T	p.(T100I)	0	96	0	4	4402	0.0908	1.000
	41756846	missense	possibly-damaging:0.873	c.311A>G	p.(N104S)	0	96	0	1	4405	0.0227	1.000

	41756906	missense	benign:0.094	c.371T>C	p.(V124A)	0	96	0	4	4402	0.0908	1.000
	41761293	missense	probably-damaging:0.998	c.469G>A	p.(V157M)	0	96	0	7	4399	0.1589	1.000
	41761299	missense	probably-damaging:0.972	c.475G>C	p.(D159H)	0	96	0	1	4405	0.0227	1.000
	41761311	missense	probably-damaging:1.0	c.487G>A	p.(V163M)	0	96	0	1	4405	0.0227	1.000
	41761437	missense	benign:0.133	c.613G>A	p.(V205I)	0	96	0	1	4405	0.0227	1.000
	41761449	missense	probably-damaging:1.0	c.625G>A	p.(D209N)	0	96	0	1	4405	0.0227	1.000
	41761498	missense	benign:0.029	c.674G>A	p.(R225H)	0	96	0	1	4405	0.0227	1.000
	41762068	missense	possibly-damaging:0.803	c.803A>T	p.(Q268L)	0	96	0	2	4404	0.0454	1.000
	41762092	missense	probably-damaging:0.999	c.827C>T	p.(T276M)	1	95	1.04166667	8	4398	0.1816	0.177
	41762148	missense	probably-damaging:0.999	c.883G>A	p.(G295R)	0	96	0	1	4405	0.0227	1.000
	41762184	missense	probably-damaging:1.0	c.919G>T	p.(G307C)	0	96	0	2	4404	0.0454	1.000
	41762212	missense	possibly-damaging:0.662	c.947T>A	p.(I316N)	0	96	0	1	4405	0.0227	1.000
	41766218	missense	benign:0.081	c.994G>A	p.(G332S)	0	96	0	2	4404	0.0454	1.000
CEACAM7: NM_006890	41677443	missense	benign:0.045	c.767T>C	p.(I256T)	0	96	0.000	1	4405	0.023	1.000
	41683797	missense	probably-damaging:0.997	c.694C>G	p.(L232V)	0	96	0.000	1	4405	0.023	1.000
	41683859	missense	benign:0.0	c.632T>G	p.(I211R)	0	96	0.000	1	4405	0.023	1.000
	41683943	missense	benign:0.019	c.548A>G	p.(N183S)	0	96	0.000	1	4405	0.023	1.000
	41686913	missense	probably-damaging:0.999	c.373G>T	p.(V125F)	0	96	0.000	32	4374	0.726	1.000
	41686921	missense	probably-damaging:1.0	c.365C>T	p.(T122I)	0	96	0.000	1	4405	0.023	1.000
	41686972	missense	probably-damaging:0.986	c.314A>G	p.(N105S)	0	96	0.000	1	4405	0.023	1.000
	41686986	missense	probably-damaging:1.0	c.300G>C	p.(E100D)	0	96	0.000	1	4405	0.023	1.000
	41686991	nonsense	.	c.295C>T	p.(R99*)	0	96	0.000	1	4405	0.023	1.000
	41687016	frameshift	.	c.269del1	p.(N90Mfs*20)	0	96	0.000	1	4261	0.023	1.000
	41687036	missense	benign:0.058	c.250A>G	p.(K84E)	0	96	0.000	1	4405	0.023	1.000
	41687155	missense	possibly-damaging:0.585	c.131A>G	p.(N44S)	0	96	0.000	1	4405	0.023	1.000
	41687161	missense	probably-damaging:1.0	c.125C>T	p.(P42L)	0	96	0.000	2	4404	0.045	1.000
	41687165	missense	benign:0.156	c.121G>A	p.(V41M)	0	96	0.000	0	4406	0.000	1.000

	41688101	splice	.	c.64+1G>T		0	96	0.000	1	4405	0.023	1.000
	41688164	missense	probably-damaging:1.0	c.2T>G	p.(M1R)	0	96	0.000	7	4399	0.159	1.000
CEACAM8: NM_001816	42588820	missense	benign:0.005	c.922C>T	p.(R308C)	0	96	0	1	4405	0.0227	1.000
	42588859	missense	probably-damaging:0.96	c.883G>A	p.(G295R)	0	96	0	2	4404	0.0454	1.000
	42588874	missense	probably-damaging:0.997	c.868A>G	p.(T290A)	0	96	0	1	4405	0.0227	1.000
	42588969	missense	benign:0.015	c.773C>T	p.(S258F)	0	96	0	1	4405	0.0227	1.000
	42589003	missense	benign:0.001	c.739A>G	p.(T247A)	4	92	4.16666667	41	4365	0.9305	0.015
	42589020	missense	benign:0.001	c.722C>T	p.(T241I)	0	96	0	1	4405	0.0227	1.000
	42589528	missense	possibly-damaging:0.498	c.632G>A	p.(G211E)	0	96	0	4	4402	0.0908	1.000
	42589540	missense	benign:0.027	c.620G>A	p.(R207K)	0	96	0	1	4405	0.0227	1.000
	42589549	missense	benign:0.001	c.611G>A	p.(S204N)	0	96	0	1	4405	0.0227	1.000
	42589608	frameshift	.	c.550_551del2	p.(L185Pfs*24)	0	96	0	9	4255	0.2111	1.000
	42589612	missense	probably-damaging:1.0	c.548A>C	p.(Q183P)	0	96	0	2	4404	0.0454	1.000
	42589636	missense	possibly-damaging:0.836	c.524C>T	p.(T175I)	0	96	0	1	4405	0.0227	1.000
	42589735	missense	benign:0.005	c.425C>T	p.(P142L)	0	96	0	2	4404	0.0454	1.000
	42593603	missense	probably-damaging:0.995	c.362C>A	p.(T121N)	0	96	0	4	4402	0.0908	1.000
	42593636	missense	benign:0.0	c.329G>A	p.(R110Q)	0	96	0	34	4372	0.7717	1.000
	42593675	missense	benign:0.043	c.290A>T	p.(N97I)	0	96	0	1	4405	0.0227	1.000
CEACAM16: NM_001039213	44703354	missense	benign:0.0	c.43T>C	p.(F15L)	0	96	0	33	3975	0.8234	1.000
	44703427	missense	probably-damaging:1.0	c.116T>A	p.(L39Q)	0	96	0	1	4149	0.0241	1.000
	44703445	missense	benign:0.013	c.134C>T	p.(S45L)	0	96	0	3	4147	0.0723	1.000
	44703484	missense	possibly-damaging:0.473	c.173C>T	p.(T58I)	0	96	0	1	4173	0.0240	1.000
	44703586	frameshift	.	c.276del1	p.(L93Wfs*125)	0	96	0	15	3941	0.3792	1.000
	44703597	frameshift	.	c.287_290del4	p.(Q96Pfs*121)	0	96	0	7	3937	0.1775	1.000
	44703663	missense	probably-damaging:0.973	c.352G>A	p.(E118K)	0	96	0	4	4144	0.0964	1.000
	44704035	missense	benign:0.086	c.400A>G	p.(T134A)	0	96	0	1	4043	0.0247	1.000
	44704143	missense	benign:0.122	c.508G>A	p.(A170T)	0	96	0	3	3879	0.0773	1.000

	44704224	missense	probably-damaging:1.0	c.589G>A	p.(G197S)	0	96	0	1	3815	0.0262	1.000
	44705685	missense	probably-damaging:0.995	c.757G>A	p.(V253M)	0	96	0	1	4213	0.0237	1.000
	44705709	missense	possibly-damaging:0.824	c.781G>A	p.(E261K)	0	96	0	11	4209	0.2607	1.000
	44705784	frameshift	.	c.857del1	p.(Q287Rfs*34)	0	96	0	11	4101	0.2675	1.000
	44705784	missense	benign:0.0	c.856G>A	p.(A286T)	0	96	0	1	4281	0.0234	1.000
	44705797	missense	probably-damaging:1.0	c.869C>T	p.(T290M)	1	95	1.04166667	5	4289	0.1164	1.000
	44707884	missense	probably-damaging:1.0	c.964G>A	p.(V322M)	0	96	0	1	4333	0.0231	1.000
	44708011	missense	probably-damaging:0.997	c.1091A>T	p.(Q364L)	0	96	0	1	3967	0.0252	1.000
CEACAM18: NM_001080405	51478692	missense	benign:0.003	c.233T>C	p.(M78T)	0	96	0	31	3875	0.7937	1.000
	51480485	missense	benign:0.011	c.388G>T	p.(A130S)	0	96	0	1	4067	0.0246	1.000
	51480639	missense	benign:0.11	c.542G>A	p.(G181D)	0	96	0	5	3941	0.1267	1.000
	51480667	nonsense	.	c.570G>A	p.(W190*)	0	96	0	1	3913	0.0255	1.000
	51481473	missense	benign:0.243	c.664A>G	p.(T222A)	2	94	2.08333333	30	4110	0.7246	0.163
	51481516	missense	benign:0.0	c.707C>T	p.(T236I)	0	96	0	1	4123	0.0242	1.000
	51481530	missense	probably-damaging:1.0	c.721C>T	p.(R241W)	0	96	0	2	4072	0.0491	1.000
	51481531	missense	probably-damaging:0.999	c.722G>A	p.(R241Q)	0	96	0	1	4059	0.0246	1.000
	51481645	missense	benign:0.037	c.836G>A	p.(R279H)	0	96	0	1	3931	0.0254	1.000
	51483053	missense	benign:0.013	c.893A>T	p.(D298V)	0	96	0	1	3929	0.0254	1.000
	51483079	missense	benign:0.025	c.919G>A	p.(E307K)	0	96	0	1	3957	0.0253	1.000
	51483194	missense	benign:0.167	c.1034C>T	p.(S345L)	0	96	0	2	4104	0.0487	1.000
	51483197	missense	benign:0.115	c.1037G>A	p.(S346N)	0	96	0	5	4133	0.1208	1.000
	51483217	missense	probably-damaging:1.0	c.1057G>A	p.(G353S)	0	96	0	1	4177	0.0239	1.000
	51483229	missense	probably-damaging:1.0	c.1069T>G	p.(C357G)	0	96	0	1	4189	0.0239	1.000
CEACAM19: NM_020219	44672697	missense	benign:0.051	c.157G>C	p.(V53L)	0	96	0	2	4404	0.0454	1.000
	44672757	missense	probably-damaging:1.0	c.217G>A	p.(G73R)	0	96	0	3	4403	0.0681	1.000
	44672796	missense	possibly-damaging:0.607	c.256C>T	p.(R86W)	0	96	0	14	4392	0.3177	1.000
	44672829	nonsense	.	c.289C>T	p.(R97*)	0	96	0	1	4405	0.0227	1.000

	44672872	missense	benign:0.015	c.332G>A	p.(R111H)	0	96	0	1	4405	0.0227	1.000
	44676299	missense	benign:0.0	c.453C>A	p.(H151Q)	0	96	0	1	4405	0.0227	1.000
	44676336	missense	probably-damaging:0.969	c.490A>T	p.(I164F)	0	96	0	1	4405	0.0227	1.000
	44680325	missense	benign:0.087	c.697C>T	p.(H233Y)	0	96	0	18	4388	0.4085	1.000
	44681259	missense	probably-damaging:0.989	c.739C>G	p.(P247A)	0	96	0	1	4405	0.0227	1.000
	44682595	missense	benign:0.273	c.821C>T	p.(A274V)	0	96	0	3	4393	0.0682	1.000
	44683452	missense	probably-damaging:1.0	c.865G>A	p.(D289N)	0	96	0	2	4392	0.0455	1.000
CEACAM20: NM_001102597	44511133	missense	benign:0.017	c.1635G>A	p.(R545H)	0	96	0	1	3837	0.0261	1.000
	44511144	frameshift	.	c.1623del1	p.(F542Sfs*56)	0	96	0	2	3710	0.0539	1.000
	44511154	missense	benign:0.106	c.1614C>T	p.(T538M)	0	96	0	2	3852	0.0519	1.000
	44512044	missense	possibly-damaging:0.61	c.1549G>C	p.(Q516H)	0	96	0	7	4027	0.1735	1.000
	44512056	nonsense	.	c.1537C>A	p.(C512*)	0	96	0	11	3985	0.2753	1.000
	44512868	missense	benign:0.028	c.1514G>A	p.(E505K)	0	96	0	2	3798	0.0526	1.000
	44516976	missense	possibly-damaging:0.459	c.1279C>T	p.(R427C)	0	96	0	12	4074	0.2937	1.000
	44520480	missense	benign:0.166	c.1024A>G	p.(I342V)	0	96	0	1	3897	0.0257	1.000
	44520501	missense	probably-damaging:0.999	c.1003A>G	p.(S335G)	0	96	0	1	3869	0.0258	1.000
	44520634	missense	benign:0.005	c.870T>G	p.(S290R)	0	96	0	1	4247	0.0235	1.000
	44520729	missense	possibly-damaging:0.911	c.775G>A	p.(V259M)	0	96	0	1	4169	0.0240	1.000
	44520741	missense	benign:0.035	c.763A>G	p.(M255V)	0	96	0	1	4149	0.0241	1.000
	44522793	missense	possibly-damaging:0.789	c.592G>A	p.(A198T)	0	96	0	7	4209	0.1660	1.000
	44522819	missense	benign:0.438	c.566C>T	p.(A189V)	0	96	0	1	4243	0.0236	1.000
	44524198	missense	probably-damaging:0.971	c.260C>T	p.(T87I)	0	96	0	6	4238	0.1414	1.000
	44525131	missense	benign:0.0	c.166A>G	p.(R56G)	1	95	1.04166667	7	3995	0.1749	0.173
	44525199	missense	benign:0.275	c.98C>T	p.(T33I)	0	96	0	2	4186	0.0478	1.000
	44529503	missense	benign:0.027	c.7C>T	p.(P3S)	0	96	0	1	4181	0.0239	1.000
CEACAM21: NM_001098506	41576318	nonsense	.	c.44G>A	p.(W15*)	0	96	0	1	4405	0.023	1.000
	41577200	missense	benign:0.384	c.65C>T	p.(A22V)	0	96	0	1	4199	0.024	1.000

	41577225	frameshift	.	c.91del1	p.(T32Pfs*47)	1	95	1.042	24	3970	0.601	0.449
	41577397	missense	possibly-damaging:0.744	c.262G>A	p.(V88I)	0	96	0	7	4319	0.162	1.000
	41577485	missense	probably-damaging:0.994	c.350C>T	p.(T117M)	0	96	0	1	4365	0.023	1.000
	41577545	missense	benign:0.395	c.410A>G	p.(H137R)	0	96	0	1	4387	0.023	1.000
	41577560	splice	.	c.424+1G>A	.,.	0	96	0	1	4391	0.023	1.000
	41579398	frameshift	.	c.471_472del2	p.(K159Gfs*11)	0	96	0	5	3949	0.126	1.000
	41579410	missense	probably-damaging:0.992	c.482C>T	p.(S161F)	0	96	0	1	4175	0.024	1.000
	41579476	missense	probably-damaging:0.994	c.548G>A	p.(R183H)	1	95	1.042	16	4062	0.392	0.327
	41579527	missense	probably-damaging:0.988	c.599C>T	p.(T200I)	0	96	0	1	4085	0.024	1.000
	41584379	missense	benign:0.007	c.733G>C	p.(V245L)	0	96	0	2	4346	0.046	1.000
	41585445	missense	benign:0.066	c.800C>T	p.(A267V)	0	96	0	2	4370	0.046	1.000
	41585449	missense	probably-damaging:0.993	c.804C>A	p.(S268R)	0	96	0	19	4353	0.435	1.000

Supplemental Table 4.2: Full list of rare (MAF <1%) Stop Gain, Frameshifting, splice-site and missense mutations identified in the "White or Caucasian" TCGA-COAD cohort and the CEACAM EVS EA cohort

Gene	Chr 19 Position	Function	Functonal Prediction - PH	cDNA	Protein	TCGA EA Colon			EVS EA			EA Indivdual P-values
						Alt Allele Frequency	Ref Allele Frequency	MAF (%)	Alt Allele Frequency	Ref Allele Frequency	MAF (%)	
CEACAM1: NM_001184815	42509115	frameshift	.	c.1379del1	p.(K460Sfs*23)	0	398	0	35	8219	0.4240	0.408
	42509177	missense	benign:0.0	c.1318C>A	p.(Q440K)	0	398	0	1	8599	0.0116	1.000
	42511629	splice	.	c.1182-1G>A	.,.,.,.,.,.,.	0	398	0	1	8599	0.0116	1.000
	42521279	missense	probably-damaging:0.984	c.946A>C	p.(I316L)	0	398	0	1	8599	0.0116	1.000
	42521281	missense	possibly-damaging:0.602	c.944C>T	p.(T315M)	0	398	0	1	8599	0.0116	1.000
	42521369	missense	probably-damaging:0.983	c.856A>G	p.(I286V)	0	398	0	1	8599	0.0116	1.000

	42521433	frameshift	.	c.791_792insG	p.(N264Kfs*25)	0	398	0	1	8253	0.0121	1.000
	42521953	missense	benign:0.095	c.674G>A	p.(R225H)	0	398	0	1	8599	0.0116	1.000
	42521959	missense	benign:0.444	c.668C>T	p.(A223V)	0	398	0	1	8599	0.0116	1.000
	42521970	missense	probably-damaging:1.0	c.657C>G	p.(N219K)	0	398	0	1	8599	0.0116	1.000
	42522070	missense	possibly-damaging:0.924	c.557C>A	p.(P186Q)	0	398	0	2	8598	0.0233	1.000
	42522073	frameshift	.	c.553_554insAGGC	p.(L185Qfs*26)	0	398	0	11	8241	0.1333	1.000
	42522185	missense	possibly-damaging:0.956	c.442T>C	p.(S148P)	0	398	0	4	8596	0.0465	1.000
	42522202	missense	benign:0.034	c.425C>T	p.(P142L)	0	398	0	1	8599	0.0116	1.000
	42527041	missense	benign:0.307	c.424C>T	p.(P142S)	0	398	0	1	8599	0.0116	1.000
	42527096	missense	benign:0.387	c.369A>C	p.(Q123H)	3	395	0.75376884	28	8572	0.3256	0.156
	42527176	missense	possibly-damaging:0.917	c.289G>A	p.(G97S)	0	398	0	2	8598	0.0233	1.000
	42527203	missense	benign:0.053	c.262C>A	p.(Q88K)	0	398	0	3	8597	0.0349	1.000
	42527217	missense	benign:0.001	c.248C>T	p.(A83V)	3	395	0.75376884	22	8578	0.2558	0.096
	42527262	missense	probably-damaging:1.0	c.203A>G	p.(Y68C)	2	396	0.50251256	6	8594	0.0698	0.046
	42527331	missense	probably-damaging:1.0	c.134T>A	p.(V45D)	0	398	0	1	8599	0.0116	1.000
	42527362	missense	benign:0.078	c.103C>A	p.(Q35K)	3	395	0.75376884	22	8578	0.2558	0.096
	42528358	missense	probably-damaging:0.985	c.17C>T	p.(A6V)	0	398	0	5	8595	0.0581	1.000
CEACAM3: NM_001815	41797704	missense	probably-damaging:0.998	c.180G>C	p.(Q60H)	0	398	0	7	8593	0.0814	1.000
	41797730	missense	possibly-damaging:0.928	c.206A>G	p.(K69R)	0	398	0	1	8599	0.0116	1.000
	41797813	missense	possibly-damaging:0.587	c.289G>A	p.(G97S)	0	398	0	2	8598	0.0233	1.000
	41797831	missense	benign:0.0	c.307A>C	p.(T103P)	0	398	0	5	8595	0.0581	1.000
	41797922	missense	probably-damaging:1.0	c.398A>G	p.(E133G)	0	398	0	4	8596	0.0465	1.000

	41797924	missense	probably-damaging:0.998	c.400G>A	p.(A134T)	0	398	0	1	8599	0.0116	1.000
	41808857	missense	probably-damaging:0.968	c.469G>A	p.(V157M)	0	398	0	1	8599	0.0116	1.000
	41808931	splice	.	c.542+1G>T		0	398	0	1	8599	0.0116	1.000
	41809979	missense	benign:0.0	c.557G>A	p.(R186H)	0	398	0	4	8596	0.0465	1.000
	41810353	missense	benign:0.0	c.626C>T	p.(S209L)	0	398	0	1	8599	0.0116	1.000
CEACAM4: NM_001817	41619340	missense	possibly-damaging:0.691	c.725T>C	p.(V242A)	0	398	0.000	2	8598	0.023	1.000
	41620613	missense	benign:0.026	c.557G>T	p.(R186L)	0	398	0.000	1	8599	0.012	1.000
	41621655	missense	possibly-damaging:0.756	c.538G>A	p.(G180R)	0	398	0.000	1	8599	0.012	1.000
	41621736	missense	benign:0.019	c.457G>A	p.(V153I)	0	398	0.000	1	8599	0.012	1.000
	41621760	missense	benign:0.0	c.433G>A	p.(V145I)	0	398	0.000	1	8599	0.012	1.000
	41625607	missense	probably-damaging:0.989	c.418G>A	p.(V140I)	1	397	0.251	19	8581	0.221	0.596
	41625631	missense	possibly-damaging:0.835	c.394G>A	p.(D132N)	0	398	0.000	1	8599	0.012	1.000
	41625645	missense	benign:0.27	c.380C>T	p.(A127V)	0	398	0.000	1	8599	0.012	1.000
	41625651	missense	benign:0.09	c.374T>C	p.(I125T)	0	398	0.000	2	8598	0.023	1.000
	41625657	missense	possibly-damaging:0.896	c.368G>T	p.(R123L)	1	397	0.251	7	8593	0.081	0.304
	41625657	missense	benign:0.325	c.368G>A	p.(R123E)	2	396	0.503	0	8588	0.000	0.002
	41625708	missense	probably-damaging:1.0	c.317C>A	p.(S106Y)	0	398	0.000	10	8590	0.116	1.000
	41625714	missense	probably-damaging:0.997	c.311A>G	p.(N104S)	0	398	0.000	1	8599	0.012	1.000
	41625895	missense	possibly-damaging:0.761	c.130A>G	p.(S44G)	0	398	0.000	2	8598	0.023	1.000
	41625900	missense	probably-damaging:1.0	c.125C>T	p.(P42L)	0	398	0.000	12	8588	0.140	1.000
	41626899	splice	.	c.64+1G>C	.	0	398	0.000	1	8599	0.012	1.000

	41626951	frameshift	.	c.12_13insC	p.(S5Lfs*35)	0	398	0.000	9	8245	0.109	1.000
	41626957	missense	probably-damaging:0.996	c.7C>G	p.(P3A)	0	398	0.000	1	8599	0.012	1.000
CEACAM5: NM_004363	41709698	nonsense	.	c.83G>A	p.(W28*)	0	398	0	1	8599	0.01162791	1.000
	41709737	missense	benign:0.135	c.122C>T	p.(T41M)	0	398	0	1	8599	0.012	1.000
	41709854	missense	benign:0.101	c.239T>C	p.(I80T)	0	398	0	1	8599	0.012	1.000
	41709914	missense	benign:0.009	c.299T>C	p.(I100T)	0	398	0	3	8597	0.035	1.000
	41709917	missense	possibly-damaging:0.733	c.302T>C	p.(I101T)	0	398	0	1	8599	0.012	1.000
	41709926	missense	possibly-damaging:0.934	c.311A>G	p.(N104S)	0	398	0	1	8599	0.012	1.000
	41709932	missense	probably-damaging:0.996	c.317C>G	p.(S106C)	1	397	0.25125628	5	8595	0.058	1.000
	41709984	missense	benign:0.058	c.369C>A	p.(H123Q)	0	398	0	4	8596	0.047	1.000
	41710040	splice	.	c.424+1G>A		0	398	0	1	8599	0.012	1.000
	41714989	missense	benign:0.18	c.443C>T	p.(S148F)	0	398	0	1	8593	0.012	1.000
	41715015	missense	probably-damaging:0.993	c.469G>A	p.(V157M)	0	398	0	2	8594	0.023	1.000
	41715201	missense	probably-damaging:1.0	c.655A>C	p.(N219H)	0	398	0	3	8597	0.035	1.000
	41715219	missense	probably-damaging:0.973	c.673C>A	p.(R225S)	0	398	0	1	8599	0.012	1.000
	41715712	missense	possibly-damaging:0.94	c.766A>T	p.(N256Y)	0	398	0	1	8599	0.012	1.000
	41715713	missense	benign:0.018	c.767A>G	p.(N256S)	0	398	0	1	8599	0.012	1.000
	41715740	missense	probably-damaging:1.0	c.794C>T	p.(P265L)	0	398	0	7	8593	0.081	1.000
	41715763	missense	benign:0.024	c.817G>A	p.(V273I)	0	398	0	1	8599	0.012	1.000
	41715848	missense	benign:0.179	c.902C>T	p.(A301V)	0	398	0	5	8595	0.058	1.000
	41715865	missense	possibly-damaging:0.83	c.919G>T	p.(G307C)	0	398	0	1	8599	0.012	1.000

	41715904	missense	possibly-damaging:0.762	c.958G>T	p.(A320S)	0	398	0	1	8599	0.012	1.000
	41717499	missense	probably-damaging:0.993	c.1003G>A	p.(V335M)	0	398	0	7	8593	0.081	1.000
	41717512	missense	probably-damaging:1.0	c.1016A>T	p.(D339V)	0	398	0	1	8599	0.012	1.000
	41717515	missense	possibly-damaging:0.597	c.1019C>A	p.(A340D)	0	398	0	3	8597	0.035	1.000
	41717635	missense	probably-damaging:0.998	c.1139T>C	p.(L380P)	0	398	0	1	8599	0.012	1.000
	41717709	missense	probably-damaging:0.997	c.1213G>A	p.(D405N)	0	398	0	2	8598	0.023	1.000
	41717719	missense	benign:0.0	c.1223T>C	p.(I408T)	0	398	0	1	8599	0.012	1.000
	41718139	missense	benign:0.072	c.1249G>A	p.(D417N)	0	398	0	1	8599	0.012	1.000
	41718172	missense	probably-damaging:1.0	c.1282C>T	p.(R428C)	0	398	0	1	8599	0.012	1.000
	41718173	missense	benign:0.158	c.1283G>A	p.(R428H)	0	398	0	1	8599	0.012	1.000
	41718340	missense	benign:0.297	c.1450A>G	p.(S484G)	0	398	0	1	8599	0.012	1.000
	41719974	missense	benign:0.406	c.1537G>A	p.(V513M)	0	398	0	1	8595	0.012	1.000
	41719987	missense	benign:0.345	c.1550A>T	p.(D517V)	1	397	0.25125628	11	8589	0.128	1.000
	41720133	missense	benign:0.018	c.1696G>A	p.(A566T)	0	398	0	3	8597	0.035	1.000
	41720178	missense	possibly-damaging:0.83	c.1741C>T	p.(R581C)	0	398	0	1	8599	0.012	1.000
	41720179	missense	benign:0.001	c.1742G>A	p.(R581H)	0	398	0	7	8593	0.081	1.000
	41720951	missense	benign:0.005	c.1801C>G	p.(P601A)	0	398	0	1	8599	0.012	1.000
	41720957	missense	benign:0.015	c.1807T>G	p.(S603A)	0	398	0	1	8599	0.012	1.000
	41721053	missense	benign:0.316	c.1903C>G	p.(Q635E)	0	398	0	1	8599	0.012	1.000
	41721140	missense	benign:0.035	c.1990C>A	p.(R664S)	0	398	0	2	8598	0.023	1.000
	41727275	missense	possibly-damaging:0.884	c.2068G>A	p.(G690S)	0	398	0	1	8599	0.012	1.000
CEACAM6: NM_002483	41756630	missense	probably-damaging:1.0	c.95C>T	p.(T32I)	0	398	0	2	8598	0.023	1.000

	41756657	missense	benign:0.013	c.122C>T	p.(T41M)	0	398	0	13	8587	0.151	1.000
	41756666	missense	benign:0.0	c.131A>G	p.(N44S)	0	398	0	1	8599	0.012	1.000
	41756671	missense	possibly-damaging:0.454	c.136G>A	p.(A46T)	0	398	0	1	8599	0.012	1.000
	41756720	missense	probably-damaging:0.971	c.185G>A	p.(R62H)	0	398	0	1	8599	0.012	1.000
	41756746	missense	benign:0.043	c.211G>A	p.(E71K)	0	398	0	1	8599	0.012	1.000
	41756789	missense	benign:0.0	c.254G>A	p.(G85E)	1	397	0.25125628	7	8589	0.081	1.000
	41756810	missense	probably-damaging:0.991	c.275G>T	p.(G92V)	0	398	0	1	8599	0.012	1.000
	41756812	missense	possibly-damaging:0.873	c.277C>G	p.(P93A)	0	398	0	1	8599	0.012	1.000
	41756834	missense	benign:0.011	c.299C>T	p.(T100I)	0	398	0	5	8595	0.058	1.000
	41756848	missense	probably-damaging:0.971	c.313G>A	p.(A105T)	0	398	0	1	8599	0.012	1.000
	41756923	missense	benign:0.082	c.388G>C	p.(V130L)	0	398	0	1	8599	0.012	1.000
	41756941	missense	possibly-damaging:0.517	c.406G>A	p.(G136R)	0	398	0	1	8599	0.012	1.000
	41761293	missense	probably-damaging:0.998	c.469G>A	p.(V157M)	0	398	0	1	8595	0.012	1.000
	41761299	missense	probably-damaging:0.972	c.475G>C	p.(D159H)	0	398	0	7	8593	0.081	1.000
	41761311	missense	probably-damaging:1.0	c.487G>A	p.(V163M)	0	398	0	1	8599	0.012	1.000
	41761354	missense	probably-damaging:1.0	c.530T>C	p.(L177P)	0	398	0	1	8599	0.012	1.000
	41761359	missense	probably-damaging:1.0	c.535T>A	p.(W179R)	0	398	0	2	8598	0.023	1.000
	41761449	missense	probably-damaging:1.0	c.625G>A	p.(D209N)	0	398	0	14	8586	0.163	1.000
	41761498	missense	benign:0.029	c.674G>A	p.(R225H)	0	398	0	1	8599	0.012	1.000
	41762076	missense	probably-damaging:1.0	c.811T>C	p.(W271R)	0	398	0	1	8599	0.012	1.000
	41762148	missense	probably-damaging:0.999	c.883G>A	p.(G295R)	0	398	0	1	8599	0.012	1.000

	41762184	missense	probably-damaging:1.0	c.919G>T	p.(G307C)	1	397	0.25125628	7	8593	0.081	0.304
	41762200	missense	probably-damaging:0.999	c.935C>G	p.(T312R)	0	398	0	1	8599	0.012	1.000
	41762206	missense	benign:0.177	c.941C>T	p.(T314M)	0	398	0	1	8599	0.012	1.000
	41762212	missense	possibly-damaging:0.662	c.947T>A	p.(I316N)	1	397	0.25125628	10	8590	0.116	0.392
CEACAM7: NM_006890	41683785	missense	benign:0.061	c.706T>C	p.(Y236H)	3	395	0.754	25	8575	0.291	0.125
	41683788	missense	probably-damaging:0.978	c.703C>T	p.(R235C)	0	398	0.000	1	8599	0.012	1.000
	41683815	missense	benign:0.318	c.676C>T	p.(R226C)	0	398	0.000	1	8599	0.012	1.000
	41683820	missense	benign:0.082	c.671C>T	p.(A224V)	0	398	0.000	1	8599	0.012	1.000
	41683859	missense	benign:0.0	c.632T>G	p.(I211R)	0	398	0.000	1	8599	0.012	1.000
	41684021	missense	possibly-damaging:0.73	c.470C>T	p.(P157L)	0	398	0.000	1	8599	0.012	1.000
	41686865	missense	probably-damaging:0.996	c.421G>A	p.(V141I)	0	398	0.000	2	8598	0.023	1.000
	41686889	nonsense	.	c.397G>T	p.(E133*)	0	398	0.000	1	8599	0.012	1.000
	41686940	missense	benign:0.188	c.346A>G	p.(N116D)	3	395	0.754	23	8577	0.267	0.106
	41686990	missense	probably-damaging:1.0	c.296G>A	p.(R99Q)	0	398	0.000	1	8599	0.012	1.000
	41686991	nonsense	.	c.295C>T	p.(R99*)	0	398	0.000	1	8599	0.012	1.000
	41687016	frameshift	.	c.269del1	p.(N90Mfs*20)	0	398	0.000	4	8240	0.049	1.000
	41687165	missense	benign:0.156	c.121G>A	p.(V41M)	0	398	0.000	1	8599	0.012	1.000
	41688101	splice	.	c.64+1G>T		2	396	0.503	36	8564	0.419	0.685
	41688114	missense	possibly-damaging:0.839	c.52C>T	p.(L18F)	0	398	0.000	2	8598	0.023	1.000
CEACAM8: NM_001816	42583276	missense	probably-damaging:0.999	c.1020T>G	p.(I340M)	2	396	0.50251256	18	8582	0.2093	0.221
	42583314	frameshift	.	c.981_982insT	p.(P328Sfs*6)	0	398	0	3	8251	0.0363	1.000
	42583326	nonsense	.	c.970C>T	p.(Q324*)	0	398	0	1	8599	0.0116	1.000

	42588822	missense	probably-damaging:1.0	c.920G>T	p.(G307V)	0	398	0	1	8599	0.0116	1.000
	42588849	missense	benign:0.069	c.893C>T	p.(A298V)	0	398	0	7	8593	0.0814	1.000
	42588859	missense	probably-damaging:0.96	c.883G>A	p.(G295R)	0	398	0	1	8599	0.0116	1.000
	42588873	missense	probably-damaging:1.0	c.869C>T	p.(T290I)	0	398	0	1	8599	0.0116	1.000
	42588874	missense	probably-damaging:0.997	c.868A>G	p.(T290A)	0	398	0	10	8590	0.1163	1.000
	42589498	missense	benign:0.0	c.662C>T	p.(A221V)	0	398	0	1	8599	0.0116	1.000
	42589503	missense	probably-damaging:0.989	c.657C>G	p.(N219K)	0	398	0	3	8597	0.0349	1.000
	42589522	missense	probably-damaging:1.0	c.638A>G	p.(Y213C)	0	398	0	1	8599	0.0116	1.000
	42589525	missense	possibly-damaging:0.892	c.635C>G	p.(P212R)	0	398	0	2	8598	0.0233	1.000
	42589608	frameshift	.	c.550_551del2	p.(L185Pfs*24)	0	398	0	11	8241	0.1333	1.000
	42589675	missense	benign:0.007	c.485C>T	p.(A162V)	0	398	0	1	8599	0.0116	1.000
	42589694	missense	probably-damaging:0.967	c.466C>G	p.(P156A)	0	398	0	1	8599	0.0116	1.000
	42589735	missense	benign:0.005	c.425C>T	p.(P142L)	2	396	0.50251256	1	8599	0.0116	0.006
	42593541	missense	benign:0.056	c.424C>A	p.(P142T)	1	397	0.25125628	5	8595	0.0581	0.238
	42593547	missense	probably-damaging:0.986	c.418G>A	p.(V140I)	0	398	0	1	8599	0.0116	1.000
	42593600	frameshift	.	c.364del1	p.(L122Yfs*4)	0	398	0	2	8252	0.0242	1.000
	42593732	missense	benign:0.111	c.233G>A	p.(R78Q)	0	398	0	1	8599	0.0116	1.000
	42593754	missense	benign:0.037	c.211G>A	p.(E71K)	0	398	0	1	8599	0.0116	1.000
	42593777	missense	benign:0.071	c.188G>A	p.(R63H)	0	398	0	1	8599	0.0116	1.000
	42594777	missense	benign:0.024	c.52C>T	p.(L18F)	0	398	0	1	8599	0.0116	1.000
	42594779	missense	possibly-damaging:0.698	c.50G>T	p.(G17V)	2	396	0.50251256	18	8582	0.2093	0.221

	42594794	missense	benign:0.001	c.35G>A	p.(R12H)	0	398	0	1	8599	0.0116	1.000
CEACAM16: NM_001039213	44701461	missense	benign:0.006	c.5C>T	p.(A2V)	0	398	0	1	8397	0.0119	1.000
	44703372	missense	possibly-damaging:0.922	c.61G>A	p.(E21K)	0	398	0	1	8331	0.0120	1.000
	44703406	missense	benign:0.009	c.95G>T	p.(S32I)	0	398	0	11	8321	0.1320	1.000
	44703420	missense	benign:0.159	c.109G>A	p.(V37I)	0	398	0	1	8375	0.0119	1.000
	44703445	missense	benign:0.013	c.134C>T	p.(S45L)	0	398	0	1	8393	0.0119	1.000
	44703486	missense	possibly-damaging:0.898	c.175C>T	p.(L59F)	0	398	0	1	8419	0.0119	1.000
	44703508	missense	benign:0.002	c.197C>T	p.(A66V)	0	398	0	2	8402	0.0238	1.000
	44703519	missense	probably-damaging:0.973	c.208G>A	p.(V70M)	0	398	0	1	8393	0.0119	1.000
	44703573	missense	probably-damaging:1.0	c.262C>T	p.(R88C)	0	398	0	1	8329	0.0120	1.000
	44703586	frameshift	.	c.276del1	p.(L93Wfs*125)	0	398	0	19	7945	0.2386	1.000
	44703597	frameshift	.	c.287_290del4	p.(Q96Pfs*121)	0	398	0	14	7964	0.1755	1.000
	44703663	missense	probably-damaging:0.973	c.352G>A	p.(E118K)	2	396	0.50251256	23	8371	0.2740	0.314
	44704068	missense	probably-damaging:0.999	c.433C>T	p.(R145C)	0	398	0	1	8337	0.0120	1.000
	44704143	missense	benign:0.122	c.508G>A	p.(A170T)	0	398	0	19	8183	0.2317	1.000
	44704212	missense	probably-damaging:1.0	c.577C>T	p.(R193W)	0	398	0	1	8069	0.0124	1.000
	44704224	missense	probably-damaging:1.0	c.589G>A	p.(G197S)	0	398	0	1	7963	0.0126	1.000
	44704290	missense	probably-damaging:1.0	c.655G>A	p.(V219M)	0	398	0	1	7283	0.0137	1.000
	44705590	missense	benign:0.001	c.662T>A	p.(F221Y)	0	398	0	2	8352	0.0239	1.000
	44705626	missense	benign:0.166	c.698C>G	p.(T233S)	0	398	0	1	8401	0.0119	1.000
	44705709	missense	possibly-damaging:0.824	c.781G>A	p.(E261K)	0	398	0	1	8477	0.0118	1.000
	44705711	missense	possibly-damaging:0.625	c.783G>C	p.(E261D)	0	398	0	1	8477	0.0118	1.000

	44705784	frameshift	.	c.857del1	p.(Q287Rfs*34)	0	398	0	47	8067	0.5792	0.172
	44707867	missense	benign:0.068	c.947C>T	p.(A316V)	0	398	0	1	8507	0.0118	1.000
	44707876	missense	benign:0.002	c.956C>T	p.(T319M)	0	398	0	1	8505	0.0118	1.000
	44707906	missense	benign:0.039	c.986C>T	p.(T329M)	0	398	0	1	8493	0.0118	1.000
	44707983	missense	possibly-damaging:0.822	c.1063G>A	p.(A355T)	0	398	0	1	8291	0.0121	1.000
	44708044	missense	probably-damaging:1.0	c.1124C>T	p.(A375V)	0	398	0	1	8301	0.0120	1.000
	44708169	missense	benign:0.011	c.1249G>C	p.(V417L)	0	398	0	9	8445	0.1065	1.000
CEACAM18: NM_001080405	51478668	missense	benign:0.014	c.209G>A	p.(S70N)	0	398	0	1	8313	0.0120	1.000
	51480390	missense	benign:0.005	c.293C>T	p.(T98I)	0	398	0	1	8227	0.0122	1.000
	51480402	missense	possibly-damaging:0.668	c.305A>C	p.(K102T)	0	398	0	1	8245	0.0121	1.000
	51480428	missense	probably-damaging:0.965	c.331G>T	p.(D111Y)	0	398	0	1	8327	0.0120	1.000
	51480437	missense	benign:0.213	c.340C>T	p.(P114S)	0	398	0	1	8343	0.0120	1.000
	51480462	nonsense	.	c.365G>A	p.(W122*)	0	398	0	1	8337	0.0120	1.000
	51480479	missense	benign:0.126	c.382G>A	p.(D128N)	0	398	0	1	8387	0.0119	1.000
	51480564	missense	benign:0.069	c.467A>G	p.(N156S)	0	398	0	1	8339	0.0120	1.000
	51481406	missense	benign:0.003	c.597T>G	p.(N199K)	0	398	0	1	8305	0.0120	1.000
	51481419	missense	benign:0.32	c.610G>A	p.(V204I)	1	397	0.25125628	3	8345	0.0359	1.000
	51481563	missense	benign:0.022	c.754G>A	p.(V252I)	0	398	0	21	8333	0.2514	0.623
	51481644	missense	possibly-damaging:0.955	c.835C>T	p.(R279C)	0	398	0	1	8317	0.0120	1.000
	51483073	missense	benign:0.013	c.913A>T	p.(T305S)	0	398	0	2	8358	0.0239	1.000
	51483111	missense	benign:0.432	c.951C>G	p.(I317M)	0	398	0	1	8365	0.0120	1.000
	51483125	missense	benign:0.0	c.965T>C	p.(L322P)	0	398	0	5	8349	0.0599	1.000
	51483160	missense	benign:0.071	c.1000C>G	p.(L334V)	0	398	0	2	8366	0.0239	1.000
	51483185	missense	benign:0.255	c.1025T>C	p.(M342T)	0	398	0	1	8387	0.0119	1.000

	51483197	missense	benign:0.115	c.1037G>A	p.(S346N)	1	397	0.25125628	45	8359	0.5355	0.723
	51483221	missense	benign:0.304	c.1061G>A	p.(R354H)	0	398	0	1	8441	0.0118	1.000
	51483227	missense	probably-damaging:0.994	c.1067G>A	p.(R356Q)	0	398	0	1	8457	0.0118	1.000
	51483229	missense	probably-damaging:1.0	c.1069T>G	p.(C357G)	2	396	0.50251256	5	8455	0.0591	0.036
	51483274	missense	benign:0.102	c.1114G>A	p.(V372I)	0	398	0	2	8456	0.0236	1.000
	51483284	missense	benign:0.013	c.1124A>G	p.(Q375R)	2	396	0.50251256	5	8453	0.0591	0.036
CEACAM19: NM_020219	44672678	missense	possibly-damaging:0.611	c.138C>G	p.(N46K)	0	398	0	9	8591	0.1047	1.000
	44672796	missense	possibly-damaging:0.607	c.256C>T	p.(R86W)	0	398	0	1	8599	0.0116	1.000
	44672797	missense	probably-damaging:1.0	c.257G>A	p.(R86Q)	0	398	0	2	8598	0.0233	1.000
	44672851	missense	probably-damaging:1.0	c.311A>T	p.(N104I)	0	398	0	1	8599	0.0116	1.000
	44672868	missense	probably-damaging:1.0	c.328C>T	p.(R110C)	0	398	0	3	8597	0.0349	1.000
	44672869	missense	benign:0.095	c.329G>A	p.(R110H)	0	398	0	1	8599	0.0116	1.000
	44672874	missense	possibly-damaging:0.742	c.334G>A	p.(A112T)	0	398	0	1	8599	0.0116	1.000
	44672923	frameshift	.	c.384_387del4	p.(E129Gfs*3)	0	398	0	2	8252	0.0242	1.000
	44676373	missense	benign:0.087	c.527G>A	p.(C176Y)	1	397	0.25125628	3	8597	0.0349	1.000
	44676394	missense	probably-damaging:0.966	c.548C>T	p.(T183I)	0	398	0	4	8596	0.0465	1.000
	44678878	missense	benign:0.0	c.601T>C	p.(S201P)	0	398	0	1	8599	0.0116	1.000
	44680331	missense	benign:0.004	c.703G>T	p.(A235S)	1	397	0.25125628	6	8594	0.0698	0.272
	44681254	missense	probably-damaging:0.997	c.734C>G	p.(P245R)	0	398	0	1	8599	0.0116	1.000
	44681293	missense	benign:0.01	c.773G>C	p.(R258T)	4	394	1.00502513	8	8592	0.0930	0.001
	44681313	splice	.	c.792+1G>C	.	0	398	0	1	8599	0.0116	1.000

	44683459	missense	probably-damaging:0.982	c.872C>T	p.(A291V)	0	398	0	1	8573	0.0117	1.000
CEACAM20: NM_001102597	44511077	missense	possibly-damaging:0.791	c.1691C>A	p.(L564I)	0	398	0	2	8226	0.0243	1.000
	44511134	missense	probably-damaging:0.985	c.1634C>T	p.(R545C)	0	398	0	7	8233	0.0850	1.000
	44511144	frameshift	.	c.1623del1	p.(F542Sfs*56)	1	397	0.25125628	2	7894	0.0253	0.137
	44511154	missense	benign:0.106	c.1614C>T	p.(T538M)	0	398	0	3	8249	0.0364	1.000
	44512936	missense	benign:0.062	c.1445C>T	p.T482I	2	396	0.50251256	0	8070	0.0000	0.002
	44513242	missense	benign:0.024	c.1357A>G	p.(I453V)	0	398	0	1	8347	0.0120	1.000
	44513290	splice	.	c.1310-1G>C	.,.,.,.	0	398	0	3	8337	0.0360	1.000
	44516957	missense	probably-damaging:0.993	c.1298T>G	p.(V433G)	0	398	0	1	8357	0.0120	1.000
	44516978	missense	probably-damaging:0.972	c.1277C>T	p.(A426V)	0	398	0	7	8357	0.0837	1.000
	44520501	missense	probably-damaging:0.999	c.1003A>G	p.(S335G)	0	398	0	7	8257	0.0847	1.000
	44520503	missense	probably-damaging:1.0	c.1001G>A	p.(R334Q)	1	397	0.25125628	1	8251	0.0121	0.090
	44520546	missense	probably-damaging:1.0	c.958G>A	p.(G320R)	0	398	0	2	8286	0.0241	1.000
	44520548	missense	probably-damaging:0.999	c.956C>T	p.(T319M)	0	398	0	1	8317	0.0120	1.000
	44520581	missense	probably-damaging:0.971	c.923C>T	p.(T308I)	0	398	0	3	8423	0.0356	1.000
	44520647	missense	benign:0.011	c.857A>G	p.(Q286R)	0	398	0	1	8483	0.0118	1.000
	44520672	missense	possibly-damaging:0.945	c.832A>G	p.(T278A)	0	398	0	1	8483	0.0118	1.000
	44520729	missense	possibly-damaging:0.911	c.775G>A	p.(V259M)	0	398	0	1	8445	0.0118	1.000
	44522643	nonsense	.	c.742C>T	p.(R248*)	0	398	0	1	8301	0.0120	1.000
	44522910	missense	probably-damaging:1.0	c.475G>A	p.(G159S)	0	398	0	1	8435	0.0119	1.000

	44524010	missense	probably-damaging:0.999	c.448G>A	p.(D150N)	0	398	0	1	8193	0.0122	1.000
	44524039	missense	probably-damaging:1.0	c.419C>G	p.(A140G)	0	398	0	1	8391	0.0119	1.000
	44524120	missense	probably-damaging:1.0	c.338G>A	p.(R113H)	0	398	0	3	8457	0.0355	1.000
	44524198	missense	probably-damaging:0.971	c.260C>T	p.(T87I)	1	397	0.25125628	30	8436	0.3544	1.000
	44525109	missense	possibly-damaging:0.881	c.188G>A	p.(R63K)	1	397	0.25125628	1	8313	0.0120	0.089
	44525164	missense	probably-damaging:0.996	c.133G>A	p.(E45K)	0	398	0	1	8413	0.0119	1.000
	44525166	missense	possibly-damaging:0.895	c.131G>A	p.(S44N)	0	398	0	1	8413	0.0119	1.000
	44525185	missense	probably-damaging:0.97	c.112C>G	p.(P38A)	0	398	0	1	8429	0.0119	1.000
	44525230	missense	benign:0.291	c.67G>A	p.(V23I)	0	398	0	1	8425	0.0119	1.000
CEACAM21: NM_001098506	41577225	frameshift	.	c.91del1	p.(T32Pfs*47)	0	398	0.000	2	8096	0.02	1.000
	41577290	missense	benign:0.007	c.155A>T	p.(H52L)	0	398	0.000	1	8391	0.01	1.000
	41577397	missense	possibly-damaging:0.744	c.262G>A	p.(V88I)	0	398	0.000	1	8569	0.01	1.000
	41577427	nonsense	.	c.292C>T	p.(R98*)	0	398	0.000	1	8585	0.01	1.000
	41579398	frameshift	.	c.471_472del2	p.(K159Gfs*11)	0	398	0.000	5	8071	0.06	1.000
	41579407	missense	benign:0.274	c.479G>T	p.(G160V)	0	398	0.000	1	8461	0.01	1.000
	41579476	missense	probably-damaging:0.994	c.548G>A	p.(R183H)	0	398	0.000	1	8377	0.01	1.000
	41579527	missense	probably-damaging:0.988	c.599C>T	p.(T200I)	1	397	0.251	16	8378	0.19	0.545
	41579602	missense	probably-damaging:1.0	c.674G>A	p.(S225N)	0	398	0.000	1	8315	0.01	1.000
	41585450	missense	probably-damaging:0.966	c.805G>A	p.(D269N)	0	398	0.000	1	8591	0.01	1.000
	41585872	splice	.	c.882+1G>A	.	4	394	1.005	58	8460	0.68	0.358

Chapter 5: Conclusion for a comparative genomics approach to identify novel inherited cancer risk variants in dogs and humans

This dissertation highlights the usefulness of comparative oncology and the ability of dogs to serve as a model of naturally occurring human cancers,[71] due to the high degree of genetics similarity and disease presentations among dogs and humans.[55; 58] Dogs allow for distinctive opportunity in inherited disease studies. Due to the breeding for specific traits, specific dog breeds are extremely homogenous populations with a high degree of linkage disequilibrium (LD).[7] This breeding practice has resulted in an enrichment of disease influencing mutations, and the LD greatly aids in the identification of disease causing mutations. Also heavily emphasized is the power of whole genome sequencing (WGS) and its ability to identify mutations influencing both dog and human diseases,[49; 53; 66] including cancers.[80]

Canine mammary tumors (CMTs) are the dog comparable cancer type to human breast cancer.[80] These cancers both have a similar presentation and progression pattern; along with similar risk influences including hormone, age, and familial history.[74-79] Previous studies have found mutations in genes that influence both breast cancer and CMT risk.[80; 95; 121; 122; 124] An initial pedigree analysis of purebred dogs previously affected with CMT led to the identification of breed-specific common ancestors, highlighting that most dogs within a specific breed are descending from a small number within a closed breeding population.[1; 3] Analysis into breast cancer susceptibility genes was carried out within the samples to determine the similarities of risk mutations between this cohort and previous CMT and breast cancer cohorts. From this work, mutations within *BRCA2* and *STK11*, both clinically relevant breast cancer susceptibility risk genes, were associated with CMT risk among the cohort. The majority of *BRCA2* and *STK11* variants were most significantly associated with the Siberian Huskies investigated. The *BRCA2* variants identified as significant do correspond to human residues as variants of unknown significance, highlighting the need for further analysis in both human and dog cohorts on these mutations. This work highlighted a first investigation to determine what influence orthologs of human breast cancer susceptibility genes played on the CMT-affected dogs through WGS. It was a gene exclusion effort, before more exploratory analyses began. This dataset allows for novel discovery of risk genes and mutations benefitting both CMT and breast cancer research.

Most previous studies have focused on orthologs of breast cancer for CMT-risk;[80; 137] however, mutations within those genes did not explain the disease prevalence within this CMT-affected Golden Retriever cohort.[137] To further elucidate variants influencing risk within the Golden Retrievers of the CMT-affected cohort, a whole genome breed-specific analysis was carried out. The five WGS Golden Retriever samples were investigated for protein-truncating variants (PTVs) found in all five dogs. A single frameshifting variant found in the dog *CEACAM24* was validated and then investigated in the remaining 13 CMT-affected Golden Retrievers within the cohort. This analysis found a cohort frequency of approximately 67%, which was significant when compared to general European dog population controls (*p-value* 1.52×10^{-8}). However, in United States purebred dog populations, the mutation was found to be present in a similar frequency, 68%, in the control Golden Retriever population, and have approximately 22% allele frequency in other dog breeds, with the allele frequency ranging from 0% to 80% allele frequency in some breeds. Interestingly, the breeds with the highest frequency do tend to have higher rates of cancer affection.[88] Due to the possibility of this variant as a low penetrant CMT susceptibility mutation and the similarities between CMT and breast cancer, along with the higher degree of homology between the dog CEACAM24 protein and the human CEACAM proteins, human breast cancer cases were investigated to determine the influence of mutations within the *CEACAM* gene family on inherited breast cancer risk. While no inherited mutations within the *CEACAM* family of genes have been previously been associated with breast cancer or any cancer, alterations in protein expression and function have been associated with the progression and development of many different cancers of the years.[198-200] Rare (<1% minor allele frequency) protein truncating variants (PTVs), including nonsense, frameshifting, and slice-site variants, within the *CEACAM* genes in The Cancer Genome Atlas (TCGA) breast cancer cases were investigated and found an overall association between rare PTVs within the entire gene family and breast cancer risk. Previously, splice variants within *CEACAM* genes have been suggested to play a role in breast cancer tumorigenesis.[215; 216] Within the TCGA breast cancer cohort, *CEACAM6/7/&8* were all associated as individual genes with European American breast cancer risk, while only *CEACAM7* was associated with African American breast cancer risk. In previous analysis of breast cancer cells, *CEACAM6* and *CEACAM8* co-expression was determined to inhibit proliferation and invasiveness.[214] Furthermore, a loss of heterozygosity of *CEACAM1*, *CEACAM3*, *CEACAM5*, *CEACAM6*, *CEACAM7* and *CEACAM8* in breast cancer tumors was associated with metastasis. This could indicate the synergistic way the gene family regulates tumorigenesis.[217]

Interestingly, both colorectal cancer (CRC) and breast cancer share many risk factors, including an increased risk of both cancers in certain hereditary cancer syndromes (i.e., Lynch syndrome and *BRCA1/2* mutations).[231; 237; 238] Due to the previous association of rare PTVs in the CEACAM gene family with inherited breast cancer (Chapter 3), extensive history of the *CEACAM* gene family and colorectal cancer progression[187; 244; 259] and the clinical use of the CEACAM5 and CEACAM6 proteins as biomarkers for CRC,[260; 261] TCGA colorectal adenocarcinoma cases were investigated to determine the possible influence of mutations within the CEACAM gene family on inherited risk for colorectal cancer. From this, a limited association was found between individual mutations within the *CEACAM* gene family and inherited CRC risk. There were no rare PTVs identified as significant, which potentially further ties the gene family into a role of mostly increased expression influencing colorectal cancer risk, as is linked to somatic influences of the gene family.[190; 218; 232; 243; 262; 263] Additionally, there was no gene specific aggregation analysis that was found to be significant with rare missense mutations. While no large significance was determined between mutations in the *CEACAM* family of genes and CRC risk, 9 total different mutations were determined to individually be associated with risk. These mutations do not yet have clear indicators of what their impact could be on protein function or expression. However, minimal genetic changes are known to potentially have very large effects on the function of *CEACAM* genes.[223] Three of the significant mutations were identified in the Ig V-set domains of their individual proteins, CEACAM1, CEACAM3, and CEACAM4. This domain is known to be important in the dimerization of many CEACAM proteins,[245; 246] and this dimerization is often crucial to their downstream functions.[246; 249; 250] Two additional mutations within *CEACAM8* were individually significant and occur in between domain regions of the protein; however, they could affect the ability of those domains to properly function. CEACAM8 is known to heterodimerize with both CEACAM6 and CEACAM1,[246; 249] which both have previous CRC associations.[206; 207; 243] Individual mutations in *CEACAM18* (two mutations), *CEACAM19*, and *CEACAM20* were also associated, but occur after the functional domains of the protein with unknown significance on their impact.

The size of the TCGA CRC cohort is not very large, as compared to the TCGA breast cancer cohort, with only 48 African American samples and 199 European American CRC samples, which can contribute to limited identification of mutations. Additionally, the TCGA CRC cohort does not represent an exclusively inherited CRC cohort, and while some cases are likely inherited, due to lower age of onset (associated with inherited CRC syndromes), the lack of clear inherited cases does limit the ability to identify mutations influencing inherited CRC risk. This work highlights the limitations and benefits of public repositories. While the *CEACAM* gene family has been known to have and influence in CRC development and progression, no previous inherited links had been investigated within this cohort and previous inheritance links of the gene family have been limited to a breast cancer susceptibility analysis (Chapter 3). Overall, this dissertation highlights the usefulness and far reaching effects of WGS dogs for comparative oncology research. WGS efforts provided a survey of the entire genome for genetic study and this was helpful in identifying mutations not only in dog studies, but also human.[5] By WGS a select group of dogs with CMT, a gene exclusion analysis could be carried out in breast cancer susceptibility genes and following that a breed specific genome analysis was successful to identify mutations possibly influencing both CMT risk and inherited breast cancer risk. This finding resulted in the *CEACAM* gene family that has been known to influence a multitude of cancer to be investigated in a new light, as this was the first inherited analysis of the gene family. Furthermore, the significance of the entire *CEACAM* gene family in a breast cancer cohort lead to an investigation into a CRC cohort for potential risk influences within that cohort.

Section 6: Complete Reference List

1. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., 3rd, Zody, M.C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803-819.
2. Ostrander, E.A., and Wayne, R.K. (2005). The canine genome. Genome Res 15, 1706-1716.
3. Lewis, T.W., Abhayaratne, B.M., and Blott, S.C. (2015). Trends in genetic diversity for all Kennel Club registered pedigree dog breeds. Canine Genet Epidemiol 2, 13.
4. Wayne, R.K. (1986). Cranial Morphology of Domestic and Wild Canids: The Influence of Development on Morphological Change. Evolution 40, 243-261.
5. Shaffer, L.G. (2019). Special issue on canine genetics: animal models for human disease and gene therapies, new discoveries for canine inherited diseases, and standards and guidelines for clinical genetic testing for domestic dogs. Hum Genet 138, 437-440.
6. Chen, W.K., Swartz, J.D., Rush, L.J., and Alvarez, C.E. (2009). Mapping DNA structural variation in dogs. Genome Res 19, 500-509.
7. Goldstein, O., Zangerl, B., Pearce-Kelling, S., Sidjanin, D.J., Kijas, J.W., Felix, J., Acland, G.M., and Aguirre, G.D. (2006). Linkage disequilibrium mapping in domestic dog breeds narrows the progressive rod-cone degeneration interval and identifies ancestral disease-transmitting chromosome. Genomics 88, 541-550.
8. Goode, E.L. (2011). Linkage Disequilibrium. In Encyclopedia of Cancer, M. Schwab, ed. (Berlin, Heidelberg, Springer Berlin Heidelberg), pp 2043-2048.
9. Sutter, N.B., Eberle, M.A., Parker, H.G., Pullar, B.J., Kirkness, E.F., Kruglyak, L., and Ostrander, E.A. (2004). Extensive and breed-specific linkage disequilibrium in Canis familiaris. Genome Res 14, 2388-2396.
10. Norrgard, K. (2008). Genetic variation and disease:GWAS. Nature Education 1.
11. Chang, M., He, L., and Cai, L. (2018). An Overview of Genome-Wide Association Studies. Methods Mol Biol 1754, 97-108.
12. Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. PLoS Comput Biol 8, e1002822.
13. Yuzbasiyan-Gurkan, V., Blanton, S.H., Cao, Y., Ferguson, P., Li, J., Venta, P.J., and Brewer, G.J. (1997). Linkage of a microsatellite marker to the canine copper toxicosis locus in Bedlington terriers. Am J Vet Res 58, 23-27.
14. Kolicheski, A. (2017). Discovering disease causing variants in dogs through whole genome sequencing. In. (
15. Wood, S.H., Clements, D.N., Ollier, W.E., Nuttall, T., McEwan, N.A., and Carter, S.D. (2009). Gene expression in canine atopic dermatitis and correlation with clinical severity scores. J Dermatol Sci 55, 27-33.
16. Mellersh, C.S., Langston, A.A., Acland, G.M., Fleming, M.A., Ray, K., Wiegand, N.A., Francisco, L.V., Gibbs, M., Aguirre, G.D., and Ostrander, E.A. (1997). A linkage map of the canine genome. Genomics 46, 326-336.
17. Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M., et al. (2003). The dog genome: survey sequencing and comparative analysis. Science 301, 1898-1903.
18. Hoeppner, M.P., Lundquist, A., Pirun, M., Meadows, J.R., Zamani, N., Johnson, J., Sundstrom, G., Cook, A., FitzGerald, M.G., Swofford, R., et al. (2014). An improved

canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One 9, e91172.

19. Behjati, S., and Tarpey, P.S. (2013). What is next generation sequencing? Arch Dis Child Educ Pract Ed 98, 236-238.
20. Kanzi, A.M., San, J.E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., and de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Front Genet 11, 544162.
21. Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13, 36-46.
22. Bishop, M.R., Huskey, A.L.W., Hetzel, J., and Merner, N.D. (2019). A research-based gene panel to investigate breast, ovarian and prostate cancer genetic risk. PLoS One 14, e0220929.
23. Gulilat, M., Lamb, T., Teft, W.A., Wang, J., Dron, J.S., Robinson, J.F., Tirona, R.G., Hegele, R.A., Kim, R.B., and Schwarz, U.I. (2019). Targeted next generation sequencing as a tool for precision medicine. BMC Med Genomics 12, 81.
24. McDonald, J.T., Kritharis, A., Beheshti, A., Pilichowska, M., Burgess, K., Ricks-Santi, L., McNiel, E., London, C.A., Ravi, D., and Evens, A.M. (2018). Comparative oncology DNA sequencing of canine T cell lymphoma via human hotspot panel. Oncotarget 9, 22693-22702.
25. Petrackova, A., Vasinek, M., Sedlarikova, L., Dyskova, T., Schneiderova, P., Novosad, T., Papajik, T., and Kriegova, E. (2019). Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Mutations in Cancer Diagnostics. Front Oncol 9, 851.
26. Wang, X., Shen, X., Fang, F., Ding, C.H., Zhang, H., Cao, Z.H., and An, D.Y. (2018). Phenotype-Driven Virtual Panel Is an Effective Method to Analyze WES Data of Neurological Disease. Front Pharmacol 9, 1529.
27. Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., and Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. G3 (Bethesda) 5, 1543-1550.
28. Sun, Y., Ruivenkamp, C.A., Hoffer, M.J., Vrijenhoek, T., Kriek, M., van Asperen, C.J., den Dunnen, J.T., and Santen, G.W. (2015). Next-generation diagnostics: gene panel, exome, or whole genome? Hum Mutat 36, 648-655.
29. Broeckx, B.J., Coopman, F., Verhoeven, G.E., Bavegems, V., De Keulenaer, S., De Meester, E., Van Niewerburgh, F., and Deforce, D. (2014). Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1). Sci Rep 4, 5597.
30. Broeckx, B.J., Hitte, C., Coopman, F., Verhoeven, G.E., De Keulenaer, S., De Meester, E., Derrien, T., Alfoldi, J., Lindblad-Toh, K., Bosmans, T., et al. (2015). Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes. Sci Rep 5, 12810.
31. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11, 415-425.
32. NG, P.C., and Kirness, E.F. (2010). Whole Genome Sequencing. In Barnes, M.; Breen, G. (eds) Genetic Variation. Methods in Molecular Biology (Methods and Protocols) 628.
33. Meienberg, J., Bruggmann, R., Oexle, K., and Matyas, G. (2016). Clinical sequencing: is WGS the better WES? Human Genetics 135, 359-362.

34. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.L., and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A 112, 5473-5478.
35. Sayyab, S., Viluma, A., Bergvall, K., Brunberg, E., Jagannathan, V., Leeb, T., Andersson, G., and Bergstrom, T.F. (2016). Whole-Genome Sequencing of a Canine Family Trio Reveals a FAM83G Variant Associated with Hereditary Footpad Hyperkeratosis. G3 (Bethesda) 6, 521-527.
36. Bianchi, M., Rafati, N., Karlsson, A., Muren, E., Rubin, C.J., Sundberg, K., Andersson, G., Kampe, O., Hedhammar, A., Lindblad-Toh, K., et al. (2020). Whole-genome genotyping and resequencing reveal the association of a deletion in the complex interferon alpha gene cluster with hypothyroidism in dogs. BMC Genomics 21, 307.
37. Mansour, T.A., Woolard, K.D., Vernau, K.L., Ancona, D.M., Thomasy, S.M., Sebbag, L., Moore, B.A., Knipe, M.F., Seada, H.A., Cowan, T.M., et al. (2020). Whole genome sequencing for mutation discovery in a single case of lysosomal storage disease (MPS type 1) in the dog. Sci Rep 10, 6558.
38. Viluma, A., Sayyab, S., Mikko, S., Andersson, G., and Bergstrom, T.F. (2015). Evaluation of whole-genome sequencing of four Chinese crested dogs for variant detection using the ion proton system. Canine Genet Epidemiol 2, 16.
39. Gerber, M., Fischer, A., Jagannathan, V., Drogemuller, M., Drogemuller, C., Schmidt, M.J., Bernardino, F., Manz, E., Matiasek, K., Rentmeister, K., et al. (2015). A deletion in the VLDLR gene in Eurasier dogs with cerebellar hypoplasia resembling a Dandy-Walker-like malformation (DWLM). PLoS One 10, e0108917.
40. Kyostila, K., Syrja, P., Jagannathan, V., Chandrasekar, G., Jokinen, T.S., Seppala, E.H., Becker, D., Drogemuller, M., Dietschi, E., Drogemuller, C., et al. (2015). A missense change in the ATG4D gene links aberrant autophagy to a neurodegenerative vacuolar storage disease. PLoS Genet 11, e1005169.
41. Owczarek-Lipska, M., Jagannathan, V., Drogemuller, C., Lutz, S., Glanemann, B., Leeb, T., and Kook, P.H. (2013). A frameshift mutation in the cubilin gene (CUBN) in Border Collies with Imerslund-Grasbeck syndrome (selective cobalamin malabsorption). PLoS One 8, e61144.
42. Hytonen, M.K., and Lohi, H. (2019). A frameshift insertion in SGK3 leads to recessive hairlessness in Scottish Deerhounds: a candidate gene for human alopecia conditions. Hum Genet 138, 535-539.
43. Bauer, A., Waluk, D.P., Galichet, A., Timm, K., Jagannathan, V., Sayar, B.S., Wiener, D.J., Dietschi, E., Muller, E.J., Roosje, P., et al. (2017). A de novo variant in the ASPRV1 gene in a dog with ichthyosis. PLoS Genet 13, e1006651.
44. Kessler, M.D., Loesch, D.P., Perry, J.A., Heard-Costa, N.L., Taliun, D., Cade, B.E., Wang, H., Daya, M., Ziniti, J., Datta, S., et al. (2020). De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. Proc Natl Acad Sci U S A 117, 2560-2569.
45. Acuna-Hidalgo, R., Veltman, J.A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome Biol 17, 241.
46. Makelainen, S., Godia, M., Hellsand, M., Viluma, A., Hahn, D., Makdoumi, K., Zeiss, C.J., Mellersh, C., Ricketts, S.L., Narfstrom, K., et al. (2019). An ABCA4 loss-of-function mutation causes a canine form of Stargardt disease. PLoS Genet 15, e1007873.

47. Murgiano, L., Becker, D., Spector, C., Carlin, K., Santana, E., Niggel, J.K., Jagannathan, V., Leeb, T., Pearce-Kelling, S., Aguirre, G.D., et al. (2020). CCDC66 frameshift variant associated with a new form of early-onset progressive retinal atrophy in Portuguese Water Dogs. Sci Rep 10, 21162.
48. Drogemuller, M., Jagannathan, V., Howard, J., Bruggmann, R., Drogemuller, C., Ruetten, M., Leeb, T., and Kook, P.H. (2014). A frameshift mutation in the cubilin gene (CUBN) in Beagles with Imerslund-Grasbeck syndrome (selective cobalamin malabsorption). Anim Genet 45, 148-150.
49. Gilliam, D., O'Brien, D.P., Coates, J.R., Johnson, G.S., Johnson, G.C., Mhlanga-Mutangadura, T., Hansen, L., Taylor, J.F., and Schnabel, R.D. (2014). A homozygous KCNJ10 mutation in Jack Russell Terriers and related breeds with spinocerebellar ataxia with myokymia, seizures, or both. J Vet Intern Med 28, 871-877.
50. Meurs, K.M., Friedenberg, S.G., Kolb, J., Saripalli, C., Tonino, P., Woodruff, K., Olby, N.J., Keene, B.W., Adin, D.B., Yost, O.L., et al. (2019). A missense variant in the titin gene in Doberman pinscher dogs with familial dilated cardiomyopathy and sudden cardiac death. Hum Genet 138, 515-524.
51. Alfoldi, J., and Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. Genome Res 23, 1063-1068.
52. Kolicheski, A., Barnes Heller, H.L., Arnold, S., Schnabel, R.D., Taylor, J.F., Knox, C.A., Mhlanga-Mutangadura, T., O'Brien, D.P., Johnson, G.S., Dreyfus, J., et al. (2017). Homozygous PPT1 Splice Donor Mutation in a Cane Corso Dog With Neuronal Ceroid Lipofuscinosis. J Vet Intern Med 31, 149-157.
53. Kolicheski, A.L., Johnson, G.S., Mhlanga-Mutangadura, T., Taylor, J.F., Schnabel, R.D., Kinoshita, T., Murakami, Y., and O'Brien, D.P. (2017). A homozygous PIGN missense mutation in Soft-Coated Wheaten Terriers with a canine paroxysmal dyskinesia. Neurogenetics 18, 39-47.
54. Kolicheski, A., Johnson, G.S., O'Brien, D.P., Mhlanga-Mutangadura, T., Gilliam, D., Guo, J., Anderson-Sieg, T.D., Schnabel, R.D., Taylor, J.F., Lebowitz, A., et al. (2016). Australian Cattle Dogs with Neuronal Ceroid Lipofuscinosis are Homozygous for a CLN5 Nonsense Mutation Previously Identified in Border Collies. J Vet Intern Med 30, 1149-1158.
55. Switonski, M. (2014). Dog as a model in studies on human hereditary diseases and their gene therapy. Reprod Biol 14, 44-50.
56. Moreno, C., Lazar, J., Jacob, H.J., and Kwitek, A.E. (2008). Comparative genomics for detecting human disease genes. Adv Genet 60, 655-697.
57. Meadows, J.R.S., and Lindblad-Toh, K. (2017). Dissecting evolution and disease using comparative vertebrate genomics. Nat Rev Genet 18, 624-636.
58. Lindblad-Toh, K. (2020). What animals can teach us about evolution, the human genome, and human disease. Ups J Med Sci 125, 1-9.
59. Baird, A.E., Carter, S.D., Innes, J.F., Ollier, W.E., and Short, A.D. (2014). Genetic basis of cranial cruciate ligament rupture (CCLR) in dogs. Connect Tissue Res 55, 275-281.
60. Tsai, K.L., Clark, L.A., and Murphy, K.E. (2007). Understanding hereditary diseases using the dog and human as companion model systems. Mamm Genome 18, 444-451.
61. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.

62. Murphy, S.C., Recio, A., de la Fuente, C., Guo, L.T., Shelton, G.D., and Clark, L.A. (2019). A glycine transporter SLC6A5 frameshift mutation causes startle disease in Spanish greyhounds. Hum Genet 138, 509-513.
63. Mhlanga-Mutangadura, T., Johnson, G.S., Schnabel, R.D., Taylor, J.F., Johnson, G.C., Katz, M.L., Shelton, G.D., Lever, T.E., Giuliano, E., Granger, N., et al. (2016). A mutation in the Warburg syndrome gene, RAB3GAP1, causes a similar syndrome with polyneuropathy and neuronal vacuolation in Black Russian Terrier dogs. Neurobiol Dis 86, 75-85.
64. Guo, J., O'Brien, D.P., Mhlanga-Mutangadura, T., Olby, N.J., Taylor, J.F., Schnabel, R.D., Katz, M.L., and Johnson, G.S. (2015). A rare homozygous MFSD8 single-base-pair deletion and frameshift in the whole genome sequence of a Chinese Crested dog with neuronal ceroid lipofuscinosis. BMC Vet Res 10, 960.
65. Gilliam, D., Kolicheski, A., Johnson, G.S., Mhlanga-Mutangadura, T., Taylor, J.F., Schnabel, R.D., and Katz, M.L. (2015). Golden Retriever dogs with neuronal ceroid lipofuscinosis have a two-base-pair deletion and frameshift in CLN5. Mol Genet Metab 115, 101-109.
66. Guo, J., Johnson, G.S., Brown, H.A., Provencher, M.L., da Costa, R.C., Mhlanga-Mutangadura, T., Taylor, J.F., Schnabel, R.D., O'Brien, D.P., and Katz, M.L. (2014). A CLN8 nonsense mutation in the whole genome sequence of a mixed breed dog with neuronal ceroid lipofuscinosis and Australian Shepherd ancestry. Mol Genet Metab 112, 302-309.
67. Sharp, N.J., Kornegay, J.N., Van Camp, S.D., Herbstreith, M.H., Secore, S.L., Kettle, S., Hung, W.Y., Constantinou, C.D., Dykstra, M.J., Roses, A.D., et al. (1992). An error in dystrophin mRNA processing in golden retriever muscular dystrophy, an animal homologue of Duchenne muscular dystrophy. Genomics 13, 115-121.
68. Lin, L., Faraco, J., Li, R., Kadotani, H., Rogers, W., Lin, X., Qiu, X., de Jong, P.J., Nishino, S., and Mignot, E. (1999). The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. Cell 98, 365-376.
69. Wolf, Z.T., Brand, H.A., Shaffer, J.R., Leslie, E.J., Arzi, B., Willet, C.E., Cox, T.C., McHenry, T., Narayan, N., Feingold, E., et al. (2015). Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate. PLoS Genet 11, e1005059.
70. Rowell, J.L., McCarthy, D.O., and Alvarez, C.E. (2011). Dog models of naturally occurring cancer. Trends Mol Med 17, 380-388.
71. Garden, O.A., Volk, S.W., Mason, N.J., and Perry, J.A. (2018). Companion animals in comparative oncology: One Medicine in action. Vet J 240, 6-13.
72. Lingaas, F., Comstock, K.E., Kirkness, E.F., Sorensen, A., Aarskaug, T., Hitte, C., Nickerson, M.L., Moe, L., Schmidt, L.S., Thomas, R., et al. (2003). A mutation in the canine BHD gene is associated with hereditary multifocal renal cystadenocarcinoma and nodular dermatofibrosis in the German Shepherd dog. Hum Mol Genet 12, 3043-3053.
73. Jiang, W., Fujii, H., Matsumoto, T., Ohtsuji, N., Tsurumaru, M., and Hino, O. (2007). Birt-Hogg-Dube (BHD) gene mutations in human gastric cancer with high frequency microsatellite instability. Cancer Lett 248, 103-111.
74. Liu, D., Xiong, H., Ellis, A.E., Northrup, N.C., Rodriguez, C.O., Jr., O'Regan, R.M., Dalton, S., and Zhao, S. (2014). Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. Cancer Res 74, 5045-5056.

75. Ettlin, J., Clementi, E., Amini, P., Malbon, A., and Markkanen, E. (2017). Analysis of Gene Expression Signatures in Cancer-Associated Stroma from Canine Mammary Tumours Reveals Molecular Homology to Human Breast Carcinomas. Int J Mol Sci 18.
76. Lee, K.H., Park, H.M., Son, K.H., Shin, T.J., and Cho, J.Y. (2018). Transcriptome Signatures of Canine Mammary Gland Tumors and Its Comparison to Human Breast Cancers. Cancers (Basel) 10.
77. Lee, K.H., Hwang, H.J., Noh, H.J., Shin, T.J., and Cho, J.Y. (2019). Somatic Mutation of PIK3CA (H1047R) Is a Common Driver Mutation Hotspot in Canine Mammary Tumors as Well as Human Breast Cancers. Cancers (Basel) 11.
78. Kim, K.K., Seung, B.J., Kim, D., Park, H.M., Lee, S., Song, D.W., Lee, G., Cheong, J.H., Nam, H., Sur, J.H., et al. (2019). Whole-exome and whole-transcriptome sequencing of canine mammary gland tumors. Sci Data 6, 147.
79. Gray, M., Meehan, J., Martinez-Perez, C., Kay, C., Turnbull, A.K., Morrison, L.R., Pang, L.Y., and Argyle, D. (2020). Naturally-Occurring Canine Mammary Tumors as a Translational Model for Human Breast Cancer. Front Oncol 10, 617.
80. Goebel, K., and Merner, N.D. (2017). A monograph proposing the use of canine mammary tumours as a model for the study of hereditary breast cancer susceptibility genes in humans. Vet Med Sci 3, 51-62.
81. American Kennel Club. Breeds by Year Recognized. In. (
82. Parker, H.G., Shearin, A.L., and Ostrander, E.A. (2010). Man's best friend becomes biology's best in show: genome analyses in the domestic dog. Annu Rev Genet 44, 309-336.
83. Ostrander, E.A. (2012). Franklin H. Epstein Lecture. Both ends of the leash--the human links to good dogs with bad genes. N Engl J Med 367, 636-646.
84. Evans, J.P., Brinkhous, K.M., Brayer, G.D., Reisner, H.M., and High, K.A. (1989). Canine hemophilia B resulting from a point mutation with unusual consequences. Proc Natl Acad Sci U S A 86, 10095-10099.
85. Grall, A., Guaguere, E., Planchais, S., Grond, S., Bourrat, E., Hausser, I., Hitte, C., Le Gallo, M., Derbois, C., Kim, G.J., et al. (2012). PNPLA1 mutations cause autosomal recessive congenital ichthyosis in golden retriever dogs and humans. Nat Genet 44, 140-147.
86. van De Sluis, B., Rothuizen, J., Pearson, P.L., van Oost, B.A., and Wijmenga, C. (2002). Identification of a new copper metabolism gene by positional cloning in a purebred dog population. Hum Mol Genet 11, 165-173.
87. Jonasdottir, T.J., Mellersh, C.S., Moe, L., Heggebo, R., Gamlem, H., Ostrander, E.A., and Lingaas, F. (2000). Genetic mapping of a naturally occurring hereditary renal cancer syndrome in dogs. Proc Natl Acad Sci U S A 97, 4132-4137.
88. Dobson, J.M. (2013). Breed-predispositions to cancer in pedigree dogs. ISRN Vet Sci 2013, 941275.
89. C**handler, M.R.**, Bilgili, E.P., and Merner, N.D. (2016). A Review of Whole-Exome Sequencing Efforts Toward Hereditary Breast Cancer Susceptibility Gene Discovery. Hum Mutat 37, 835-846.
90. Schmidt, L.S., Warren, M.B., Nickerson, M.L., Weirich, G., Matrosova, V., Toro, J.R., Turner, M.L., Duray, P., Merino, M., Hewitt, S., et al. (2001). Birt-Hogg-Dube syndrome, a genodermatosis associated with spontaneous pneumothorax and kidney neoplasia, maps to chromosome 17p11.2. Am J Hum Genet 69, 876-882.
91. Nickerson, M.L., Warren, M.B., Toro, J.R., Matrosova, V., Glenn, G., Turner, M.L., Duray, P., Merino, M., Choyke, P., Pavlovich, C.P., et al. (2002). Mutations in a novel gene lead

to kidney tumors, lung wall defects, and benign tumors of the hair follicle in patients with the Birt-Hogg-Dube syndrome. Cancer Cell 2, 157-164.

92. Egenvall, A., Bonnett, B.N., Ohagen, P., Olson, P., Hedhammar, A., and von Euler, H. (2005). Incidence of and survival after mammary tumors in a population of over 80,000 insured female dogs in Sweden from 1995 to 2002. Prev Vet Med 69, 109-127.
93. Jitpean, S., Hagman, R., Strom Holst, B., Hoglund, O.V., Pettersson, A., and Egenvall, A. (2012). Breed variations in the incidence of pyometra and mammary tumours in Swedish dogs. Reprod Domest Anim 47 Suppl 6, 347-350.
94. Rivera, P., Melin, M., Biagi, T., Fall, T., Haggstrom, J., Lindblad-Toh, K., and von Euler, H. (2009). Mammary tumor development in dogs is associated with BRCA1 and BRCA2. Cancer Res 69, 8770-8774.
95. Borge, K.S., Melin, M., Rivera, P., Thoresen, S.I., Webster, M.T., von Euler, H., Lindblad-Toh, K., and Lingaas, F. (2013). The ESR1 gene is associated with risk for canine mammary tumours. BMC Vet Res 9, 69.
96. Melin, M., Rivera, P., Arendt, M., Elvers, I., Muren, E., Gustafson, U., Starkey, M., Borge, K.S., Lingaas, F., Haggstrom, J., et al. (2016). Genome-Wide Analysis Identifies Germ-Line Risk Factors Associated with Canine Mammary Tumours. PLoS Genet 12, e1006029.
97. Fyfe, J.C., Hemker, S.L., Frampton, A., Raj, K., Nagy, P.L., Gibbon, K.J., and Giger, U. (2018). Inherited selective cobalamin malabsorption in Komondor dogs associated with a CUBN splice site variant. BMC Vet Res 14, 418.
98. Plassais, J., Rimbault, M., Williams, F.J., Davis, B.W., Schoenebeck, J.J., and Ostrander, E.A. (2017). Analysis of large versus small dogs reveals three genes on the canine X chromosome associated with body weight, muscling and back fat thickness. PLoS Genet 13, e1006661.
99. Liu, Y.H., Wang, L., Xu, T., Guo, X., Li, Y., Yin, T.T., Yang, H.C., Hu, Y., Adeola, A.C., Sanke, O.J., et al. (2018). Whole-Genome Sequencing of African Dogs Provides Insights into Adaptations against Tropical Parasites. Mol Biol Evol 35, 287-298.
100. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43, 11 10 11-11 10 33.
101. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.
102. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
103. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
104. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.
105. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44, D862-868.

106. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res 46, D754-D761.
107. Lemke, J.R., Lal, D., Reinthaler, E.M., Steiner, I., Nothnagel, M., Alber, M., Geider, K., Laube, B., Schwake, M., Finsterwalder, K., et al. (2013). Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes. Nat Genet 45, 1067-1072.
108. Exome Variant Server. (2019). NHLBI GO Exome Sequencing Project (ESP). In. (Seattle, WA.
109. Guo, M.H., Plummer, L., Chan, Y.M., Hirschhorn, J.N., and Lippincott, M.F. (2018). Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. Am J Hum Genet 103, 522-534.
110. Smith, B.N., Ticozzi, N., Fallini, C., Gkazi, A.S., Topp, S., Kenna, K.P., Scotter, E.L., Kost, J., Keagle, P., Miller, J.W., et al. (2014). Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. Neuron 84, 324-331.
111. Olesen, M.S., Andreasen, L., Jabbari, J., Refsgaard, L., Haunso, S., Olesen, S.P., Nielsen, J.B., Schmitt, N., and Svendsen, J.H. (2014). Very early-onset lone atrial fibrillation patients have a high prevalence of rare variants in genes previously associated with atrial fibrillation. Heart Rhythm 11, 246-251.
112. Ryman, N., and Jorde, P.E. (2001). Statistical power when testing for genetic differentiation. Mol Ecol 10, 2361-2373.
113. Kim, H.Y. (2016). Statistical notes for clinical researchers: Sample size calculation 2. Comparison of two independent proportions. Restor Dent Endod 41, 154-156.
114. Pritzlaff, M., Summerour, P., McFarland, R., Li, S., Reineke, P., Dolinsky, J.S., Goldgar, D.E., Shimelis, H., Couch, F.J., Chao, E.C., et al. (2017). Male breast cancer in a multi-gene panel testing cohort: insights and unexpected results. Breast Cancer Res Treat 161, 575-586.
115. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12, 745-755.
116. Frischknecht, M., Niehof-Oellers, H., Jagannathan, V., Owczarek-Lipska, M., Drogemuller, C., Dietschi, E., Dolf, G., Tellhelm, B., Lang, J., Tiira, K., et al. (2013). A COL11A2 mutation in Labrador retrievers with mild disproportionate dwarfism. PLoS One 8, e60149.
117. Jagannathan, V., Bannoehr, J., Plattet, P., Hauswirth, R., Drogemuller, C., Drogemuller, M., Wiener, D.J., Doherr, M., Owczarek-Lipska, M., Galichet, A., et al. (2013). A mutation in the SUV39H2 gene in Labrador Retrievers with hereditary nasal parakeratosis (HNPK) provides insights into the epigenetics of keratinocyte differentiation. PLoS Genet 9, e1003848.
118. Ahonen, S.J., Arumilli, M., and Lohi, H. (2013). A CNGB1 frameshift mutation in Papillon and Phalene dogs with progressive retinal atrophy. PLoS One 8, e72122.
119. van Steenbeek, F.G., Hytonen, M.K., Leegwater, P.A., and Lohi, H. (2016). The canine era: the rise of a biomedical model. Anim Genet 47, 519-527.
120. Forman, O.P., Hitti, R.J., Boursnell, M., Miyadera, K., Sargan, D., and Mellersh, C. (2016). Canine genome assembly correction facilitates identification of a MAP9 deletion as a potential age of onset modifier for RPGRIP1-associated canine retinal degeneration. Mamm Genome 27, 237-245.

121. Yoshikawa, Y., Morimatsu, M., Ochiai, K., Nagano, M., Yamane, Y., Tomizawa, N., Sasaki, N., and Hashizume, K. (2005). Insertion/deletion polymorphism in the BRCA2 nuclear localization signal. Biomed Res 26, 109-116.
122. Yoshikawa, Y., Morimatsu, M., Ochiai, K., Nagano, M., Tomioka, Y., Sasaki, N., Hashizume, K., and Iwanaga, T. (2008). Novel variations and loss of heterozygosity of BRCA2 identified in a dog with mammary tumors. Am J Vet Res 69, 1323-1328.
123. Enginler, S.O., Ates, A., Diren Sigirci, B., Sontas, B.H., Sonmez, K., Karacam, E., Ekici, H., Evkuran Dal, G., and Gurel, A. (2014). Measurement of C-reactive protein and prostaglandin F2alpha metabolite concentrations in differentiation of canine pyometra and cystic endometrial hyperplasia/mucometra. Reprod Domest Anim 49, 641-647.
124. Borge, K.S., Borresen-Dale, A.L., and Lingaas, F. (2011). Identification of genetic variation in 11 candidate genes of canine mammary tumour. Vet Comp Oncol 9, 241-250.
125. Murray, M.L., Cerrato, F., Bennett, R.L., and Jarvik, G.P. (2011). Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions. Genet Med 13, 998-1005.
126. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods 7, 248-249.
127. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet Chapter 7, Unit7 20.
128. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32, 358-368.
129. Wu, K., Hinson, S.R., Ohashi, A., Farrugia, D., Wendt, P., Tavtigian, S.V., Deffenbaugh, A., Goldgar, D., and Couch, F.J. (2005). Functional evaluation and cancer risk assessment of BRCA2 unclassified variants. Cancer Res 65, 417-426.
130. So, M.K., Jeong, T.D., Lim, W., Moon, B.I., Paik, N.S., Kim, S.C., and Huh, J. (2019). Reinterpretation of BRCA1 and BRCA2 variants of uncertain significance in patients with hereditary breast/ovarian cancer using the ACMG/AMP 2015 guidelines. Breast Cancer.
131. Canadas, A., Santos, M., Nogueira, A., Assis, J., Gomes, M., Lemos, C., Medeiros, R., and Dias-Pereira, P. (2018). Canine mammary tumor risk is associated with polymorphisms in RAD51 and STK11 genes. J Vet Diagn Invest 30, 733-738.
132. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.H., Wang, Q., Bolla, M.K., et al. (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 104, 21-34.
133. Godet, I., and Gilkes, D.M. (2017). BRCA1 and BRCA2 mutations and treatment strategies for breast cancer. Integr Cancer Sci Ther 4.
134. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68, 394-424.
135. American Cancer Society. (2020). Cancer Facts & Figures 2020. In. (
136. Cardoso, F., Harbeck, N., Barrios, C.H., Bergh, J., Cortes, J., El Saghir, N., Francis, P.A., Hudis, C.A., Ohno, S., Partridge, A.H., et al. (2017). Research needs in breast cancer. Ann Oncol 28, 208-217.

137. Huskey, A.L.W., Goebel, K., Lloveras-Fuentes, C., McNeely, I., and Merner, N.D. (2020). Whole genome sequencing for the investigation of canine mammary tumor inheritance - an initial assessment of high-risk breast cancer genes reveal BRCA2 and STK11 variants potentially associated with risk in purebred dogs. Canine Medicine and Genetics 7, 8.
138. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47, W636-W641.
139. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the integrative protein signature database. Nucleic Acids Res 37, D211-215.
140. Kumar, M., Gouw, M., Michael, S., Samano-Sanchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J.A., Bukirova, D., Calyseva, J., et al. (2020). ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res 48, D296-D306.
141. Sprent, P. (2011). Fisher Exact Test. In International Encyclopedia of Statistical Science, M. Lovric, ed. (Berlin, Heidelberg, Springer Berlin Heidelberg), pp 524-525.
142. Fisher's Exact Test for Single Variant Analysis. In. (
143. Fisher, R.A. (1925). Statistical methods for research workers.(Oliver and Boyd: Edinburgh).
144. Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Song, F. (2000). Methods for meta-analysis in medical research.(Wiley: Chichester).
145. Wildeman, M., van Ophuizen, E., den Dunnen, J.T., and Taschner, P.E. (2008). Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29, 6-13.
146. Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res 48, D127-D131.
147. Chen, F., Zhou, H., Wu, C., and Yan, H. (2018). Identification of miRNA profiling in prediction of tumor recurrence and progress and bioinformatics analysis for patients with primary esophageal cancer: Study based on TCGA database. Pathol Res Pract 214, 2081-2086.
148. Zhang, S., Chen, H., Liu, W., Fang, L., Qian, Z., Kong, R., Zhang, Q., Li, J., and Cao, X. (2020). miR-766-3p Targeting BCL9L Suppressed Tumorigenesis, Epithelial-Mesenchymal Transition, and Metastasis Through the beta-Catenin Signaling Pathway in Osteosarcoma Cells. Front Cell Dev Biol 8, 594135.
149. Chen, C., Xue, S., Zhang, J., Chen, W., Gong, D., Zheng, J., Ma, J., Xue, W., Chen, Y., Zhai, W., et al. (2017). DNA-methylation-mediated repression of miR-766-3p promotes cell proliferation via targeting SF2 expression in renal cell carcinoma. Int J Cancer 141, 1867-1878.
150. You, Y., Que, K., Zhou, Y., Zhang, Z., Zhao, X., Gong, J., and Liu, Z. (2018). MicroRNA-766-3p Inhibits Tumour Progression by Targeting Wnt3a in Hepatocellular Carcinoma. Mol Cells 41, 830-841.
151. Liu, S., Lin, Z., Zheng, Z., Rao, W., Lin, Y., Chen, H., Xie, Q., Chen, Y., and Hu, Z. (2020). Serum exosomal microRNA-766-3p expression is associated with poor prognosis of esophageal squamous cell carcinoma. Cancer Sci 111, 3881-3892.
152. Alshamrani, A.A. (2020). Roles of microRNAs in Ovarian Cancer Tumorigenesis: Two Decades Later, What Have We Learned? Front Oncol 10, 1084.
153. Tuncer, S.B., Erdogan, O.S., Erciyas, S.K., Saral, M.A., Celik, B., Odemis, D.A., Turkcan, G.K., and Yazici, H. (2020). miRNA expression profile changes in the peripheral blood

of monozygotic discordant twins for epithelial ovarian carcinoma: potential new biomarkers for early diagnosis and prognosis of ovarian carcinoma. J Ovarian Res 13, 99.

154. Wang, Q., Selth, L.A., and Callen, D.F. (2017). MiR-766 induces p53 accumulation and G2/M arrest by directly targeting MDM4. Oncotarget 8, 29914-29924.

155. Ge, S., Sun, C., Hu, Q., Guo, Y., Xia, G., Mi, Y., and Zhu, L. (2020). Differential expression profiles of circRNAs in human prostate cancer based on chip and bioinformatic analysis. Int J Clin Exp Pathol 13, 1045-1052.

156. Ren, H., Liu, Z., Liu, S., Zhou, X., Wang, H., Xu, J., Wang, D., and Yuan, G. (2018). Profile and clinical implication of circular RNAs in human papillary thyroid carcinoma. PeerJ 6, e5363.

157. Yasui, T., Yanagida, T., Ito, S., Konakade, Y., Takeshita, D., Naganawa, T., Nagashima, K., Shimada, T., Kaji, N., Nakamura, Y., et al. (2017). Unveiling massive numbers of cancer-related urinary-microRNA candidates via nanowires. Sci Adv 3, e1701133.

158. Zhang, C., Xue, Q., Xu, Z., and Lu, C. (2018). MiR-5702 suppresses proliferation and invasion in non-small-cell lung cancer cells via posttranscriptional suppression of ZEB1. J Biochem Mol Toxicol, e22163.

159. Mou, E., and Wang, H. (2019). LncRNA LUCAT1 facilitates tumorigenesis and metastasis of triple-negative breast cancer through modulating miR-5702. Biosci Rep 39.

160. Liu, F., Cai, Y., Rong, X., Chen, J., Zheng, D., Chen, L., Zhang, J., Luo, R., Zhao, P., and Ruan, J. (2017). MiR-661 promotes tumor invasion and metastasis by directly inhibiting RB1 in non small cell lung cancer. Mol Cancer 16, 122.

161. Hoffman, Y., Bublik, D.R., Pilpel, Y., and Oren, M. (2014). miR-661 downregulates both Mdm2 and Mdm4 to activate p53. Cell Death Differ 21, 302-309.

162. Wang, S., Li, Q., Wang, Y., Li, X., Wang, R., Kang, Y., Xue, X., Meng, R., Wei, Q., and Feng, X. (2018). Upregulation of circ-UBAP2 predicts poor prognosis and promotes triple-negative breast cancer progression through the miR-661/MTA1 pathway. Biochem Biophys Res Commun 505, 996-1002.

163. Sun, Y., Li, X., Chen, A., Shi, W., Wang, L., Yi, R., and Qiu, J. (2019). circPIP5K1A serves as a competitive endogenous RNA contributing to ovarian cancer progression via regulation of miR-661/IGFBP5 signaling. J Cell Biochem 120, 19406-19414.

164. Zhu, T., Yuan, J., Wang, Y., Gong, C., Xie, Y., and Li, H. (2015). MiR-661 contributed to cell proliferation of human ovarian cancer cells by repressing INPP5J expression. Biomed Pharmacother 75, 123-128.

165. Vetter, G., Saumet, A., Moes, M., Vallar, L., Le Bechec, A., Laurini, C., Sabbah, M., Arar, K., Theillet, C., Lecellier, C.H., et al. (2010). miR-661 expression in SNAI1-induced epithelial to mesenchymal transition contributes to breast cancer cell invasion by targeting Nectin-1 and StarD10 messengers. Oncogene 29, 4436-4448.

166. Shu, L., Wang, Z., Wang, Q., Wang, Y., and Zhang, X. (2018). Signature miRNAs in peripheral blood monocytes of patients with gastric or breast cancers. Open Biol 8.

167. Wang, D.X., Zou, Y.J., Zhuang, X.B., Chen, S.X., Lin, Y., Li, W.L., Lin, J.J., and Lin, Z.Q. (2017). Sulforaphane suppresses EMT and metastasis in human lung cancer through miR-616-5p-mediated GSK3beta/beta-catenin signaling pathways. Acta Pharmacol Sin 38, 241-251.

168. Bai, Q.L., Hu, C.W., Wang, X.R., Shang, J.X., and Yin, G.F. (2017). MiR-616 promotes proliferation and inhibits apoptosis in glioma cells by suppressing expression of SOX7 via the Wnt signaling pathway. Eur Rev Med Pharmacol Sci 21, 5630-5637.

169. Chen, Z., Zhu, J., Zhu, Y., and Wang, J. (2018). MicroRNA-616 promotes the progression of ovarian cancer by targeting TIMP2. Oncol Rep 39, 2960-2968.

170. Zhu, L.M., and Li, N. (2020). Downregulation of long noncoding RNA TUSC7 promoted cell growth, invasion and migration through sponging with miR-616-5p/GSK3beta pathway in ovarian cancer. Eur Rev Med Pharmacol Sci 24, 7253-7265.

171. Li, W., Li, Y., Ma, W., Zhou, J., Sun, Z., and Yan, X. (2020). Long noncoding RNA AC114812.8 promotes the progression of bladder cancer through miR-371b-5p/FUT4 axis. Biomed Pharmacother 121, 109605.

172. Luo, X., Zhang, X., Peng, J., Chen, Y., Zhao, W., Jiang, X., Su, L., Xie, M., and Lin, B. (2020). miR-371b-5p promotes cell proliferation, migration and invasion in non-small cell lung cancer via SCAI. Biosci Rep 40.

173. Cartier, F., Indersie, E., Lesjean, S., Charpentier, J., Hooks, K.B., Ghousein, A., Desplat, A., Dugot-Senant, N., Trezeguet, V., Sagliocco, F., et al. (2017). New tumor suppressor microRNAs target glypican-3 in human liver cancer. Oncotarget 8, 41211-41226.

174. Jiang, X., Jiang, M., Xu, M., Xu, J., and Li, Y. (2019). Identification of diagnostic utility and molecular mechanisms of circulating miR-551b-5p in gastric cancer. Pathol Res Pract 215, 900-904.

175. Yokoi, A., Matsuzaki, J., Yamamoto, Y., Tate, K., Yoneoka, Y., Shimizu, H., Uehara, T., Ishikawa, M., Takizawa, S., Aoki, Y., et al. (2019). Serum microRNA profile enables preoperative diagnosis of uterine leiomyosarcoma. Cancer Sci 110, 3718-3726.

176. Guan, H., Liu, C., Fang, F., Huang, Y., Tao, T., Ling, Z., You, Z., Han, X., Chen, S., Xu, B., et al. (2017). MicroRNA-744 promotes prostate cancer progression through aberrantly activating Wnt/beta-catenin signaling. Oncotarget 8, 14693-14707.

177. Shimojo, M., Kasahara, Y., Inoue, M., Tsunoda, S.I., Shudo, Y., Kurata, T., and Obika, S. (2019). A gapmer antisense oligonucleotide targeting SRRM4 is a novel therapeutic medicine for lung cancer. Sci Rep 9, 7618.

178. Kent, M.S., Burton, J.H., Dank, G., Bannasch, D.L., and Rebhun, R.B. (2018). Association of cancer-related mortality, age and gonadectomy in golden retriever dogs at a veterinary academic center (1989-2016). PLoS One 13, e0192578.

179. Salas, Y., Marquez, A., Diaz, D., and Romero, L. (2015). Epidemiological Study of Mammary Tumors in Female Dogs Diagnosed during the Period 2002-2012: A Growing Animal Health Problem. PLoS One 10, e0127381.

180. Brackman, J. (2020). Large-Scale Cancer Study of Golden Retrievers Holds Hope For All Dogs. In. (

181. Hayward, J.J., Castelhano, M.G., Oliveira, K.C., Corey, E., Balkman, C., Baxter, T.L., Casal, M.L., Center, S.A., Fang, M., Garrison, S.J., et al. (2016). Complex disease and phenotype mapping in the domestic dog. Nat Commun 7, 10460.

182. Tonomura, N., Elvers, I., Thomas, R., Megquier, K., Turner-Maier, J., Howald, C., Sarver, A.L., Swofford, R., Frantz, A.M., Ito, D., et al. (2015). Genome-wide association study identifies shared risk loci common to two malignancies in golden retrievers. PLoS Genet 11, e1004922.

183. Petmed. (2014). Pet Health Report: Australian Cattle Dog. In. (

184. Smith, D.K., and Xue, H. (1997). Sequence profiles of immunoglobulin and immunoglobulin-like domains. J Mol Biol 274, 530-545.

185. Kuespert, K., Pils, S., and Hauck, C.R. (2006). CEACAMs: their role in physiology and pathophysiology. Curr Opin Cell Biol 18, 565-571.

186. Beauchemin, N., and Arabzadeh, A. (2013). Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis. Cancer Metastasis Rev 32, 643-671.

187. Han, Z.W., Lyv, Z.W., Cui, B., Wang, Y.Y., Cheng, J.T., Zhang, Y., Cai, W.Q., Zhou, Y., Ma, Z.W., Wang, X.W., et al. (2020). The old CEACAMs find their new role in tumor immunotherapy. Invest New Drugs 38, 1888-1898.

188. Kammerer, R., Popp, T., Hartle, S., Singer, B.B., and Zimmermann, W. (2007). Species-specific evolution of immune receptor tyrosine based activation motif-containing CEACAM1-related immune receptors in the dog. BMC Evol Biol 7, 196.

189. Kammerer, R., and Zimmermann, W. (2010). Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families. BMC Biol 8, 12.

190. Weichselbaumer, M., Willmann, M., Reifinger, M., Singer, J., Bajna, E., Sobanov, Y., Mechtcherikova, D., Selzer, E., Thalhammer, J.G., Kammerer, R., et al. (2011). Phylogenetic discordance of human and canine carcinoembryonic antigen (CEA, CEACAM) families, but striking identity of the CEA receptors will impact comparative oncology studies. PLoS Curr 3, RRN1223.

191. Rockenbauer, E., Bendixen, M.H., Bukowy, Z., Yin, J., Jacobsen, N.R., Hedayati, M., Vogel, U., Grossman, L., Bolund, L., and Nexo, B.A. (2002). Association of chromosome 19q13.2-3 haplotypes with basal cell carcinoma: tentative delineation of an involved region using data for single nucleotide polymorphisms in two cohorts. Carcinogenesis 23, 1149-1153.

192. Yin, J., Rockenbauer, E., Hedayati, M., Jacobsen, N.R., Vogel, U., Grossman, L., Bolund, L., and Nexo, B.A. (2002). Multiple single nucleotide polymorphisms on human chromosome 19q13.2-3 associate with risk of Basal cell carcinoma. Cancer Epidemiol Biomarkers Prev 11, 1449-1453.

193. Nexo, B.A., Vogel, U., Olsen, A., Ketelsen, T., Bukowy, Z., Thomsen, B.L., Wallin, H., Overvad, K., and Tjonneland, A. (2003). A specific haplotype of single nucleotide polymorphisms on chromosome 19q13.2-3 encompassing the gene RAI is indicative of post-menopausal breast cancer before age 55. Carcinogenesis 24, 899-904.

194. Vogel, U., Laros, I., Jacobsen, N.R., Thomsen, B.L., Bak, H., Olsen, A., Bukowy, Z., Wallin, H., Overvad, K., Tjonneland, A., et al. (2004). Two regions in chromosome 19q13.2-3 are associated with risk of lung cancer. Mutat Res 546, 65-74.

195. Nexo, B.A., Vogel, U., Olsen, A., Nyegaard, M., Bukowy, Z., Rockenbauer, E., Zhang, X., Koca, C., Mains, M., Hansen, B., et al. (2008). Linkage disequilibrium mapping of a breast cancer susceptibility locus near RAI/PPP1R13L/iASPP. BMC Med Genet 9, 56.

196. Amin Al Olama, A., Kote-Jarai, Z., Schumacher, F.R., Wiklund, F., Berndt, S.I., Benlloch, S., Giles, G.G., Severi, G., Neal, D.E., Hamdy, F.C., et al. (2013). A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. Hum Mol Genet 22, 408-415.

197. Gao, P., Xia, J.H., Sipeky, C., Dong, X.M., Zhang, Q., Yang, Y., Zhang, P., Cruz, S.P., Zhang, K., Zhu, J., et al. (2018). Biology and Clinical Implications of the 19q13 Aggressive Prostate Cancer Susceptibility Locus. Cell 174, 576-589 e518.

198. Zheng, J., Miller, K.K., Yang, T., Hildebrand, M.S., Shearer, A.E., DeLuca, A.P., Scheetz, T.E., Drummond, J., Scherer, S.E., Legan, P.K., et al. (2011). Carcinoembryonic antigen-related cell adhesion molecule 16 interacts with alpha-tectorin and is mutated in autosomal dominant hearing loss (DFNA4). Proc Natl Acad Sci U S A 108, 4218-4223.

199. Wang, H., Wang, X., He, C., Li, H., Qing, J., Grati, M., Hu, Z., Li, J., Hu, Y., Xia, K., et al. (2015). Exome sequencing identifies a novel CEACAM16 mutation associated with autosomal dominant nonsyndromic hearing loss DFNA4B in a Chinese family. J Hum Genet 60, 119-126.

200. Kammerer, R., Ruttiger, L., Riesenberg, R., Schauble, C., Krupar, R., Kamp, A., Sunami, K., Eisenried, A., Hennenberg, M., Grunert, F., et al. (2012). Loss of mammal-specific tectorial membrane component carcinoembryonic antigen cell adhesion molecule 16 (CEACAM16) leads to hearing impairment at low and high frequencies. J Biol Chem 287, 21584-21598.

201. Yang, C., He, P., Liu, Y., He, Y., Yang, C., Du, Y., Zhou, M., Wang, W., Zhang, G., Wu, M., et al. (2015). Down-regulation of CEACAM1 in breast cancer. Acta Biochim Biophys Sin (Shanghai) 47, 788-794.

202. Busch, C., Hanssen, T.A., Wagener, C., and B, O.B. (2002). Down-regulation of CEACAM1 in human prostate cancer: correlation with loss of cell polarity, increased proliferation rate, and Gleason grade 3 to 4 transition. Hum Pathol 33, 290-298.

203. Liu, J., Muturi, H.T., Khuder, S.S., Helal, R.A., Ghadieh, H.E., Ramakrishnan, S.K., Kaw, M.K., Lester, S.G., Al-Khudhair, A., Conran, P.B., et al. (2020). Loss of Ceacam1 promotes prostate cancer progression in Pten haploinsufficient male mice. Metabolism 107, 154215.

204. Bamberger, A.M., Riethdorf, L., Nollau, P., Naumann, M., Erdmann, I., Gotze, J., Brummer, J., Schulte, H.M., Wagener, C., and Loning, T. (1998). Dysregulated expression of CD66a (BGP, C-CAM), an adhesion molecule of the CEA family, in endometrial cancer. Am J Pathol 152, 1401-1406.

205. Takeuchi, A., Yokoyama, S., Nakamori, M., Nakamura, M., Ojima, T., Yamaguchi, S., Mitani, Y., Shively, J.E., and Yamaue, H. (2019). Loss of CEACAM1 is associated with poor prognosis and peritoneal dissemination of patients with gastric cancer. Scientific Reports 9, 12702.

206. Fournes, B., Sadekova, S., Turbide, C., Letourneau, S., and Beauchemin, N. (2001). The CEACAM1-L Ser503 residue is crucial for inhibition of colon cancer cell tumorigenicity. Oncogene 20, 219-230.

207. Song, J.H., Cao, Z., Yoon, J.H., Nam, S.W., Kim, S.Y., Lee, J.Y., and Park, W.S. (2011). Genetic alterations and expression pattern of CEACAM1 in colorectal adenomas and cancers. Pathol Oncol Res 17, 67-74.

208. Powell, E., Shao, J., Picon, H.M., Bristow, C., Ge, Z., Peoples, M., Robinson, F., Jeter-Jones, S.L., Schlosberg, C., Grzeskowiak, C.L., et al. (2018). A functional genomic screen in vivo identifies CEACAM5 as a clinically relevant driver of breast cancer metastasis. npj Breast Cancer 4, 9.

209. Maraqa, L., Cummings, M., Peter, M.B., Shaaban, A.M., Horgan, K., Hanby, A.M., and Speirs, V. (2008). Carcinoembryonic antigen cell adhesion molecule 6 predicts breast cancer recurrence following adjuvant tamoxifen. Clin Cancer Res 14, 405-411.

210. Rizeq, B., Zakaria, Z., and Ouhtit, A. (2018). Towards understanding the mechanisms of actions of carcinoembryonic antigen-related cell adhesion molecule 6 in cancer progression. Cancer Sci 109, 33-42.

211. Tsang, J.Y., Kwok, Y.K., Chan, K.W., Ni, Y.B., Chow, W.N., Lau, K.F., Shao, M.M., Chan, S.K., Tan, P.H., and Tse, G.M. (2013). Expression and clinical significance of

carcinoembryonic antigen-related cell adhesion molecule 6 in breast cancers. Breast Cancer Res Treat 142, 311-322.

212. Michaelidou, K., Tzovaras, A., Missitzis, I., Ardavanis, A., and Scorilas, A. (2013). The expression of the CEACAM19 gene, a novel member of the CEA family, is associated with breast cancer progression. Int J Oncol 42, 1770-1777.

213. Estiar, M.A., Esmaeili, R., Zare, A.A., Farahmand, L., Fazilaty, H., Zekri, A., Jafarbeik-Iravani, N., and Majidzadeh, A.K. (2017). High expression of CEACAM19, a new member of carcinoembryonic antigen gene family, in patients with breast cancer. Clin Exp Med 17, 547-553.

214. Iwabuchi, E., Miki, Y., Onodera, Y., Shibahara, Y., Takagi, K., Suzuki, T., Ishida, T., and Sasano, H. (2019). Co-expression of carcinoembryonic antigen-related cell adhesion molecule 6 and 8 inhibits proliferation and invasiveness of breast carcinoma cells. Clin Exp Metastasis 36, 423-432.

215. Gaur, S., Shively, J.E., Yen, Y., and Gaur, R.K. (2008). Altered splicing of CEACAM1 in breast cancer: identification of regulatory sequences that control splicing of CEACAM1 into long or short cytoplasmic domain isoforms. Mol Cancer 7, 46.

216. Zisi, Z., Adamopoulos, P.G., Kontos, C.K., and Scorilas, A. (2020). Identification and expression analysis of novel splice variants of the human carcinoembryonic antigen-related cell adhesion molecule 19 (CEACAM19) gene using a high-throughput sequencing approach. Genomics 112, 4268-4276.

217. Li, H., Yang, B., Xing, K., Yuan, N., Wang, B., Chen, Z., He, W., and Zhou, J. (2014). A preliminary study of the relationship between breast cancer metastasis and loss of heterozygosity by using exome sequencing. Sci Rep 4, 5460.

218. Scholzel, S., Zimmermann, W., Schwarzkopf, G., Grunert, F., Rogaczewski, B., and Thompson, J. (2000). Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas. Am J Pathol 156, 595-605.

219. Raj, D., Nikolaidi, M., Garces, I., Lorizio, D., Castro, N.M., Caiafa, S.G., Moore, K., Brown, N.F., Kocher, H.M., Duan, X., et al. (2021). CEACAM7 Is an Effective Target for CAR T-cell Therapy of Pancreatic Ductal Adenocarcinoma. Clin Cancer Res 27, 1538-1552.

220. Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science 347, 1260419.

221. Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F., Aittomaki, K., Andrulis, I.L., Anton-Culver, H., Arason, A., Arndt, V., Aronson, K.J., Arun, B.K., et al. (2019). Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat Commun 10, 1741.

222. Balogh, G.A., Russo, J., Mailo, D.A., Heulings, R., Russo, P.A., Morrison, P., Sheriff, F., and Russo, I.H. (2007). The breast of parous women without cancer has a different genomic profile compared to those with cancer. Int J Oncol 31, 1165-1175.

223. Naghibalhossaini, F., and Stanners, C.P. (2004). Minimal mutations are required to effect a radical change in function in CEA family members of the Ig superfamily. J Cell Sci 117, 761-769.

224. Kuroki, M., Abe, H., Imakiirei, T., Liao, S., Uchida, H., Yamauchi, Y., Oikawa, S., and Kuroki, M. (2001). Identification and comparison of residues critical for cell-adhesion

activities of two neutrophil CD66 antigens, CEACAM6 and CEACAM8. J Leukoc Biol 70, 543-550.

225. Villullas, S., Hill, D.J., Sessions, R.B., Rea, J., and Virji, M. (2007). Mutational analysis of human CEACAM1: the potential of receptor polymorphism in increasing host susceptibility to bacterial infection. Cell Microbiol 9, 329-346.
226. Markel, G., Gruda, R., Achdout, H., Katz, G., Nechama, M., Blumberg, R.S., Kammerer, R., Zimmermann, W., and Mandelboim, O. (2004). The critical role of residues 43R and 44Q of carcinoembryonic antigen cell adhesion molecules-1 in the protection from killing by human NK cells. J Immunol 173, 3732-3739.
227. Gu, S., Zaidi, S., Hassan, M.I., Mohammad, T., Malta, T.M., Noushmehr, H., Nguyen, B., Crandall, K.A., Srivastav, J., Obias, V., et al. (2020). Mutated CEACAMs Disrupt Transforming Growth Factor Beta Signaling and Alter the Intestinal Microbiome to Promote Colorectal Carcinogenesis. Gastroenterology 158, 238-252.
228. Eslami, S.Z., Majidzadeh, A.K., Halvaei, S., Babapirali, F., and Esmaeili, R. (2020). Microbiome and Breast Cancer: New Role for an Ancient Population. Front Oncol 10, 120.
229. Fernandez, M.F., Reina-Perez, I., Astorga, J.M., Rodriguez-Carrillo, A., Plaza-Diaz, J., and Fontana, L. (2018). Breast Cancer and Its Relationship with the Microbiota. Int J Environ Res Public Health 15.
230. Tchoupa, A.K., Schuhmacher, T., and Hauck, C.R. (2014). Signaling by epithelial members of the CEACAM family - mucosal docking sites for pathogenic bacteria. Cell Commun Signal 12, 27.
231. Society, A.C. (2020). Colorectal Cancer Facts & Figures 2020-2022. In. (Atlanta, GA, American Cancer Society.
232. Calvert, P.M., and Frucht, H. (2002). The genetics of colorectal cancer. Ann Intern Med 137, 603-612.
233. Kastrinos, F., and Syngal, S. (2011). Inherited colorectal cancer syndromes. Cancer J 17, 405-415.
234. Haraldsdottir, S., Rafnar, T., Frankel, W.L., Einarsdottir, S., Sigurdsson, A., Hampel, H., Snaebjornsson, P., Masson, G., Weng, D., Arngrimsson, R., et al. (2017). Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2. Nat Commun 8, 14755.
235. Jasperson, K.W., Tuohy, T.M., Neklason, D.W., and Burt, R.W. (2010). Hereditary and familial colon cancer. Gastroenterology 138, 2044-2058.
236. Bucksch, K., Zachariae, S., Aretz, S., Buttner, R., Holinski-Feder, E., Holzapfel, S., Huneburg, R., Kloor, M., von Knebel Doeberitz, M., Morak, M., et al. (2020). Cancer risks in Lynch syndrome, Lynch-like syndrome, and familial colorectal cancer type X: a prospective cohort study. BMC Cancer 20, 460.
237. Lynch, H.T., Snyder, C.L., Shaw, T.G., Heinen, C.D., and Hitchins, M.P. (2015). Milestones of Lynch syndrome: 1895-2015. Nat Rev Cancer 15, 181-194.
238. Scott, R.J., and Ashton, K.A. (2004). Familial breast and bowel cancer: does it exist? Hered Cancer Clin Pract 2, 25-29.
239. Shin, D.W., Choi, Y.J., Kim, H.S., Han, K.D., Yoon, H., Park, Y.S., Kim, N., and Lee, D.H. (2018). Secondary Breast, Ovarian, and Uterine Cancers After Colorectal Cancer: A Nationwide Population-Based Cohort Study in Korea. Dis Colon Rectum 61, 1250-1257.

240. Yang, C., He, P., Liu, Y., He, Y., Yang, C., Du, Y., Zhou, M., Wang, W., Zhang, G., Wu, M., et al. (2015). Assay of serum CEACAM1 as a potential biomarker for breast cancer. Clin Chim Acta 450, 277-281.
241. Gold, P., and Freedman, S.O. (1965). Specific carcinoembryonic antigens of the human digestive system. J Exp Med 122, 467-481.
242. Gold, P., and Freedman, S.O. (1965). Demonstration of Tumor-Specific Antigens in Human Colonic Carcinomata by Immunological Tolerance and Absorption Techniques. J Exp Med 121, 439-462.
243. Kim, K.S., Kim, J.T., Lee, S.J., Kang, M.A., Choe, I.S., Kang, Y.H., Kim, S.Y., Yeom, Y.I., Lee, Y.H., Kim, J.H., et al. (2013). Overexpression and clinical significance of carcinoembryonic antigen-related cell adhesion molecule 6 in colorectal cancer. Clin Chim Acta 415, 12-19.
244. Messick, C.A., Sanchez, J., Dejulius, K.L., Hammel, J., Ishwaran, H., and Kalady, M.F. (2010). CEACAM-7: a predictive marker for rectal cancer recurrence. Surgery 147, 713-719.
245. Gandhi, A.K., Sun, Z.J., Kim, W.M., Huang, Y.H., Kondo, Y., Bonsor, D.A., Sundberg, E.J., Wagner, G., Kuchroo, V.K., Petsko, G.A., et al. (2021). Structural basis of the dynamic human CEACAM1 monomer-dimer equilibrium. Commun Biol 4, 360.
246. Bonsor, D.A., Gunther, S., Beadenkopf, R., Beckett, D., and Sundberg, E.J. (2015). Diverse oligomeric states of CEACAM IgV domains. Proc Natl Acad Sci U S A 112, 13561-13566.
247. Kim, W.M., Huang, Y.H., Gandhi, A., and Blumberg, R.S. (2019). CEACAM1 structure and function in immunity and its therapeutic implications. Semin Immunol 42, 101296.
248. Zhuo, Y., Yang, J.Y., Moremen, K.W., and Prestegard, J.H. (2016). Glycosylation Alters Dimerization Properties of a Cell-surface Signaling Protein, Carcinoembryonic Antigen-related Cell Adhesion Molecule 1 (CEACAM1). J Biol Chem 291, 20085-20095.
249. Skubitz, K.M., and Skubitz, A.P. (2008). Interdependency of CEACAM-1, -3, -6, and -8 induced human neutrophil adhesion to endothelial cells. J Transl Med 6, 78.
250. Rueckschloss, U., Kuerten, S., and Ergun, S. (2016). The role of CEA-related cell adhesion molecule-1 (CEACAM1) in vascular homeostasis. Histochem Cell Biol 146, 657-671.
251. Kelleher, M., Singh, R., O'Driscoll, C.M., and Melgar, S. (2019). Carcinoembryonic antigen (CEACAM) family members and Inflammatory Bowel Disease. Cytokine Growth Factor Rev 47, 21-31.
252. Hollandsworth, H.M., Amirfakhri, S., Filemoni, F., Schmitt, V., Wennemuth, G., Schmidt, A., Hoffman, R.M., Singer, B.B., and Bouvet, M. (2020). Anti-carcinoembryonic antigen-related cell adhesion molecule antibody for fluorescence visualization of primary colon cancer and metastases in patient-derived orthotopic xenograft mouse models. Oncotarget 11, 429-439.
253. Wakabayashi-Nakao, K., Hatakeyama, K., Ohshima, K., Ken Yamaguchi, K., and Mochizuki, T. (2014). Carcinoembryonic antigen-related cell adhesion molecule 4 (CEACAM4) is specifically expressed in medullary thyroid carcinoma cells. Biomed Res 35, 237-242.
254. Zhao, H., Xu, J., Wang, Y., Jiang, R., Li, X., Zhang, L., and Che, Y. (2018). Knockdown of CEACAM19 suppresses human gastric cancer through inhibition of PI3K/Akt and NF-kappaB. Surg Oncol 27, 495-502.

255. Zhang, H., Eisenried, A., Zimmermann, W., and Shively, J.E. (2013). Role of CEACAM1 and CEACAM20 in an in vitro model of prostate morphogenesis. PLoS One 8, e53359.
256. Kitamura, Y., Murata, Y., Park, J.H., Kotani, T., Imada, S., Saito, Y., Okazawa, H., Azuma, T., and Matozaki, T. (2015). Regulation by gut commensal bacteria of carcinoembryonic antigen-related cell adhesion molecule expression in the intestinal epithelium. Genes Cells 20, 578-589.
257. Murata, Y., Kotani, T., Supriatna, Y., Kitamura, Y., Imada, S., Kawahara, K., Nishio, M., Daniwijaya, E.W., Sadakata, H., Kusakari, S., et al. (2015). Protein tyrosine phosphatase SAP-1 protects against colitis through regulation of CEACAM20 in the intestinal epithelium. Proc Natl Acad Sci U S A 112, E4264-4271.
258. Augustus, G.J., and Ellis, N.A. (2018). Colorectal Cancer Disparity in African Americans: Risk Factors and Carcinogenic Mechanisms. Am J Pathol 188, 291-303.
259. Han, Z.M., Huang, H.M., and Sun, Y.W. (2018). Effect of CEACAM-1 knockdown in human colorectal cancer cells. Oncol Lett 16, 1622-1626.
260. Hauptman, N., and Glavac, D. (2017). Colorectal Cancer Blood-Based Biomarkers. Gastroenterol Res Pract 2017, 2195361.
261. Ru, G.Q., Han, Y., Wang, W., Chen, Y., Wang, H.J., Xu, W.J., Ma, J., Ye, M., Chen, X., He, X.L., et al. (2017). CEACAM6 is a prognostic biomarker and potential therapeutic target for gastric carcinoma. Oncotarget 8, 83673-83683.
262. Minton, J.P., and Martin, E.W., Jr. (1978). The use of serial CEA determinations to predict recurrence of colon cancer and when to do a second-look operation. Cancer 42, 1422-1427.
263. Moertel, C.G., Fleming, T.R., Macdonald, J.S., Haller, D.G., Laurie, J.A., and Tangen, C. (1993). An evaluation of the carcinoembryonic antigen (CEA) test for monitoring patients with resected colon cancer. JAMA 270, 943-947.