

**Genomics underlying
a canine hereditary
thyroid follicular cell carcinoma**

Yun Yu 喻运

Propositions

1. Eradication of hereditary thyroid cancer from German Longhaired Pointer dogs requires using a genetic test (this thesis).
2. Dogs are undervalued as models for studying genetic architectures of hereditary diseases (this thesis).
3. Barriers to scientific collaboration between countries cause waste of scientific resources.
4. Free trading is the primary factor to glue the world together.
5. Media abuse is the biggest threat to world peace.
6. Poverty kills more human lives than climate change.

Propositions belonging to the thesis entitled

Genomics underlying a canine hereditary thyroid follicular cell carcinoma

Yun Yu

Wageningen 11 October 2022

Genomics underlying a canine hereditary thyroid follicular cell carcinoma

Yun Yu

Thesis Committee

Promotor

Prof. Dr MAM Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Copromotor

Dr RPMA Crooijmans
Assistant professor, Animal Breeding and Genomics
Wageningen University & Research

Other members

Prof. Dr JL van Leeuwen, Wageningen University & Research
Prof. Dr A. Gröne, University of Utrecht, the Netherlands
Dr FJB van Duijnhoven, Wageningen University & Research
Dr J. Ubels, Oncode Institute, Prinses Maxima Centrum, Utrecht

This research was conducted under the auspices of the Graduate School Wageningen Institute of Animal Sciences.

Genomics underlying a canine hereditary thyroid follicular cell carcinoma

Yun Yu

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University,

by the authority of the Rector Magnificus,

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Tuesday 11 October 2022

at 4 p.m. in the Omnia Auditorium.

Yun Yu

Genomics underlying a canine hereditary thyroid follicular cell carcinoma

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)

With references, with summary in English

ISBN: 978-94-6447-279-0

DOI: <https://doi.org/10.18174/571907>

Abstract

Yu, Y. (2022). Genomics underlying a canine thyroid follicular cell carcinoma. PhD thesis, Wageningen University, the Netherlands

Intense human selection on domestic dogs over the past ~200 years has created a variety of pure dog breeds, but also resulted in breed-specific predispositions to many hereditary diseases/disorders. The incidence of thyroid cancer (TC) in the Dutch population of German Longhaired Pointer (GLP) dogs has been extremely high over the past approximately 20 years, indicating a population predisposition to TC. To help the breeders to eradicate TCs from Dutch GLPs, I performed a series of analyses to decode TCs that occurred in those Dutch GLPs. Firstly, we determined the histological subtype of identified TCs and revealed that these TCs are a familial disease according to consanguinity of affected GLPs. I investigated the effect of inbreeding on the incidence of the familial TCs in these Dutch GLPs based on both pedigree and genotype data and revealed that inbreeding contributed to the high incidence of the familial TCs in these GLPs. Furthermore, I identified germline risk mutations for familial TC using a combination of a genome-wide association study and homozygosity mapping. The identified germline risk mutation is used in a genetic test that identifies GLPs at a high risk for familial TC. To further understand the molecular mechanism underlying familial TC initiation and development, I profiled the somatic mutation landscape of 7 familial TCs and identified a recurrent missense mutation that very likely drives tumorigenesis. Furthermore, I genetically characterized the GLP breed and identified a specific selection signature that might contribute to hunting performance of GLPs. Lastly, I tested a novel approach to predict driver mutations using prior signaling pathway knowledge. Together I comprehensively decoded the familial TCs in the Dutch GLPs and developed a genetic test to identify dogs at a high risk for familial TCs.

Contents

9	General introduction
33	Familial follicular cell thyroid carcinomas in a large number of Dutch German longhaired pointers
53	Deleterious Mutations in the <i>TPO</i> Gene Associated with Familial Thyroid Follicular Cell Carcinoma in Dutch German Longhaired Pointers
85	A recurrent somatic missense mutation in the <i>GNAS</i> gene identified in familial thyroid follicular cell carcinomas in German longhaired pointer dogs
121	Unique genetic signature and selection footprints in Dutch population of German Longhaired Pointer dogs
149	A cancer gene score based on pathways and its application in driver mutation prediction using machine learning approach
163	General discussion
189	Summary
191	Appendices

1

General introduction

1.1. Cancer

Cancer is a disease in which some of the body's cells grow uncontrollably in a tissue and may spread to other parts of the body (NCI 2021). Cancer is a leading cause of death in humans worldwide. In 2020, more than 19 million new cancer cases appeared and nearly 10 million people died of a variety of cancers (Sung et al. 2021). Cancer can start almost everywhere in the body, such as the brain, bone marrow, and blood or lymph vessels, and spread to other parts through the bloodstream or lymphatic system. According to cell origin, cancer can be divided into 6 major categories. Carcinoma is the cancer derived from epithelial cells, sarcoma from connective tissue (i.e. bone, cartilage, fat, nerve), lymphoma from cells in the lymphatic system, leukemia from the bone marrow, blastoma from immature "precursor" cells or embryonic tissue, and myeloma originating from plasma cells of bone marrow (Linares-Clemente et al. 2017; Barwick et al. 2019; Chopra and Bohlander 2019; Carbone 2020). The place where the cancer starts is called the primary site with the tumor as primary tumor. The tumor resulting from spreading is called secondary tumor or metastasis. Like in human most cancers also appear in animals. The oldest evidence of a cancer was found in fossilized dinosaurs that lived approximately 77.0 – 75.5 million years ago (Ekhtiari et al. 2020). In companion animals, dogs and cats, cancer is one of the leading causes of death (Zappulli et al. 2005; Davis and Ostrander 2014).

Although the number of deaths due to cancer is still high, cancer survival has increased over decades. For instance, cancer survival for ten or more years in the UK has increased from 24% in the 1970s to 50% in 2011 (Quaresma et al. 2015). Cancer is still hard to cure, especially for certain specific cancer types, such as pancreatic and lung cancer. Cancer survival increased due to a variety of treatments that has been developed and improved, such as surgery, chemotherapy, radiotherapy, targeted medication, and immunotherapy (Esfahani et al. 2020). Still more research is needed to really beat cancer.

Cancer is a disease that is caused by genetic changes to genes, especially genes that control cell growth and division. These genetic changes are so called causal mutations. These causal mutations can occur and accumulate in the cells by several ways: 1) replication errors that occur spontaneously as cells divide; 2) damage to DNA by some environmental harmful substances, such as chemicals (in tobacco smoke) and ultraviolet rays (from the sun); 3) inheritance from parents (Martincorena and Campbell 2015). In addition, viruses are known causes of some cancers, for example, the papillomaviruses can cause cervical cancer in humans and oral carcinomas in dogs (Munday et al. 2015). Cancer can also be contagious. For

instance, canine transmissible venereal tumor, which can be transmitted by sexual intercourse and cause genital cancer in dogs (Pimentel et al. 2021).

Cancer results from the clonal expansion of a single abnormal cell (Martincorena and Campbell 2015). To become a cancerous cell from a normal cell, several abilities must be gained. According to Hanahan and Weinberg (2000), these abilities include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis (Figure 1.1). In their follow-up paper in 2011 (Hanahan and Weinberg 2011), they described two other enabling characteristics underlying those hallmarks, i.e., genome instability and inflammation. Genome instability generates the genetic diversity that expedites their acquisition and inflammation fosters multiple hallmark functions. In addition, two emerging hallmarks are also important, i.e., reprogramming of energy metabolism and evading immune destruction (Hanahan and Weinberg 2011).

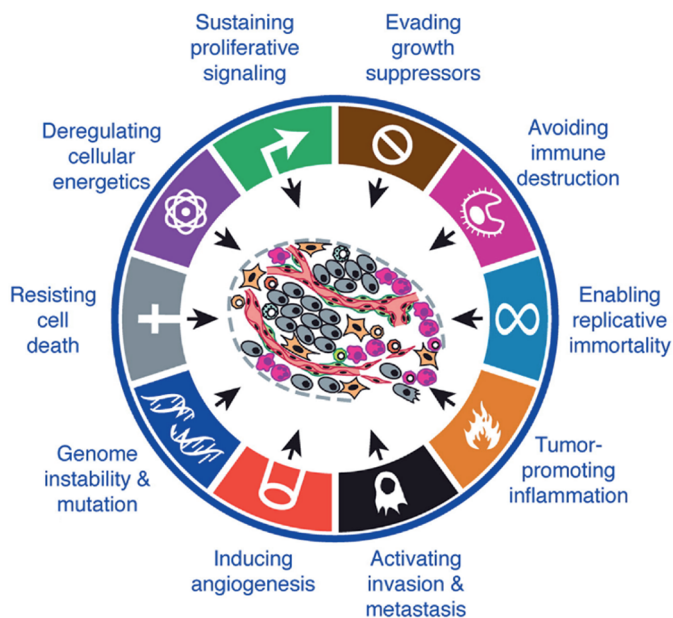


Figure 1.1. The hallmarks of cancer. Reprinted from “Hallmarks of Cancer: The Next Generation” of Hanahan and Weinberg (2011) with permission.

1.2. Familial and sporadic cancer

Based on how the causal mutation is obtained, cancer can be of two categories, the familial (hereditary) and sporadic (non-hereditary) types (Anderson 1992). In

familial cancer the causal germline mutation(s), is(are) inherited from one or both parents. Familial cancer usually presents some specific features, such as, development of cancer at a young age and occurrence in both of a pair of organs (e.g. kidney and thyroid gland) (Winship and Dudding 2008). Sporadic cancer has causal mutation(s) that occurs during somatic cell division or is a *de novo* mutation arising in the germ cells or during early embryogenesis after fertilization (Acuna-Hidalgo et al. 2016). Most of these mutations are acquired by random DNA replication error that occurred by chance or as a result of exposure to exogenous carcinogens that can increase the risk of cancer, such as smoking and X-rays, or endogenous factors, such as reactive oxygen species and aldehydes (Martincorena and Campbell 2015). These mutations occur in somatic cells, and thus are called somatic mutations (Miles and Tadi 2022). The term “somatic mutation” was first used by Ernest E Tyzzer in 1916 (Tyzzer 1916; Wunderlich 2007). Somatic mutations can be passed to the progeny of the mutated cell during cell division but are not usually able to be transmitted to descendants. Most human cancers are sporadic. Familial cancer accounts for only around 5% to 10% of all cancers, which amounts to 0.95-1.9 million hereditary cancer patients in 2020 (Rahner and Steinke 2008; AlHarthi et al. 2020).

1.2.1. Inheritance pattern

The familial cancer can be monogenic or polygenic according to the number of genes involved in the disease predisposition. A monogenic disease results from a genetic alteration in a single gene, while a polygenic disease (also called complex disease) results from combined effects of multiple mutations in multiple genes (Debniak and Lubinski 2008; Crouch and Bodmer 2020). A monogenic familial cancer can have different inheritance patterns depending on type of chromosome where the gene is on and required number of causal alleles according to classic mendelian genetics, such as autosomal recessive inheritance, autosomal dominant inheritance, X-linked dominant, X-linked recessive, and mitochondrial. A polygenic cancer does not have a clear-cut pattern of inheritance (Crouch and Bodmer 2020). The pedigree can be used to analyze the inheritance pattern of a mendelian disease segregating in a family before mapping the genetic cause of the disease using genomic data. For a dominant disease, at least one of the parents of the affected individual must be affected. For an autosomal recessive disease, unaffected parents can have affected offspring and different from X-linked recessive diseases, male and female offspring are equally affected. A mitochondrial disease is inherited through the maternal line because paternal mitochondrial DNA from sperm is believed not to enter the fertilized egg (Schapira 2006). The thyroid tumor in the Dutch German Longhaired Pointers (GLPs) we studied in this thesis is most likely an autosomal recessive disease according to the pedigree (described in detail in Chapter 2). Most pedigreed dogs have a pedigree that can be tracked to several generations back

because a dog can become a registered member of a breed only if both its father and mother are registered members. With pedigree and affection information of related dogs, it is much easier to determine if the cancer is a sporadic cancer or a hereditary disease and what is the most likely inheritance pattern before a further genetic study is done.

Autosomal recessive: Mutation in both copies of the gene is needed to lead to tumor formation.

Autosomal dominant: Mutation in only one copy of the gene is sufficient to cause the cancer.

X-linked dominant: Cancer is caused by a mutation in the gene located on the X-chromosome. In males, the mutation in the only copy of the gene will cause the cancer. In females, mutation in any copy of the gene is sufficient to cause the cancer.

X-linked recessive: Cancer is caused by a mutation in the gene located on the X-chromosome. In males, a mutation in the only copy of the gene will cause the cancer. While in females, mutations in both copy of the gene are needed to cause the cancer.

Mitochondrial: Cancer caused by mutations in the gene on the mitochondrial genome.

1.2.2. Genes involved in cancer

There are two main types of genes, i.e., proto-oncogene and tumor suppressor gene (TSG), that can play a role in carcinogenesis. A proto-oncogene is generally involved in signaling pathways that promote cell growth and division. Gain-of-function mutations in a proto-oncogene can turn it into a malfunctioning gene, i.e. oncogene, which stimulates cell growth, division and survival (Malebary et al. 2021). A TSG normally helps prevent unrestrained cellular growth and promotes DNA repair and cell cycle checkpoint activation (Lee and Muller 2010). Loss-of-function mutations in TSG result in inactivation of the gene, leading to carcinogenesis. In addition to loss-of-function mutations, epigenetic silencing, proteasomal degradation by ubiquitination, abnormal cellular localization, and transcriptional regulation can also result in inactivation of a TSG (Wang et al. 2018). All these mutations, both gain-of-function and loss-of-function mutations, can be either germline or somatic.

The causal mutation in the genome, not only a germline mutation but also a somatic mutation, can be of different types, including single nucleotide variant (SNV, also known as single nucleotide polymorphism, SNP), small insertion and deletion (InDel), copy number alteration (CNA), and other structural variants (SVs). Among

them, SNV is the most frequent type among all identified germline risk factors (introduced in the next section) and somatic mutations. In general, all these different types of mutations can either have an impact on expression or function. SNVs can occur in non-coding or coding sequences and most disease associated SNVs (approximately 90%) are identified in non-coding regions (Farh et al. 2015). SNVs that occur in the coding region can be silent (synonymous) or can change the encoded protein sequence (non-synonymous). To predict the consequence of non-synonymous SNVs in the coding region, pathogenicity prediction tools can be used, such as SIFT and PolyPhen-2 (Ng and Henikoff 2003; Adzhubei et al. 2010). These can predict if a SNV is likely to have any functional impact, either being deleterious or tolerated. SNVs that occur in the non-coding region can also affect the expression or structure of the product. For instance, SNVs in the promoter and enhancer region can up or down regulate the expression of gene (Rojano et al. 2018). SNVs in a splice site may result in alternative splicing, therefore changing the sequence of protein (Hsiao et al. 2016). An InDel in a coding region can result in a frameshift or non-frameshift alteration, both leading to a changed protein (Lin et al. 2017). A CNA can affect the gene expression level through dosage effects (Shao et al. 2019). SVs can cause various types of dysfunctions, such as deletions or rearrangements truncating genes, amplification of genes resulting in overexpression, gene fusions, and a change of location of gene regulatory elements, leading to changes in gene expression (Mahmoud et al. 2019).

SNV: Single Nucleotide Variant, substitution of one single nucleotide for another (Shastri 2009).

InDel: Insertion-Deletion, additions or deletions of one or more nucleotides in DNA sequence (Lin et al. 2017).

CNA: Copy Number Alteration, structural alterations to chromosomes of >1000 bases in length that can intersect multiple genes (Lee and Scherer 2010). It is a particular subtype of SV that can increase or decrease the copy number of a given region.

SV: Structural Variant, large genomic alterations with a length from ~50 bp to well over megabases of sequence. It has a variety of subclasses, including inversion, interchromosomal and intrachromosomal translocations, and CNAs (Ho et al. 2020).

Coding region: The proportion of a gene's DNA that codes for protein.

1.2.3. Germline genetic risk factor

A germline genetic risk factor is a germline mutation that is associated with an increased risk of specific cancer development. Identification of a germline risk

factor is critical for understanding the molecular mechanism underlying the inherited disease development. Moreover, these germline risk factors may be used in genetic tests to support disease diagnosis and more importantly, they enable early prevention before formation of the cancer based on a positive genetic test. A well well-known example is the mutations in the *BRCA1* or *BRCA2* tumor suppressor genes, which lead to increased risk of breast, ovarian, and prostate cancers (Levy-Lahad and Friedman 2007). A woman can have prophylactic bilateral mastectomies and reconstruction if the genetic test shows that she carries the *BRCA1* or *BRCA2* gene mutation (Padamsee et al. 2017). In most cases, inherited causal mutations only increase the likelihood that the person will develop a certain cancer, but this doesn't mean that the person will absolutely develop the cancer, just like the causal mutation in the *BRCA1* gene. This is recognized as incomplete penetrance of the germline causal variant.

1.2.4. Cancer driver gene

Many somatic mutations (1000 - 2000 in most human tumors) can accumulate in the tumor cells during further cell divisions, but only a few of them, known as “driver mutations”, contribute to cancer initiation and progression. The majority of somatic mutations, referred to as “passenger mutations”, are harmless and have no contribution to tumor growth (Martincorena and Campbell 2015). It is believed that cancer is an evolutionary process (Lipinski et al. 2016). In many aspects, selection in a tumor mirrors species selection. Somatic mutations occur and accumulate in tumor cells as the tumor continues to grow, yielding a heterogeneous tumor which is composed of multiple subclones of tumor cells with combinations of different somatic mutations. Driver mutations are under positive selection in the tumorigenesis process and confer a growth and survival advantage to cells, leading to preferential growth and expansion of a clone (Martincorena and Campbell 2015; Martínez-Jiménez et al. 2020). Identification of driver mutations from somatic mutations is still one of the major challenges in oncogenic research that aims to understand the molecular mechanism of cancer development and develop effective therapies to cure cancer.

1.3. Thyroid gland

Thyroid cancer is a type of cancer that occurs in the thyroid gland, a butterfly shaped endocrine gland lying in the front of the neck in a position just below the Adam's apple in humans. It has two lobes, the right and left lobe, on each side of windpipe, joined by a small bridge called isthmus. It produces thyroid hormones T4 (thyroxine) and T3 (triiodothyronine) and a peptide hormone, calcitonin. Thyroid hormones are responsible for regulating the metabolism in the body and help to regulate heart rate, blood pressure and body temperature, and in children, growth and development (van

der Spek et al. 2017). Calcitonin regulates calcium homeostasis (Felsenfeld and Levine 2015). There are several steps in the thyroid hormone synthesis (Figure 1.2): 1) iodine uptake into thyroid follicular cells by the sodium/iodide symporter (NIS); 2) synthesis of two key proteins, Thyroid Peroxidase (TPO) and thyroglobulin (TG), and secretion of TG into the follicular lumen; 3) iodide transport into the follicular lumen; 4) iodide oxidation to form iodine by TPO; 5) iodination of TG tyrosine residues to generate monoiodotyrosine (MIT) and diiodotyrosine (DIT) by TPO; 6) coupling of iodotyrosines to form thyroxine (T4) and triiodothyronine (T3) by TPO; 7) endocytosis of TG-thyroid hormone complex and T3 and T4 cleaved from it by proteases in the lysosomes (Rousset et al. 2000; Koibuchi 2012). TPO is involved in steps 4, 5 and 6, and it needs H_2O_2 which is generated by dual oxidase (DUOX), a NADPH oxidase, to catalyze those reactions (Rousset et al. 2000). The thyroid hormones synthesis is regulated by thyroid-stimulating hormone (TSH) produced by the pituitary gland, which is in turn regulated by thyrotropin-releasing hormone (TRH) secreted by the hypothalamus (Shahid et al. 2022). Almost each key gene involved in thyroid hormone synthesis has been associated with increased thyroid cancer risk in humans, such as *TG* (Akdi et al. 2011), *DUOX2* (Bann et al. 2019), *TPO* (Cipollini et al. 2013), and *FOXO1* (Landa et al. 2009; Chen and Zhang 2018).

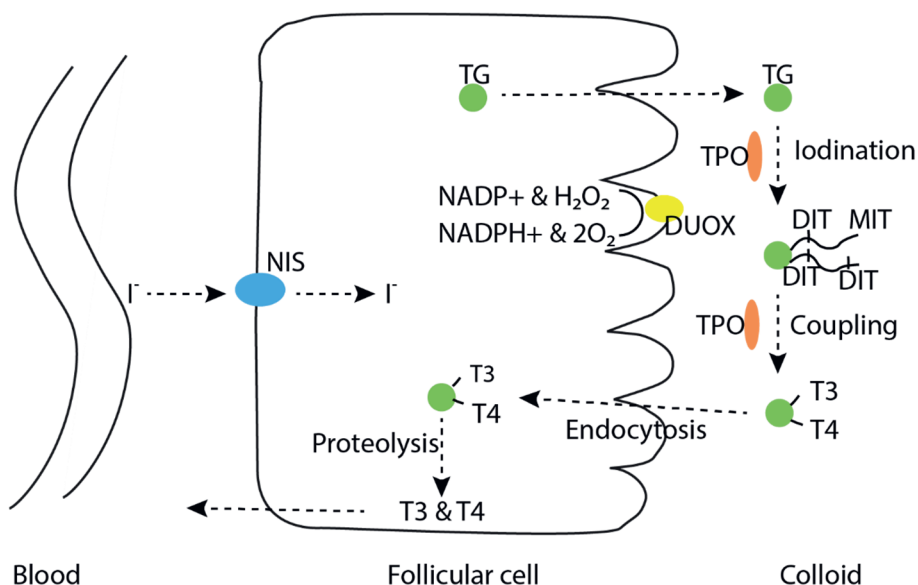


Figure 1.2. Schematic of thyroid hormone synthesis in the thyroid gland. The key players involved in the process include sodium/iodide symporter (NIS), dual oxidase (DUOX), thyroid peroxidase (TPO), and thyroglobulin (TG).

1.3.1. Thyroid cancer in humans

In 2020, thyroid cancer occurrence ranked 11th among 36 most dominant cancers in humans, with 586,202 new cases and 43,646 deaths (Sung et al. 2021). The number of deaths ranks 25th among 36 cancers. Females have a 3-fold higher risk for thyroid cancer than males (Sung et al. 2021). There are 5 main types of thyroid cancer classified according to morphological features in histology by the World Health Organization (WHO) (Klöppel et al. 2017). 1) papillary thyroid carcinoma (PTC, 85-90%), 2) follicular thyroid carcinoma (FTC, 5-10%), 3) poorly differentiated thyroid carcinoma (PDTC, 2%), 4) anaplastic thyroid carcinoma (ATC, 2%), and 5) medullary thyroid cancer (MTC, 2%) (Katoh et al. 2015; Ibrahimasic et al. 2019; Aziz et al. 2021; Vural et al. 2021). A differentiated cancer is the cancer in which the cells are mature and represent many features from the tissue cells it arose from. In contrast, undifferentiated or poorly differentiated cancer cells look and behave very different from normal tissue derived cells. Both PTC and FTC are well-differentiated cancers of follicular cell origin and belong to the differentiated thyroid carcinomas (DTC).

In humans, familial thyroid cancer can originate either from follicular cells, which is called familial non-medullary thyroid carcinoma (FNMTC), or from calcitonin-producing C cells, i.e., familial MTC (FMTC) (Guilmette and Nosé 2018). FNMTC is defined by the presence of at least two first-degree relatives when other known familial syndromes are not present (Capezzone et al. 2021). FNMTC can be either syndromic or non-syndromic (Hińcza et al. 2019). The known syndromic FNMTC includes familial adenomatous polyposis (FAP), Cowden syndrome (CS), Werner syndrome (WS), Carney complex (CNC), McCune-Albright syndrome, Pendred syndrome, ataxia-telangiectasia syndrome, Li-Fraumeni syndrome, DICER 1 syndrome, and Peutz-Jeghers syndrome (Khan et al. 2017; Robertson et al. 2018; Kamilaris et al. 2019; Capezzone et al. 2021).

Only 5% of syndromic form of FNMTC have well-documented germline risk factors (Guilmette and Nosé 2018). In non-syndromic FNMTC, several germline susceptibility genes have been identified including *SRGAP1*, *SRRM2*, *MYO1F*, *FTEN*, *MAP2K5*, *FOXE1*, *DIRC3*, and *CHEK2* (Landa et al. 2009; He et al. 2013; Köhler et al. 2013; Beg et al. 2015; Siołek et al. 2015; Tomsic et al. 2015; Chen and Zhang 2018; Diquigiovanni et al. 2018; Ye et al. 2019). In addition, also microRNAs (miRNA) have been identified as predisposition factor, such as, miR886-3p and miR-20a (Gudmundsson et al. 2017). The driver mutations most frequently identified in human thyroid tumors are the *BRAF*^{V600E} mutation, and mutations in the RAS gene family (*KRAS*, *NRAS*, *HRAS*). Other driver genes identified include DNA repair genes (*CHEK2* and *PPM1D*), PI3K/AKT pathway

genes (*PTEN* and *AKT1/2*), tumor suppressor/checkpoint genes (*TP53*, *RBI*, *NFI* and *MEN1*), the translation-associated factor *EIF1AX*, epigenetic regulators (*MLL*, *MLL3* and *ARID1B*), and the *TERT* gene (promoter region) (Cancer Genome Atlas Research 2014).

1.3.2. Thyroid cancer in dogs

Dogs have a pair of thyroid glands located on each side of the windpipe in the neck (Figure 1.3). Canine thyroid tumors are the most common endocrine neoplasm, accounting for 1% to 4% of all neoplasms in dogs (Hassan et al. 2020). A typical sign of a thyroid tumor is the mass in the neck at the thyroid gland and therefore mostly detected by physical examination. Canine thyroid tumors can be divided in benign (adenoma) or malignant (carcinoma). The carcinomas are mostly represented, approximately between 60% and 90% of all thyroid tumors (Wucherer and Wilke 2010). A canine thyroid tumor can be unilateral (67% - 75%) or bilateral (25% - 35%) (Liptak 2007). The histological growth patterns of canine thyroid carcinomas are largely similar to those in humans. According to classification by the WHO (Kiupel et al. 2008), canine thyroid tumors originating from follicular cells have mainly 7 histological subtypes, 1) follicular thyroid carcinoma (FTC), 2) compact thyroid carcinoma (CTC), 3) follicular-compact thyroid carcinoma (FCTC), 4) papillary thyroid carcinoma (PTC), 5) poorly differentiated thyroid carcinoma, 6) undifferentiated thyroid carcinoma, and 7) carcinosarcoma. All these 7 types of thyroid tumor originate from follicular cells in the thyroid gland and therefore fall under the group of follicular cell carcinoma (FCC). In addition, thyroid tumors can originate from parafollicular cells, called C-cell carcinoma or medullary thyroid carcinoma (MTC). In dogs, FCC is more common than MTC and account for approximately 64% - 71% of all thyroid tumors (Liptak 2007; Campos et al. 2014). Metastasis is relatively common in dogs with malignant thyroid carcinoma, where around one third of all affected dogs have metastases at the time of diagnosis (Hassan et al. 2020). Canine thyroid carcinoma, similar to human thyroid carcinoma, commonly metastasizes to the lung and regional lymph nodes (Hassan et al. 2020).

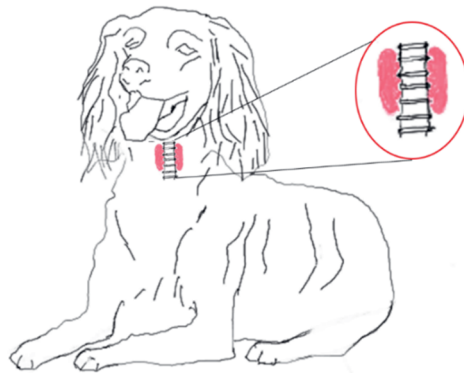


Figure 1.3. Sketch of a German Longhaired Pointer and a pair of thyroid glands in its neck.

Golden Retrievers, Beagles and Siberian Huskies were found to be predisposed to thyroid tumors more often than other dog breeds (Wucherer and Wilke 2010). There is no detected sex predisposition for canine thyroid tumors. The risk for thyroid tumors increases with age, where the mean age at diagnosis is between 9 and 10 years whereas dogs between 10 and 15 years of age have a significantly higher risk of developing thyroid tumors than younger dogs (Wucherer and Wilke 2010; Hassan et al. 2020).

The report on canine familial thyroid tumors is limited. Besides the familial FCC in the Dutch GLPs in our study, only a canine familial medullary thyroid carcinoma (MTC) was reported (Lee et al. 2006). The genetics of familial canine thyroid cancer are poorly investigated yet. The genetic basis of the canine familial MTC is poorly studied, thus not yet clear. Likewise, the somatic mutation landscape of canine thyroid cancers is unclear.

1.4. Cancer cause research in the genomic era

Traditional methods to investigate cancer include histopathological and cytogenetic studies (Gurcan et al. 2009; Wan 2014). The former investigates the changes in tissues/cells. Cytogenetics studies the structure of DNA within the cell nucleus, including the number and morphology of chromosomes. These two methods shed very little light on the genetic origin of the tumor and the molecular mechanism of the tumorigenesis. Investigating the disease at the genomics level will apprehend the essence of the cancer better. Coming into the era of high-throughput sequencing, this becomes possible.

1.4.1. Germline causal mutation mapping

Generating high-throughput molecular data is becoming cheaper and easier to perform and facilitates the cancer causal mutation identification. Omics data have been used in the identification of cancer causal mutations, including both the germline and somatic mutations. SNP array genotyping, whole-genome sequencing (WGS), and whole-exome sequencing (WES) are technologies to identify different types of genetic variants across the genome. Following the genome-wide germline variants identification using either SNP arrays, WGS or WES, a genome wide association study (GWAS) can be performed to identify variants that are associated with the disease. A GWAS contrasts the affected individuals and unaffected individuals to identify the mutations that present a significantly higher frequency in affected samples than in the healthy population (Cano-Gamez and Trynka 2020). To achieve sufficient power of the GWAS, a rather large number of individuals (samples) are usually needed. Also, a careful selection of these individuals is important to avoid population stratification. WGS is an option but mostly too expensive for a GWAS. Therefore, a good alternative for performing GWAS are SNP arrays. A SNP array is a type of DNA microarray containing designed probes harboring the SNP positions, which are hybridized with DNA to determine the specific alleles of every SNP on the array. Because the SNP array only genotypes a selected number of SNPs on the genome, the causal variant related to the trait of interest is most likely missing. However, neighboring variants are often correlated with one another, as they tend to be inherited together due to co-segregation during meiotic recombination, which is called linkage disequilibrium (LD) (Cano-Gamez and Trynka 2020). Thanks to this LD in the genome, the association between causal variant and disease can be captured by so-called tag SNPs, the variants in close LD with the causal variant. We therefore can identify a target region associated with the trait by a GWAS based on SNP array data. This target region is usually surrounding the causal variant.

For a hereditary disease with a recessive inheritance pattern, homozygosity mapping (also called autozygosity mapping) is another commonly used and powerful method to identify the target region that links to the disease, especially in consanguineous families or isolated populations (Alkuraya 2010). For rare autosomal recessive diseases, the chromosomal segments surrounding the causal variant (haplotype) is rendered homozygous in the patients because they were received from a pair of parents who in turn inherited it from a common ancestor and recombination by crossing over has not taken place at this position (Gershlick et al. 2016). Homozygosity mapping identifies and scores the presence of consecutive homozygous genotypes, also called runs of homozygosity (ROH), in each individual's genome (Quinodoz et al. 2021). ROH can be detected using SNP

genotypes derived from SNP arrays or WGS data. Beside the application in identification of target regions harboring causal variants for recessive diseases, the number and length of ROH reflect individual demographic history and inbreeding (Ceballos et al. 2018). Different methods can be used together, and derived results can be integrated to exclude possible false positives, such as noise resulting from specific population structure in the GWAS.

GWAS and homozygosity mapping identify a target region. Subsequently, to identify the causal variant in the target region, which is called “fine-mapping”, we can take use of WGS data of a small number of individuals or sequence the candidate region (Pollott 2018). By these sequencing methods, many variants of different types can be identified in the target region. Due to LD, often multiple variants in the region are associated with the disease and the variant with the strongest association may not be the causal variant (Hutchinson et al. 2020). Currently, most studies only investigate the causality of variants in coding regions. However, over 90% of GWAS variants fall in non-coding regions (e.g. intronic region and intergenic region) of the genome and can also be causal by disrupting the proper regulation of the expression of genes e.g. by disrupting the binding sites for transcription factors (Cano-Gamez and Trynka 2020). Identifying such causal variants is more challenging. To fine-map the causal variants in non-coding regions, one approach is to integrate GWAS or homozygosity mapping results with functional genomic datasets, such as transcriptome and chromatin annotations (e.g. open chromatin regions, histone modifications, DNA methylation) (Cano-Gamez and Trynka 2020). To accurately map causal variants in coding regions, other characteristics of mutations must be taken into account together, such as, mutation frequency in the study population and a general population that is not related to the study population, pathogenicity prediction, and the function of the gene (MacArthur et al. 2014). Meanwhile, segregation of the variant must fit the inheritance pattern of the disease. While it is noteworthy that causal variants are not necessarily fully penetrant, even for monogenic diseases (MacArthur et al. 2014), because biases associated with sample ascertainment often exist in studies. The incomplete penetrance of the causal variants makes the identification of actual causal variants more difficult because it makes the association between appearance of causal variants and the cancer imperfect. Finally, further experimental validation is necessary to confirm the causality of the candidate variant by investigating the impact of a variant on gene function, or cell or organism phenotype, if it has not been performed (MacArthur et al. 2014).

1.4.2. Recurrent somatic mutation and driver mutation prediction

To identify somatic mutations in the tumor at a whole-genome scale, we have to obtain the WGS of the tumor and matched normal tissue. Whole-exome sequencing can also be used at a lower cost but does not allow the identification of somatic mutations in non-coding regions or detection of somatic SVs. The matched normal tissue could be the normal tissue adjacent to the tumor, or other healthy tissues available from the same individual (mostly blood sample). Sequencing depth can influence sensitivity and specificity of both germline and somatic mutation variant detection (Brastianos et al. 2013; Wilmott et al. 2015). Due to the potential heterogeneity (subclones) within the tumor and potential contamination from the normal cell, sequencing at a high depth is needed to be able to reliably identify somatic mutations. Tumor heterogeneity means that tumor cells might differ from each other within the tumor in morphology, phenotype, and molecular signature. To be able to accurately, and confidently, identify somatic mutations, a minimum depth of coverage of 30x for the normal tissue and 60x for the tumor tissue is recommended (Brastianos et al. 2013; Wilmott et al. 2015). A few large-scale human sequencing projects have sequenced tumor and matched normal tissues of more than 20,000 individuals spanning 33 different types of cancer (Cancer Genome Atlas Research 2008; International Cancer Genome et al. 2010). These projects yielded many driver mutations and further our understanding about the genetic basis of tumorigenesis.

Following the identification of somatic mutations in tumor tissues, the next step is to identify driver events that are responsible for tumor initiation and development. Many tools have been developed to identify driver events at a gene or mutation level from somatic mutations identified in the tumor. These prediction tools usually use the typical characteristics of a driver gene/mutation to predict driver events. The most important characteristic is the recurrence of driver events. Recurrence can be at mutation level (the identical mutation), gene level (somatic mutations in the same gene), or even at the pathway level (somatic mutations in multiple genes involved in the same signaling pathway). Each tumor has undergone an independent evolutionary process, but they converged to a common cancerous phenotype. The driver event provides a selective advantage that results in clonal expansion of its lineage. Driver mutations are expected to appear more frequently across tumors (Raphael et al. 2014). Ultimately, to determine the role of predicted driver events in tumorigenesis requires experimental validations, using *in vitro* or *in vivo* models to demonstrate that the potential driver event leads to at least one of the characteristics of the tumor, such as increased proliferation, or deficiency in DNA damage repair (Raphael et al. 2014).

Before our study, the somatic mutation profile of thyroid tumors in dogs at a whole-genome scale had not been investigated. Several other types of canine sporadic tumors, such as canine bladder cancer, lymphomas, osteosarcomas, gliomas and melanomas, have been studied and potential driver genes were identified (these are introduced in chapter 5) (Decker et al. 2015; Elvers et al. 2015; Sakthikumar et al. 2018; Wong et al. 2019; Amin et al. 2020; Alsaihati et al. 2021). These studies revealed the same driver genes between canine and human cancers.

1.4.3. Mutational signatures

A mutational signature is a characteristic combination of somatic mutation types arising from a specific mutational process (Alexandrov et al. 2013). Deciphering mutational signatures in a tumor can provide insights into the mutagens that result in tumor initiation and development. Mutational signatures can be constructed using somatic single base substitutions (SBSs), or InDels, SVs, and CNAs. SBS mutational signature is the most well investigated. Current SBS mutational signatures have been identified using 96 different trinucleotide changes, considering the mutated base, plus the flanking bases immediately 5' and 3' (Van Hoeck et al. 2019). There are 60 SBS mutational signatures that are categorized in the COSMIC database (<https://cancer.sanger.ac.uk/signatures/sbs/>) to date, of which some have been linked to specific mutational processes (or mutagens), while others have unclear etiology. To identify SBS mutational signatures from somatic mutation profiles, it is possible to either extract novel mutational signatures using Non-negative Matrix Factorization (NMF), which is a type of dimensionality reduction, or to fit previous defined mutational signatures using a non-negative least-squares optimization approach (Manders et al. 2022).

1.5. Disease susceptibility in breed dogs

Dog is the first animal to be domesticated by humans. It was domesticated from grey wolf around 15,000 years ago (Larson et al. 2012; Frantz et al. 2020). In the past ~200 years, approximately since the Victorian era, intensive human selection was applied on dogs and more than 400 different pure dog breeds with specific characteristics, recognized by different kennel clubs worldwide, were created (Shannon et al. 2015). Each dog breed has undergone strong artificial selection to fix certain desirable attributes. Therefore, pedigree dogs underwent at least two bottleneck events in history, which are domestication and breed formation. These bottleneck events, together with subsequent continuous human selection, have resulted in limited genetic diversity (extensive inbreeding), extensive LD across the genome (Lindblad-Toh et al. 2005), and strong phenotypic homogeneity within each breed. The breed formation process over just the past 2-3 centuries has resulted in seven-fold reduction in genetic diversity compared to the dog domestication

thousands of years ago (Schiffman and Breen 2015). Moreover, extensive use of popular sires also lowers the genetic diversity within each pure breed (Lewis et al. 2015). Each breed has a unique breed standard that guides the breeding practice to keep the key phenotypes that are characteristic to that breed. Intense and diverse human selection created a big difference in the variety of phenotypes between breeds, such as height, body size, ear shape, coat color, fur length, leg height, skull shape, and moustache. Probably the most impressive example is the body size where the shortest dog is a Chihuahua (9.6 cm in height) and the tallest dog is a Great Dane (111.8 cm in height) according to Guinness World Records (Millward 2019). Likewise, intense human selection exaggerates many adorable traits and makes dogs perform excellently in specific fields, such as hunting, life-searching, and herding. From the genetic perspective, these abundant phenotypes make dogs ideal subjects to investigate the genetic architecture of different traits. In addition, a limited genetic diversity within the breed also makes genetic mapping more powerful to identify association between genotype and phenotype by linkage or association studies. Because of extensive LD in pure breed dogs, the number of SNPs needed for a GWAS in dogs is much smaller than that needed for a comparable human study (Davis and Ostrander 2014).

Besides those diverse phenotypes, intense human selection, not surprisingly, also results in higher risks for certain diseases and disorders, including cancers, in purebred dogs compared to mixed breed dogs, due to increased inbreeding. There are numerous examples. For instance, Rottweiler, Rhodesian Ridgeback and Great Dane are predisposed to osteosarcoma (Edmunds et al. 2021). Labrador Retriever, Rottweiler, German Shepherd, and Staffordshire Bull Terrier are predisposed to elbow dysplasia. Chihuahua is predisposed to medial patellar luxation and Rottweiler are predisposed to cranial cruciate ligament disease (Boge et al. 2019). Flatcoated Retriever is predisposed to histiocytic sarcomas (Boerkamp et al. 2013). Weimaraner and Irish Setter are predisposed to hypertrophic osteodystrophy. In a study of Farrell et al. (2015), 396 inherited disorders were identified in 215 dog breeds officially recognized by the Kennel Club in the UK.

1.6. Inbreeding contributes to inherited disease

Inbreeding is the production of offspring from the mating or breeding of individuals or organisms that are closely related genetically. An inbred individual has lower genome-wide heterozygosity because parts of its genome is identical-by-descent (IBD). Pure breed dogs usually have extensive inbreeding (Lewis et al. 2015; Wijnrocx et al. 2016) which can cause inbreeding depression (Ujvari et al. 2018). In livestock industry, inbreeding decreases the production of products and the performance of animals. Meanwhile, it increases the incidence of inherited

diseases/disorders in the population, especially those with a recessive inheritance pattern. Individuals that have a closer relationship have a higher chance carrying a copy of a recessive allele coming from a common ancestor. Inbreeding can increase the chance of such a combination of two IBD alleles. Therefore, inbred animals have a higher chance to carry two IBD recessive alleles.

The inbreeding coefficient of an individual is the probability that two alleles at a locus in that individual are IBD (Gomez-Raya et al. 2015). The inbreeding coefficient can be estimated based on pedigree using path coefficients (Fped) (Wright 1922), or genomic data, such as SNP array data and WGS. The accuracy to estimate inbreeding based on pedigree is limited by completeness of pedigree and unintentional errors in the pedigree. Moreover, pedigree based inbreeding estimation is not able to consider the stochastic recombination events that occurred during meiosis. Genomic based inbreeding is usually more accurate than pedigree based inbreeding when a sufficient number of markers (e.g. thousands) is used in the estimation (Kardos et al. 2015). Moreover, in some cases, a pedigree is not available, such as for an animal in the wild. In this case, genomic measure of inbreeding can still be used. There are a few different approaches to estimate inbreeding using genomic data, such as ROH based F_{ROH} , and excess homozygosity based F_{HOM} (Kardos et al. 2015).

F_{HOM} : The increase in genome-wide homozygosity of an individual relative to Hardy–Weinberg expected homozygosity.

F_{ROH} : Runs of homozygosity based inbreeding coefficient. It's the proportion of autosomal genome covered by runs of homozygosity.

Fped: Inbreeding coefficient estimated using path-counting algorithm based on pedigree. It is a function of the number and location of the common ancestors of both parents of an individual in the given pedigree.

1.7. Aim and outline of this thesis

The major goal of the studies presented in this thesis is to unravel the genomics underlying the familial thyroid cancer identified in the Dutch GLPs. In **Chapter 2**, we described the clinical and histopathological findings regarding the thyroid tumors identified in the Dutch GLPs. In addition, with the availability of the pedigree, I determined the inheritance pattern of the thyroid cancer in the Dutch GLPs and investigated the contribution of inbreeding to the thyroid cancer in these dogs. In **Chapter 3**, I identified the germline risk factors using a combination of GWAS and homozygosity mapping based on SNP array and WGS data. In **Chapter 4**, I investigated the somatic mutation landscape of these FCCs and identified a promising driver mutation. Herein, both the germline risk mutation and somatic

driver mutation were investigated. In **Chapter 5**, I investigated the genetic diversity of Dutch GLPs and genetic relationship between GLP and several other pointer setter breeds. Meanwhile, I also identified genomic selection signatures in these Dutch GLPs and revealed one selected region that might contribute to good athletic performance. In **Chapter 6**, with a question how the abundant molecular signaling pathway knowledge can be used in driver mutation prediction, I proposed a cancer gene score and showed that this cancer gene score can improve driver mutation prediction. Lastly, in **Chapter 7**, I summarize the major findings in previous chapters and discussed them in the context of current literature. Furthermore, recommendations for further studies are suggested and discussed in a broad context.

1.8. Reference

- Acuna-Hidalgo R, Veltman JA, Hoischen A. 2016. New insights into the generation and role of de novo mutations in health and disease. *Genome Biology* **17**: 241.
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7**: 248-249.
- Akdi A, Pérez G, Pastor S, Castell J, Biarnés J, Marcos R, Velázquez A. 2011. Common variants of the thyroglobulin gene are associated with differentiated thyroid cancer risk. *Thyroid* **21**: 519-525.
- Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**: 415-421.
- AlHarthi FS, Qari A, Edress A, Abedalthagafi M. 2020. Familial/inherited cancer syndrome: a focus on the highly consanguineous Arab population. *npj Genomic Medicine* **5**: 3.
- Alkuraya FS. 2010. Homozygosity mapping: One more tool in the clinical geneticist's toolbox. *Genetics in Medicine* **12**: 236-239.
- Alsaihati BA, Ho K-L, Watson J, Feng Y, Wang T, Dobbin KK, Zhao S. 2021. Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. *Nature Communications* **12**: 4670.
- Amin SB, Anderson KJ, Boudreau CE, Martinez-Ledesma E, Kocakavuk E, Johnson KC, Barthel FP, Varn FS, Kassab C, Ling X et al. 2020. Comparative Molecular Life History of Spontaneous Canine and Human Gliomas. *Cancer Cell* **37**: 243-257.e247.
- Anderson DE. 1992. Familial versus sporadic breast cancer. *Cancer* **70**: 1740-1746.
- Aziz A, Masood MQ, Sattar S, Fatima S, Islam N. 2021. Follicular Thyroid Carcinoma in a Developing Country: A 10-Year Retrospective Study. *Cureus* **13**: e16594-e16594.
- Bann DV, Jin Q, Sheldon KE, Houser KR, Nguyen L, Warrick JI, Baker MJ, Broach JR, Gerhard GS, Goldenberg D. 2019. Genetic Variants Implicate Dual Oxidase-2 in Familial and Sporadic Nonmedullary Thyroid Cancer. *Cancer Research* **79**: 5490-5499.
- Barwick BG, Gupta VA, Vertino PM, Boise LH. 2019. Cell of Origin and Genetic Alterations in the Pathogenesis of Multiple Myeloma. *Front Immunol* **10**: 1121-1121.
- Beg S, Siraj AK, Jehan Z, Prabakaran S, Al-Sobhi SS, Al-Dawish M, Al-Dayel F, Al-Kuraya KS. 2015. PTEN loss is associated with follicular variant of Middle Eastern papillary thyroid carcinoma. *British Journal of Cancer* **112**: 1938-1943.
- Boerkamp KM, van der Kooij M, van Steenbeek FG, van Wolferen ME, Groot Koerkamp MJA, van Leenen D, Grinwis GCM, Penning LC, Wiemer EAC, Rutteman GR. 2013. Gene Expression Profiling of Histiocytic Sarcomas in a Canine Model: The Predisposed Flatcoated Retriever Dog. *PLOS ONE* **8**: e71094.

- Boge GS, Moldal ER, Dimopoulou M, Skjerve E, Bergström A. 2019. Breed susceptibility for common surgically treated orthopaedic diseases in 12 dog breeds. *Acta Veterinaria Scandinavica* **61**: 19.
- Brastianos PK, Horowitz PM, Santagata S, Jones RT, McKenna A, Getz G, Ligon KL, Palescandolo E, Van Hummelen P, Ducar MD et al. 2013. Genomic sequencing of meningiomas identifies oncogenic SMO and AKT1 mutations. *Nature Genetics* **45**: 285-289.
- Campos M, Ducatelle R, Rutteman G, Kooistra HS, Duchateau L, de Rooster H, Peremans K, Daminet S. 2014. Clinical, pathologic, and immunohistochemical prognostic factors in dogs with thyroid carcinoma. *Journal of veterinary internal medicine* **28**: 1805-1813.
- Cancer Genome Atlas Research N. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**: 1061-1068.
- Cancer Genome Atlas Research N. 2014. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**: 676-690.
- Cano-Gamez E, Trynka G. 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11**.
- Capezzone M, Robenshtok E, Cantara S, Castagna MG. 2021. Familial non-medullary thyroid cancer: a critical review. *Journal of Endocrinological Investigation* **44**: 943-950.
- Carbone A. 2020. Cancer Classification at the Crossroads. *Cancers* **12**: 980.
- Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics* **19**: 220-234.
- Chen Y-H, Zhang Y-Q. 2018. Exploration of the association between FOXE1 gene polymorphism and differentiated thyroid cancer: a meta-analysis. *BMC Medical Genetics* **19**: 83.
- Chopra M, Bohlander SK. 2019. The cell of origin and the leukemia stem cell in acute myeloid leukemia. *Genes, Chromosomes and Cancer* **58**: 850-858.
- Cipollini M, Pastor S, Gemignani F, Castell J, Garritano S, Bonotti A, Biarnés J, Figlioli G, Romei C, Marcos R et al. 2013. TPO genetic variants and risk of differentiated thyroid carcinoma in two European populations. *International Journal of Cancer* **133**: 2843-2851.
- Crouch DJM, Bodmer WF. 2020. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences* **117**: 18924-18933.
- Davis BW, Ostrander EA. 2014. Domestic dogs and cancer research: a breed-based genomics approach. *ILAR J* **55**: 59-68.
- Debniak T, Lubinski J. 2008. Principles of genetic predisposition to malignancies. *Hered Cancer Clin Pract* **6**: 69-72.
- Decker B, Parker HG, Dhawan D, Kwon EM, Karlins E, Davis BW, Ramos-Vara JA, Bonney PL, McNiel EA, Knapp DW et al. 2015. Homologous Mutation to Human BRAF V600E Is Common in Naturally Occurring Canine Bladder Cancer--Evidence for a Relevant Model System and Urine-Based Diagnostic Test. *Mol Cancer Res* **13**: 993-1002.
- Diquigiovanni C, Bergamini C, Evangelisti C, Isidori F, Vettori A, Tiso N, Argenton F, Costanzini A, Iommarini L, Anbunathan H et al. 2018. Mutant MYO1F alters the mitochondrial network and induces tumor proliferation in thyroid cancer. *International Journal of Cancer* **143**: 1706-1719.
- Edmunds GL, Smalley MJ, Beck S, Errington RJ, Gould S, Winter H, Brodbelt DC, O'Neill DG. 2021. Dog breeds and body conformations with predisposition to osteosarcoma in the UK: a case-control study. *Canine Medicine and Genetics* **8**: 2.
- Ekhtiari S, Chiba K, Popovic S, Crowther R, Wohl G, Kin On Wong A, Tanke DH, Dufault DM, Geen OD, Parasu N et al. 2020. First case of osteosarcoma in a dinosaur: a multimodal diagnosis. *The Lancet Oncology* **21**: 1021-1022.
- Elvers I, Turner-Maier J, Swofford R, Koltookian M, Johnson J, Stewart C, Zhang CZ, Schumacher SE, Beroukhir R, Rosenberg M et al. 2015. Exome sequencing of lymphomas from three dog

- breeds reveals somatic mutation patterns reflecting genetic background. *Genome Research* **25**: 1634-1645.
- Esfahani K, Roudaia L, Buhlaiga N, Del Rincon SV, Papneja N, Miller WH. 2020. A Review of Cancer Immunotherapy: From the Past, to the Present, to the Future. *Current Oncology* **27**: 87-97.
- Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shores N, Whittom H, Ryan RJH, Shishkin AA et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**: 337-343.
- Farrell LL, Schoenebeck JJ, Wiener P, Clements DN, Summers KM. 2015. The challenges of pedigree dog health: approaches to combating inherited disease. *Canine Genetics and Epidemiology* **2**: 3.
- Felsburg PJ. 2002. Overview of immune system development in the dog: comparison with humans. *Human & Experimental Toxicology* **21**: 487-492.
- Felsenfeld AJ, Levine BS. 2015. Calcitonin, the forgotten hormone: does it deserve to be forgotten? *Clin Kidney J* **8**: 180-187.
- Frantz LAF, Bradley DG, Larson G, Orlando L. 2020. Animal domestication in the era of ancient genomics. *Nature Reviews Genetics* **21**: 449-460.
- Gardner HL, Fenger JM, London CA. 2016. Dogs as a Model for Cancer. *Annual review of animal biosciences* **4**: 199-222.
- Gershlick DC, Schindler C, Chen Y, Bonifacio JS. 2016. TSSC1 is novel component of the endosomal retrieval machinery. *Mol Biol Cell* **27**: 2867-2878.
- Gomez-Raya L, Rodríguez C, Barragán C, Silió L. 2015. Genomic inbreeding coefficients based on the distribution of the length of runs of homozygosity in a closed line of Iberian pigs. *Genet Sel Evol* **47**: 81-81.
- Gudmundsson J, Thorleifsson G, Sigurdsson JK, Stefansdottir L, Jonasson JG, Gudjonsson SA, Gudbjartsson DF, Masson G, Johannsdottir H, Halldorsson GH et al. 2017. A genome-wide association study yields five novel thyroid cancer risk loci. *Nature Communications* **8**: 14517.
- Guilmette J, Nosé V. 2018. Hereditary and familial thyroid tumours. *Histopathology* **72**: 70-81.
- Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. 2009. Histopathological image analysis: a review. *IEEE Rev Biomed Eng* **2**: 147-171.
- Hanahan D, Weinberg RA. 2000. The Hallmarks of Cancer. *Cell* **100**: 57-70.
- Hanahan D, Weinberg Robert A. 2011. Hallmarks of Cancer: The Next Generation. *Cell* **144**: 646-674.
- Hassan BB, Altstadt LA, Dirksen WP, Elshafae SM, Rosol TJ. 2020. Canine Thyroid Cancer: Molecular Characterization and Cell Line Growth in Nude Mice. *Veterinary Pathology* **57**: 227-240.
- He H, Bronisz A, Liyanarachchi S, Nagy R, Li W, Huang Y, Akagi K, Saji M, Kula D, Wojcicka A et al. 2013. SRGAP1 Is a Candidate Gene for Papillary Thyroid Carcinoma Susceptibility. *The Journal of Clinical Endocrinology & Metabolism* **98**: E973-E980.
- Hińcza K, Kowalik A, Kowalska A. 2019. Current Knowledge of Germline Genetic Risk Factors for the Development of Non-Medullary Thyroid Cancer. *Genes* **10**: 482.
- Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nature Reviews Genetics* **21**: 171-189.
- Hsiao Y-HE, Bahn JH, Lin X, Chan T-M, Wang R, Xiao X. 2016. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome research* **26**: 440-450.
- Hutchinson A, Asimit J, Wallace C. 2020. Fine-mapping genetic associations. *Human Molecular Genetics* **29**: R81-R88.
- Ibrahimasic T, Ghossein R, Shah JP, Ganly I. 2019. Poorly Differentiated Carcinoma of the Thyroid Gland: Current Status and Future Prospects. *Thyroid : official journal of the American Thyroid Association* **29**: 311-321.

- International Cancer Genome C Hudson TJ Anderson W Artez A Barker AD Bell C Bernabé RR Bhan MK Calvo F Eerola I et al. 2010. International network of cancer genome projects. *Nature* **464**: 993-998.
- Kamilaris CDC, Faucz FR, Voutetakis A, Stratakis CA. 2019. Carney Complex. *Exp Clin Endocrinol Diabetes* **127**: 156-164.
- Kardos M, Luikart G, Allendorf FW. 2015. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity* **115**: 63-72.
- Katoh H, Yamashita K, Enomoto T, Watanabe M. 2015. Classification and general considerations of thyroid cancer. *Ann Clin Pathol* **3**: 1045.
- Khan NE, Bauer AJ, Schultz KAP, Doros L, Decastro RM, Ling A, Lodish MB, Harney LA, Kase RG, Carr AG. 2017. Quantification of thyroid cancer and multinodular goiter risk in the DICER1 syndrome: a family-based cohort study. *The Journal of Clinical Endocrinology & Metabolism* **102**: 1614-1622.
- Kiupel M, Capen C, Miller M, Smedley R. 2008. *Histological classification of tumors of the endocrine system of domestic animals*. Armed Forces Inst. of Pathology.
- Klöppl G, Couvelard A, Hruban R, Klimstra D, Komminoth P, Osamura R, Perren A, Rindi G. 2017. WHO classification of tumours of endocrine organs. *Lyon, France: World Health Organization*.
- Knowledge Gf. 2016. Man's best friend: global pet ownership and feeding trends. Vol 2022.
- Köhler A, Chen B, Gemignani F, Elisei R, Romei C, Figlioli G, Cipollini M, Cristaudo A, Bambi F, Hoffmann P et al. 2013. Genome-Wide Association Study on Differentiated Thyroid Cancer. *The Journal of Clinical Endocrinology & Metabolism* **98**: E1674-E1681.
- Koibuchi N. 2012. [Molecular mechanisms of thyroid hormone synthesis and secretion]. *Nihon Rinsho* **70**: 1844-1848.
- Landa I, Ruiz-Llorente S, Montero-Conde C, Inglada-Pérez L, Schiavi F, Leskelä S, Pita G, Milne R, Maravall J, Ramos I et al. 2009. The Variant rs1867277 in FOXE1 Gene Confers Thyroid Cancer Susceptibility through the Recruitment of USF1/USF2 Transcription Factors. *PLOS Genetics* **5**: e1000637.
- Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, Stahl PW, Piper PJ, Lingaas F, Fredholm M et al. 2012. Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proceedings of the National Academy of Sciences* **109**: 8878-8883.
- Lee C, Scherer SW. 2010. The clinical context of copy number variation in the human genome. *Expert Reviews in Molecular Medicine* **12**: e8.
- Lee EYHP, Muller WJ. 2010. Oncogenes and tumor suppressor genes. *Cold Spring Harb Perspect Biol* **2**: a003236-a003236.
- Lee J-J, Larsson C, Lui W-O, Höög A, Von Euler H. 2006. A dog pedigree with familial medullary thyroid cancer. *International journal of oncology* **29**: 1173-1182.
- Levy-Lahad E, Friedman E. 2007. Cancer risks among BRCA1 and BRCA2 mutation carriers. *Br J Cancer* **96**: 11-15.
- Lewis TW, Abhayaratne BM, Blott SC. 2015. Trends in genetic diversity for all Kennel Club registered pedigree dog breeds. *Canine Genetics and Epidemiology* **2**: 13.
- Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo J-T. 2017. Effects of short indels on protein structure and function in human genomes. *Scientific reports* **7**: 9313-9313.
- Linares-Clemente P, Aguilar-Morante D, Rodríguez-Prieto I, Ramírez G, de Torres C, Santamaría V, Pascual-Vaca D, Colmenero-Repiso A, Vega FM, Mora J et al. 2017. Neural crest derived progenitor cells contribute to tumor stroma and aggressiveness in stage 4/M neuroblastoma. *Oncotarget* **8**: 89775-89792.
- Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803-819.

- Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. 2016. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer* **2**: 49-63.
- Liptak JM. 2007. Canine Thyroid Carcinoma. *Clinical Techniques in Small Animal Practice* **22**: 75-81.
- MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**: 469-476.
- Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biology* **20**: 246.
- Mak IW, Evaniew N, Ghert M. 2014. Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res* **6**: 114-118.
- Malebary SJ, Khan R, Khan YD. 2021. ProtoPred: Advancing Oncological Research Through Identification of Proto-Oncogene Proteins. *IEEE Access* **9**: 68788-68797.
- Manders F, Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, van der Roest B, van Hoeck A, Cuppen E, van Boxtel R. 2022. MutationalPatterns: the one stop shop for the analysis of mutational processes. *BMC Genomics* **23**: 134.
- Martincorena I, Campbell PJ. 2015. Somatic mutation in cancer and normal cells. *Science* **349**: 1483-1489.
- Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H et al. 2020. A compendium of mutational cancer driver genes. *Nature Reviews Cancer* **20**: 555-572.
- Miles B, Tadi P. 2022. Genetics, Somatic Mutation. In *StatPearls*. StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC., Treasure Island (FL).
- Millward A. 2019. A history of the biggest and smallest dog breeds – from giant Great Danes to tiny Chihuahuas. Vol 2022.
- Munday JS, Tucker RS, Kiupel M, Harvey CJ. 2015. Multiple oral carcinomas associated with a novel papillomavirus in a dog. *Journal of Veterinary Diagnostic Investigation* **27**: 221-225.
- NCI. 2021. What Is Cancer? , Vol 2022.
- Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**: 3812-3814.
- Padamsee TJ, Wills CE, Yee LD, Paskett ED. 2017. Decision making for breast cancer prevention among women at elevated risk. *Breast Cancer Research* **19**: 34.
- Pimentel PAB, Oliveira CSF, Horta RS. 2021. Epidemiological study of canine transmissible venereal tumor (CTVT) in Brazil, 2000–2020. *Preventive Veterinary Medicine* **197**: 105526.
- Pollott GE. 2018. Invited review: Bioinformatic methods to discover the likely causal variant of a new autosomal recessive genetic condition using genome-wide data. *Animal* **12**: 2221-2234.
- Quaresma M, Coleman MP, Rachet B. 2015. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *The Lancet* **385**: 1206-1218.
- Quinodoz M, Peter VG, Bedoni N, Royer Bertrand B, Cisarova K, Salmaninejad A, Sepahi N, Rodrigues R, Piran M, Mojarrad M et al. 2021. AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. *Nature Communications* **12**: 518.
- Rahner N, Steinke V. 2008. Hereditary cancer syndromes. *Dtsch Arztebl Int* **105**: 706-714.
- Raphael BJ, Dobson JR, Oesper L, Vandin F. 2014. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine* **6**: 5.
- Robertson JC, Jorcyk CL, Oxford JT. 2018. DICER1 Syndrome: DICER1 Mutations in Rare Cancers. *Cancers* **10**: 143.
- Rojano E, Seoane P, Ranea JAG, Perkins JR. 2018. Regulatory variants: from detection to predicting impact. *Briefings in Bioinformatics* **20**: 1639-1654.

- Rousset B, Dupuy C, Miot F, Dumont J. 2000. Chapter 2 Thyroid Hormone Synthesis And Secretion. In *Endotext*, (ed. KR Feingold, et al.). MDText.com, Inc. Copyright © 2000-2022, MDText.com, Inc., South Dartmouth (MA).
- Sakthikumar S, Elvers I, Kim J, Arendt ML, Thomas R, Turner-Maier J, Swofford R, Johnson J, Schumacher SE, Alföldi J et al. 2018. SETD2 Is Recurrently Mutated in Whole-Exome Sequenced Canine Osteosarcoma. *Cancer Res* **78**: 3421-3431.
- Schapira AHV. 2006. Mitochondrial disease. *The Lancet* **368**: 70-82.
- Schiffman JD, Breen M. 2015. Comparative oncology: what dogs and other species can teach us about humans with cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**: 20140231.
- Shahid MA, Ashraf MA, Sharma S. 2022. Physiology, Thyroid Hormone. In *StatPearls*. StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC., Treasure Island (FL).
- Shannon LM, Boyko RH, Castelhana M, Corey E, Hayward JJ, McLean C, White ME, Said MA, Anita BA, Bondjengo NI et al. 2015. Genetic structure in village dogs reveals a Central Asian domestication origin. *Proceedings of the National Academy of Sciences* **112**: 13639-13644.
- Shao X, Lv N, Liao J, Long J, Xue R, Ai N, Xu D, Fan X. 2019. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genetics* **20**: 175.
- Shastri BS. 2009. SNPs: Impact on Gene Function and Phenotype. In *Single Nucleotide Polymorphisms: Methods and Protocols*, doi:10.1007/978-1-60327-411-1_1 (ed. AA Komar), pp. 3-22. Humana Press, Totowa, NJ.
- Siołek M, Cybulski C, Gąsior-Perczak D, Kowalik A, Kozak-Klonowska B, Kowalska A, Chłopek M, Kluźniak W, Wokołorczyk D, Pałyga I et al. 2015. CHEK2 mutations and the risk of papillary thyroid cancer. *International Journal of Cancer* **137**: 548-552.
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71**: 209-249.
- Tomsic J, He H, Akagi K, Liyanarachchi S, Pan Q, Bertani B, Nagy R, Symer DE, Blencowe BJ, Chapelle Adl. 2015. A germline mutation in SRRM2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Scientific Reports* **5**: 10566.
- Tyzzer EE. 1916. Tumor immunity. *The Journal of Cancer Research* **1**: 125-156.
- Ujvari B, Klaassen M, Raven N, Russell T, Vittecoq M, Hamede R, Thomas F, Madsen T. 2018. Genetic diversity, inbreeding and cancer. *Proceedings of the Royal Society B: Biological Sciences* **285**: 20172589.
- van der Spek AH, Fliers E, Boelen A. 2017. The classic pathways of thyroid hormone metabolism. *Molecular and Cellular Endocrinology* **458**: 29-38.
- Van Hoeck A, Tjoonk NH, van Boxtel R, Cuppen E. 2019. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **19**: 457.
- Vural Ç, Kiraz U, Turan G, Özkara SK, Sözen M, Çetinarslan B. 2021. Solid variant of papillary thyroid carcinoma: An analysis of 28 cases with current literature. *Annals of Diagnostic Pathology* **52**: 151737.
- Wan TSK. 2014. Cancer cytogenetics: methodology revisited. *Ann Lab Med* **34**: 413-425.
- Wang LH, Wu CF, Rajasekaran N, Shin YK. 2018. Loss of Tumor Suppressor Gene Function in Human Cancer: An Overview. *Cellular Physiology and Biochemistry* **51**: 2647-2693.
- Wijnrocx K, François L, Stinckens A, Janssens S, Buys N. 2016. Half of 23 Belgian dog breeds has a compromised genetic diversity, as revealed by genealogical and molecular data analysis. *Journal of Animal Breeding and Genetics* **133**: 375-383.
- Wilmott JS, Field MA, Johansson PA, Kakavand H, Shang P, De Paoli-Iseppi R, Vilain RE, Pupo GM, Tembe V, Jakrot V et al. 2015. Tumour procurement, DNA extraction, coverage analysis and

- optimisation of mutation-detection algorithms for human melanoma genomes. *Pathology* **47**: 683-693.
- Winship IM, Dudding TE. 2008. Lessons from the skin—cutaneous features of familial cancer. *The Lancet Oncology* **9**: 462-472.
- Wong K, van der Weyden L, Schott CR, Foote A, Constantino-Casas F, Smith S, Dobson JM, Murchison EP, Wu H, Yeh I et al. 2019. Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. *Nature Communications* **10**: 353.
- Wright S. 1922. Coefficients of inbreeding and relationship. *The American Naturalist* **56**: 330-338.
- Wucherer KL, Wilke V. 2010. Thyroid Cancer in Dogs: An Update Based on 638 Cases (1995–2005). *Journal of the American Animal Hospital Association* **46**: 249-254.
- Wunderlich V. 2007. Early references to the mutational origin of cancer. *International Journal of Epidemiology* **36**: 246-247.
- Ye F, Gao H, Xiao L, Zuo Z, Liu Y, Zhao Q, Chen H, Feng W, Fu B, Sun L et al. 2019. Whole exome and target sequencing identifies MAP2K5 as novel susceptibility gene for familial non-medullary thyroid carcinoma. *International Journal of Cancer* **144**: 1321-1330.
- Zappulli V, De Zan G, Cardazzo B, Bargelloni L, Castagnaro M. 2005. Feline mammary tumours in comparative oncology. *Journal of Dairy Research* **72**: 98-106.

2

Familial follicular cell thyroid carcinomas in a large number of Dutch German longhaired pointers

Yun Yu¹, Adriana Krupa², Rebekah I Keesler^{3, 4}, Guy C M Grinwis³, Mariska de Ruijsscher², Johan de Vos^{2†}, Martien A. M. Groenen¹, Richard P. M. A. Crooijmans¹

¹ Wageningen University & Research, Animal Breeding and Genomics, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

² AniCura Dierenziekenhuis Zeeuws-Vlaanderen, van Diemenstraat 83, 4535 AR Terneuzen, The Netherlands

³ Department of Biomolecular Health Sciences, Division of Pathology, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 1, Utrecht, The Netherlands

[†] Passed away April 6, 2017.

⁴ Current affiliate: Charles River Laboratories, 6995 Longley Lane, Reno, Nevada 89521, United States

Abstract

Thyroid carcinomas (TCs) originating from follicular cells of the thyroid gland occur in both humans and dogs, and they have highly similar histomorphologic patterns. In dogs, TCs have not been extensively investigated, especially concerning the familial origin of TCs. Here, we report familial thyroid follicular cell carcinomas (FCCs) confirmed by histology in 54 Dutch origin German longhaired pointers. From the pedigree, 45 of 54 histopathologically confirmed cases are closely related to a pair of first-half cousins in the past, indicating a familial disease. In addition, genetics contributed more to the thyroid FCC than other factors by an estimated heritability of 0.62 based on pedigree. The age of diagnosis ranged between 4.5 and 13.5 years, and 76% of cases were diagnosed before 10 years of age, implying an early onset of disease. We observed a significant higher pedigree-based inbreeding coefficient in the affected dogs (mean F , 0.23) compared to unaffected dogs (mean F , 0.14), suggesting the contribution of inbreeding to tumor development. The unique occurrence of familial thyroid FCC in this dog population and the large number of affected dogs make this population an important model to identify the genetic basis of familial thyroid FCC in this breed and may contribute to the research into pathogenesis, prevention and treatment in humans.

2.1. INTRODUCTION

Many dog breeds are predisposed to a variety of specific cancers due to consanguinity and inbreeding [1]. According to researches cancer is one of the major cause of death in dogs, accounting for 8.7%–27% of all deaths [2–4]. Skin and soft tissues were the most common sites for tumor development, followed by alimentary, mammary, urogenital, lymphoid, endocrine, and oropharyngeal [2]. Within the tumors in the endocrine organs, thyroid carcinoma (TC) is the most common type, which represents 1.2%–3.8% of all canine tumors and accounts for 90% of thyroid tumors [5–7]. TC can originate from either follicular cells (follicular cell carcinoma [FCC]) or parafollicular cells (C-cell carcinoma). Within FCC, four main histological subtypes of differentiated TCs (dTCs) are described: follicular TC (FTC), compact TC (CTC), follicular-CTC (FCTC) and papillary TC (PTC) with FTC and CTC the most frequent [6, 8]. Furthermore, poorly differentiated, and undifferentiated carcinomas and thyroid carcinosarcomas (TCs) are also recognized [8]. In humans, TC is the ninth most common type of cancer and accounts for approximately 3.1% of all cancers [9]. The histologic growth patterns in humans are largely similar to those in dogs. Additionally, TC shows no sex preference in dogs, although in humans, females have a three-fold higher risk than males [7, 10]. The prevalence of TC in older dogs (between 10 and 15 years old) is significantly higher compared to earlier onset [7].

Thyroid tumors can be of familial or spontaneous origin. In humans, the majority of TCs are sporadic, and approximately, 5%–15% of them are considered to be of familial origin [11, 12]. Due to the relatively low prevalence of familial TCs, the genetic causes are less investigated than sporadic types, thus are still poorly understood [13]. To the authors' knowledge, in dogs, there has only been one pedigree of apparent familial medullary TC reported [14]. Investigations and reports of familial thyroid tumors in dogs have been limited.

Over a period of more than 21 years, a relatively large number of TCs were diagnosed in the German longhaired pointers (GLPs) born in the Netherlands (Dutch GLPs). In this retrospective study, we review clinical and histopathological assessments of the GLPs with thyroid tumors and present genetic assessment including the inbreeding and heritability estimation based on pedigree.

2.2. MATERIALS AND METHODS

2.2.1. Study population

Medical records of the clinics belonging to Dutch and Belgian collaborating veterinary cancer centres and the database of two Dutch veterinary diagnostic pathology laboratories were searched for client owned GLPs diagnosed with thyroid

tumors between 1996 and 2017. Additionally, the owners of GLPs registered in the database of the Dutch GLP association were contacted to identify any dogs with a history of thyroid tumor. Once the dog was diagnosed with a thyroid tumor, the primary or referring veterinarian was contacted to obtain relevant information. If more than one dog was affected in the litter, the owners of the remaining littermates as well as dogs related to each of the parents were identified and contacted. Pedigree records were provided by GLP “Langhaar” association (www.germanlonghair.com) in order to perform a pedigree analysis.

Only GLPs with histopathologically confirmed follicular cell TC were included as cases in this study and used in genetic analysis. Surgical removal of the affected thyroid glands, when feasible, was centralized in one clinic. Tumor tissue obtained either at surgery or necropsy was collected from affected dogs, and part of the sample was formalin-fixed for histopathology, and the remaining sample was stored in RNAlater (RNA stabilization reagent: Qiagen, Hilden, Germany).

Cases were excluded if owners rejected to participate in the research. All the data and samples in this research were permitted to be used for scientific purpose and in publication.

2.2.2.Clinical data

The following information was retrieved from the medical records, if available: signalment, physical examination findings including tumor size (longest diameter), location and mobility (determined by palpation), clinical signs, time to presentation and date of diagnosis.

Whenever performed, the results of additional diagnostic tests, including blood tests and imaging tests, were recorded. Blood tests included complete blood cell count, serum biochemistry profiles, basal circulating total thyroxine (TT4) and thyroid stimulating hormone (TSH) concentrations. Staging was performed using diagnostic imaging (thoracic radiographs, cervical ultrasonography and computed tomography [CT]). If available, the presence of ectopic thyroid tumor was recorded.

2.2.3.Histopathology analysis

For histopathological evaluation, tissues harvested during surgery or necropsy were fixed in 10% neutral buffered formalin. Representative sections were routinely embedded in paraffin and sectioned at 4 µm and stained with haematoxylin and eosin (H&E) and examined via immunohistochemistry for thyroglobulin and calcitonin expression.

Thyroglobulin and calcitonin immunohistochemistry (IHC). Four-micron tissue sections of the formalin-fixed and paraffin-embedded tumor tissue were dried

overnight, deparaffinized and rehydrated with xylene (2×5 min) and 100% alcohol (2×3 min). Endogenous peroxidase activity was blocked with 1% H_2O_2 in methanol for 30 min. After rinsing in 1% Tween20 in PBS, the slides were treated with 1:10 normal goat serum in PBS for 15 min and incubated for 60 min with 1:200 diluted Rabbit anti-human thyroglobulin (Dako, Denmark) or 1:400 diluted Rabbit anti-human calcitonin (Dako, Denmark) at room temperature. After rinsing in PBS/Tween, the slides were then incubated with Goat anti-rabbit/biotin (Vector Labs) secondary antibody (dilution 1:250 in PBS) for 30 min at room temperature. The slides were rinsed with 1% Tween20 in PBS, incubated with ABC/PO complex (Vector Labs) for 30 min and rinsed with PBS. Finally, the slides were incubated with DAB solution for 25 min and counter stained with haematoxylin for 30–60 s at room temperature. For negative controls, the primary antibody was omitted.

Tissues were evaluated by two veterinary pathologists and classified according to the World Health Organization (WHO) classification of tumors of the endocrine system scheme [8]. If a tumor had multiple growth patterns, then classification was based on the most predominant pattern. If capsular penetration of the neoplasm was unclear, additional H&E sections were cut for additional evaluation.

2.2.4. Genetic analysis

To assess the genetic relationship between dogs collected, the family tree of all unaffected and affected (suspected and histopathologically diagnosed) dogs were constructed using Kinship2 package in R [15]. Pedigree-based inbreeding coefficients (F) of all the dogs were estimated using the CFC (Coancestry, inbreeding (F) and Contribution) program [16] based on the whole pedigree of GLPs. To evaluate the contribution of inbreeding to the incidence of the thyroid tumors in the population, the rank sum test of F between affected dogs histopathologically confirmed and unaffected dogs born before 2007 was done in R using Wilcoxon test. We excluded 86 unaffected dogs born after 2007, because many of these dogs are closely related to the affected dogs, and they could be highly susceptible to FCC. Although they are unaffected at the time of analysis, they could become affected later in their lives, thus biasing the result.

2.2.5. Heritability estimation

Heritability was estimated using ASReml 4.1 based on the pedigree relationship between the unaffected dogs and cases histopathologically confirmed [17]. Unaffected dogs born after 2007 were also excluded from the estimation. The model used is as follows.

$$y \sim \mu + \alpha + \delta + e$$

where y is the phenotype, which is a binary trait, affected status coded as 1 and unaffected status coded as 0. α is the fixed effect of gender, female or male. δ is the random animal effect. e is the random residual.

Heritability calculation equation is:

$$h^2 = \frac{v_\delta}{v_y}$$

Where v_δ is the variance of the random animal effect, v_y is the variance of FCC phenotype.

2.3. RESULTS

In total, 264 GLPs born between 1991 and 2017 were identified (supplementary Table S2.1). One hundred eighty dogs were unaffected and had no signs of thyroid tumor at the time of entering the study, data analysis or during follow-up (1996–2019). Twenty-nine dogs were suspected of thyroid neoplasia based on typical clinical signs like the presence of cervical mass, but no further diagnostics have been performed. These dogs were suspected cases in this study. Fifty-four dogs met the inclusion criteria of real cases given the histopathological diagnosis of FCC. One dog was additionally diagnosed with thyroid adenoma. Among the 54 cases, 34 (63%) were male (four castrated, 30 intact) and 20 (37%) were female (seven spayed, 13 intact). The median age was of 8 years (range, 4.5–13.5 years). Forty-one dogs (76%) developed thyroid tumor before reaching the age of 10 years.

2.3.1. Clinical complaints

Forty-four of 54 dogs (81%) had information regarding clinical complaints related to thyroid tumor recorded. Duration from the onset of clinical signs to the presentation ranged from 61 to 732 days. Detection of palpable thyroid mass without any other concurrent signs was reported in the majority of dogs (37). Seven dogs (13%) demonstrated additional clinical signs that included: intermittent cough (three dogs), alopecia (three dogs), polyuria (two dogs) polydipsia (two dogs), weight loss (one dog) and lethargy (one dog).

One dog was asymptomatic with the diagnosis of the first thyroid tumor but developed clinical signs at the time of contralateral tumor. In contrast, another dog presented with complaints related to the first thyroid tumor, while no clinical signs were recorded at diagnosis of the second tumor.

2.3.2. Tumor details

Bilateral tumors were identified in 35 dogs, and unilateral tumors in 19 dogs. Eleven tumors were left-sided, six right-sided, and for two, the site of involvement was not mentioned. Three dogs were suspected of having ectopic tumors: two in the cranial mediastinum, and one at the base of the heart.

Of the 23 tumors for which information regarding the palpable mobility of the mass was available, 13 were described as moveable, whereas 10 were described as fixed. Mobility of the remaining tumors on palpation was not specified in the medical record.

Information regarding tumor size was most available in the form of the maximum dimension. Estimated tumor size based on physical examination was available for 33 dogs. Median maximal tumor diameter was 5 cm (range 2–12 cm).

2.3.3. Diagnostic findings

Thirty-three of 54 dogs (61%) underwent at least one diagnostic imaging, including CT of the cervical region and thorax (13 dogs), cervical ultrasonography (three dogs), thoracic radiographs (22 dogs) and abdominal ultrasonography (four dogs). Six of these 33 dogs had more than one test performed. Sixteen of 54 dogs (30%) had no imaging, while in five dogs (9%), required data were missing.

Based on diagnostic imaging, four dogs had involvement of the regional lymph nodes: two dogs ipsilateral retropharyngeal lymph node, one dog ipsilateral mandibular and retropharyngeal lymph node and one dog ipsilateral cervical superficial lymph node. Histopathology confirmed metastatic disease in three dogs. One dog underwent post-mortem examination, but the suspected lymph node was not evaluated.

Distant metastases were suspected in only one dog (pulmonary nodules); however, further diagnostics were not performed to confirm this.

2.3.4. Clinical pathology

On presentation, TT4 (total T4) was measured in 30 dogs and TSH in 11 dogs. Four dogs with elevated TT4 and decreased TSH showed clinical signs compatible with hyperthyroidism. Seventeen dogs had TT4 within normal limits, while in nine dogs, it was below the lower end of the reference interval. Four dogs had elevated TSH, while their TT4 was also increased (three dogs) or within the reference ranges (one dog). Three dogs had unremarkable TSH and TT4.

Other clinical pathological abnormalities were sporadic and mild, including anaemia (three dogs), leukocytosis (one dog), hypocalcaemia (one dog), alkaline phosphatase elevation (one dog) and hypercholesterolemia (four dogs).

2.3.5. Histopathology

Thyroid FCCs were diagnosed in 54 dogs. Bilateral neoplasms were diagnosed in 29 dogs. The majority of the 83 carcinomas showed a follicular growth pattern ($n = 37$; Figure 2.1A), whereas compact (solid) ($n = 15$; Figure 2.1B), follicular-compact ($n = 16$) and papillary ($n = 9$; Figure 2.1C) growth patterns were seen in the other carcinomas. In three dogs, a carcinosarcoma, characterized by osteosarcoma and carcinoma (Figure 2.1D), was diagnosed. In two dogs, a carcinoma not otherwise specified (NOS) was diagnosed. In one dog, diagnosed in 1996, which was the first case we found, the diagnosis was only thyroid tumor with signs of malignancy. In four carcinomas, well-differentiated bone tissue was seen (metaplastic bone formation). An ectopic compact FCC was found at the heart-base during necropsy in one dog that also had follicular-compact type carcinoma in both thyroid glands.

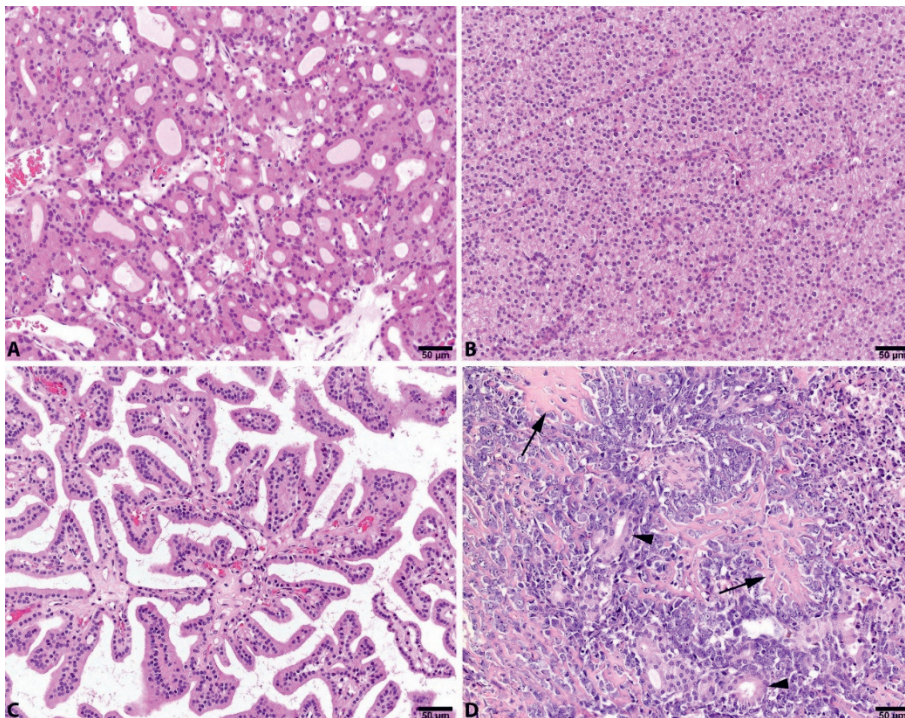


Figure 2.1. Histological pictures of different histological types of thyroid follicle cell carcinomas in German longhaired pointers. (A) Follicular, (B) compact, and (C)

papillary growth pattern of neoplastic cells. (D) A carcinosarcoma with osteoid (arrows) producing mesenchymal neoplastic cells and cattered neoplastic follicular structures (arrowheads). H&E.

Immunohistochemistry was performed on the neoplasms of 40 dogs. The neoplastic cells were vaguely to markedly positive for thyroglobulin in all tumors. The strongest immunoreactivity was typically noted in the colloid with lower staining intensity in the neoplastic cells. All neoplasms were negative for calcitonin.

2.3.6. Heritability

For the heritability estimation besides the 54 histologically confirmed cases, 94 unaffected GLP dogs born before 2007 were incorporated in the analysis. Heritability of the FCC in these dogs was estimated to be 0.62 (± 0.19).

2.3.7. Inbreeding

The complete GLP pedigree registered worldwide used for inbreeding estimation included 58,634 GLPs. The 17,786 Dutch GLPs have higher inbreeding coefficient (average $F = 0.19$) compared to GLPs born in other countries (average $F = 0.10$) with a p-value of $2.2e-16$ (Figure 2.2A). Based on this complete GLP pedigree, the inbreeding coefficients of 52 of 54 histologically confirmed affected dogs were 0.23 where in the unaffected dogs born before 2007, it was 0.14. Affected dogs are more inbred than unaffected dogs (p-value of $2.473e-08$) (Figure 2.2B).

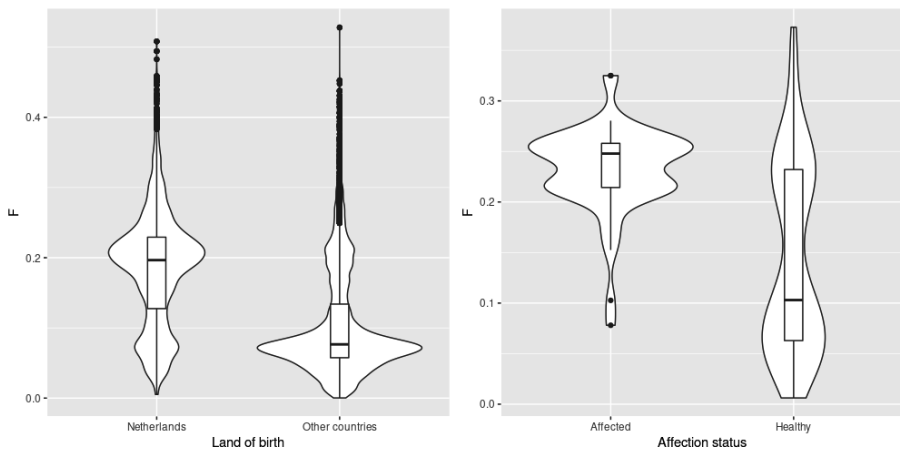


Figure 2.2. (A) Inbreeding of dogs born in the Netherlands and other countries. (B) Inbreeding of histopathologically confirmed affected GLPs and unaffected GLPs born before 2007.

2.4. DISCUSSION

In humans, familial TC is diagnosed when two or more first-degree relatives are affected [12]. Here, we showed that the incidence of FCC is strikingly high in some families of Dutch GLPs, like in the pedigrees of GLP52 and GLP905. These two dogs have a most recent common ancestor, GLP306 (Figure S2.2), born in 1989, with the F of 20.57%. Furthermore, 78 probably affected GLPs (26 suspected and 52 histopathologically confirmed FCC cases) can be traced back to a common cross of six generations prior to GLP52, the cross between GLP319 and GLP296 (Figure S2.3). With such close relationships between the majority of the affected dogs, the FCC in these dogs is considered to be a familial disease.

In this study, besides the 54 histopathologically confirmed cases, twenty-nine dogs were suspected to be affected by thyroid tumor based on clinical findings (e.g., presence of a mass lesion at the location of the thyroid gland), but because no histological assessment was performed, these suspicions could not be confirmed. Interestingly, these suspected cases are very closely related to the most affected GLPs with diagnosis (Figure S2.5). Among them, twenty-two are closely related to the two prominent spreaders of the disease, GLP52 and GLP905 (Figure 2.3), as either the siblings or direct descendants. These suspected dogs are very likely affected by the same familial FCC.

Familial cancers usually occur at a relatively young age. TC normally occurs at the median age of 9–10 years in dogs, and its occurrence increases with age [5]. In a previous study, approximately 57% of FTC in dogs occurred between 10 and 15 years [7], while in our cohort of Dutch GLPs, the FCC showed early onset with 76% of cases occurring before 10 years of age. However, some cases can have very late onset, as there are 10 dogs with an age at diagnosis of >10 years, which could represent spontaneous cases within our cohort.

In humans, different inheritance patterns of familial thyroid follicular cell cancer have been postulated, including an autosomal dominant inheritance pattern with incomplete penetrance [18, 19] and a polygenic (oligogenic) inheritance [20]. Genetic heterogeneity has been proved. However, FCC in these Dutch GLPs is likely a recessive trait, according to the occurrence of FCC in the family of GLP160 and GLP124 (Figure 2.4). GLP124 is the offspring of a half-sibling of GLP52, and GLP124 and GLP160 have a common ancestor with GLP52 and GLP905, a male dog born in 1971. Both GLP124 and GLP160 were unaffected, while one of their five offspring was confirmed to be affected, and one was a suspected case. This strongly suggests the recessive behaviour of the trait, although considering the possible incomplete penetrance of this disease, there is a small chance that unaffected parents could be carriers of a dominant causal gene but do not show the

phenotype. Therefore, to determine completely whether the TC in this study is recessive, dominant or polygenic, further genetic analysis is needed.

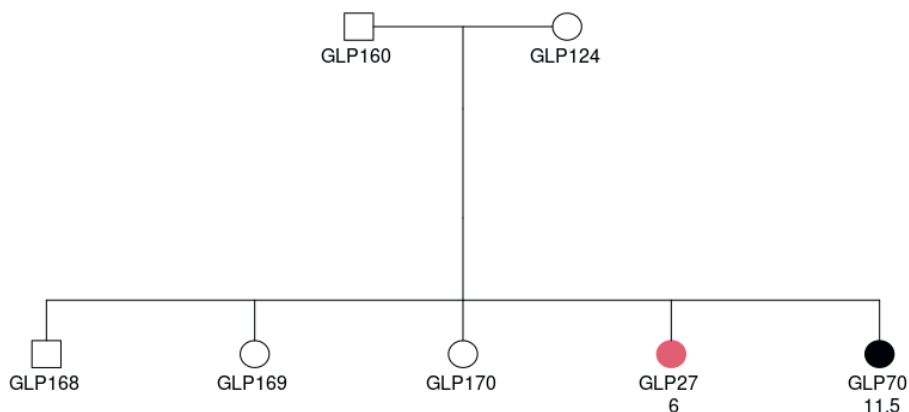


Figure 2.4. Affected status of a cross between unaffected individuals GLP160 and GLP124. Square denotes male, and circle represents female. The individual in red was confirmed to be affected by histopathology. Black colour indicates a suspected case based on clinical signs. The two rows of texts below the circles or squares represent the ID and diagnosis age (in years), respectively.

One reason why thyroid tumors occurred in so many GLPs is that the age of detection is higher than the typical breeding age. This reason should be true for many other tumors in dogs as well. Before any signs of the thyroid tumors were noticed, the dogs produced offspring, like GLP52. GLP52 was a dog affected at 12 years of age, but that had already been crossed with GLP905 and GLP333 and produced 37 affected offspring, many years before the first case was diagnosed in the offspring generation. In addition, intensive use of few dogs in the breeding programs also contributed to the high incidence of TC in these dogs. In total, GLP52 and GLP905 have 602 and 512 descendants, respectively, demonstrating how a few dogs were used intensively. These dogs and their future offspring all have high susceptibility to TC because of the consanguinity.

Inbreeding is an important tool used in dog breeding programs to fix desirable traits in a population. However, harmful side effects, such as inbreeding depression, could decrease animal performance and result in a high risk of propagation of recessive diseases or defects [21-23], as demonstrated in this study. Inbreeding contributed to the high incidence of FCC in this dog population, because we found a significant

higher F in the affected GLPs compared to the unaffected GLPs (Figure 2.2B). Moreover, the two prominent spreaders, GLP52 and GLP905, are highly inbred, with inbreeding coefficients of 0.21 and 0.24, respectively. Both parents of GLP52 are from inbred crosses between half-siblings. We also see other extreme inbreeding examples, which produced affected dogs. For instance, GLP905 was crossed with its half-sibling GLP1119 and produced two affected dogs (one confirmed and one suspected).

Cancer incidence is complex and is determined by a combination of many factors, including genetic make-up, the environment, and the lifestyle of the carrier, with genetics playing a large role. In humans, TC has the strongest genetic component among all the cancers, with genetic contribution exceeding other factors [24]. In these GLPs with TC, genetic factors may contribute more than environmental factors as well, with a heritability estimated to be 0.62.

The genetic basis of familial thyroid cancer is poorly defined in humans, as only 5% of familial FCC cases have well-defined germline mutations [13, 20]. Research of TC in dogs can contribute to the knowledge of corresponding TC in humans. Dogs have been proposed as an ideal model for human cancer research, because many cancers have strong similarity in histological appearances, genetic causes, biological behaviours, and response to conventional therapy [25]. Additionally, dogs share their environments with human pet owners, thus are partly exposed to similar risk factors, which can be exploited for epidemiological studies of cancers common in humans and dogs [26]. The affected GLPs we reported here can serve as a model to identify the genetic basis of FCC. We have a uniquely large number of affected dogs from one breed, and they are inbred (average F 0.23) and very likely share common genetic mutations that are associated with carcinogenesis. The large sample size gives more possibility and power to further define the underlying mutation(s) of this disease by genetic and genomic techniques, like, e.g., whole genome association analyses.

ACKNOWLEDGEMENTS

The authors thank the members of the SKK working group of the Langhaar association for providing access to the dog breeders and owners including financial support for the research. They especially thank Annie van der Sluis for her help in the pedigree data collection. Yun Yu's PhD study was supported by scholarship #201806760044 from China Scholarship Council (CSC).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Adriana Krupa and Johan de Vos performed clinical diagnosis and analysis. Rebekah I. Keesler and Guy C. M. Grinwis performed histopathological analysis. Mariska de Ruijscher and Johan de Vos collected the data. Richard P. M. A. Crooijmans and Yun Yu designed the study. Yun Yu wrote the manuscript with input from all authors. Johan de Vos and Richard P. M. A. Crooijmans set up the early ideas of this research.

2.5. Reference

1. Schiffman J, Breen M. Comparative oncology: What dogs and other species can teach us about humans with cancer. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2015;370.
2. Dobson JM, Samuel S, Milstein H, Rogers K, Wood JL. Canine neoplasia in the UK: estimates of incidence rates from a population of insured dogs. *J Small Anim Pract*. 2002;43(6):240-246.
3. Proschowsky HF, Rugbjerg H, Ersbøll AK. Mortality of purebred and mixed-breed dogs in Denmark. *Prev Vet Med*. 2003;58(1-2):63-74.
4. Lewis TW, Wiles BM, Llewellyn-Zaidi AM, Evans KM, O'Neill DG. Longevity and mortality in Kennel Club registered dog breeds in the UK in 2014. *Canine Genet Epidemiol*. 2018;5:10.
5. Barber LG. Thyroid Tumors in Dogs and Cats. *Veterinary Clinics of North America: Small Animal Practice*. 2007;37(4):755-773.
6. Liptak JM. Canine Thyroid Carcinoma. *Clinical Techniques in Small Animal Practice*. 2007;22(2):75-81.
7. Wucherer KL, Wilke V. Thyroid Cancer in Dogs: An Update Based on 638 Cases (1995–2005). *Journal of the American Animal Hospital Association*. 2010;46(4):249-254.
8. Kiupel M, Capen C, Miller M, Smedley R. Tumors of the thyroid. *Histological Classification of Tumors of the Endocrine System of Domestic Animals 2nd Series (Schulman, FY ed)*, Armed Forces Institute of Pathology, Washington, DC. 2008:25-39.
9. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res*. 2014;74(11):2913-2921.
10. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 0(0).
11. Alzahrani AS, Kannan MA, Ebtesam Q, Hindi A-H. HAP2 gene mutations do not cause familial or sporadic non-medullary thyroid cancer in a highly inbred middle eastern population. *Thyroid*. 2016; 26(5): 667- 671.
12. Nosé V. Familial thyroid cancer: a review. *Modern Pathology*. 2011;24:S19.
13. Ngeow J, Eng C. HAP2 in Familial Nonmedullary Thyroid Cancer: Will the Real Mutation Please Stand Up? *JNCI: Journal of the National Cancer Institute*. 2016;108(6).
14. Lee J-J, Larsson C, Lui W-O, Höög A, Von Euler H. A dog pedigree with familial medullary thyroid cancer. *International journal of oncology*. 2006;29(5):1173-1182.
15. Sinnwell JP, Therneau TM, Schaid DJ. The kinship2 R package for pedigree data. *Human heredity*. 2014;78(2):91-93.
16. Sargolzaei M, Iwaisaki H, Colleau JJ. CFC: A tool for monitoring genetic diversity. *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*. 2006:27-28.
17. Gilmour AR, Gogel, B. J., Cullis, B. R., Welham, S. J. and Thompson, R. *ASReml User Guide Release 4.1 Structural Specification*, H. VSN International Ltd. 2015.

18. Kebebew E. Hereditary Non-medullary Thyroid Cancer. *World Journal of Surgery*. 2008;32(5):678-682.
19. Saporito D, Brock P, Hampel H, et al. Penetrance of a rare familial mutation predisposing to papillary thyroid cancer. *Familial Cancer*. 2018;17(3):431-434.
20. Diquigiovanni C, Bonora E. Genetics of Familial Non-Medullary Thyroid Carcinoma (FNMTc). *Cancers*. 2021; 13(9):2178. <https://doi.org/10.3390/cancers13092178>.
21. Doekes HP, Veerkamp RF, Bijma P, de Jong G, Hiemstra SJ, Windig JJ. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein–Friesian dairy cattle. *Genetics Selection Evolution*. 2019;51(1):54.
22. Ujvari B, Klaassen M, Raven N, et al. Genetic diversity, inbreeding and cancer. *Proceedings of the Royal Society B: Biological Sciences*. 2018;285(1875):20172589.
23. Abadie J, Hédan B, Cadieu E, De Brito C, Devauchelle P, Bourgain C, et al. Epidemiology, Pathology, and Genetics of Histiocytic Sarcoma in the Bernese Mountain Dog Breed. *Journal of Heredity*. 2009;100(suppl_1):S19-S27.
24. Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer*. 2002;99(2):260-266.
25. Khanna C, Lindblad-Toh K, Vail D, et al. The dog as a cancer model. *Nature Biotechnology*. 2006;24(9):1065-1066.
26. Rowell JL, McCarthy DO, Alvarez CE. Dog models of naturally occurring cancer. *Trends Mol Med*. 2011;17(7):380-388.

2.6. Supplementary materials

Supplementary **Table S2.1** can be found through this link

https://github.com/YunYu93/Data-depository/blob/main/Supplementary_Table_S2.1.xlsx

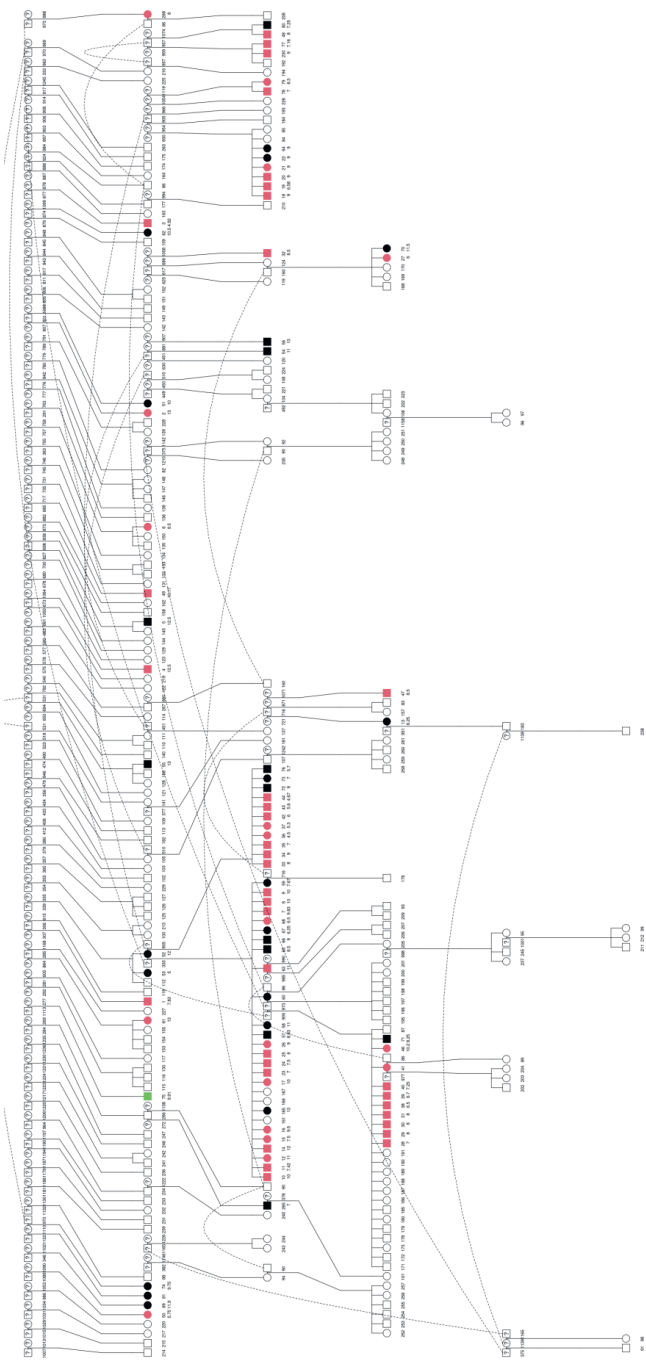


Figure S2.1. The pedigree of all 264 GLPs with affection information. Circles represent females, squares represent males. Dot line show identical dogs. Dogs with FCC histologically diagnosed are highlighted in red, the dog with follicular thyroid adenoma is highlighted in green, and suspected affected dogs are in black, whereas unaffected dogs remain white. A question mark represents the dogs with unknown status. The 2 rows of texts below the circles or squares represent the ID and diagnosis age (in years), respectively.

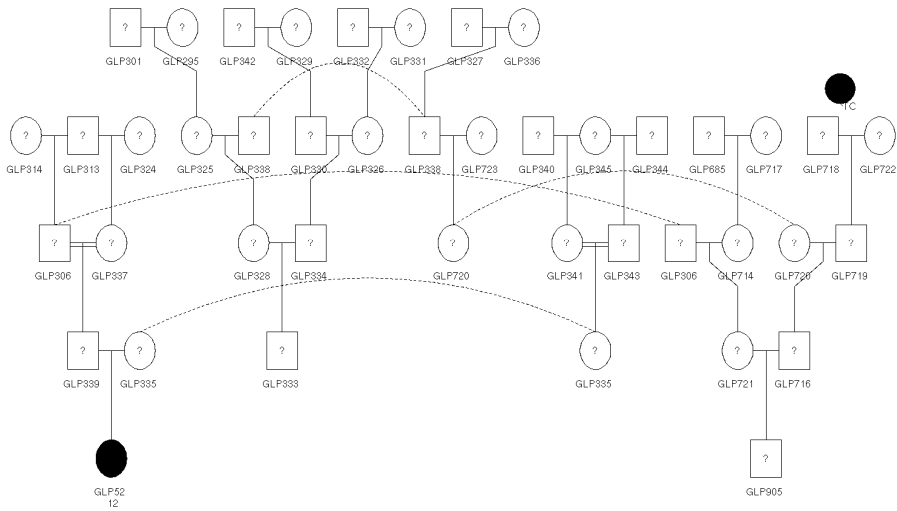


Figure S2.2. Ancestry family tree of GLP52 and GLP905. These two dogs are half-first cousins with a common grandfather GLP306. GLP52 was a suspected case. Circles represent females, squares represent males. Dot line show identical dogs. The 2 rows of texts below the circles or squares represent the ID and diagnosis age (in years), respectively.

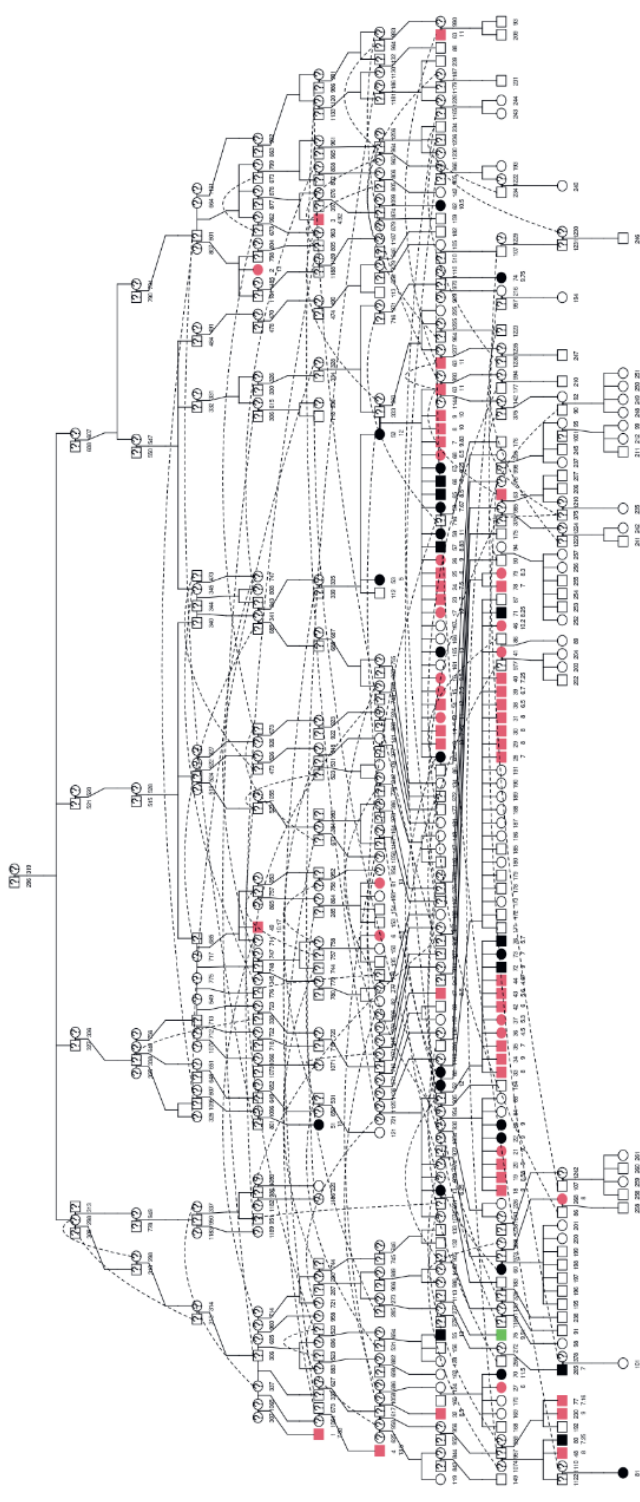


Figure S2.3. Pedigree of GLPs that could be traced back to the cross between GLP296 and GLP319. Circles represent females, squares represent males. Dot line show identical dogs. The 2 rows of texts below the circles or squares represent the ID and diagnosis age (in years), respectively. Dogs with FCC histologically diagnosed are highlighted in red, the dog with follicular thyroid adenoma is highlighted in green, and suspected affected dogs are in black, whereas unaffected dogs remain white. A question mark represents the dogs with unknown status.

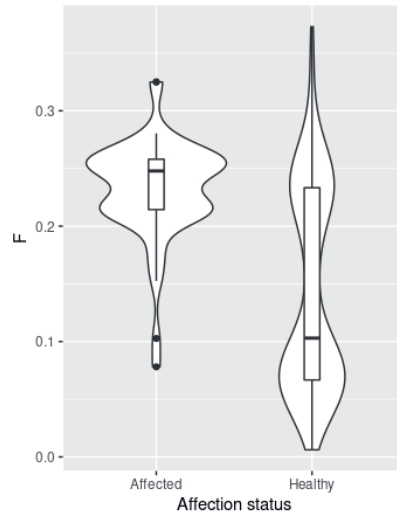


Figure S2.4. F of affected and unaffected GLPs including the dogs born after 2007 in our dataset (54 cases and 177 controls) (Wilcoxon test, p -value= $4.317\text{e-}10$). Affected dogs are more inbred than unaffected dogs.

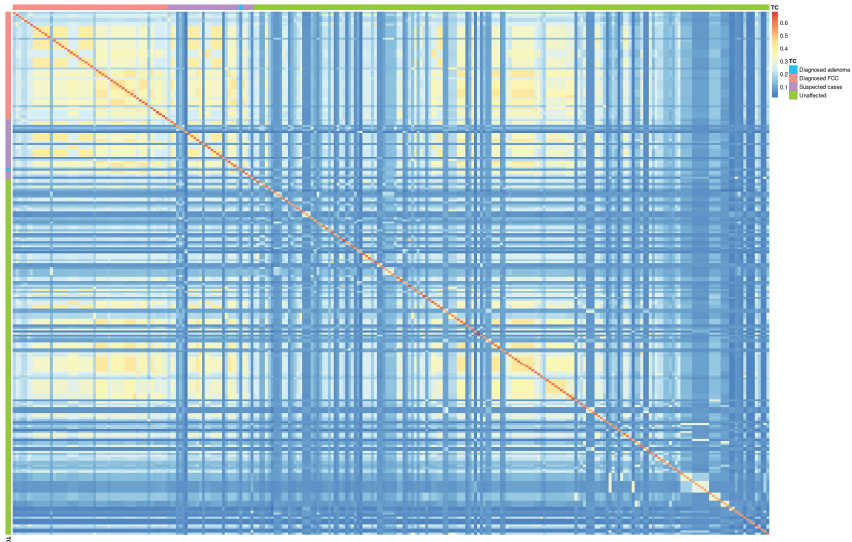


Figure S2.5. Kinship matrix between 54 FCC cases, 1 dog with adenoma, 29 suspected cases, and 180 unaffected dogs. Kinship matrix was estimated using kinship2 package in R. The 54 FCC cases are closed related to each other. Meanwhile, most suspected cases are closely related to the histologically diagnosed FCC cases.

3

Deleterious Mutations in the *TPO* Gene Associated with Familial Thyroid Follicular Cell Carcinoma in Dutch German Longhaired Pointers

Yun Yu, Henk Bovenhuis, Zhou Wu, Kimberley Laport, Martien A. M. Groenen
and Richard P. M. A. Crooijmans

Animal Breeding and Genomics, Wageningen University & Research,
Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

Genes. 2021; 12(7):997. doi.org/10.3390/genes12070997

Abstract

Familial thyroid cancer originating from follicular cells accounts for 5–15% of all the thyroid carcinoma cases in humans. Previously, we described thyroid follicular cell carcinomas in a large number of the Dutch German longhaired pointers (GLPs) with a likely autosomal recessive inheritance pattern. Here, we investigated the genetic causes of the disease using a combined approach of genome-wide association study and runs of homozygosity (ROH) analysis based on 170k SNP array genotype data and whole-genome sequences. A region 0–5 Mb on chromosome 17 was identified to be associated with the disease. Whole-genome sequencing revealed many mutations fitting the recessive inheritance pattern in this region including two deleterious mutations in the *TPO* gene, chr17:800788G>A (686F>V) and chr17:805276C>T (845T>M). These two SNP were subsequently genotyped in 186 GLPs (59 affected and 127 unaffected) and confirmed to be highly associated with the disease. The recessive genotypes had higher relative risks of 16.94 and 16.64 compared to homozygous genotypes for the reference alleles, respectively. This study provides novel insight into the genetic causes leading to the familial thyroid follicular cell carcinoma, and we were able to develop a genetic test to screen susceptible dogs.

Keywords: dog; thyroid carcinoma; mutation; *TPO*; GWAS

3.1. Introduction

In humans, thyroid cancer constitutes 3.4% of cancers diagnosed worldwide annually [1]. Thyroid carcinoma (TC) originating from follicular cells have two main types: follicular thyroid carcinoma (FTC) and papillary thyroid carcinoma (PTC). FTC accounts for 14% and PTC accounts for 81% [2] of thyroid carcinomas. In dogs, thyroid follicular cell carcinomas (FCC) are mainly classified into four types: FTC, PTC, compact thyroid carcinoma (CTC), and follicular-compact thyroid carcinoma (FCTC). These FCCs in dogs are remarkably similar in histology and biological behavior to thyroid carcinoma with follicular origin in humans [3]. Similarity in cell origin and histology of FCC indicates that dogs might be able to serve as a thyroid cancer model for research and treatments development.

Thyroid cancer can be of either familial or spontaneous origin, caused by heritable germline risk factor and sporadic somatic mutations, respectively. In humans, the genetics of TC were studied extensively. Genetic mutations are a major contributor to thyroid cancer [4]. Many germline genetic mutations were reported to be associated with familial TC, including mutations in *APC*, *PTEN*, *SDHB-D*, *PIK3CA*, *AKT1*, *SEC23B*, *WRN*, and *PRKAR1α*, which cause syndromic TC [5]. While most of these germline genetic mutations cause TC through a dominant mode. *WRN* gene mutations cause TC through an autosomal recessive mode [6]. Moreover, genome-wide association studies approaches identified many germline genetic mutations associated with familial TC. These include the genes *FOXE1* [7], *SRGAP1* [8], *HABP2* [9], *BRCA1* [10], *CHEK2* [11], *ATM* [12], *RASAL1* [13], *SRRM2* [14], *XRCC1* [15], and *PTCSC3* [16]. Most of these genes also cause TC through a dominant mode. Whole genome sequencing of thyroid tumor tissues identified many somatic mutations driving the initiation and the progression of TC. The type and the number of somatic mutations between cases with familial and spontaneous TC are similar [17]. *BRAF* (V600E) is the most common somatic mutation associated with PTC [18]. *RAS* somatic mutations are the second most common type of mutations found in fine needle aspiration of thyroid nodules [19]. Somatic *RAS* mutations are present in 15–30% of TC [20]. In addition, *PAX8/PPARG* was also identified to be an oncogenic driver for TC [19, 21]. Somatic *RET/PTC* rearrangement associates with PTC [19]. *TERT* promoter somatic mutations were identified in follicular carcinoma specimens and may serve as a marker for the aggressive form of TC with lethal consequences [22, 23]. TC in dogs can have the same genetic causes as TC in humans. For example, *K-RAS* somatic mutations were found in dogs with FTC and medullary thyroid carcinoma (MTC). Besides, different genetic causes were also identified. Germline mutations in the *RET* oncogene on chromosome 10q11.2 underlie most hereditary forms of MTC in humans with an autosomal dominant inheritance pattern [24], while, in dogs, the

RET mutations were not found in hereditary MTC [25]. Compared to genetic research of TC in humans, research on the genetic background of TC in dogs is limited.

Familial thyroid carcinoma can be defined when two or more first-degree relatives are affected in the absence of other cancer predisposition syndromes [26]. Previously, we reported a large number of familial FCC in the Dutch German longhaired pointers (GLPs). The pedigree suggests a recessive mode of inheritance [27]. The aim of the current study was to identify the germline causal gene(s) of familial FCC in the GLP population. A variety of approaches was used, including a genome-wide association study (GWAS) and ROH analyses based on SNP array data. In addition, whole genome sequences of affected and unaffected GLPs were obtained to identify the potential causal/susceptible gene/variant in the candidate region, followed by validation of the candidate variants through PCR-RFLP in a larger number of Dutch GLPs.

3.2. Materials and Methods

3.2.1. Animal and Diagnosis

All the German longhaired pointers used in this research were from the dataset described previously [27]. Briefly, in total, 264 GLPs were examined, and 84 cases were identified, of which 54 were histopathologically confirmed FCC cases, 1 was a follicular adenoma case, and 29 were suspected of thyroid neoplasia based on typical clinical signs such as the presence of cervical mass, but no further diagnostics were performed. Clinical examinations were performed by the veterinary oncology center “AniCural”. Blood of 186 GLPs and tumor samples 36 GLPs were collected by veterinarians at the time of the diagnosis. The owners of the dogs gave permission for the tissues to be used for research purposes. The histology assessments were performed by the Department of Pathology, Utrecht University. Detailed description of samples and diagnosis procedures can be found in the previous study [27].

3.2.2. Genotyping

DNA was extracted from animals genotyped and sequenced using Gentra Puregene Blood Kit (Qiagen, Hilden, Germany). Twenty-five affected and twenty-six unaffected GLPs were genotyped. The age at diagnosis of the genotyped dogs ranged between 4.5 and 9.8 years with an average of 7.3 years. All unaffected dogs had ages >13 years. Forty-three of the dogs were genotyped using the 170 k canineHD SNP beadchip array (Illumina Inc., San Diego, CA, USA). The remaining 8 dogs were genotyped using the 230 k canineHD SNP beadchip array (Illumina Inc., San Diego, CA, USA), which is the extended version of the 170 k canineHD SNP beadchip array (Illumina Inc., San Diego, CA, USA). Detailed information about the

samples is in Supplementary Table S3.1. The 173,662 SNPs shared between these two SNP arrays were extracted for each genotyped dog and used in subsequent analyses. All the SNPs were remapped to the canine reference genome CanFam3.1 using the NCBI Remap tool. Quality control was performed using the following criteria: minor allele frequency > 0.05, maximum missing call rate per variant 0.1, and maximum missing genotype rate per individual 0.1. A total of 7,398 variants were removed due to missing call rate.

To detect the genetic relationship between these GLPs genotyped, principal component analysis (PCA) was performed using Plink v1.9 [28], and the first two principal components were plotted using R. Inbreeding coefficient (F) based on the difference between observed and expected counts of autosomal homozygous genotypes was calculated using Plink v1.9 by --het command.

3.2.3. Whole Genome Sequencing (WGS)

To identify the causal variant(s) in the candidate region identified by GWAS study, 22 GLPs (11 affected and 11 unaffected) were whole genome sequenced. DNA samples were used for library construction following the manufacture's recommendations using NEB Next® Ultra™ DNA Library Prep Kit (Cat No. E7370L) (NEB, Ipswich, MA, USA). Index codes were added to each sample. Briefly, the genomic DNA was randomly fragmented to an average size of 350 bp. DNA fragments were end polished, A-tailed, ligated with adapters, size selected, and further PCR enriched. Then PCR products were purified (AMPure XP system, Beckman Coulter, Indianapolis, IN, USA), followed by size distribution by Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and quantification using real-time PCR. Libraries were sequenced on a NovaSeq 6000 S4 flow cell with PE150 strategy. Three dogs were sequenced at a depth of 30x, 3 dogs at a depth of 60x, and 16 dogs at a depth of 10x. All the healthy dogs had an age above 13 years by the time of the study. Detailed information about the samples can be found in Supplementary Table S3.1. FastQC [29] was used to evaluate the quality of the sequences. Sickle was used to trim the reads using default settings. Sequences were aligned to the CanFam 3.1 reference genome using BWA-MEM algorithm (version 0.7.15) [30], then samtools 1.9 [31] was used to sort the aligned reads and to remove duplications. GATK3.5 [32] was used to perform indel-based re-alignment.

Freebayes [33] was used to call single-nucleotide variants (SNPs) and small insertions or deletions (InDels) from WGS for each dog. Filtering was performed using bcftools v1.9 [34]. Loci covered by less than 4 reads were removed. In addition, variants with a calling quality less than 20 were also discarded. Structural variants were called using Manta [35]. Nine WGSs (GLP77, GLP44, GLP39, GLP84, GLP85, GLP169, GLP82, GLP04, GLP25) were used for SV calling. The

other WGSs were discarded due to uneven coverage across the genome. Variants were annotated and analyzed for predicted effects using VEP [36] program and were visually confirmed in Jbrowse [37].

3.2.4.RNA-Sequencing and Data Processing

Tumor tissues of left thyroid gland from 7 affected dogs were sampled at the time of diagnosis and stored in RNAlater RNA stabilization reagent (Qiagen, Hilden, Germany). RNA was extracted from the tumor tissue using AllPrep RNA Mini Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. The RNA samples were used for library preparation. The directional libraries were prepared using NEBNext® Ultra TM Directional RNA Library Prep Kit for Illumina® (NEB, Ipswich, MA, USA) following manufacturer's protocol. Indices were included to multiplex multiple samples. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesized using random hexamer primers followed by the second strand cDNA synthesis. The strand-specific library was ready after end repair, A-tailing, adapter ligation, size selection, and USER enzyme digestion. After amplification and purification, insert size of the library was validated on an Agilent 2100 and quantified using quantitative PCR (Q-PCR). Libraries were then sequenced on the Illumina NovaSeq 6000 S4 flowcell with PE150 according to results from library quality control and expected data volume. FastQC were used to check the read quality. Hisat2 [38] was used to map the reads to reference genome CanFam3.1 with --dta option. Then, FeatureCounts [39] was used to quantify mapped reads to genomic features such as genes, exons, gene bodies, genomic bins, and chromosomal locations. Alignments were visually inspected in IGV [40].

3.2.5.GWAS

GEMMA 0.98.1 [41] was used to perform the genome wide association analysis using a univariate linear-mixed model, correcting for population stratification by accounting for family relationships among dogs by incorporating a standardized genomic relationship matrix, calculated from SNP array data by GEMMA. There were 45,819 SNPs discarded from the analysis by the default filtering of GEMMA. Manhattan plots were generated by plotting p-value of Wald test using qqman package in R.

3.2.6.Runs of Homozygosity

Previous analyses suggest a recessive mode of inheritance. In that case, affected dogs are expected to carry two copies of the causal allele. Therefore, we expected, in the region carrying the causal mutation, a run of homozygosity (ROH) in affected dogs while expecting it would be absent in unaffected dogs. Runs of homozygosity

across the genome were detected using PLINK v1.9. ROHs were defined according to the following criteria: (i) the minimum count of SNPs in a sliding window was 15; (ii) the minimum ROH length was set to 1 Mb; (iii) the maximum inverse density was 100 Kb per SNP; (iv) to avoid the effects of low SNP density region, the maximum gap length between consecutive SNPs was 1 Mb; (v) the minimum hit rate of all scanning windows containing the SNP was set to 0.05. The ROH autozygosity on each chromosome was plotted contrasting affected and unaffected dogs using in-house R script.

3.2.7.Candidate SNPs PCR-RFLP Genotyping

PCR assay was done using 60 ng of genomic DNA with 0.4 μ M of each primer and 5 \times FIREPol® Master Mix, 7.5 mM MgCl₂ (Solis BioDyne, Estonia) in a final volume of 12 μ L. PCR primers for chr17:800788G>A were Forward 5'-CAGGTTACAACGCGTGGAG -3' and Reverse 5'-TCCCTCAGAGCCTTCATCTG -3' to generate a 232 bp amplicon. The PCR primers for chr17:805276C>T were Forward 5'-AGGGTGGTTTCAGGTGTGAG -3' and Reverse 5'-GTGAGGACACGGCAAGAGAT -3' to generate a 172 bp amplicon. The PCR reaction was carried out in a T100 Thermal Cycler (BioRad, CA, USA) and included an initial denaturation for 1 min at 95 °C was followed by 35 cycles of 95 °C for 30 s, 55 °C for 45 s, and 72 °C for 90 s, followed by a 5 min extension at 72 °C. The electrophoresis of PCR products was performed in 1.5% agarose gel containing Stain G (Serva, Germany) together with a 100 bp DNA ladder (New England Biolabs, Ipswich, MA, USA) and photographed using a gel documentation imaging system (BioRad, Hercules, CA, USA). Regarding the RFLP of TPO gene PCR product, the 232 base pair product of chr17:800788G>A was digested with BssHII restriction enzyme (New England Biolabs, Ipswich, MA, USA) according to manufacturer's instructions to generate fragments (5 h at 50 °C followed by 20 min at 65 °C). Digestions were carried out in a total volume of 10 μ L. The reaction mixture consisted of 5 μ L of PCR product, 5 U of restriction enzyme/Cutsmart Buffer, and volume adjusted with sterile distilled water. The 172 base pair product of chr17:805276C>T was digested with Hpy99I restriction enzyme (New England Biolabs) according to manufacturer's instructions to generate fragments (5 h at 37 °C followed by 20 min at 65 °C). Digestions were carried out in a total volume of 10 μ L. The reaction mixture consisted of 5 μ L of PCR product, 2 U of restriction enzyme/Cutsmart Buffer, and volume adjusted with sterile distilled water. The digest was electrophoresed in 3% agarose with Stain G (Serva, Germany) together with a 100 bp DNA ladder (New England Biolabs, Ipswich, MA, USA) and photographed using a gel documentation imaging system (BioRad, Hercules, CA, USA).

3.2.8. Criteria for Candidate Variants

We called variants in the candidate region using whole genome sequencing. The genotypes of the candidate variants associated with the familial FCC in affected and unaffected dogs were to fit the following pattern: affected dogs (excluding GLP04 and GLP60, reason is shown in the result Section 3.2) should be homozygous, and all unaffected dogs should be heterozygous or homozygous for the alternative allele; the alternative allele frequency in NCBI dbSNP should be lower than 0.05. Additionally, for the mutations in the exonic region, we focused on the mutations predicted to be deleterious by SIFT [42], PROVEAN [43], PANTHER-PSEP [44], and PolyPhen-2 [45].

3.2.9. Amino Acid Conservation between Species

TPO amino acid sequences of six species (Human, Dog, Pig, Chicken, Mouse, Rhesus macaque) were obtained from NCBI and aligned using Clustal Omega from EMBL-EBI.

3.3. Results

3.3.1. Study Population

From the affected GLPs we previously described, 25 FCC affected GLPs were selected for SNP array genotyping. Affection status of 23 GLPs was confirmed by histology, while 2 cases were suspected based on typical clinical signs, e.g., the cervical mass (Supplementary Table S3.1). Of the 22 dogs that were sequenced, 11 were affected, of which 3 were suspected based on clinical signs (Supplementary Table S3.1).

The GLPs were highly inbred with a mean inbreeding coefficient estimated based on difference between observed and expected homozygotes of 0.50. Inbreeding was higher in affected (0.51) than in unaffected dogs (0.48) (Welch two sample test: p-value of 0.002) (Figure 3.1a), which is in agreement with our previous analysis based on inbreeding coefficients estimated from the pedigree [27]. In the genotyped affected dogs, 72% were male (18 out of 25), and in the unaffected dogs, this was 38% (10 out of 26).

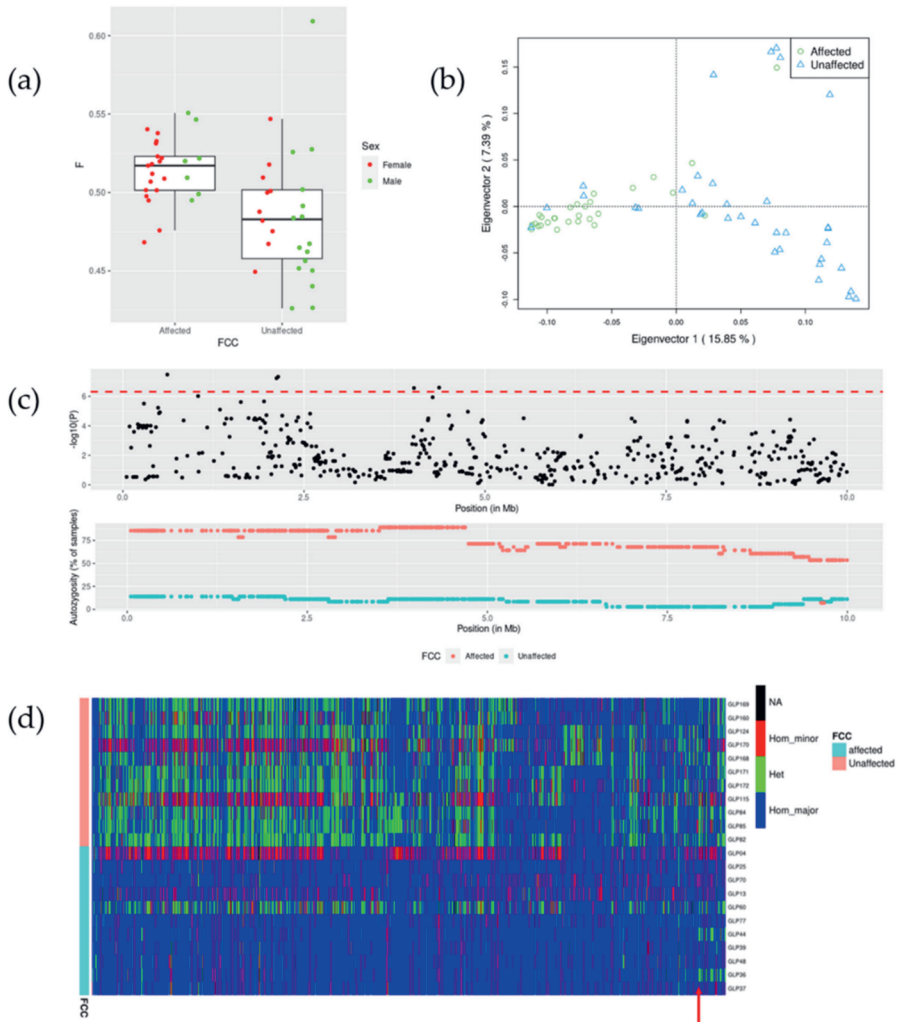


Figure 3.1. (a) Inbreeding coefficient of affected and unaffected dogs based on SNP chip genotype data. (b) PCA plot. First and second components are plotted for 28 affected and 36 unaffected GLPs used in the GWAS analysis. Most affected dogs are clustered apart from unaffected dogs, indicating clear population stratification. (c) Manhattan plot and autozygosity of ROH segments in affected and unaffected GLPs in the region between 1–10 Mb on chr17. The red dashed line in the Manhattan plot denotes the Bonferroni corrected significance threshold. (d) Genotypes of the variants between 0 and 5 Mb on chromosome 17 identified by WGS from 22 dogs. Colors blue, green, red, and black denote homozygous for major allele, heterozygous, homozygous for minor allele, and missing genotype, respectively.

As reported earlier [27], the FCC in these Dutch GLPs is a heritable disease and very likely a recessive trait. Most affected dogs are closely related. Here, the principal component analysis shows that affected dogs were clearly separated from unaffected dogs (Figure 3.1b). This emphasizes the importance of accounting for family relations in the association analysis in order to control for false positive discoveries.

3.3.2. Genomic Region Associated with FCC

To identify the genomic region responsible for FCC, a combination of two different methods was used: a GWAS and an ROH analysis. The GWAS analysis identified the region associated with the disease. The ROH analysis identified the homozygous genomic region present in the affected dogs, while it was absent from the unaffected dogs. To increase the statistic power of the GWAS analysis, we combined SNP array genotype data and WGS data to obtain a larger sample size for a total of 103,744 SNPs shared by the two methods. Three whole-genome sequenced cases (GLP25, GLP60, GLP77) and 10 controls (GLP84, GLP85, GLP115, GLP124, GLP160, GLP168, GLP169, GLP170, GLP171, GLP172) were added to the panel to achieve a sample size of 64 (28 cases, 36 controls). The other GLPs were discarded from the analysis because of either a late age at diagnosis (>10 years) or being already genotyped with the SNP array. The signal on chromosome 17 was captured by GWAS. The Manhattan plot of the GWAS result (Supplementary Figure S3.1) across the whole genome and the qq plot (Supplementary Figure S3.2) is shown in the Supplementary Material. Furthermore, a long ROH segment (Figure 3.1c) overlapping the location of the signal of GWAS was present in the affected dogs, while it was absent from the unaffected dogs. The autozygosity of ROH across the whole chromosome 17 is shown in Supplementary Figure S3.3 and autozygosity on each other chromosome is shown in Supplementary Figure S3.4. Starting from the position chr17:4741065, the autozygosity in affected dogs dropped from 86% to 71% and continued to drop further, while, in unaffected dogs, the autozygosity remained low at 8% to 14%. Therefore, we set the region 0–5 Mb as the candidate region, which was somewhat wider and covered the homozygous region. This region was also supported by the haplotypes of dogs with WGS data (Figure 3.1d), where the long ROH segment broke at a position close to 5 Mb (indicated by a red arrow in Figure 3.1d), as two dogs (GLP44 and GLP36) appeared to be heterozygous from thereon. In the candidate region, five SNPs surpassed the Bonferroni corrected significance threshold ($-\log_{10}(p) = 6.3$) in the GWAS analysis. There were 13 known protein coding genes in the region: *SNTG2*, *TPO*, *PXDN*, *MYTIL*, *EIPRI*, *TRAPPC12*, *RPS7*, *RNASEH1*, *COLEC11*, *DCDC2C*, *ALLC*, *RSAD2*, and *RNF144A*.

In the candidate region, 2 of 11 sequenced affected GLPs (GLP04 and GLP60) showed distinct haplotypes from the other 9 affected GLPs (Figure 3.1d). GLP04 has a very late onset age (13.5 years) and could be a spontaneous case with different genetic or environmental causes. Another dog, GLP60, was a suspected case without histology confirmation. It could be actually affected by other thyroid diseases which show similar clinical signs rather than FCC. Therefore, these two GLPs were excluded from cases when using WGS to identify case-specific variants.

The WGS analysis revealed 23,338 variants (SNPs and InDels) in this 5 Mb candidate region. Among them, 2,374 variants were case-specific (excluding GLP04 and GLP60) intronic, intergenic, synonymous, and nonsynonymous homozygous mutations (SNP and InDel) (Table 3.1) in the region of 0–5 Mb chr17. Three mutations (two in the *TPO* gene, one in the *SNTG2* gene) were predicted to be deleterious by in silico pathogenic prediction tools. SVs were also called in the candidate region, and case-specific SVs are shown in Table 3.2. RNA-seq of FCC tumor from seven affected GLPs was used to check the mRNA expression and architecture by visual inspection in IGV. No SV was found to change the mRNA structure of the corresponding genes according to the inspection of the mRNA expression. They were excluded as candidate causal variants.

Table 3.1. Variants identified via the whole genome sequence of GLPs.

Total number of variants	12,248,323
Variants in the candidate region	23,338
Of which homozygous	6,171
Of which private for cases	2,374
Of which exonic	18
Of which missense	7
Of which deleterious	3

Table 3.2. Private SV variants for cases.

Chromosome coordination	SV-type	Location	SV-length
chr17:370832-370908	deletion	intergenic	77
chr17:379867-380195	deletion	intergenic	328
chr17:491919-492126	deletion	intergenic	208
chr17:503763-503968	deletion	intergenic	206
chr17:533005-533231	deletion	intergenic	227
chr17:551908-551961	deletion	intergenic	54
chr17:626018	insertion	intergenic	95
chr17:731556	insertion	Intron- <i>SNTG2</i>	55
chr17:885408-885477	deletion	downstream	70
chr17:900381-900486	deletion	intergenic	106
chr17:1633933	insertion	intergenic	57
chr17:1859631	insertion	intergenic	214
chr17:1938475-1938524	deletion	Intron- <i>EIPRI</i>	50
chr17:2136862-2137405	deletion	upstream	544

3.3.3. Deleterious Mutations in the *TPO* Gene

In the *TPO* gene, two missense mutations, chr17:800788G>A (686F>V) and chr17:805276C>T (845T>M), were identified in the WGS data from 22 GLPs (11 affected and 11 unaffected) which were exclusively found in affected dogs. These mutations were predicted to be deleterious by several pathogenic prediction tools (SIFT, PROVEAN, PANTHER, PolyPhen-2) (Table 3.3). Variant chr17:800788G>A (686F>V) is not present in 722 canine genomes [46] from over 144 modern breeds, 54 wild canids and 100 village dogs. It is a novel mutation that is not yet annotated in NCBI dbSNP. Variant chr17:805276C>T (845T>M) was

detected at a very low allele frequency of 2%, with only 6 homozygotes and 15 heterozygotes in the 722 dogs. In our sequenced GLPs, 9 of the 11 affected dogs were homozygous for both variants. The other two exceptions were GLP04 (homozygous for reference allele) and GLP60 (heterozygous). No unaffected dogs were homozygous for the two variants (nine heterozygotes and two homozygotes for the reference allele). The genotypes for the two sites fit the autosomal recessive inheritance pattern.

Table 3.3. Genotype counts of candidate SNPs.

Genomic coordinates	Gene	Amino acid change	SIFT	PROVEAN SCORE (cutoff = -0.25)	PANTHER	PolyPhen-2	Genotype counts ¹		
							Affected GLPs	Unaffected GLPs	722 dogs ²
Chr17:800788G>A	<i>TPO</i>	686F>V	Deleterious (0)	-6.775	0.89	0.999	9/1/1	0/9/2	-
Chr17:805276C>T	<i>TPO</i>	845T>M	Tolerated (0.06)	-4.042	0.89	1	9/1/1	0/9/2	6/15/637
Chr17:743943T>C	<i>SNTG2</i>	360F>S	Deleterious (0)	-1.901	0.27	0.337	9/1/1	0/9/2	4/16/615

Note: ¹ counts of recessive homozygotes/heterozygotes/homozygotes for the reference allele. ² 722 dogs covering 144 modern breeds, 54 wild canids and a hundred village dogs.

Structural variants were also noticed but no structural variants were found in the *TPO* gene. According to the Sashimi plot from RNA of tumor tissues of seven affected dogs in the IGV (Supplementary Figure S3.5), the mRNA structure of the gene did not change without alternative splicing events or gene fusion.

The *TPO* gene encodes an enzyme named thyroid peroxidase, which is a poorly glycosylated membrane-bound enzyme. It is involved in thyroid hormone synthesis and a target autoantigen in autoimmune thyroid disorders. TPO oxidizes iodide ions to form iodine atoms for addition onto tyrosine residues on thyroglobulin for the production of thyroxine (T4) or triiodothyronine (T3), the thyroid hormones [47].

Both variant locations are conserved across species (Figure 3.2). Canine *TPO* c.F686 corresponds to human *TPO* h.678 located in the MPO-like domain of the protein. The MPO-like domain consists of two immunodominant regions. Canine *TPO* c.T845 corresponds to human *TPO* h.837 located in a conserved calcium-binding EGF-like domain. The EGF-like domain is involved in ligand recognition and protein–protein interaction. The amino acid changes in the region may change the three-dimensional structure, which could impact the catalytic activity or the autoimmunity of *TPO* [48].

tr F1NN54 F1NN54_CHICK	DVNLGGLVEDFLPGARTGPLFACLIQKQMKALRDGDRFHWENDNVFTDAQKHELKHSLS	707
sp Q8HYB7 PERT_CANLF	DVNLGGLAEPLPRARTGPLFACLIQKQMKALRDGDRFWESSGVFTDEQRRELARHSLS	700
sp P09933 PERT_PIG	DVNLGGLAEFLPGARTGPLFACLIQKQMKALRDGDRFWENPGVFTEAQRRELSRHSMS	690
sp P35419 PERT_MOUSE	DVNLGGLAEKFLPGARTGPLFACLIQKQMKALRDGDRFWENTNVFTDAQRRELEKHSLS	680
sp P07202 PERT_HUMAN	DVNLGGLAENFLPRARTGPLFACLIQKQMKALRDGDRFWENSHVFTDAQRRELEKHSLS	692
tr F6ZA88 F6ZA88_MACMU	DVNLGGLAENFLPRARTGPLFACLIQKQMKALRDGDRFWENSHVFTDAQRRELEKHSLS	690
	*****.* : ** *****:**:** ***** ***, ***: *:* ** :	
tr F1NN54 F1NN54_CHICK	KCINTKGSYKCFCTEPYKLAEDGRITCIDSIREPAVH-----	863
sp Q8HYB7 PERT_CANLF	RCRNTKGGFRCECTDPAVLGEDGRTICVDSGRLPKASLVSIAGIVLVVGLAGLTNTLVCR	880
sp P09933 PERT_PIG	RCKNTKGGVLCESDPLVLGEDGRTICVDAGRLPRASVVSIALGAVLVCGLAGLANTVVC	870
sp P35419 PERT_MOUSE	QCKNTKGSFQCVCTDPYVLGEDEKRTICVDSGRLPASVWSIALGALLIGGLASLTWIVIC	860
sp P07202 PERT_HUMAN	RCRNTKGGFQCLCADPYELGDDGRTICVDSGRLPRTWISLSLAALLIGGFAGLTSTVIC	872
tr F6ZA88 F6ZA88_MACMU	RCRNTKGGFQCLCADPYELGDDGRTICVDSGRLPRTWISLSLAALLIGGLAGLTSTVIC	870
	: * ****. * *:* * . * **:** * *	

Figure 3.2. Conservation of the two amino acids of the *TPO* corresponding to the two deleterious mutations between six species. The two deleterious mutation loci (indicated in red box) are very conserved across species.

Except for the 2 deleterious mutations, 4 synonymous mutations and 31 intronic variants were also identified in the *TPO* gene of affected dogs. These variants were in strong LD with the two deleterious mutations and were less likely to be the causal mutations and therefore were not investigated further.

3.3.4.Deleterious Mutation in the *SNTG2* Gene

In the *SNTG2* gene, the WGS analysis revealed a case-specific variant 743943T>C (360F>S) (Table 3.3), which is also a rare mutation with an alternative allele frequency of 0.02 in the 722 dogs [46]. This variant was predicted to be deleterious by SIFT, while it was predicted to be a neutral mutation by PROVEAN, probably benign by PANTHER, and benign by PolyPhen-2. Similar as for the two deleterious mutations found in the *TPO* gene, 9 of 11 affected dogs were homozygous for this mutation, while GLP60 was heterozygous and GLP04 was homozygous for the reference allele. None of the unaffected dogs were homozygous for this mutation.

3.3.5.Variants in the *EIPRI* Gene

The *EIPRI* gene (also named *TSSC1*, tumor-suppressing subchromosomal transferable fragment candidate gene 1) codes for a protein that acts as a specific

interactor of both GARP (Golgi-associated retrograde protein) and EARP (endosome-associated recycling protein), playing a critical role in endosomal retrieval pathways [49]. The *EIPRI* gene harbors 10 variants that fit a recessive inheritance pattern (Table 3.4). All the cases, including GLP04 and GLP60, were homozygous for these mutations, while controls were heterozygous or homozygous for the reference alleles. All 10 mutations were located in introns or the downstream region of the gene. None of these variants were predicted to affect the mRNA structure of the gene. Additionally, within the 722 dogs, homozygotes for the alternative alleles at these loci were relatively common (Table 3.4). Therefore, these mutations were unlikely to play a critical role in thyroid tumor development and were excluded as potential candidate causal variants for FCC in this study.

Table 3.4. Case-specific SNPs found in the *EIPRI* gene.

Chromosome	Position	Gene	Element	Genotype counts ¹		
				Case	Control	722 dogs ²
17	1869833	<i>EIPRI</i>	Downstream gene	11/0/0	0/10/1	38/93/536
17	1898881	<i>EIPRI</i>	4th intron	11/0/0	0/10/1	169/190/318
17	1898831	<i>EIPRI</i>	4th intron	11/0/0	0/10/1	37/117/528
17	1898919	<i>EIPRI</i>	4th intron	11/0/0	0/9/2	163/190/326
17	1901542	<i>EIPRI</i>	4th intron	11/0/0	0/10/1	112/170/376
17	1901551	<i>EIPRI</i>	4th intron	11/0/0	0/10/1	102/166/384
17	1901564	<i>EIPRI</i>	4th intron	11/0/0	0/10/1	104/172/373
17	1905133	<i>EIPRI</i>	4th intron	11/0/0	0/10/1	182/214/315
17	1933629	<i>EIPRI</i>	3rd intron	11/0/0	0/10/1	174/222/315
17	1948595	<i>EIPRI</i>	3rd intron	11/0/0	0/10/1	161/182/338

Note: ¹ count of recessive homozygotes, heterozygotes and homozygotes for the reference allele. ² 722 dogs from over 144 modern breeds, 54 wild canids and a hundred village dogs.

3.3.6. Validation by PCR-RFLP

To confirm the association between the two mutations in the *TPO* gene and the familial FCC, we genotyped chr17:800788G>A in a further 59 cases and 123 controls and chr17:805276C>T in 59 cases and 127 controls using PCR-RFLP (Table 3.5). Genotype of each dog can be found in Supplementary Figure S3.6. For both variants, 45 of 59 cases (76%) were homozygous and 9 (7%) and 10 (8%) controls were homozygous for the two variants, respectively. Totals of 83% and 82% of dogs with homozygous variant for the two mutations were affected, respectively. These totals suggest that these two mutations in the *TPO* gene were highly associated with the FCC with a p-value $< 2.2 \times 10^{-16}$ for the Fisher's exact test. Furthermore, chr17:800788G>A had lower SIFT and PROVEAN scores than chr17:805276C>T, and it is a novel mutation. Therefore, chr17:800788G>A was of more interest than chr17:805276C>T. Homozygous mutants of both variants had extremely high relative risk (16.94 and 16.64, respectively) compared to the homozygous genotype for the reference alleles. The heterozygous genotypes had a higher relative risk, but the differences were not significant. According to affection status and genotypes of the dogs in the pedigree in Figure 3.3, along with the long ROH segment in affected dogs, these two mutations were associated with the FCC in dogs in an autosomal recessive inheritance pattern. Among those 14 non-homozygous cases, 5 dogs were suspected cases without histology confirmation, which could be mis-diagnosed. In the remaining 9 cases with histology diagnosis, 6 dogs had ages at diagnosis beyond 10 years. One dog had unknown age at diagnosis. Only 2 dogs had age at diagnosis less than 10 years. The affected dogs with old age at diagnosis (we used a threshold of 10 years in this study) possibly had a somatic genetic causal mutation due to an environmental risk factor or aging. Ten of 124 unaffected dogs were homozygous. Among them, seven dogs were born after 2007 (four dogs after 2012). They were fewer than 12 years old at the moment of the data collection and could be affected at an older age.

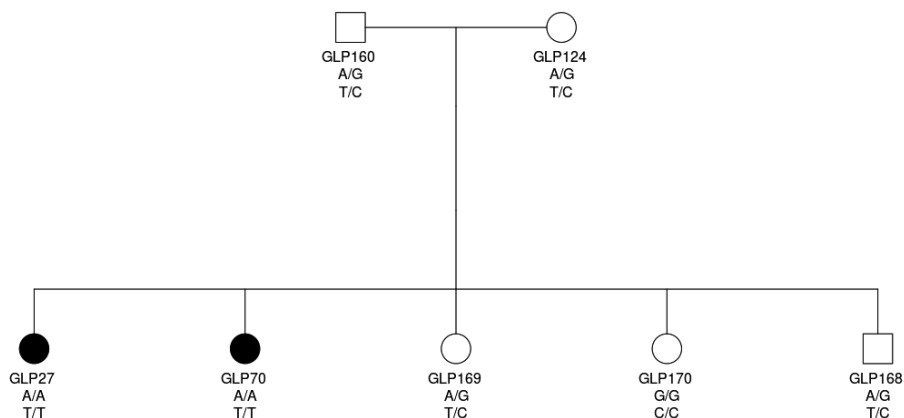


Figure 3.3. Genotypes of dogs suggesting a recessive trait of the disease. A circle denotes a female dog and a square denotes a male dog. Black background indicates that the dog is affected.

Table 3.5. Genotypes of the two deleterious SNPs in the *TPO* gene in the Dutch GLPs.

SNP	AA	AR	RR	Relative risk of AA	Relative risk of AR
chr17:800788G>A	45/9	11/56	3/58	16.94 (<i>p</i> -value 2.20e-16)	3.34 (<i>p</i> -value 0.07)
chr17:805276C>T	45/10	11/59	3/58	16.64 (<i>p</i> -value 2.24e-16)	3.20 (<i>p</i> -value 0.09)

Note: A represent alternative allele, and R represent reference allele. Numbers in the cell are the counts of affected and unaffected GLPs. The *p*-value derived from chi-square test.

3.4. Discussion

Many dog breeds experienced considerable inbreeding and show comparable diversity loss compared to other domestic species due to artificial selection, management in closed populations, and historical bottlenecks [50]. Inbreeding depression in the form of a variety of diseases and disorders is seen in many dog

breeds due to this loss of genetic diversity [51]. Based on pedigree data, we previously showed that affected dogs are more inbred than unaffected dogs [27]. This was confirmed by the inbreeding coefficients estimated from SNP array genotype data. Likewise, affected dogs have more ROH segments above 2 Mb in length (Supplementary Figure S3.7), which also implies that affected dogs exhibit more inbreeding.

We identified a region located between positions 0 and 5 Mb on chromosome 17 that is associated with FCC in the Dutch GLPs using a combination of GWAS and ROH analyses. In the affected dogs, this region showed a loss of diversity. The causal mutation located in a long ROH segment resulting from inbreeding was captured by ROH autozygosity analysis.

Using whole genome sequencing, we identified three rare deleterious mutations of interest within *SNTG2* and *TPO* genes, located in the long ROH region on chr17. Mutation Chr17:800788G>A, located in the *TPO* gene, was never reported. The other two have very low allele frequency in dogs from a variety of breeds. *SNTG2* is a syntrophin gene. The SNTG2 protein binds to components of mechanosensitive sodium channels and to the C termini of dystrophin, α -dystrobrevin, and β -dystrobrevin [52]. The SNTG2 gene is expressed in various tissues in humans and was reported to be associated with osteoporotic vertebral fracture [53] and autism [54]. These give the gene an unlikely role in the development of TC. Additionally, no common structure change of SNTG2 mRNA was found within the cases from the RNA-seq (Supplementary Figures S3.8 and S3.9). The deleterious mutations in the *TPO* gene are highly associated with the familial FCC in these dogs. However, the high linkage (5 Mb) in this study opens the possibility that other mutations in the noncoding regions could also explain the association. Nonetheless, *TPO* mutation chr17:800788G>A is of more interest with the lowest SIFT and PROVEAN scores and the aspect of it being a novel mutation.

The association between *TPO* gene mutations including intronic variants and missense variants and thyroid carcinoma was also seen in humans [55,56]. Mutations in the *TPO* gene also cause congenital goitrous primary hypothyroidism. Inactivating mutations in *TPO* gene were shown to cause the autosomal recessive trait congenital hypothyroidism in humans and dogs [57].

TPO is expressed specifically in thyroid gland tissue. Germline genetic alterations in other thyroid-specific genes were also associated with thyroid carcinoma. In humans, the rs965513[A] allele, which confers the greatest relative risk for the development of thyroid cancer, is located within an enhancer element controlling expression of *FOXE1*. *FOXE1* is a thyroid-specific transcription factor that regulates several genes involved in thyroid hormone production, including *TPO*, thyroglobulin (*TG*),

sodium-iodide symporter (*SLC5A5*), and dual oxidase (*DUOX2*). Furthermore, *TG* [58], *FOXE1* [59], and *DUOX2* [60] were also shown to be associated with thyroid cancer. Mutations in *SLC5A5* are associated with congenital hypothyroidism [61].

Although the *TPO* gene was identified to be associated with thyroid cancer and many other thyroid disorders, how *TPO* influences the risk of TC is still unclear. Up to 70% of TCs are caused by somatic mutations that activate the RAS/ERK mitogenic signaling pathway (MAPK/ERK) [21]. Upregulation of mitogen-activated protein kinase (MAPK) and phosphatidylinositol-3-kinase (PI3K)/Akt signaling pathways was reported to cause the thyroid gland tumorigenesis in dogs and humans [3]. Here, in dogs, the mechanism through which we identified deleterious mutations in the *TPO* gene influence risk of FCC may be more similar to that of the mutations associated with TC found in other thyroid-specific genes, for example, dysregulated hydrogen peroxide metabolism [60], because these genes work closely together to synthesize thyroid hormones.

We sequenced the RNA derived from the tumor tissue of seven affected dogs. The mRNA of the *TPO* gene was not interrupted (Supplementary Figure S3.6). However, it is not known whether the *TPO* mRNA expression in the tumor changed compared to the expression in normal canine thyroid gland. Likewise, this is not known for other genes in the long ROH within this region. However, mutations in the regulatory region, e.g., promoter and enhancer, could alter the expression level of the gene and thereby induce the tumor. Thus, further studies focusing on gene expression differences between affected and unaffected dogs are needed. Moreover, non-coding RNA, including lncRNA, microRNA, and circRNA, were shown to be involved in the tumorigenesis of thyroid tumor [1]. The lack of RNA-seq data from the normal thyroid tissue prohibited investigating the expression of the non-coding RNA genes.

Animal models for specific human diseases can contribute to research and treatment development. Many mouse models for thyroid cancer with varied genetic causes regarding different types of thyroid cancer were induced [62]. However, skepticism about their relevance with human thyroid cancer and their value in clinical translation is also presented mainly due to the difference between mice and humans and the fact that there is a very low translational rate of approximately 4–5% [63]. Dog models for some diseases were already introduced. For instance, a dog model for Alzheimer's disease was generated by overexpressing a mutated human amyloid precursor protein [64]. Likewise, a canine model of glycogen storage disease type Ia (GSDIa) was also described with causal mutations in the same gene [65]. Compared to rodent models, dog models have many advantages. For instance, dogs are more similar to human in genetics, physiology, and living environment compared to

rodents. Dogs receive good medical care, especially in developed countries; therefore, diseases in dogs are easily identified. Many dog breeds are predisposed to specific diseases, which can provide sufficient numbers of naturally affected dogs for research. Dogs can also benefit from research and treatment development using dog models. The treatment successfully developed from the model can also cure the disease in dogs.

3.5. Conclusions

In conclusion, we identified two deleterious recessive mutations in the *TPO* gene which are highly associated with the familial FCC in the Dutch GLPs. These findings provide a novel candidate gene and new insights to the tumorigenesis of FCC. A genetic test can be developed for veterinary diagnostic and selective breeding to eradicate this disease from the population. The dogs closely related to the affected dogs diagnosed are valuable to serve as a disease model for research or treatment development of TC caused by the alteration in genes involved in the thyroid function molecular pathway.

Author Contributions

Conceptualization, R.P.M.A.C.; data curation, Y.Y. and R.P.M.A.C.; formal analysis, Y.Y. and K.L.; funding acquisition, M.A.M.G. and R.P.M.A.C.; investigation, Y.Y.; methodology, Y.Y., H.B., Z.W. and R.P.M.A.C.; project administration, M.A.M.G. and R.P.M.A.C.; supervision, M.A.M.G. and R.P.M.A.C.; writing—original draft, Y.Y.; writing—review & editing, H.B., Z.W., M.A.M.G. and R.P.M.A.C. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by “Nederlands Kankerfonds voor Dieren”.

Data Availability Statement

Sequencing data presented in this study are openly available at EMBL-EBI ENA database with reference number PRJEB43017. SNP array genotype data are available through ArrayExpress (accession number E-MTAB-10241).

Acknowledgments

We acknowledge the financial support of the “Nederlands Kankerfonds voor Dieren”. We also thank the breeder association involved for their support of this study. We thank Mariska de Ruijscher and Johan de Vos for their contribution in sampling. Library preparation/sequencing were performed by Novogene (UK) Company Limited. We acknowledge support Xi Wan from the Novogene (UK)

Company Limited. Yun's PhD study was supported by scholarship #201806760044 from China Scholarship Council (CSC).

Conflicts of Interest

The authors declare no conflict of interest.

3.6. References

1. Cao, J.; Zhang, M.; Zhang, L.; Lou, J.; Zhou, F.; Fang, M. Non-coding RNA in thyroid cancer—Functions and mechanisms. *Cancer Lett.* 2021, 496, 117–126.
2. Bartsch, R.; Brinkmann, B.; Jahnke, G.; Laube, B.; Lohmann, R.; Michaelsen, S.; Neumann, I.; Greim, H. Human relevance of follicular thyroid tumors in rodents caused by non-genotoxic substances. *Regul. Toxicol. Pharmacol.* 2018, 98, 199–208.
3. Campos, M.; Kool, M.M.; Daminet, S.; Ducatelle, R.; Rutteman, G.; Kooistra, H.S.; Galac, S.; Mol, J.A. Upregulation of the PI3K/Akt pathway in the tumorigenesis of canine thyroid carcinoma. *J. Vet. Intern. Med.* 2014, 28, 1814–1823.
4. Czene, K.; Lichtenstein, P.; Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *Int. J. Cancer* 2002, 99, 260–266.
5. Hińcza, K.; Kowalik, A.; Kowalska, A. Current knowledge of germline genetic risk factors for the development of non-medullary thyroid cancer. *Genes* 2019, 10, 482.
6. Yokote, K.; Chanprasert, S.; Lee, L.; Eirich, K.; Takemoto, M.; Watanabe, A.; Koizumi, N.; Lessel, D.; Mori, T.; Hisama, F.M.; et al. Wrm mutation update: Mutation spectrum, patient registries, and translational prospects. *Hum. Mutat.* 2017, 38, 7–15.
7. Pereira, J.S.; da Silva, J.G.; Tomaz, R.A.; Pinto, A.E.; Bugalho, M.J.; Leite, V.; Cavaco, B.M. Identification of a novel germline foxe1 variant in patients with familial non-medullary thyroid carcinoma (fnmtc). *Endocrine* 2015, 49, 204–214.
8. He, H.; Bronisz, A.; Liyanarachchi, S.; Nagy, R.; Li, W.; Huang, Y.; Akagi, K.; Saji, M.; Kula, D.; Wojcicka, A.; et al. Srgap1 is a candidate gene for papillary thyroid carcinoma susceptibility. *J. Clin. Endocrinol. Metab.* 2013, 98, E973–E980.
9. Gara, S.K.; Jia, L.; Merino, M.J.; Agarwal, S.K.; Zhang, L.; Cam, M.; Patel, D.; Kebebew, E. Germline habp2 mutation causing familial nonmedullary thyroid cancer. *N. Engl. J. Med.* 2015, 373, 448–455.
10. Wójcicka, A.; Czetwertyńska, M.; Świerniak, M.; Długosińska, J.; Maciąg, M.; Czajka, A.; Dymecka, K.; Kubiak, A.; Kot, A.; Płoski, R.; et al. Variants in the atm-chek2-brca1 axis determine genetic predisposition and clinical presentation of papillary thyroid carcinoma. *Genes Chromosomes Cancer* 2014, 53, 516–523.
11. Siołek, M.; Cybulski, C.; Gąsior-Perczak, D.; Kowalik, A.; Kozak-Klonowska, B.; Kowalska, A.; Chłopek, M.; Kluźniak, W.; Wokołorczyk, D.; Pałyga, I.; et al. Chek2 mutations and the risk of papillary thyroid cancer. *Int. J. Cancer* 2015, 137, 548–552.
12. Gu, Y.; Yu, Y.; Ai, L.; Shi, J.; Liu, X.; Sun, H.; Liu, Y. Association of the atm gene polymorphisms with papillary thyroid cancer. *Endocrine* 2014, 45, 454–461.
13. Ngeow, J.; Ni, Y.; Tohme, R.; Song Chen, F.; Bebek, G.; Eng, C. Germline alterations in rasal1 in cowden syndrome patients presenting with follicular thyroid cancer and in individuals with apparently sporadic epithelial thyroid cancer. *J. Clin. Endocrinol. Metab.* 2014, 99, E1316–E1321.
14. Tomsic, J.; He, H.; Akagi, K.; Liyanarachchi, S.; Pan, Q.; Bertani, B.; Nagy, R.; Symer, D.E.; Blencowe, B.J.; de la Chapelle, A. A germline mutation in srrm2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Sci. Rep.* 2015, 5, 10566.
15. Ryu, R.A.; Tae, K.; Min, H.J.; Jeong, J.H.; Cho, S.H.; Lee, S.H.; Ahn, Y.H. Xrcc1 polymorphisms and risk of papillary thyroid carcinoma in a korean sample. *J. Korean Med. Sci.* 2011, 26, 991–995.

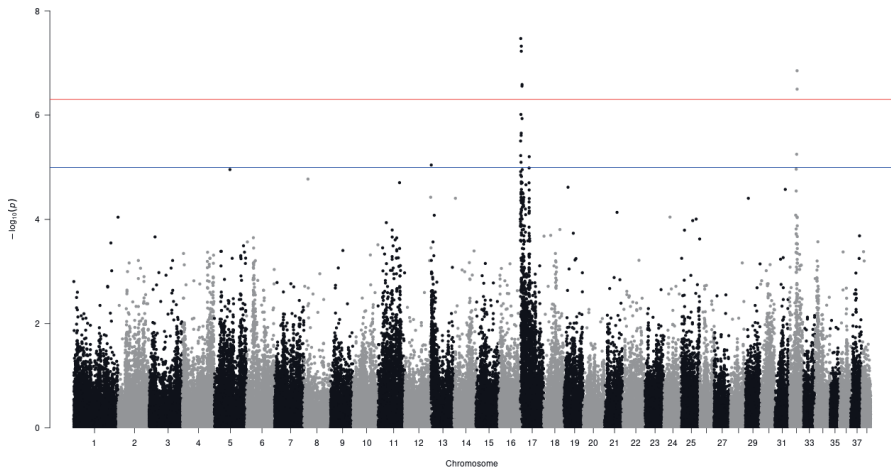
16. Jendrzewski, J.; He, H.; Radomska, H.S.; Li, W.; Tomsic, J.; Liyanarachchi, S.; Davuluri, R.V.; Nagy, R.; de la Chapelle, A. The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. USA* 2012, 109, 8646–8651.
17. Prevalence, clinicopathologic features, and somatic genetic mutation profile in familial versus sporadic nonmedullary thyroid cancer. *Thyroid* 2011, 21, 367–371.
18. Di Cristofaro, J.; Marcy, M.; Vasko, V.; Sebag, F.; Fakhry, N.; Wynford-Thomas, D.; de Micco, C. Molecular genetic study comparing follicular variant versus classic papillary thyroid carcinomas: Association of n-ras mutation in codon 61 with follicular variant. *Hum. Pathol* 2006, 37, 824–830.
19. Nikiforov, Y.E. Molecular analysis of thyroid tumors. *Mod. Pathol.* 2011, 24, S34–S43.
20. Montero-Conde, C.; Leandro-Garcia, L.J.; Chen, X.; Oler, G.; Ruiz-Llorente, S.; Ryder, M.; Landa, I.; Sanchez-Vega, F.; La, K.; Ghossein, R.A.; et al. Transposon mutagenesis identifies chromatin modifiers cooperating with Ras in thyroid tumorigenesis and detects ATXN7 as a cancer gene. *Proc. Natl. Acad. Sci. USA* 2017, 114, E4951.
21. Miguel, A.Z.; Adrián, A.-R.; Marta, M.; Piero, C.; Pilar, S. Regulators of the ras-erk pathway as therapeutic targets in thyroid cancer. *Endocr. Relat. Cancer* 2019, 26, R319–R344.
22. Liu, X.; Bishop, J.; Shan, Y.; Pai, S.; Liu, D.; Murugan, A.K.; Sun, H.; El-Naggar, A.K.; Xing, M. Highly prevalent tert promoter mutations in aggressive thyroid cancers. *Endocr. Relat. Cancer* 2013, 20, 603–610.
23. Liu, T.; Wang, N.; Cao, J.; Sofiadis, A.; Dinets, A.; Zedenius, J.; Larsson, C.; Xu, D. The age- and shorter telomere-dependent tert promoter mutation in follicular thyroid cell-derived carcinomas. *Oncogene* 2014, 33, 4978–4984.
24. Elisei, R.; Tacito, A.; Ramone, T.; Ciampi, R.; Bottici, V.; Cappagli, V.; Viola, D.; Matrone, A.; Lorusso, L.; Valerio, L.; et al. Twenty-five years experience on ret genetic screening on hereditary mtc: An update on the prevalence of germline ret mutations. *Genes* 2019, 10, 698.
25. Lee, J.J.; Larsson, C.; Lui, W.O.; Höög, A.; von Euler, H. A dog pedigree with familial medullary thyroid cancer. *Int. J. Oncol.* 2006, 29, 1173–1182.
26. Sturgeon, C.; Clark, O.H. Familial nonmedullary thyroid cancer. *Thyroid* 2005, 15, 588–593.
27. Yu, Y.; Krupa, A.; Keesler, R.; Grinwis, G.C.M.; Ruijscher, M.; Groenen, M.A.M.; Croijmans, R.P.M.A. Familial thyroid follicular cell carcinomas in a large number of dutch german longhaired pointers. *bioRxiv* 2021.
28. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007, 81, 559–575.
29. Andrews, S. Fastqc: A Quality Control Tool for High Throughput Sequence Data [online]. Available online: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 1 June 2020).
30. Li, H.; Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 2009, 25, 1754–1760.
31. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics* 2009, 25, 2078–2079.
32. 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. From fastq data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 2013, 43, 11.10.11–11.10.33.
33. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* 2012, arXiv:1207.3907.
34. Li, H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011, 27, 2987–2993.

35. Chen, X.; Schulz-Trieglaff, O.; Shaw, R.; Barnes, B.; Schlesinger, F.; Källberg, M.; Cox, A.J.; Kruglyak, S.; Saunders, C.T. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016, 32, 1220–1222.
36. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. *Genome Biol.* 2016, 17, 122.
37. Buels, R.; Yao, E.; Diesh, C.M.; Hayes, R.D.; Munoz-Torres, M.; Helt, G.; Goodstein, D.M.; Elisk, C.G.; Lewis, S.E.; Stein, L.; et al. Jbrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol.* 2016, 17, 66.
38. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nat. Biotechnol.* 2019, 37, 907–915.
39. Liao, Y.; Smyth, G.K.; Shi, W. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014, 30, 923–930.
40. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. *Nat. Biotechnol.* 2011, 29, 24–26.
41. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 2012, 44, 821–824.
42. Ng, P.C.; Henikoff, S. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003, 31, 3812–3814.
43. Choi, Y.; Chan, A.P. Provean web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015, 31, 2745–2747.
44. Tang, H.; Thomas, P.D. Panther-psep: Predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics* 2016, 32, 2230–2232.
45. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Methods* 2010, 7, 248–249.
46. Plassais, J.; Kim, J.; Davis, B.W.; Karyadi, D.M.; Hogan, A.N.; Harris, A.C.; Decker, B.; Parker, H.G.; Ostrander, E.A. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.* 2019, 10, 1489.
47. Ruf, J.; Carayon, P. Structural and functional aspects of thyroid peroxidase. *Arch. Biochem. Biophys.* 2006, 445, 269–277.
48. Williams, D.E.; Le, S.N.; Hoke, D.E.; Chandler, P.G.; Gora, M.; Godlewska, M.; Banga, J.P.; Buckle, A.M. Structural studies of thyroid peroxidase show the monomer interacting with autoantibodies in thyroid autoimmune disease. *Endocrinology* 2020, 161, bqaa016.
49. Gershlick, D.C.; Schindler, C.; Chen, Y.; Bonifacio, J.S. Tsscl1 is novel component of the endosomal retrieval machinery. *Mol. Biol Cell* 2016, 27, 2867–2878.
50. Leroy, G. Genetic diversity, inbreeding and breeding practices in dogs: Results from pedigree analyses. *Vet. J.* 2011, 189, 177–182.
51. Ujvari, B.; Klaassen, M.; Raven, N.; Russell, T.; Vittecoq, M.; Hamede, R.; Thomas, F.; Madsen, T. Genetic diversity, inbreeding and cancer. *Proc. R. Soc. B Biol. Sci.* 2018, 285, 20172589.
52. Piluso, G.; Mirabella, M.; Ricci, E.; Belsito, A.; Abbondanza, C.; Servidei, S.; Puca, A.A.; Tonali, P.; Puca, G.A.; Nigro, V. Gamma1- and gamma2-syntrophins, two novel dystrophin-binding proteins localized in neuronal cells. *J. Biol. Chem.* 2000, 275, 15851–15860.
53. Neto, L.H.J.; Wicik, Z.; Torres, G.H.F.; Takayama, L.; Caparbo, V.F.; Lopes, N.H.M.; Pereira, A.C.; Pereira, R.M.R. Overexpression of sntg2, traf3ip2, and itga6 transcripts is associated with osteoporotic vertebral fracture in elderly women from community. *Mol. Genet. Genom. Med.* 2020, 8, e1391.
54. Rosenfeld, J.A.; Ballif, B.C.; Torchia, B.S.; Sahoo, T.; Ravnan, J.B.; Schultz, R.; Lamb, A.; Bejjani, B.A.; Shaffer, L.G. Copy number variations associated with autism spectrum disorders contribute to a spectrum of neurodevelopmental disorders. *Genet. Med.* 2010, 12, 694–702.

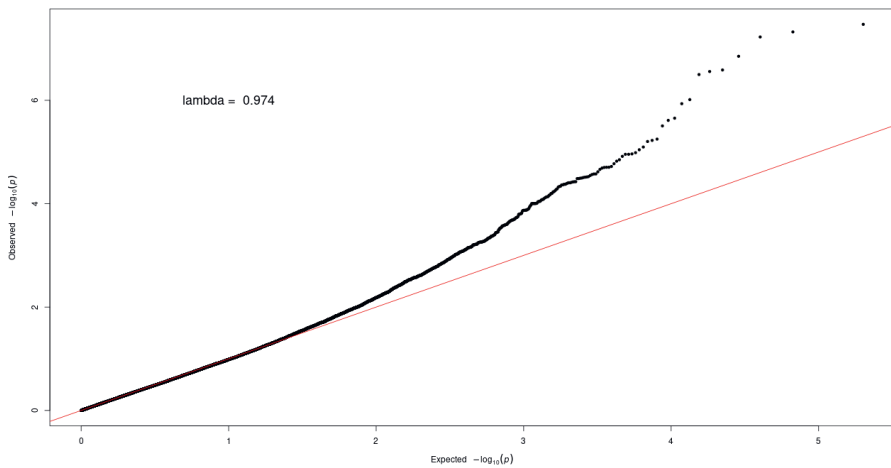
55. Cipollini, M.; Pastor, S.; Gemignani, F.; Castell, J.; Garritano, S.; Bonotti, A.; Biarnés, J.; Figlioli, G.; Romei, C.; Marcos, R.; et al. Tpo genetic variants and risk of differentiated thyroid carcinoma in two european populations. *Int. J. Cancer* 2013, 133, 2843–2851.
56. Zhu, H.; Peng, Y.G.; Ma, S.G.; Liu, H. Tpo gene mutations associated with thyroid carcinoma: Case report and literature review. *Cancer Biomark.* 2015, 15, 909–913.
57. Fyfe, J.C.; Lynch, M.; Olsen, J.; Louër, E. A thyroid peroxidase (tpo) mutation in dogs reveals a canid-specific gene structure. *Mamm. Genome* 2013, 24, 127–133.
58. Pastor, S.; Akdi, A.; González, E.R.; Castell, J.; Biarnés, J.; Marcos, R.; Velázquez, A. Common genetic variants in pituitary-thyroid axis genes and the risk of differentiated thyroid cancer. *Endocr. Connect.* 2012, 1, 68–77.
59. Son, H.-Y.; Hwangbo, Y.; Yoo, S.-K.; Im, S.-W.; Yang, S.D.; Kwak, S.-J.; Park, M.S.; Kwak, S.H.; Cho, S.W.; Ryu, J.S.; et al. Genome-wide association and expression quantitative trait loci studies identify multiple susceptibility loci for thyroid cancer. *Nat. Commun.* 2017, 8, 15966.
60. Bann, D.V.; Jin, Q.; Sheldon, K.E.; Houser, K.R.; Nguyen, L.; Warrick, J.I.; Baker, M.J.; Broach, J.R.; Gerhard, G.S.; Goldenberg, D. Genetic variants implicate dual oxidase-2 in familial and sporadic nonmedullary thyroid cancer. *Cancer Res.* 2019, 79, 5490–5499.
61. Watanabe, Y.; Ebrhim, R.S.; Abdullah, M.A.; Weiss, R.E. A novel missense mutation in the *slc5a5* gene in a sudanese family with congenital hypothyroidism. *Thyroid* 2018, 28, 1068–1070.
62. Kirschner, L.S.; Qamri, Z.; Kari, S.; Ashtekar, A. Mouse models of thyroid cancer: A 2015 update. *Mol. Cell Endocrinol.* 2016, 421, 18–27.
63. Jin, Y.; Liu, M.; Sa, R.; Fu, H.; Cheng, L.; Chen, L. Mouse models of thyroid cancer: Bridging pathogenesis and novel therapeutics. *Cancer Lett.* 2020, 469, 35–53.
64. Lee, G.S.; Jeong, Y.W.; Kim, J.J.; Park, S.W.; Ko, K.H.; Kang, M.; Kim, Y.K.; Jung, E.M.; Moon, C.; Hyun, S.H.; et al. A canine model of alzheimer's disease generated by overexpressing a mutated human amyloid precursor protein. *Int. J. Mol. Med.* 2014, 33, 1003–1012.
65. Specht, A.; Fiske, L.; Erger, K.; Cossette, T.; Verstegen, J.; Campbell-Thompson, M.; Struck, M.B.; Lee, Y.M.; Chou, J.Y.; Byrne, B.J.; et al. Glycogen storage disease type ia in canines: A model for human metabolic and genetic liver disease. *J. Biomed. Biotechnol.* 2011, 2011, 646257.

3.7. Supplementary materials

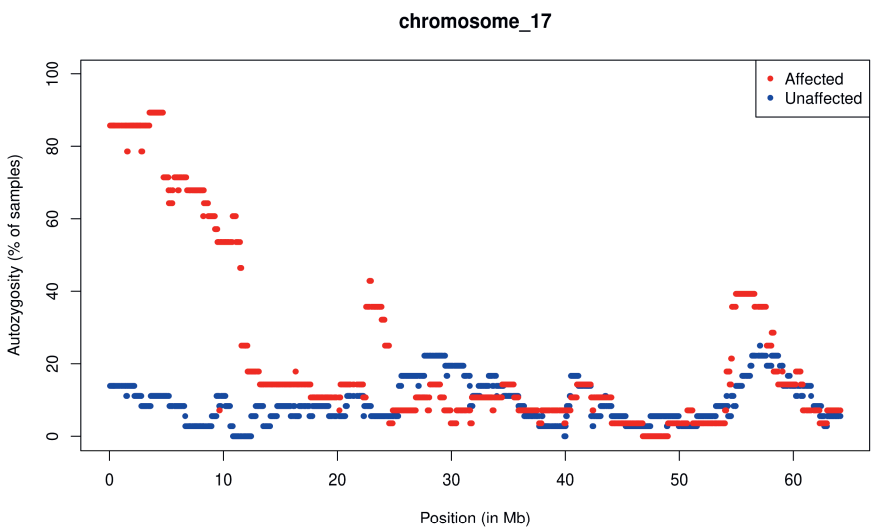
Supplementary Table S3.1 is available through this link https://github.com/YunYu93/Data-depository/blob/main/Supplementary_Table_S3.1.xlsx.



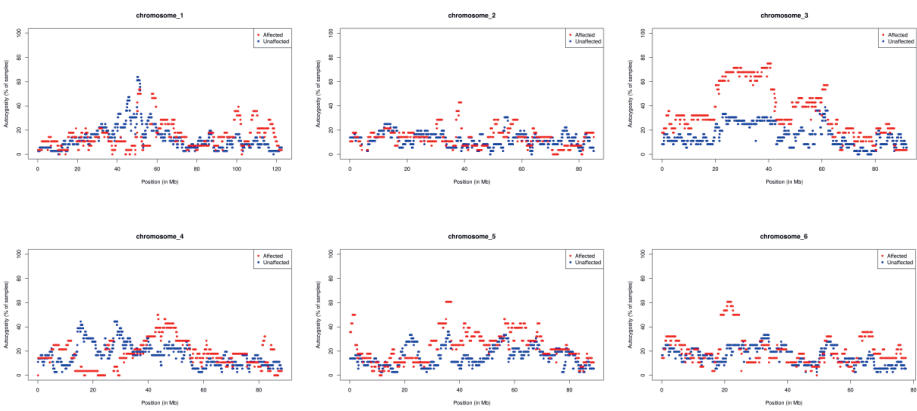
Supplementary Figure S3.1. Manhattan plot of GWAS result across the whole genome.



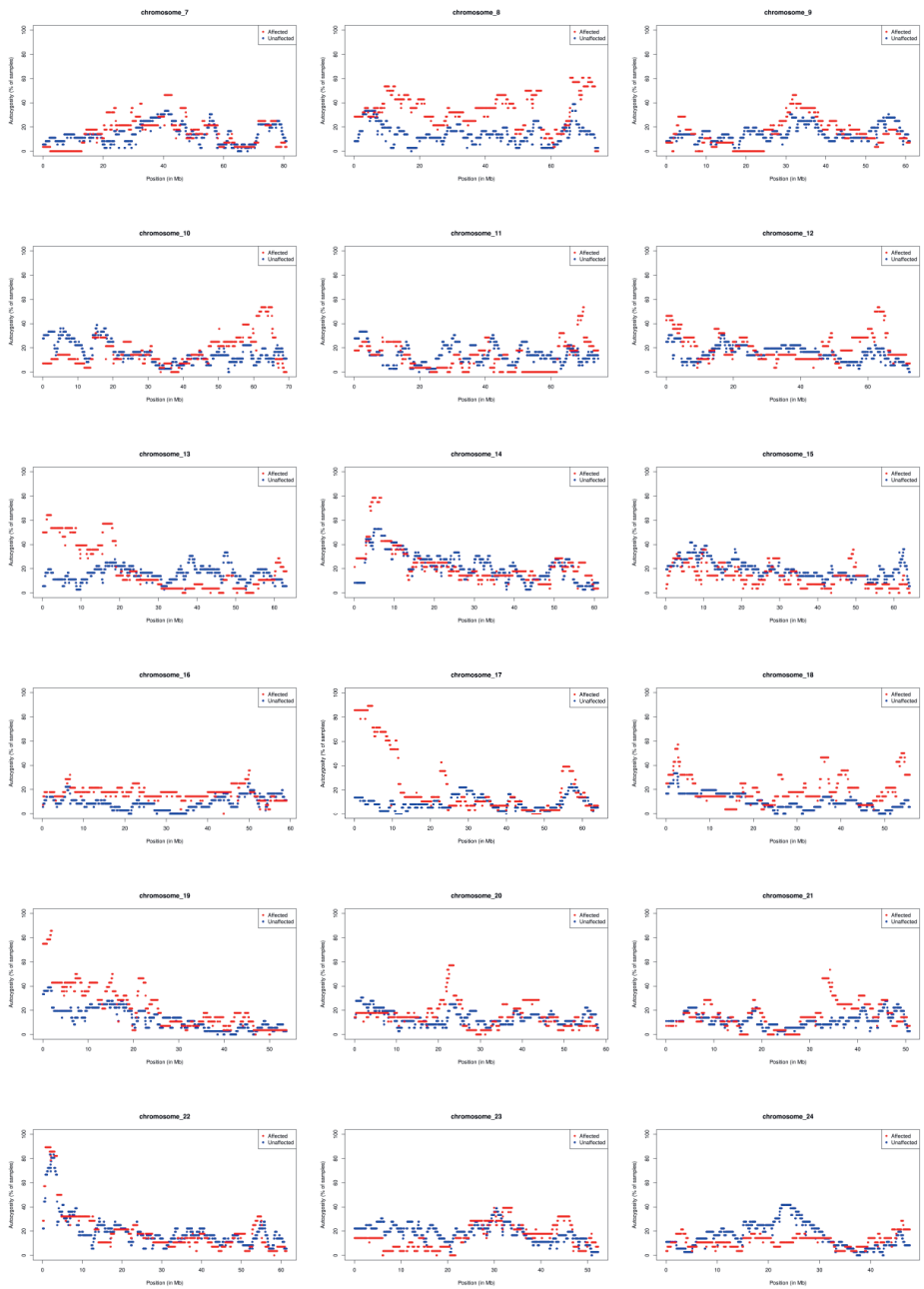
Supplementary Figure S3.2. QQ plot of the GWAS result. Inflation value is 0.974.



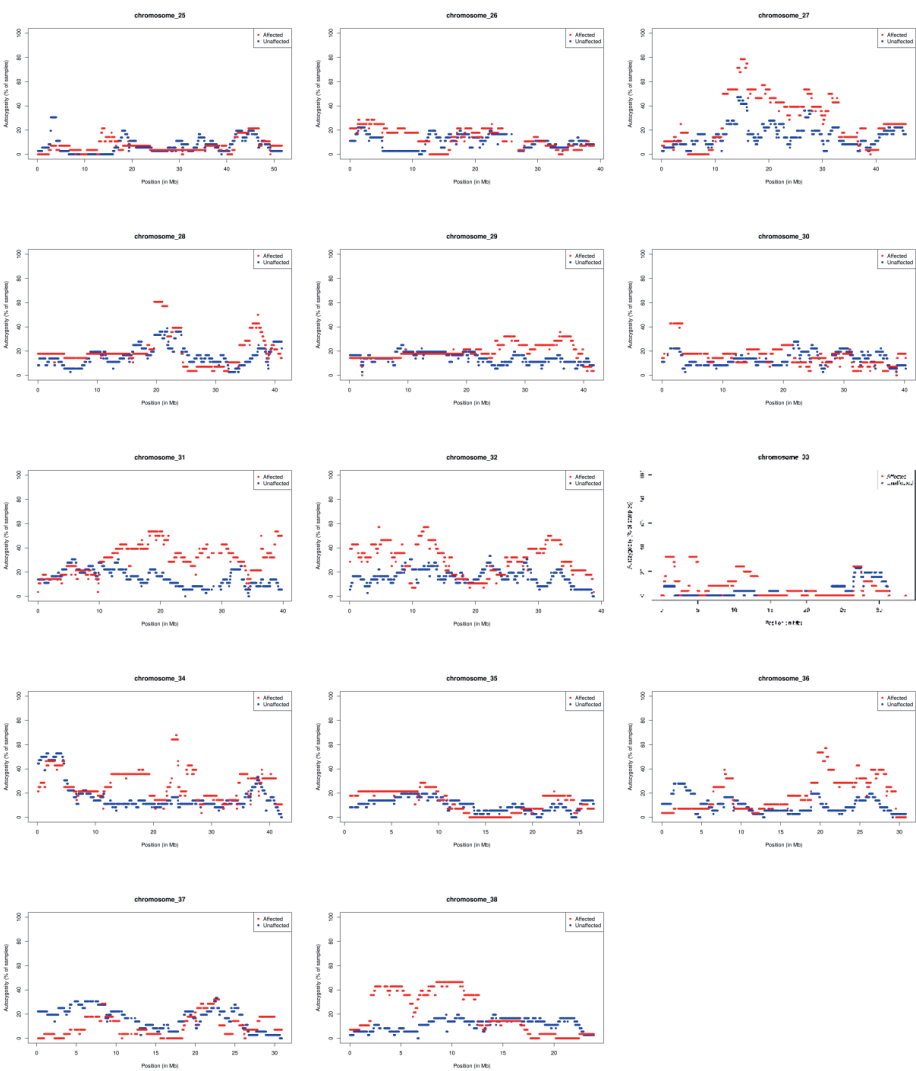
Supplementary Figure S3.3. Autozygosity of ROH segments on chromosome 17 between the affected and unaffected GLPs.



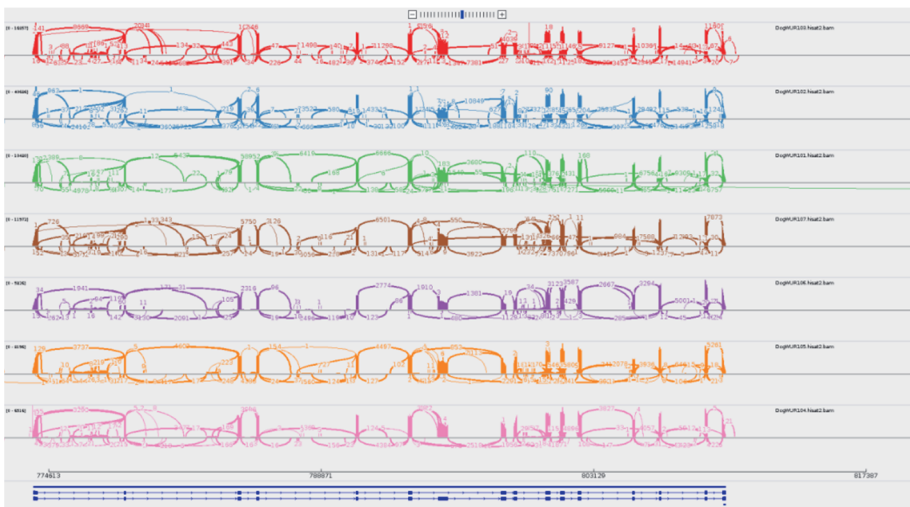
3 | Germline risk mutations in the *TPO* gene



3 | Germline risk mutations in the *TPO* gene

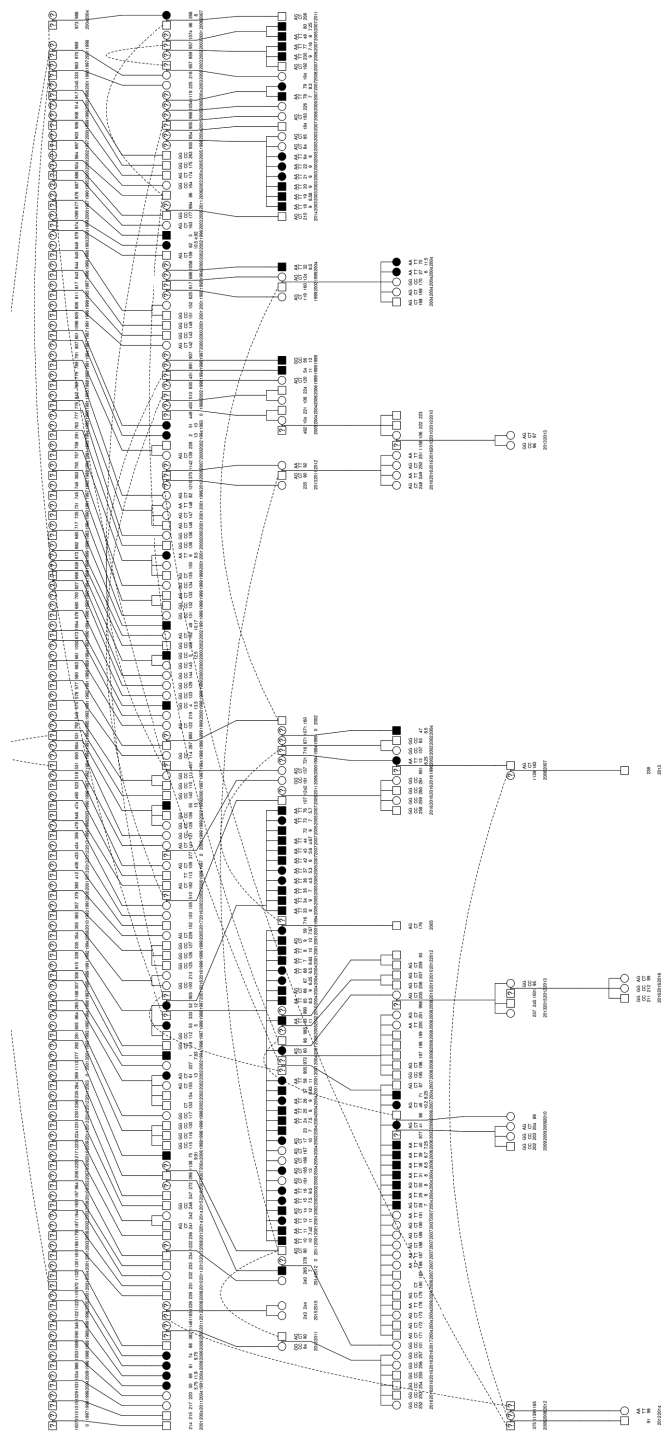


Supplementary Figure S3.4. Autozygosity of ROH segments in affected and unaffected GLPs on each autosomal chromosome (1-38).

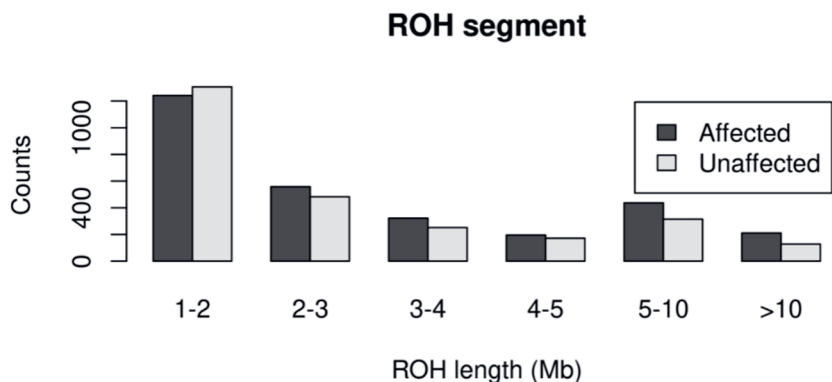


Supplementary Figure S3.5. Sashimi plot of the *TPO* mRNA in 7 thyroid tumor tissues with RNA-seq data.

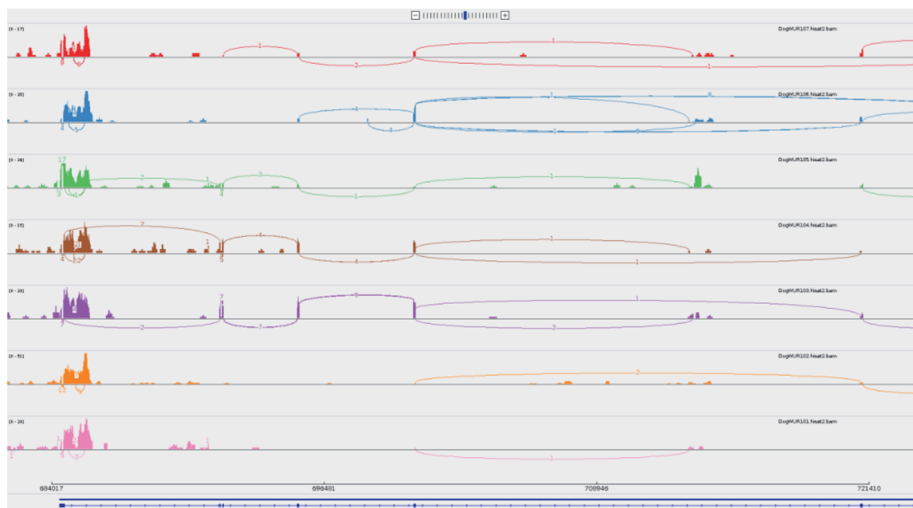
3 | Germline risk mutations in the *TPO* gene



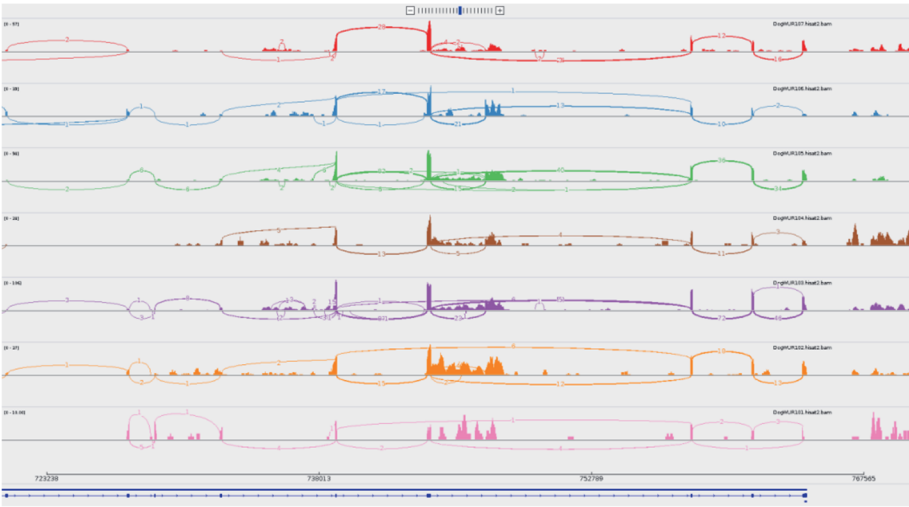
Supplementary Figure S3.6. Pedigree of dogs genotyped. Dog in black color were affected. A question mark denotes the unknown affected status. Five rows of label below the circle or square are genotype of chr17:800788G>A, genotype of chr17:800788G>T, ID of dog, diagnoses age, and year of birth, respectively.



Supplementary Figure S3.7. The Counts of ROH segments based on SNP array genotype data in different lengths between affected and unaffected GLPs. Affected dogs have more ROH segments with length above 2 Mb.



Supplementary Figure S3.8. Sashimi plot of the first 6 exons of SNTG2 mRNA in 7 thyroid tumor tissues with RNA-seq data.



Supplementary Figure S3.9. Sashimi plot of the last 9 exons of SNTG2 mRNA in 7 thyroid tumor tissues with RNA-seq data.

4

A recurrent somatic missense mutation in the *GNAS* gene identified in familial thyroid follicular cell carcinomas in German longhaired pointer dogs

Yun Yu¹, Freek Manders², Guy C M Grinwis³, Martien A.M. Groenen¹,
Richard P.M.A. Crooijmans¹

¹Wageningen University & Research, Animal Breeding and Genomics,
Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

²Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25,
3584CS Utrecht, The Netherlands

³Department of Biomolecular Health Sciences, Division of Pathology, Faculty of
Veterinary Medicine, Utrecht University, Yalelaan 1, Utrecht, The Netherlands

Abstract

Background

We previously reported a familial thyroid follicular cell carcinoma (FCC) in a large number of Dutch German longhaired pointers and identified two deleterious germline mutations in the *TPO* gene associated with disease predisposition. However, the somatic mutation profile of the FCC in dogs has not been investigated at a genome-wide scale.

Results

Herein, we comprehensively investigated the somatic mutations that potentially contribute to the inherited tumor formation and progression using high depth whole-genome sequencing. A *GNAS* A204D missense mutation was identified in 4 out of 7 FCC tumors by whole-genome sequencing and in 20 out of 32 dogs' tumors by targeted sequencing. In contrast to this, in the human TC, mutations in *GNAS* gene have lower prevalence. Meanwhile, the homologous somatic mutation in humans has not been reported. These findings suggest a difference in the somatic mutation landscape between TC in these dogs and human TC. Moreover, tumors with the *GNAS* A204D mutation had a significantly lower somatic mutation burden in these dogs. Somatic structural variant and copy number alterations were also investigated, but no potential driver event was identified.

Conclusion

This study provides novel insight in the molecular mechanism of thyroid carcinoma development in dogs. German longhaired pointers carrying *GNAS* mutations in the tumor may be used as a disease model for the development and testing of novel therapies to kill the tumor with driver mutations in the *GNAS* gene.

Key words: familial cancer, thyroid carcinoma, driver mutation, dog, *GNAS*, mutational signature, whole genome sequencing

4.1. Background

We previously reported familial thyroid follicular cell carcinomas (FCCs) in 54 Dutch German longhaired pointers (GLPs), identified by histological examination and an additional 29 dogs were suspected to be affected based on typical clinical signs [1]. The familial FCC was heterogeneous with 5 different histological subtypes: follicular thyroid carcinoma (FTC), papillary thyroid carcinoma (PTC), compact thyroid carcinoma (CTC), follicular-compact thyroid carcinoma (FCTC), and carcinosarcoma. Two homozygous deleterious mutations in the *TPO* gene were identified to be the germline risk factors of FCC predisposition in these dogs, based on a genome-wide association study (GWAS) and whole-genome sequencing (WGS) analysis [2]. However, besides the germline risk factors, key somatic mutations (driver mutations) also play an important role. These mutations can lead to uncontrolled cell division, escape from apoptosis, immune evasion and accelerated tumor growth [3, 4]. Identifying these driver mutations can contribute to unraveling the molecular mechanism of tumorigenesis.

Canine FCC is, in morphology, highly similar to human thyroid carcinomas originating from follicular cells. The GLP dogs with FCC could be an important disease model for thyroid cancer (TC) research and therapy development in dogs and humans. The somatic mutation landscapes of follicular cell thyroid carcinoma in humans have been extensively investigated [5-7]. In human thyroid cancer, the *BRAF* V600E somatic mutation is the most common driver mutation. Other frequently identified mutations in human thyroid cancers are within the *RET*, *PTEN*, and the *RAS* gene family (*KRAS*, *NRAS* and *HRAS*) [8]. In contrast to humans, the somatic mutation landscape of TC in dogs is still unclear. Other studies have identified genes with somatic mutations in a variety of canine tumors with known gene roles in human cancers. Examples of the driver genes in tumorigenesis present in both species are: the homologous mutation to the human somatic *BRAF* V600E mutation in naturally occurring canine bladder cancer [9], the *FBXW7* mutation in lymphomas [10], the recurrent somatic *SETD2* mutation in osteosarcomas [11], the somatic *TP53* and *PIK3CA* mutations in multiple cancers [12], the somatic mutations in *TP53*, *PDGFRA*, *PIK3CA*, *EGFR* and *IDH1* in sporadic gliomas [13], and the *NRAS*, *FAT4*, *PTEN* and *TP53* mutations in melanomas [14].

In this study, we generated whole genome sequencing data from both the tumor tissue and the matched animal genome. The 7 animals included in this study were closely related and are homozygous for the *TPO* variants associated with the disease we reported in our previous study [2]. Somatic single nucleotide variants (SNVs), structural variants (SVs), and copy number alterations (CNAs) were investigated. The somatic mutation landscape was further investigated, including somatic

mutational burden, driver genes or significantly mutated genes and mutational signatures. Meanwhile, several tumor tissue characteristics were also investigated, including purity, ploidy, telomere length and subclone cluster. Most interestingly, we identified a recurrent missense mutation in the *GNAS* gene, which is a novel driver mutation and correlates with a lower tumor mutational burden. Unveiling the somatic driver mutations, along with previous identification of germline risk factor in the *TPO* gene, reveals the genetic bases of this familial FCC, which could help us to understand the tumor development and enhance the use of these dogs as a disease model.

4.2. Results

4.2.1. WGS and somatic mutation landscape

Whole-genome sequencing was performed on both tumor (FCC tissues) and matched normal (derived from blood DNA) samples from 7 GLP dogs. Additionally, RNA-seq was performed on each tumor sample. The 7 GLPs are closely related (Figure 4.1). GLP36, 37 and 44 are full siblings where GLP48 and GLP77 are half siblings. GLP39 is the nephew of GLP25, and half-sibling of GLP36, 37 and 44.

The age at diagnosis of the FCC ranged between 4.5 and 8 years (Table 4.1). Five dogs were males and two were females (GLP 36 and 37). All 7 dogs were homozygous for the mutations in the *TPO* gene associated with the disease identified in our previous study [2]. The histological subtypes of FCCs include 3 FTCs, 2 FCTCs, 1 CTC, and 1 carcinosarcoma. The familial FCCs of these 7 dogs were heterogeneous in histology but were supposed to result from the same germline genetic risk factor.

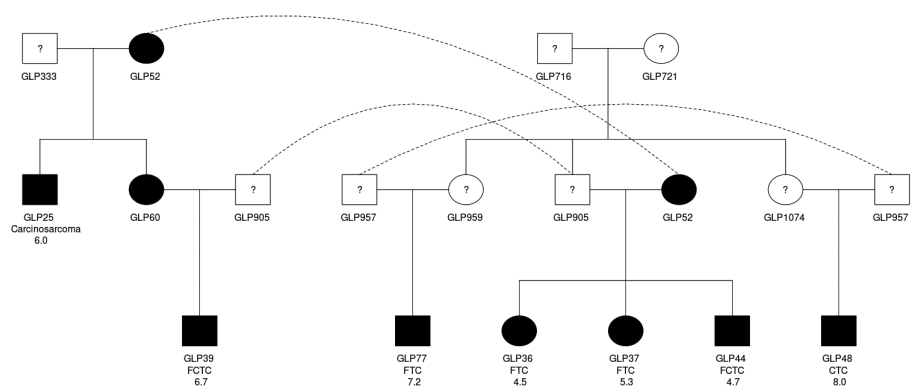


Figure 4.1. The pedigree of the 7 familial FCCs in this study. A circle and square denotes a female and a male dog respectively. Solid black indicates that the dog was affected. A question mark indicates that the disease status of the dog is unknown. The 3 rows of texts below a circle and square denote ID, thyroid cancer subtype, and age at diagnosis respectively. A dotted line indicates an identical dog.

Table 4.1. Sample information.

Animal ID	Subtype	Age at diagnosis (years)	Sex ^a	DogWUR ID - tumor (coverage)	DogWUR ID - normal (coverage)
GLP77	L:FTC; R:Adenoma	7.2	M	108 (68x)	115 (32x)
GLP48	L:CTC; R:FTC	8.0	M	109 (81x)	116 (18x)
GLP36	L:FTC; R:FCTC with bone	4.5	FS	110 (65x)	117 (17x)
GLP37	L:FTC; R:FTC	5.3	FS	111 (72x)	118 (20x)
GLP44	L:FCTC; R:FCTC	4.7	MC	112 (68x)	119 (35x)
GLP25	L:Carcinosarcoma; R:FCTC	6.0	M	113 (83x)	91 (32x)
GLP39	L:FCTC; R:FCTC	6.7	M	114 (75x)	120 (42x)

Note: ^a M denotes male, FS denotes female spayed, MC denotes male castrated.

4.2.2. Somatic mutation burden

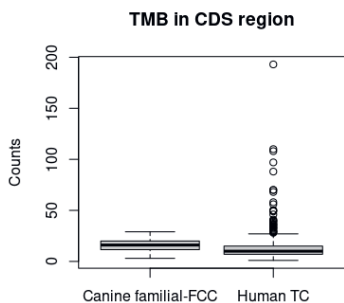
We identified 10,216 somatic SNVs, 1,034 small insertions, and 1,558 small deletions from the WGSs of the 7 GLPs, using our consensus calling method. On average, there were 10 (2 - 15) somatic mutations per sample that modified a protein, most of which were missense mutations (supplementary Figure S4.3A). The true positive calling rate among the somatic SNVs and Indels was estimated to be 0.83 by visual inspection. Furthermore, somatic SNVs have a higher true positive rate than Indels ($0.90 > 0.57$), similar to previous findings [15]. Transition to transversion ratio was between 0.91 and 1.81, with an average of 1.23 (supplementary Figure S4.3B).

Based on WGS, the average tumor mutation burden (TMB) of SNVs was estimated to be 0.58 (ranges 0.05 - 1.15) (mutations per megabase), which was estimated by the number of somatic mutations divided by the total length of the canine genome (2,500 Mb). The correlation between diagnosis age and tumor TMB was not significant ($R^2=0.04$, $p\text{-value}=0.68$, pearson correlation) (Supplementary Figure S4.4).

We calculated the number of mutations that occurred in coding regions (CDS) for all 7 tumors and compared them to the human thyroid cancer (THCA) data from The Cancer Genome Atlas (TCGA) program. We found no significant difference between humans and dogs for this tumor type (Figure 4.2A) (Welch t-test, $p\text{-value}$ 0.4932).

A

B



C

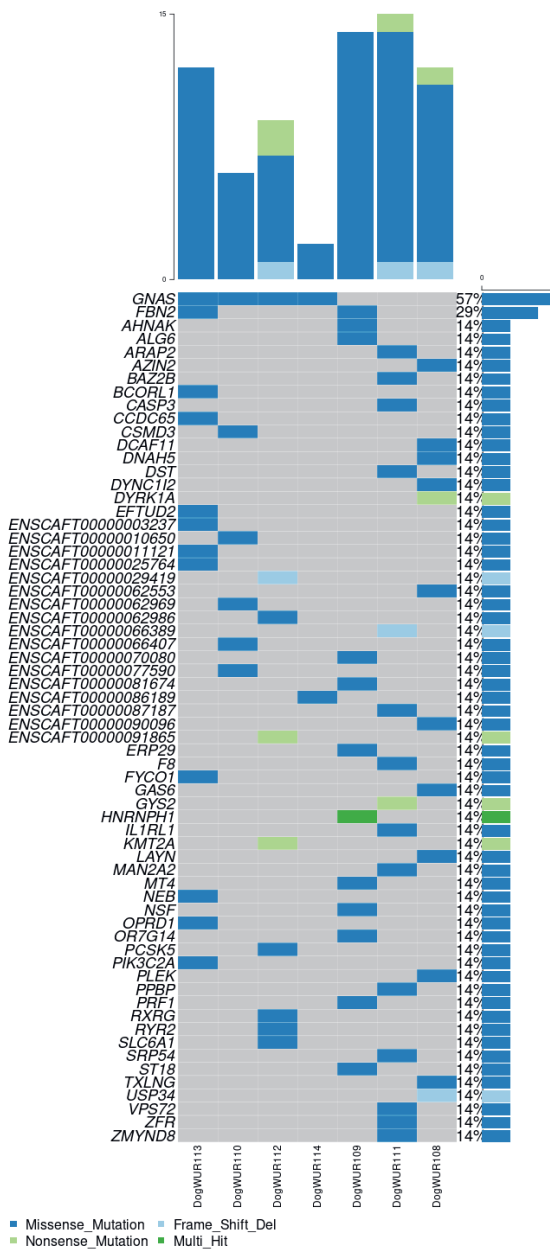
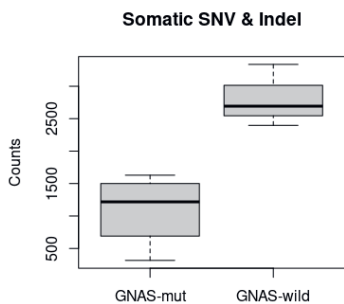


Figure 4.2. A. comparison of number of mutations identified in CDS region between canine tumors and human thyroid tumors from TCGA. B. Mutation landscape of the 7 tumors. Each column represents one sample, each row represents one somatic mutated gene. The right bar chart represents the frequencies of gene alterations across the 7 tumors. The upper bar chart represents the number of mutations in exons across the 7 tumors. C. Number of somatic SNVs and Indels identified in 4 tumors with the *GNAS* A204D mutation (*GNAS*-mut) and 3 samples without that somatic mutation (*GNAS*-wild).

4.2.3.Driver events

GNAS and *HNRNPFI* were identified as being significantly mutated genes (SMG) by MuSiC2. The *GNAS* gene was also identified by the dNdScv algorithm. There was only one missense mutation A204D (chr24:43657087C>A) in the *GNAS* gene (exon 9), which was identified in 4 of 7 tumor samples (Figure 4.2B). This mutation in the DogWUR110 sample (GLP36) was not identified by our consensus calling method but was rescued by the visual inspection. The variant allele frequency of the *GNAS* mutation was 0.28 on average (0.14 - 0.43). The RNA-seq data also supported the *GNAS* mutation in those 4 dogs. The somatic mutations identified in the *HNRNPFI* gene turned out to be false positive calls after visual inspection.

Although these familial FCCs have the same germline susceptibility, the somatic mutation landscape seems to be heterogeneous. The somatic mutation in the *GNAS* gene presented in 4 out of the 7 GLPs. The other 3 dogs (GLP37, 48, 77) had unclear driver events, suggesting heterogeneity of driver mutations among these samples. Driver mutations frequently identified in humans, such as mutations in the genes *BRAF*, *RAS*, *TP53*, *PTEN*, were not observed in the dogs used in this study (Figure 4.2B). Somatic structural variants and copy number alterations were also investigated, but no driver gene was identified from them (details in supplementary).

Canine *GNAS* A204 corresponds to human *GNAS* A249 (protein accession: P63092). The homologous mutation in humans at chr20:58909711 has not been reported in human dbSNP [16]. Canine *GNAS* A204D was predicted to be possibly damaging with a value of 0.552 by PolyPhen-2, and deleterious by PROVEAN with a score of -5.460. The *GNAS* A204 was conserved across species with a conservation score of 8 (range 1-9) according to estimation of ConSurf. Furthermore, the amino acid A (Ala) is non-polar while D (Asp) is polar and hydrophobic with a negative charge. These observations suggest that the mutation may have a big impact on the function of the *GNAS* protein. Canine *GNAS* A204 is a novel mutation in dogs, not reported either in dbSNP nor in the 722 dog genome panel.

Interestingly, we found that the tumors with the *GNAS* mutation (GNAS-mut) had significantly less somatic SNVs and Indels compared to tumors without the *GNAS* mutation (GNAS-wild) (Welch t-test, p-value 0.0082) (Figure 4.2C), likewise they also contained less protein-modifying somatic mutations.

To identify the molecular signaling pathways promoting tumorigenesis, we performed a gene expression differentiation analysis contrasting tumors with and without the somatic *GNAS* mutation. PCA analysis didn't identify a clear distinction between GNAS-mut and GNAS-wild samples (Supplementary Figure S4.5A). Moreover, there was no difference in expression level of the *GNAS* gene between these two groups (Supplementary Figure S4.5B). Differential gene expression analysis contrasting the 4 GNAS-mut samples and 3 GNAS-wild samples identified 53 differentially expressed genes by a significant threshold of 0.05 adjusted p-value. But no enriched biological process or KEGG pathway was identified by pathway enrichment analysis based on these 53 differentially expressed genes. This suggests that there is probably no difference in the molecular signaling pathways that contribute to the tumorigenesis between these two groups of dogs.

4.2.4. Association with morphological characteristics

GNAS mutation might be associated with an increased proliferation rate according to semiquantitative evaluation of the number of mitotic figures in the neoplasms. However, additional, quantitative analysis including the use of a proliferation marker such as Ki67 is necessary to determine whether there is a relationship between proliferation and the *GNAS* mutation. Other histological characteristics appear not to be associated with the mutation.

4.2.5. Prevalence of the *GNAS* mutation

To identify the prevalence of the *GNAS* A204D somatic mutation among dogs suffering from FCC, we genotyped 49 tumor samples from 34 affected dogs using Sanger sequencing (supplementary Table S4.1). Of 13 dogs, tumor tissue from both the left and right thyroid gland was genotyped successfully. However, genotyping failed in 4 samples from 4 dogs (2 of them had bilateral tumors). Finally, we obtained 45 genotypes covering 32 GLPs. We found that tumors from 20 of the 32 affected dogs had this *GNAS* somatic mutation, resulting in a prevalence of 62.5%. We also performed Sanger sequencing on normal DNA (obtained from blood) of the 7 dogs used in this study and none of them captured the *GNAS* mutation, confirming that these were somatic mutations. To ensure further that the *GNAS* A204D mutation is generally not present as a germline variant, we checked that this mutation was not present in the whole genome sequences that we previously obtained from blood DNA of 22 GLPs [2].

The germline genotype of the marker in the *TPO* gene associated with FCC was available for 31 dogs (1 of 32 GLPs with *GNAS* A204D typed failed in *TPO* germline mutation typing) (supplementary Table S4.1). We tested whether the *TPO* germline mutation and *GNAS* somatic mutation were significantly correlated, but this was not the case (Chi-square test; p -value = 0.237).

4.2.6. Telomere length

Tumor genomes have significantly shorter telomeres compared to normal genomes according to our estimation using TelSeq (Figure 4.3A), which is in concordance with our knowledge about telomere shrinkage in tumor cells [17]. There is no significant correlation between telomere length and diagnosis age of FCC (Figure 4.3B). To maintain the length of the telomeres in the tumor cells, telomerase is often activated. While in these 7 tumor samples, *TERT* expression was only detected in one dog (GLP25; DogWUR113). Additionally, somatic mutations in the *TERT* gene or its' promoter regions were not identified. Therefore, telomerase was assumed not to be activated in these tumors.

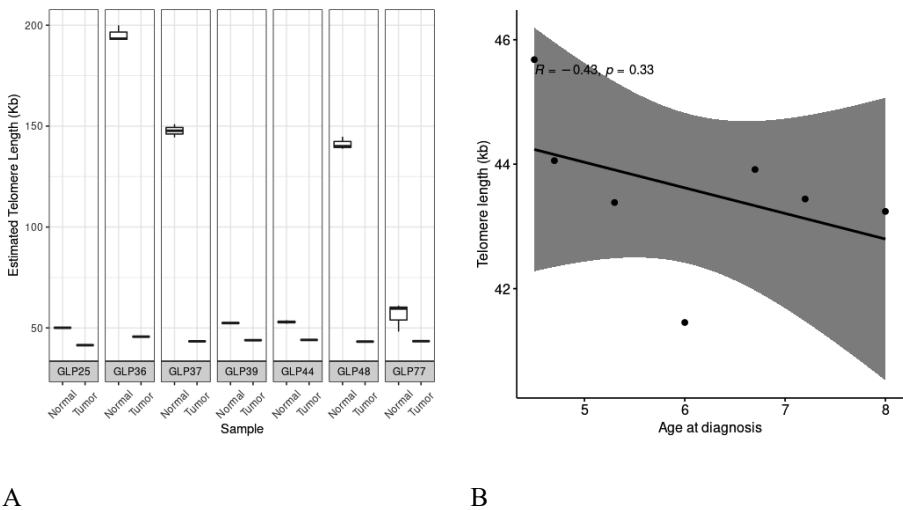


Figure 4.3. A. Boxplot of telomere length estimated from WGS of each tumor and matched normal sample. B. Correlation between the tumor telomere length and diagnosis age of FCC (years). The shaded area represents the 95% confidence interval.

4.2.7. Tumor purity, ploidy and subclone cluster

To explore the intra-tumor heterogeneity of FCC in these GLPs, we identified the subclone cluster in the 7 tumors along with purity and ploidy. According to the estimation from the TitanCNA workflow, 2 tumor samples had a good tumor cell purity of above 0.5, while 5 samples had relatively low purity, ranging between 0.3 and 0.4 (Table 4.2). Ploidy was estimated to be between 2 and 3 (Table 4.2). A subclone cluster was identified in only one tumor sample, DogWUR108, where the cancer cell fraction of the subclone was estimated to be 0.46. These tumors were supposed to have a low intra-tumor heterogeneity based on the subclone identification, which is also consistent with our previous expectation about stage from our clinical and histological examination where most tumors were supposed to be low grades.

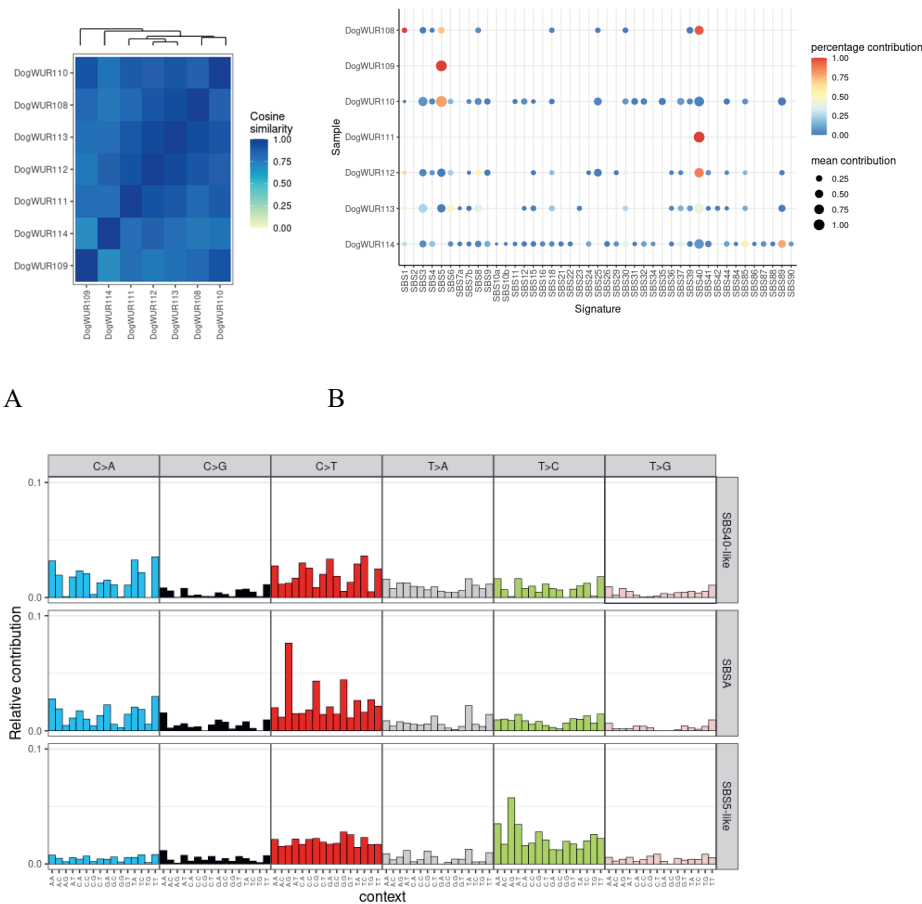
Table 4.2. Purity and ploidy estimated for each tumor sample.

	DogWUR 108	DogWUR 109	DogWUR 110	DogWUR 111	DogWUR 112	DogWUR 113	DogWUR 114
Purity	0.78	0.56	0.38	0.34	0.35	0.33	0.32
Ploidy	2.2	2.9	2.6	2.2	2.6	2.5	2.1

4.2.8. Mutational signatures

A mutational signature is the outcome of a mutagenic process comprising some form of DNA damage, subsequently acted upon by DNA repair and/or replicative machinery. Mutational signatures can reveal potential sources of mutagenesis during tumorigenesis [18]. For this, we explored the substitution mutational signatures. The mutational spectrum of the somatic substitutions is highly similar across these 7 tumors, based on their cosine similarity (Figure 4.4A). We fitted the existing COSMIC signatures to our mutational spectrum [19]. Bootstrapping was used to increase the confidence of our results. We identified that SBS1, SBS5 and SBS40 contributed most to the mutation profile (Figure 4.4B). Next, to signature refitting, three mutational signatures were extracted from somatic SNVs identified from the 7 tumors using nonnegative matrix factorization method, of which two were highly similar to known human signatures SBS40 and SBS5, and another mutational signature SBSA was most similar to SBS1 (similarity of 0.71) (Figure 4.4C). This is consistent with the results from the bootstrapping refitting. SBS5 was also the most frequently identified SBS in human THCA [7, 20]. SBS1 and SBS5 are clockwise signatures. SBS40 also correlates with patients’ ages for some types of human cancers. Moreover, C>T transitions in CpGs were the most common point alterations and correlate with age in human cancers. Enrichment of these mutations

suggests the role of age in the somatic mutation accumulation and tumorigenesis, and also indicates that the source of mutagenesis was endogenous in these dogs.



C

Figure 4.4. A. Cosine similarity of the somatic mutational spectrum across the 7 tumors. B. Contribution of known human mutational signatures to somatic mutations identified in each tumor using bootstrapping refitting in 1000 iterations. The color of the dot shows the percentage of iterations in which the signature is found (contribution > 0). The size of the dot represents the average contribution of that signature (in the iterations in which the contribution was higher than 0). C.

Mutational signatures extracted from the 7 tumors using nonnegative matrix factorization method.

4.3. Discussion

In this study, we performed comprehensive genomic analyses of familial FCCs in Dutch German longhaired pointers using whole-genome sequencing and RNA-seq data. Our somatic mutation profiling of the tumors identified a somatic missense mutation in the *GNAS* gene, which is present in 4 of 7 tumor WGS samples and 20 of 32 tumors of GLPs genotyped by sanger sequencing. It is a novel mutation identified in dogs and its' homologous mutation in humans has not been reported. Therefore, it is very likely a novel driver mutation.

The *GNAS* protein, also known as the alpha-subunit of the stimulatory G protein (Gas), normally activates adenylyl cyclase downstream of activated G-protein-coupled receptors (GPCRs), in response to hormones and a diverse number of extracellular signals. This results in the generation of a second messenger called cAMP, which activates protein kinase A (PKA). Activated PKA can phosphorylate downstream targets that are involved in many pathways and evoke downstream signaling cascades. The *GNAS* gene is included as a tier 1 oncogene in the cancer gene census (CGC) database. The mutations in genes involved in this GPCR pathway were identified in many different types of cancers, including lung adenocarcinoma and breast cancer [21]. The link between the *GNAS* mutation and tumorigenesis has been proven. The marker in the *GNAS* gene has also been included in a diagnostic panel of thyroid cancer (ThyroSeq) [22]. The activating mutations in the *GNAS* gene can result in overactivation of thyroid stimulating hormone receptor (TSHR) signaling and accumulation of cellular cAMP. The accumulation of cAMP in thyrocytes can lead to uncontrolled cell proliferation [23]. Interestingly, semiquantitative assessment of 36 FCCs in the study population suggests a possible increase in mitotic rate associated with *GNAS* A204D mutation identified in this study. However, in order to obtain more robust data on such association, more thorough, quantitative analysis of the proliferation rate of the neoplastic cells using a proliferation marker like Ki67 is necessary because the number of mitotic figures typically underestimates the actual percentage of proliferating neoplastic cells. Moreover, elevated *GNAS* expression can enhance cancer cell migration [24]. In contrast, *GNAS* knockout mice have shown reduced beta cell proliferation [25].

GNAS mutations often lead to benign thyroid cancer [26] but can also be found in malignant thyroid cancer [23]. In humans, *GNAS* mutations were also proposed to be markers of benign nodules. Interestingly, in these dogs, tumors with *GNAS* A204D mutation had lower amounts of somatic mutations (SNV, Indel, and SV). A high

tumor mutational burden (TMB) usually correlates with poor survival outcomes in humans [27, 28]. We thus suspect that patients with the *GNAS* mutation may have a better prognosis. However, we don't have survival data to validate this. This needs to be investigated further. It is not clear whether the *GNAS* mutation identified in this study is a gain-of-function or loss-of-function mutation. The impact of this mutation on the function of the *GNAS* and downstream signaling pathways is not clear. Although we performed differentially expressed gene analysis and pathway enrichment analysis based on RNA-seq data, no enriched biological process term or KEGG pathway was found. How this mutation impacts the downstream signaling pathway, such as the cAMP level, was not investigated in this study. Further experiments are needed to reveal the influence of this *GNAS* mutation on the cellular cAMP level and downstream signaling pathways.

According to studies on the somatic mutation landscape in canine cancers, sporadic canine cancers have similar driver genes to corresponding human cancers [9-14]. In a study screening 33 canine cancer cell lines, driver genes similar to human cancers were also identified [29]. However, thyroid cancer seems to be an exception. It has different somatic mutation landscape between dogs and humans.

Campos et al., investigated the somatic mutation landscapes of 43 canine FCCs and 16 canine MTCs by targeted sequencing of *H-RAS*, *N-RAS*, *PIK3CA*, *BRAF*, *RET*, and *PTEN* genes [30]. They identified 2 missense mutations in the *K-RAS* gene which have also been reported in TC of humans with a similar prevalence. No missense mutations were found in the sequenced regions of *H-RAS*, *N-RAS*, *BRAF*, *PIK3CA*, and *RET* nor in the entire coding sequence of *PTEN*. Hence, the mutations most commonly involved in thyroid tumorigenesis in humans were thought to be rare and not to play a major role in thyroid tumorigenesis in dogs. Unfortunately, the *GNAS* mutation was not investigated in that study, thus the prevalence of the *GNAS* mutation in their affected dogs is unknown.

In human thyroid cancer, driver mutations in the *GNAS* gene were also identified, but usually at a low prevalence [6, 8, 23, 31]. In 492 human thyroid carcinoma samples from the TCGA, mutations in the *GNAS* gene were detected in only 2 patients and these 2 mutations in the *GNAS* gene were at different locations from *GNAS* A204D [8]. In the COSMIC database, 73 of 3724 thyroid tumors capture mutations in the *GNAS* gene, but all at different locations. In a study including 65 human FTC samples, genetic alterations in *GNAS* were found in 5 of them (R201H in 3 samples) [6]. In another study with 154 hot thyroid nodules, 10 samples capture mutations in the *GNAS* gene [31]. The prevalence of *GNAS* mutations in TC ranges from 0.22% - 2.12% according to an investigation in 1841 human thyroid tumors [23].

In our dogs with a familial FCC, the *GNAS* mutation was the most common driver mutation identified (20 out of 32 affected dogs). The prevalence of this mutation in the affected dogs may be even higher because by chance some samples used for sequencing may contain only healthy cells but no tumor cells, resulting in false negatives. The difference in the prevalence of *GNAS* mutations in humans and our dog samples could suggest a species difference in the driver mutations for thyroid cancer between GLPs and humans. Furthermore, how this somatic mutation occurred and survived from DNA damage repair activity in that many affected dogs is also an interesting and important question. Further research is needed to answer these questions. We speculated that the germline risk factor for the FCC in some way may induce this *GNAS* mutation. We therefore tried to test the correlation between the germline risk factor in the *TPO* gene and the somatic driver mutation. However, the *GNAS* somatic mutation was also identified in the affected dogs that were supposed to be spontaneous FCC cases (3 out of 8 dogs) based on the genotype of the marker identified in the *TPO* gene. In all GLPs with the homozygous recessive genotype in the *TPO* gene, the *GNAS* somatic mutation was identified in 62.5% of them. The chi-squared test didn't show a significant correlation between the germline *TPO* mutation and the *GNAS* somatic mutation. There seems to be an interaction between the germline mutation in the *TPO* gene and the somatic mutation in the *GNAS* gene, but it is weak.

GNAS aberrations have been identified not only in thyroid tumors, but also in many other tumors in humans. Deep sequencing analysis has revealed that around 4.2% of tumors carry *GNAS* activating aberrations [32]. According to an investigation in 274,694 tumors in humans, the most common *GNAS* alterations were copy number variants (60.5%; all of which were amplifications), followed by *GNAS* codon R201 activating point mutations (34.8%). All other alterations (including the activating Q227 mutation) account for less than 5% [23]. The mutation R201C (in exon 8) is the most prevalent point mutation in *GNAS*, which was identified in many cancers, including thyroid cancer [33, 34]. The R201H mutation was also found in thyroid cancer [35].

In general, dogs have great potential as disease models of human cancers. Dogs are now increasingly highlighted as disease models for cancer research. In dogs, the report of *GNAS* mutations is still limited. According to the authors' knowledge, there was only one report where *GNAS* mutations were identified in canine adrenocortical tumors [36]. Here, we identified a familial FCC with high prevalence of *GNAS* A204D mutation in Dutch GLPs. Together with the previously identified germline risk factor in the *TPO* gene, the genetic basis of the TC development in these dogs is becoming increasingly clear. These dogs could be developed as a

disease model for research and translational medication trials for thyroid and possibly other tumor types, with driver mutations in the *GNAS* gene.

The current study has some limitations. The relatively small sample size limited the power of our analyses to identify recurrent somatic SVs and CNAs. A larger sample size may help to identify the recurrent somatic SV and CNA event in the future to elucidate whether somatic SVs or CNAs also contribute to this familial FCC. Although *GNAS* mutations were proven to be able to affect downstream signaling pathways, the potential effect of *GNAS* mutation A204D identified here was not further investigated. Further experiments are needed to elucidate how this mutation leads to tumorigenesis.

4.4. Conclusions

In this study, we profiled somatic mutation landscape of the FCC in GLP dogs at a genome-wide scale and identified a promising novel driver mutation of it, the *GNAS* A204D mutation. The prevalence of somatic mutation in the *GNAS* gene in TC is different between our dogs and humans. Our findings provide novel insights in potential molecular mechanism of thyroid carcinoma development. Moreover, our dogs with the FCC might be used as a good disease model for tumors with driver mutation in the *GNAS* gene.

4.5. Materials and Methods

4.5.1. Samples

We selected 7 GLPs affected by familial FCC from the dataset described previously, and inclusion in this study was approved by the owners [1]. Tumor tissues and blood samples were collected during the surgery or necropsy by the veterinarian of the veterinary oncology center (AniCura, the Netherlands). The tumor tissue for genetic testing was preserved in RNA-later (RNA stabilization reagent: Qiagen, Hilden, Germany) and blood was collected in K3-EDTA tubes. For histopathology, tumor tissue was fixed in 10% neutral-buffered formalin. All tumors of these dogs were evaluated histopathologically (Table 4.1). In order to correlate histomorphological characteristics of the neoplasms with mutations, other than histological pattern, the tumors were assessed semiquantitatively for several parameters (i.e. pleomorphism, anaplasia, necrosis, number of mitotic figures, infiltrative growth, aspect of cytoplasm and immunohistochemistry for thyroglobulin).

4.5.2. Whole genome sequencing

DNA from blood was extracted using the Gentra Puregene Blood Kit (Qiagen, Hilden, Germany). The DNA from the tumor tissue stored in RNAlater reagent was

extracted using Nucleospin Tissue kit (Bioke, Leiden, Netherlands). Library construction and sequencing are described in supplementary.

The tumor tissue of the left thyroid gland was sequenced to a depth of 60x. The genome of four dogs was sequenced with a coverage of 30x whereas three dogs were sequenced with a coverage of 10x. The program FastQC [37] was used to evaluate the quality of sequencing. Sickle [38] was used to trim the reads using default settings. Sequences were aligned to the CanFam 3.1 reference genome downloaded from Ensembl following the best practice guideline (<https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery>). Mapping was done using BWA-MEM algorithm (current version 0.7.15) [39], followed by sorting with samtools 1.9 [40] and marking of duplications with Picard tool [41]. Finally, base quality score recalibration was performed using GATK 4.1.8.1 [42].

4.5.3.RNA-seq

RNA-seq of tumor tissue was obtained and mapped to dog reference genome CanFam3.1, as described previously [2]. Mapping was performed using HISAT2 [43]. FeatureCounts [44] was used to quantify mapped reads to genomic features such as genes, exons, gene bodies, genomic bins, and chromosomal locations. DESeq2 package [45] was used for differential expression analysis. The clusterProfiler package [46] was used to perform the gene set enrichment analysis.

4.5.4.Somatic SNV & Indel

Three methods, Mutect2 [47], VarScan2 [48], and Strelka2 [49], were used to call the somatic SNVs in paired tumor-normal model. Firstly, Mutect2 in tumor only model was run for each normal sample, followed by GenomicsDBImport and CreateSomaticPanelOfNormals tools to create the panel of normal (PON) file. This PON file and the germline SNVs file obtained from study of Plassais et al., 2019 [50] using 722 dogs were incorporated in the somatic SNVs & Indels calling using the Mutect2 program in paired mode. The identified somatic variants were additionally filtered by CONTQ >30, MBQ >30, GERMQ >30, MMQ>30, VAF > 0.1, alternative allele count in normal sample = 0, alternative allele count > 5 in tumor sample, depth in tumor > 30, depth in normal sample > 8. The second method, VarScan2 paired mode, was run using the command: `--tumor-purity 1 --min-coverage 8 --min-coverage-normal 6 --min-coverage-tumor 8 --min-reads 5 --min-avg-qual 30`. Only the resulting high-confidence somatic variants were retained for subsequent analyses. The third model, Strelka2 paired mode, was run in default setting. Only variants passing the default filtering were used for the subsequent analyses. In addition, mutations with MQ < 30 were discarded.

A consensus approach was used to identify the reliable somatic SNVs and Indels. Bcftools (v1.9) [51] was used to intersect the variants identified by the 3 methods described above. Only the variants identified by at least two methods were considered as reliable and used for subsequent downstream analyses.

Visual inspection of somatic mutations in the genome browser Jbrowse [52], was used to detect false positive calls. We removed mutations identified in the simple repeat region and AG/TC tandem repeat regions. The simple repeat region file in correspondence with canFam3 was downloaded from the UCSC database. The AG/TC repeat regions were identified using BSgenome package [53] in R. The AG/TC repeat region was defined by 5 or more tandem AG/TC repeats. 5% Of all variants were randomly selected for visual inspection, to evaluate the true positive calling ratio among the final somatic SNVs and Indels dataset.

The VCF file containing somatic SNVs and Indels was transformed to mutation annotation format (MAF) file using vcf2maf [54] which depends on ensembl's VEP tool. The maftools package [55] was used to analyze the somatic mutations, including generating the oncoplot, which shows the mutation landscape of samples, and the lollipop, which shows the location of non-synonymous mutations in corresponding genes.

4.5.5.Somatic copy number segmentation

Somatic copy numbers were called for paired tumor-normal samples using HMMCopy tool [56] (version 1.32.0) using the author's recommendations. Briefly, GC counts and mappability files for CanFam3.1 reference genome were generated with 1000 bp window size using hmmscopy-util and GenMap [57] respectively. Read counts for each of tumor and normal bam files were generated in 1000 bp window size using readCounter in the HMMCopy package. GC counts, mappability and read counts were fed into the HMMCopy algorithm and segmentations were called using Viterbi algorithm. The segmented CNAs were fed to GISTIC2 (v2.0.22) [58] for identification of recurrent somatic CNAs. GISTIC2 identifies genomic regions that are significantly gained or lost across a set of tumors.

4.5.6.Purity, ploidy estimation

TitanCNA [59] was used to estimate the purity and ploidy of tumors. Firstly, allele counts of tumors at heterozygous sites which overlapped with germline variants identified in 722 dogs [50] were generated using the Bcftools mpileup tool. Then the allele counts and somatic CNAs were fed into the TitanCNA algorithm to infer allele-specific copy numbers, copy-number based clonality, purity, ploidy, and cellular prevalence. Cellular prevalence is the proportion of tumor cells harbouring a somatic event.

4.5.7. Significantly mutated genes

The MuSiC2 [60] and dNdScv [61] packages were used to identify the significantly mutated genes (SMG) from the somatic SNVs and InDels. MuSiC2 defined significantly mutated genes that have a significantly higher mutation ratio than the background somatic mutation burden across the tumor genomes. The dNdScv package identifies driver genes by quantifying the dN/dS ratios for missense, nonsense and essential splice mutations. Furthermore, ConSurf [62] was used to estimate the conservation score of identified amino acid changes in the SMG.

4.5.8. Mutational signature

Single base substitutions (SBS) mutational signatures were constructed using the MutationalPatterns package [63]. A high true positive ratio of somatic SNVs identified of 90% make the mutational signatures analysis reliable. SBS was classified according to the 6 possible substitutions (C>A, C>G, C>T, T>A, T>C, T>G), plus the flanking 5' and 3' bases. Non-negative matrix factorization (NMF) was run with a rank of 2-7 and a final rank of 3 was chosen. The reconstructed mutational signatures were compared to known COSMIC mutational signatures detected in humans. A signature was considered novel when its similarity to other defined COSMIC mutational signatures was less than 0.85. Next, we used the “fit_to_signatures_strict” function to fit the COSMIC signatures to the mutation profiles. This function was used to reduce overfitting. The signatures that were present according to this refitting were then used to perform bootstrapped refitting, using 1000 iterations, to determine the confidence of the refit.

4.5.9. Telomere length

Telomere length was estimated for each library using the tool TelSeq [64] along with a set of parameters to be compatible with our dataset: genome length of 2.5Gb, 78 telomere ends, and reads length of 150bp. The estimation of this tool has been shown to correlate well with Southern blot measurements based on 260 samples from the TwinsUK cohort [64].

4.5.10. Validation

PCR was done using 60 ng of genomic DNA, with 0.4 µm of each primer and FIREPol 5x Master Mix 7.5 mM Mastermix(Bio-Connect) in a final volume of 12 µl. Initial denaturation for 1 min at 95°C was followed by 35 cycles of 95°C for 30 s, 55°C for 45 s, 72°C 90 s, followed by a 5 min extension 72°C. PCR primers for somatic mutation are GCACGTTTGTCTTTTCGAT forward and TCCACAAACCTGTTGTTCCA reverse. After PCR the products were cleaned up with the use of Millipore PCR clean-up vacuum system (Multiscreen_PCR vacu 030,

Merck Millipore). Sequencing reaction was done using 10 – 20 ng of cleaned PCR product, with 5x dilution buffer, BigDye v3.1 reaction mix (Thermofisher) and 0.8 pmol/μl reverse primer. A sequencing reaction clean-up was performed with NaAc-EDTA and pure Ethanol. Sanger sequencing was performed on the ABI 3730 DNA sequencer (Applied Biosystems). SNP detection was performed using preGap and Gap (Staden Package).

Software and algorithms	Version	Identifier
bcftools	v1.10.2	http://samtools.github.io/bcftools/bcftools.html
BSgenome	v1.58.0	https://bioconductor.org/packages/release/bioc/html/BSgenome.html
BWA	v0.7.15	https://github.com/lh3/bwa
clusterProfiler	v3.18.1	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
ConSurf		https://consurf.tau.ac.il/
Delly	v0.8.3	https://github.com/dellytools/delly
DESeq2	v1.30.1	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
dNdScv	v0.0.1.0	https://github.com/im3sanger/dndscv
Featurecounts	v2.0.1	http://subread.sourceforge.net/
GATK	v4.1.8.1	https://gatk.broadinstitute.org/hc/en-us
GenMap	v1.3.0	https://github.com/cpockrandt/genmap
GISTIC2	v2.0.22	https://www.genepattern.org/modules/docs/GISTIC_2.0
GRIDSS	v2.10.2	https://github.com/PapenfussLab/gridss

HISAT2	v2.2.0	http://daehwankimlab.github.io/hisat2/#:~:text=HISAT2%20is%20a%20fast%20and,to%20a%20single%20reference%20genome .
HMMCopy	v1.32.0	https://bioconductor.org/packages/release/bioc/html/HMMcopy.html
karyoploteR	v1.16.0	http://bioconductor.org/packages/release/bioc/html/karyoploteR.html
maftools[47]	v2.6.05	https://bioconductor.org/packages/release/bioc/html/maftools.html
Manta	v1.6.0	https://github.com/Illumina/manta
MuSiC2	v0.2	https://github.com/ding-lab/MuSiC2
MutationalPatterns	v3.0.1	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html
Mutect2 - GATK4	v4.1.0.0	https://gatk.broadinstitute.org/hc/en-us
NMF	v0.23.0	https://cran.r-project.org/web/packages/NMF/index.html
picard	v2.23.8	https://broadinstitute.github.io/picard/
R	v4.0.3	https://cran.r-project.org/
Rstudio	v1.3.1093	https://www.rstudio.com/
Samplot	v1.3.0	https://github.com/ryanlayer/samplot
samtools	v1.9	http://www.htslib.org/
Strelka2	v2.9.2	https://github.com/Illumina/strelka
StructuralVariantAnnotation	v1.6.0	https://www.bioconductor.org/packages/release/bioc/html/StructuralVariantAnnotation.html
SURVIVOR	v1.0.7	https://github.com/fritzsedlazeck/SURVIVOR

SvABA	v1.1.3	https://github.com/walaj/svaba
TelSeq	v0.0.1	https://github.com/zd1/telseq
TitanCNA	v1.17.1	https://bioconductor.org/packages/release/bioc/html/TitanCNA.html
VarScan2	v2.4.4	http://varscan.sourceforge.net/
Vcf2maf	v1.6.18	https://github.com/mskcc/vcf2maf
VEP	v101.0	https://www.ensembl.org/info/docs/tools/vep/index.html

Data and code availability: Sequencing data presented in this study is freely available at the EMBI-EBL ENA database with reference number PRJEB47059.

Acknowledgement: We thank the “Nederlands Kankerfonds voor Dieren” for providing financial support for this study. We also thank the breeder association for providing pedigree information. We thank Johan de Vos[†] and Mariska de Ruijscher and for their contribution in sampling. We thank Ruben van Boxtel for his advises on mutational signatures analyses. We thank Markus J. van Roosmalen for his suggestions on somatic variants analyses. We thank Kimberley Laport for technical assistance for the RNA and DNA isolations and validation of the *GNAS* mutation. Library preparation/sequencing are performed by Novogene (UK) Company Limited. Yun’s PhD study was supported by China Scholarship Council (CSC).

Funding: This research was funded by “Nederlands Kankerfonds voor Dieren”.

Conflicts of Interest: The authors declare no potential conflict of interest.

4.6. Reference

1. Yu Y, Krupa A, Keesler RI, Grinwis GCM, de Ruijscher M, de Vos J, et al. Familial follicular cell thyroid carcinomas in a large number of Dutch German longhaired pointers. *Vet Comp Oncol*. 2021.
2. Yu Y, Bovenhuis H, Wu Z, Laport K, Groenen MAM, Crooijmans RPMA. Deleterious Mutations in the TPO Gene Associated with Familial Thyroid Follicular Cell Carcinoma in Dutch German Longhaired Pointers. *Genes*. 2021;12(7):997.
3. Pon JR, Marra MA. Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease*. 2015;10(1):25-50.
4. Hanahan D, Weinberg Robert A. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144(5):646-74.
5. Nieminen TT, Walker CJ, Olkinuora A, Genutis LK, O'Malley M, Wakely PE, et al. Thyroid Carcinomas That Occur in Familial Adenomatous Polyposis Patients Recurrently Harbor Somatic Variants in APC, BRAF, and KTM2D. *Thyroid*. 2020;30(3):380-8.
6. Pozdnyev N, Gay LM, Sokol ES, Hartmaier R, Deaver KE, Davis S, et al. Genetic Analysis of 779 Advanced Differentiated and Anaplastic Thyroid Cancers. *Clin Cancer Res*. 2018;24(13):3059-68.

7. Yoo S-K, Song YS, Lee EK, Hwang J, Kim HH, Jung G, et al. Integrative analysis of genomic and transcriptomic characteristics associated with progression of aggressive thyroid cancer. *Nature Communications*. 2019;10(1):2764.
8. Cancer Genome Atlas Research N. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*. 2014;159(3):676-90.
9. Decker B, Parker HG, Dhawan D, Kwon EM, Karlins E, Davis BW, et al. Homologous Mutation to Human BRAF V600E Is Common in Naturally Occurring Canine Bladder Cancer--Evidence for a Relevant Model System and Urine-Based Diagnostic Test. *Mol Cancer Res*. 2015;13(6):993-1002.
10. Elvers I, Turner-Maier J, Swofford R, Koltoukian M, Johnson J, Stewart C, et al. Exome sequencing of lymphomas from three dog breeds reveals somatic mutation patterns reflecting genetic background. *Genome Res*. 2015;25(11):1634-45.
11. Sakthikumar S, Elvers I, Kim J, Arendt ML, Thomas R, Turner-Maier J, et al. SETD2 Is Recurrently Mutated in Whole-Exome Sequenced Canine Osteosarcoma. *Cancer Res*. 2018;78(13):3421-31.
12. Alsaihati BA, Ho K-L, Watson J, Feng Y, Wang T, Dobbin KK, et al. Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. *Nature Communications*. 2021;12(1):4670.
13. Amin SB, Anderson KJ, Boudreau CE, Martinez-Ledesma E, Kocakavuk E, Johnson KC, et al. Comparative Molecular Life History of Spontaneous Canine and Human Gliomas. *Cancer Cell*. 2020;37(2):243-57.e7.
14. Wong K, van der Weyden L, Schott CR, Foote A, Constantino-Casas F, Smith S, et al. Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. *Nature Communications*. 2019;10(1):353.
15. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*; 2019.
16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-11.
17. Shay JW. Role of Telomeres and Telomerase in Aging and Cancer. *Cancer Discov*. 2016;6(6):584-93.
18. Koh G, Degasperis A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nature Reviews Cancer*. 2021.
19. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101.
20. Morton LM, Karyadi DM, Stewart C, Bogdanova TI, Dawson ET, Steinberg MK, et al. Radiation-related genomic profile of papillary thyroid cancer after the Chernobyl accident. *Science*. 2021:eabg2538.
21. Turan S, Bastepe M. *GNAS* Spectrum of Disorders. *Curr Osteoporos Rep*. 2015;13(3):146-58.
22. Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE. Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. *J Clin Endocrinol Metab*. 2013;98(11):E1852-E60.
23. Tirosch A, Jin DX, De Marco L, Laitman Y, Friedman E. Activating genomic alterations in the *Gs* alpha gene (*GNAS*) in 274 694 tumors. *Genes Chromosomes Cancer*. 2020;59(9):503-16.
24. Jin X, Zhu L, Cui Z, Tang J, Xie M, Ren G. Elevated expression of *GNAS* promotes breast cancer cell proliferation and migration via the PI3K/AKT/Snail1/E-cadherin axis. *Clin Transl Oncol*. 2019;21(9):1207-19.
25. Xie T, Chen M, Zhang QH, Ma Z, Weinstein LS. Beta cell-specific deficiency of the stimulatory G protein alpha-subunit *Gs*alpha leads to reduced beta cell mass and insulin-deficient diabetes. *Proc Natl Acad Sci U S A*. 2007;104(49):19601-6.
26. Alsina J, Alsina R, Gulec S. A Concise Atlas of Thyroid Cancer Next-Generation Sequencing Panel ThyroSeq v.2. *Mol Imaging Radionucl Ther*. 2017;26(Suppl 1):102-17.

27. Owada-Ozaki Y, Muto S, Takagi H, Inoue T, Watanabe Y, Fukuhara M, et al. Prognostic Impact of Tumor Mutation Burden in Patients With Completely Resected Non-Small Cell Lung Cancer: Brief Report. *J Thorac Oncol*. 2018;13(8):1217-21.
28. Xie Z, Li X, Lun Y, He Y, Wu S, Wang S, et al. Papillary thyroid carcinoma with a high tumor mutation burden has a poor prognosis. *Int Immunopharmacol*. 2020;89(Pt B):107090.
29. Das S, Idate R, Cronise KE, Gustafson DL, Duval DL. Identifying Candidate Druggable Targets in Canine Cancer Cell Lines Using Whole-Exome Sequencing. *Molecular Cancer Therapeutics*. 2019;18(8):1460.
30. Campos M, Kool MMJ, Daminet S, Ducatelle R, Rutteman G, Kooistra HS, et al. Upregulation of the PI3K/Akt Pathway in the Tumorigenesis of Canine Thyroid Carcinoma. *Journal of Veterinary Internal Medicine*. 2014;28(6):1814-23.
31. Stephenson A, Eszlinger M, Stewardson P, McIntyre JB, Boesenberg E, Bircan R, et al. Sensitive Sequencing Analysis Suggests Thyrotropin Receptor and Guanine Nucleotide-Binding Protein G Subunit Alpha as Sole Driver Mutations in Hot Thyroid Nodules. *Thyroid*. 2020;30(10):1482-9.
32. O'Hayre M, Vázquez-Prado J, Kufareva I, Stawiski EW, Handel TM, Seshagiri S, et al. The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nature Reviews Cancer*. 2013;13(6):412-24.
33. Legrand MA, Raverot G, Nicolino M, Chapurlat R. GNAS mutated thyroid carcinoma in a patient with Mc Cune Albright syndrome. *Bone Reports*. 2020;13:100299.
34. Wilson CH, McIntyre RE, Arends MJ, Adams DJ. The activating mutation R201C in GNAS promotes intestinal tumorigenesis in *Apc(Min/+)* mice through activation of Wnt and ERK1/2 MAPK pathways. *Oncogene*. 2010;29(32):4567-75.
35. Lu JY, Hung PJ, Chen PL, Yen RF, Kuo KT, Yang TL, et al. Follicular thyroid carcinoma with NRAS Q61K and GNAS R201H mutations that had a good (131)I treatment response. *Endocrinol Diabetes Metab Case Rep*. 2016;2016:150067.
36. Kool MM, Galac S, Spandauw CG, Kooistra HS, Mol JA. Activating mutations of GNAS in canine cortisol-secreting adrenocortical tumors. *J Vet Intern Med*. 2013;27(6):1486-92.
37. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformaticsbabraham.ac.uk/projects/fastqc/>. 2010.
38. Joshi NA FJ. Sickel: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. 2011.
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-60.
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
41. Institute B. Picard toolkit. <http://broadinstitute.github.io/picard/>: Broad Institute, GitHub repository; 2019.
42. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra: O'Reilly Media, Incorporated; 2020.
43. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*. 2019;37(8):907-15.
44. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-30.
45. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15(12):550.
46. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*. 2012;16(5):284-7.
47. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013;31(3):213-9.

48. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*. 2012;22(3):568-76.
49. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*. 2012;28(14):1811-7.
50. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, et al. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun*. 2019;10(1):1489.
51. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*. 2011;27(21):2987-93.
52. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology*. 2016;17(1):1-12.
53. Pagès H. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. 1.60.0 ed: R package; 2021.
54. Kandoth C. mskcc/vcf2maf: vcf2maf. v1.6.19 ed: GitHub; 2020.
55. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome research*. 2018;28(11):1747-56.
56. Daniel Lai GH, Sohrab Shah. HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data. version 1.34.0 ed: R package; 2021.
57. Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics*. 2020;36(12):3687-92.
58. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*. 2011;12(4):R41.
59. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res*. 2014;24(11):1881-93.
60. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome research*. 2012;22(8):1589-98.
61. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171(5):1029-41.e21.
62. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*. 2016;44(W1):W344-W50.
63. Manders F, Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, et al. MutationalPatterns: The one stop shop for the analysis of mutational processes. *bioRxiv*. 2021:2021.11.01.466730.
64. Ding Z, Mangino M, Aviv A, Spector T, Durbin R, Consortium UK. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res*. 2014;42(9):e75-e.

4.7. Supplementary materials

Materials and Methods

Library construction and sequencing

DNA samples were used for library construction following the manufacture's recommendations using NEB Next® Ultra™ DNA Library Prep Kit (Cat No. E7370L). Index codes were added to each sample. Briefly, the genomic DNA is randomly fragmented to an average size of 350 bp. DNA fragments were end polished, A-tailed, ligated with adapters, size selected, and further PCR enriched. Then PCR products were purified (AMPure XP system), followed by size distribution by Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA), and quantification using real-time PCR. Libraries were PE150 bp sequenced on a NovaSeq 6000 S4 flow cell.

Somatic structural variants calling

Only 4 pairs of samples (normal vs tumor) DogWUR108 vs DogWUR115, DogWUR112 vs DogWUR119, DogWUR113 vs DogWUR91 and DogWUR114 vs DogWUR120 were included in SV and CNA analyses. The other 3 pairs were not included due to the low and uneven coverage of matched normal WGS. A consensus approach was used to achieve a higher sensitivity and lower false discovery rate [1]. In this study, we used 4 somatic SV callers: GRIDSS [2], SvABA [3], Manta [4], and Delly [5]. All the callers were run using default settings. Then resulting SVs were filtered using the following criteria: SV length > 100 bp; variant supporting reads in normal sample = 0; mapping quality > 20. The filtered SVs were then merged using SURVIVOR [6] and SVs identified by 2 or more callers were retained. These consensus SVs also have to agree on the strand and have a distance of within 500 bp measured pairwise between breakpoints. Subsequently, a visual inspection was taken for each consensus SV using the Samplot package [7]. The SVs which are supported by paired reads and/or split reads present in the tumor but absent from the matched normal genome were flagged as true calls. Only the SVs passing the visual inspection were used for subsequent analyses. StructuralVariantAnnotation package [8] was used to annotate SVs, including identification of genes affected by SVs.

Results

Somatic structural variants

To investigate the role of somatic structural variants (SV) in the tumorigenesis of the FCC in these GLPs, we used 4 methods to call somatic structural variants in paired tumor-normal mode and obtained the consensus somatic variants (reported by 2 or more callers), which passed a visual inspection using the Samplot package. The

number of identified somatic SVs is diverse across samples (ranging from 6 to 32). The tumor sample DogWUR108 captured more somatic SVs than other dogs (Figure S4.1). Moreover, among conventional SV types (deletion, duplication, insertion, inversion, translocation), inter-chromosomal translocation was the most dominant type identified.

SVs can cause gene fusion thereby contributing to tumorigenesis. Gene fusions between *CNTN4* and two other genes were detected in 3 tumor samples (*CNTA4/JAK3* in DogWUR112, *CNTN4/CATSPERE* in DogWUR113 and DogWUR114). However, the expression of these fusion genes was almost 0 (Supplementary Figure S4.6 A-C), suggesting that these fusion events probably have limited consequence and do not contribute to tumorigenesis.

The consequences of other somatic SVs were also investigated in the forms of gene duplication, gene deletion, gene inversion, and gene structural interruption which was defined as the breakpoint of a somatic SV locating within the gene region. No common gene alteration across these 4 tumors was identified.

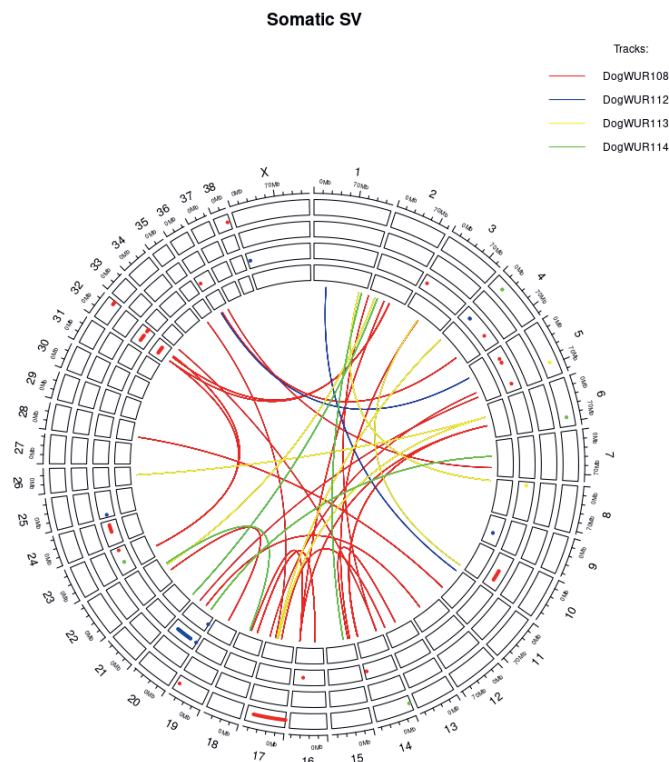


Figure S4.1. Landscape of somatic SVs identified in the tumors of 4 dogs. The circles represent deletion, insertion, inversion, duplication from the outside inwards respectively. The links inside represent the translocations. Different colors represent specific tumor sample, namely red - DogWUR108, blue - DogWUR112, yellow - DogWUR113, green - DogWUR114.

Recurrent somatic CNAs

Recurrent CNAs could play an important role in carcinogenesis by altering the gene expression. Recurrent CNA was identified using the program GISTIC2 based on segmented data derived from the TitanCNA workflow. Three focal amplification peaks passing threshold were identified, which are on chr5, chr8, and chr37 respectively (Figure S4.2A). Many recurrent copy number deletions reached statistical significance, likely because of our small sample size (Figure S4.2B). We therefore decided to focus only on the recurrent amplifications in this study. In the 3 significant recurrent amplification regions, only 1 gene was located, which is the *ERBB4* gene on chr37. However, this gene was not expressed by analyzing the

RNA-seq data. Therefore, this gene amplification likely did not contribute to tumorigenesis.

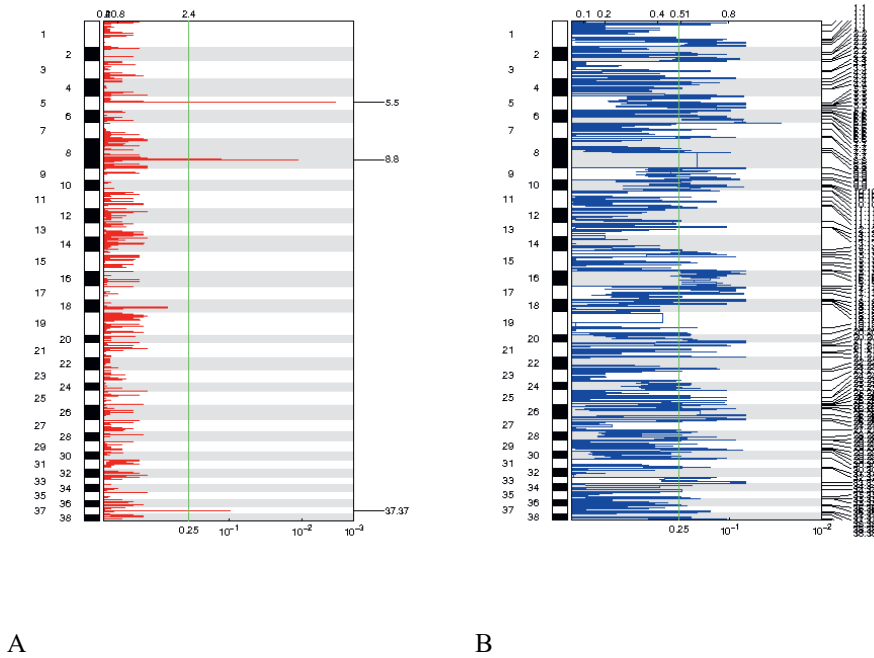
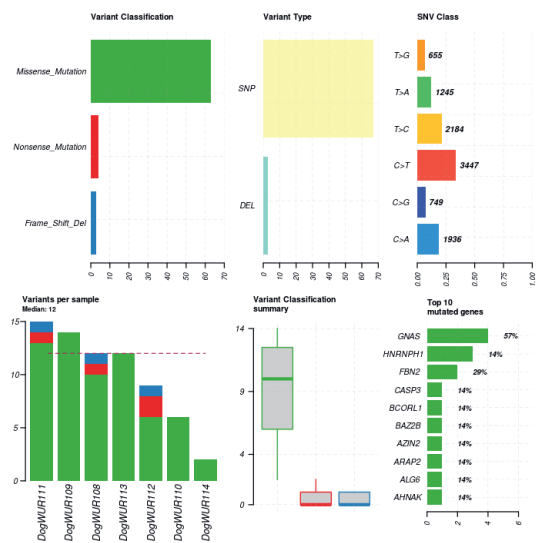


Figure S4.2. Recurrent copy number alteration identified by GISTIC2 in 4 dogs used in the analysis. A. Recurrent amplifications across canine chromosome 1 - 38. A solid green line indicates the significance threshold. B. Recurrent deletions across canine chromosome 1 - 38. A solid green line indicates significance threshold.

A



B

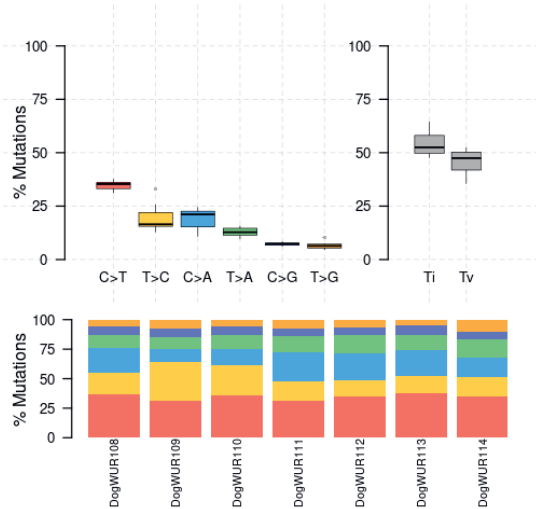


Figure S4.3. Summary of somatic SNVs and Indels identified in the 7 canine tumors. A. Summary of somatic SNVs and Indels derived from maftools package, including variant classification, variant type, SNV class, number of variants in coding region, and top 10 mutated genes. B. Somatic SNVs substitution type and transition and transversion.

Table S4.1. Genotypes of the somatic *GNAS* mutation and germline *TPO* mutation in dogs.

Dog ID	Side of tumor on thyroid gland	<i>GNAS</i>	<i>TPO</i>
GLP184	Left	-	-
GLP184	Right	CA	-
GLP16	Right	-	AA
GLP16	Left	CC	AA
GLP75	Right	CC	AG
GLP9	-	CC	AG
GLP29	Right	CA	AA
GLP29	Left	CC	AA
GLP30	-	CA	AG
GLP31	Left	CA	AA
GLP31	Right	CA	AA
GLP20	Right	CA	AA
GLP20	Left	CC	AA
GLP40	-	CC	AA
GLP22	Right	CA	AA
GLP22	Left	CC	AA
GLP230	Right	CC	AA
GLP230	Left	CA	AA
GLP63	Right	CA	AA
GLP63	Left	CA	AA

GLP35	-	CC	AA
GLP43	-	CA	AA
GLP43	Right	CC	AA
GLP15	Right	CA	AA
GLP61	-	CC	AG
GLP77	Right	CC	AA
GLP77	Left	CC	AA
GLP48	Right	CC	AA
GLP48	Left	CC	AA
GLP28	-	CA	AG
GLP127	-	CC	GG
GLP14	Right	CC	AG
GLP44	Left	CA	AA
GLP44	Right	CA	AA
GLP68	-	CA	AA
GLP17	-	CA	AG
GLP7	-	-	AA
GLP26	Left	CA	AA
GLP25	Left	CA	AA
GLP25	Right	CA	AA
GLP34	Right	CA	AA
GLP34	Left	CA	AA
GLP21	-	CC	AA

GLP24	-	CA	AA
GLP18	-	CC	AA
GLP4	-	-	GG
GLP39	Left	CA	AA
GLP39	Right	CA	AA
GLP33	Right	CA	AA

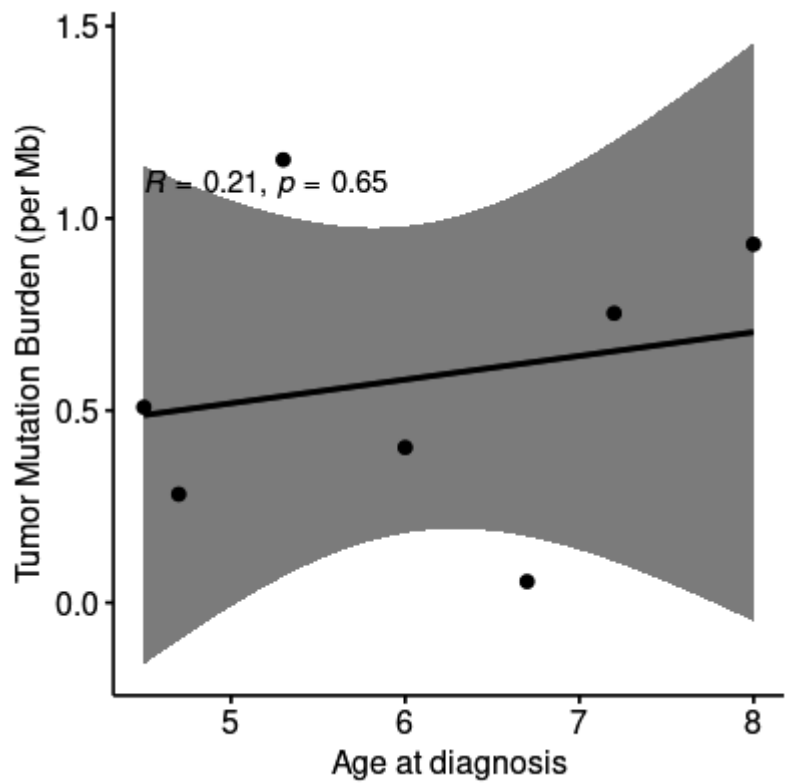


Figure S4.4. Correlation between age at diagnosis and tumor mutation burden (mutation per Mb).

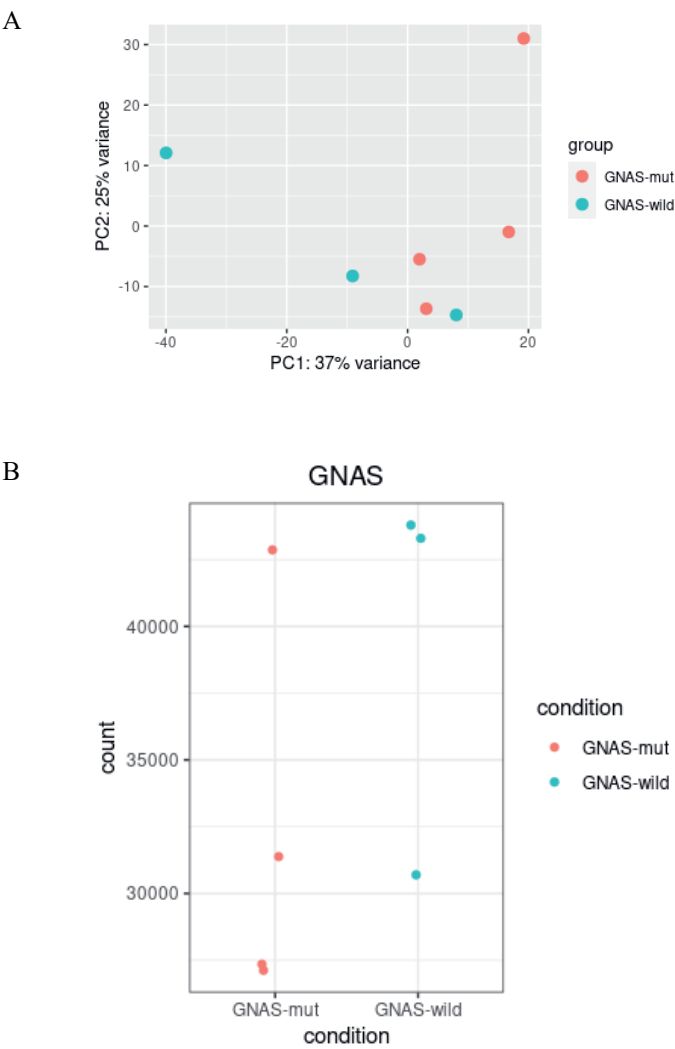


Figure S4.5. A. PCA plot of the 7 tumors based on gene expression. B. Expression level of the *GNAS* gene in tumors with and without somatic *GNAS* mutation.

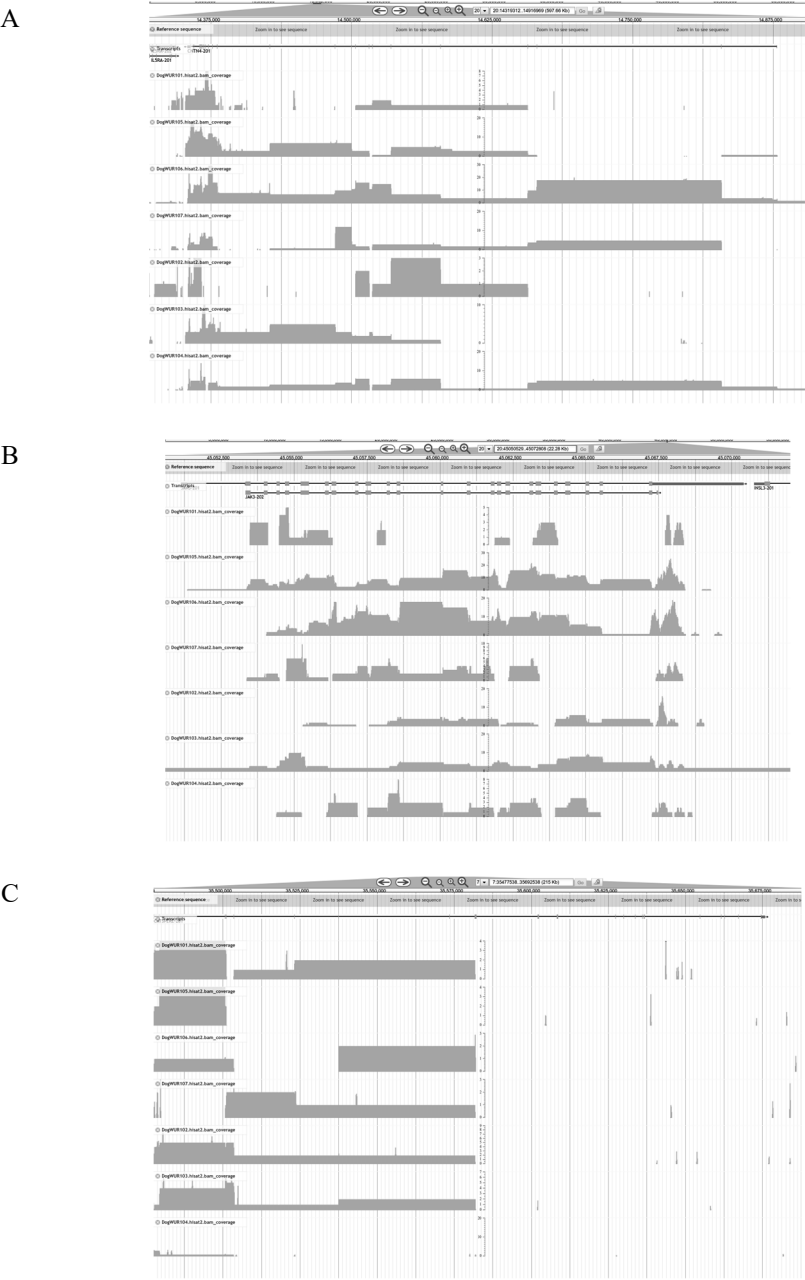


Figure S4.6. Coverage of RNA-seq reads mapped to *CNTN4* (A), *JAK3* (B), *CATSPERE* (C) gene for the 7 tumor samples.

Reference

1. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics*. 2021;22(3).
2. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome research*. 2017;27(12):2050-60.
3. Wala JA, Bandopadhyay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res*. 2018;28(4):581-91.
4. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-2.
5. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*. 2012;28(18):i333-i9.
6. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*. 2017;8(1):14061.
7. Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, et al. Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biology*. 2021;22(1):161.
8. Cameron D DR. StructuralVariantAnnotation: Variant annotations for structural variants. R package version 1.8.1; 2021.

5

Unique genetic signature and selection footprints in Dutch population of German Longhaired Pointer dogs

Yun Yu¹, Langqing Liu², Jack Windig^{1,3}, Mirte Bosse^{1,4}, Martien A.M. Groenen¹,
Richard P.M.A. Crooijmans¹

¹ Animal Breeding and Genomics, Wageningen University & Research,
Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

² Division of Evolutionary Biology, Faculty of Biology,
Ludwig-Maximilians-Universität (LMU) München, Planegg-Martinsried 82152,
Germany

³ Centre for Genetic Resources the Netherlands, Wageningen University & Research,
P.O. Box 16, 6700 AA, Wageningen, The Netherlands

⁴ Amsterdam Institute for Life and Environment (A-Life), Vrije Universiteit
Amsterdam, The Netherlands

Abstract

The German Longhaired Pointer (GLP) breed is a versatile pointer dog breed. In the current study, we investigated the genetic diversity of these dogs based on SNP array data and compared it to 11 other pointer setter breeds. The results show that GLPs have a relatively low level of inbreeding among these pointer breeds. In addition, with the availability of pedigree information of the GLPs, we demonstrate that the correlation between pedigree-based inbreeding and genotype-based inbreeding coefficients was high ($R = 0.89$ and 0.85). By investigating population structure between these 12 pointer setter breeds we showed that GLP is a breed distinct from other pointers and shares common ancestry with a few other pointing breeds. Finally, we identified selection signatures in GLPs using the runs of homozygosity (ROH) islands method. The most significant ROH island was detected on chromosome 30 harboring the genes *RYR3*, *FMN1* and *GREM1*. The *RYR3* gene plays a role in skeletal muscle contraction while the *FMN1* and *GREM1* genes are involved in limb development. The selection on these 3 genes could have contributed to the excellent athletic performance of GLPs, which is an extremely important characteristic for this hunting dog.

5.1. Introduction

The German Longhaired Pointer (GLP) is a type of multipurpose gundog, which is very active and athletic in general. GLP was assigned as a Spaniel type of continental pointer dogs by Fédération Cynologique Internationale (FCI) (<http://www.fci.be>). Pointer dogs use their instinct to point game (quarry etc.) by stopping and aiming its muzzle towards game. There are 36 pointer breeds according to FCI. These pointer dogs have one recognized ancestor, the Old Spanish Pointer, which was probably established in 100-250 BC and is now practically extinct [1]. The GLP is one of the oldest continental Pointers. The origin of the GLP dog breed is rather complex and not completely resolved. According to historical records, four hunting breeds potentially contributed to the final GLP breed, namely the German “Vogel or Habischtshund” (Quail dog), the “Wasserhund” (Water dog), “langhaarige Jagdhunde” (German Longhair) and the Spanish “Wachtelhund” (Silk dog) [2, 3]. However, whether and how much these breeds were used in the final breed formation is unknown. Meanwhile, these breeds are either extinct or rather rare, making a genomic investigation on origin of the GLP difficult. GLPs are versatile in their ability to pointing birds in the field and trace the hunted prey. In 1879 the breed characteristics for the German Longhair pointer were established, with the most important selection trait of detecting the location of shot animals and bringing them back during hunting. Currently, the GLP breed is thought to be closely related to the German Shorthaired Pointers (GSHP) and German Wirehaired Pointers (GWHP) breeds but closest to the Large Munsterlander breed (LMUN) [4].

Pedigreed dogs have high inbreeding rates [5, 6] and extensive use of popular sires is one of main causes of it [7]. Traditionally inbreeding is estimated from pedigrees, but nowadays genomics can provide additional information [8]. Besides the popularity of GLP in Germany, breeding GLPs is also popular in the Netherlands. The Dutch population of GLP was found to be predisposed to a genetic form of follicular cell thyroid carcinoma (FCC) [9]. Inbreeding analysis based on both pedigree and SNP array genotype data indicated that inbreeding contributed to the high incidence of the genetic FCC in the population [10]. The genetic FCC was found mainly in the Dutch population of GLPs in recent years with onset of cancer formation from 4.5 till 13.5 years of age. Due to cross breeding using affected or carrier Dutch GLPs, genetic FCC is also becoming a problem in GLPs in other countries.

To date, there is no report on the genetic characterization of GLP breed using genomic data. In this study, we investigated the genetic relationship between Dutch GLPs and several other pointer setter breeds by looking into the population structure. The pointer setter breeds were defined by Parker HG 2017 according to a

phylogenetic clustering [11]. Within-breed genetic diversity of these pointer setter breeds was also assessed using runs of homozygosity (ROH) and linkage disequilibrium decay. These results could be valuable to guide breeding programs of GLP. Finally, we investigated selection signatures in GLPs using the ROH islands method. Those selected genomic regions may underlie specific characteristics of GLPs.

5.2. Materials and Methods

5.2.1. Study population

We genotyped 58 Dutch GLPs in a previous study with either the 170K or 230K canine SNP array [10]. Among the 58 dogs, there were some full siblings. We therefore randomly selected only one dog from each full sibling cluster which resulted in 37 GLPs to be included in this study. All these GLPs were born in the Netherlands between 1997 and 2007, and therefore only representing the Dutch population of the GLP breed. Additionally, a publicly available dataset was obtained where 1,346 dogs from 161 breeds were genotyped with a 150K SNP array [11]. All these dogs were used in the phylogenetic analysis to determine the genetic relationship between GLP and other dog breeds, while for other analyses based on genotype data in this study, only dogs classified as pointer setter were used (Table 5.1), including 10 Brittany dogs (BRIT), 10 English Setters (ESET), 10 Gordon Setters (GORD), 10 German Shorthaired Pointers (GSHP), 2 German Wirehaired Pointers (GWHP), 9 Irish Setters (ISET), 3 Large Munsterlanders (LMUN), 2 Spinone Italiano dogs (SPIN), 7 Vizsla dogs (VIZS), 10 Weimaraner dogs (WEIM), and 6 Wirehaired Pointing Griffon dogs (WHPG). According to classification of FCI, BRIT is a Scent hound, and ESET, GORD, and ISET are Setters. Furthermore, we utilized whole genome sequences (WGS) of 22 GLPs that were generated in our previous study to validate and finemap the signatures of selection [10]. The mapping and variant calling have been described in our previous study [10].

Table 5.1. Pointer setter dogs included in this study.

Dog breed	Abbreviation	Origin country	FCI ^a classification	Number of dogs ^b
German Longhaired Pointer	GLP	Germany	Pointer dog	37
Brittany	BRIT	France	Scent hound	10
English Setter	ESET	Great Britain	Setter	10
Gordon Setter	GORD	Great Britain	Setter	10
German Shorthaired Pointer	GSHP	Germany	Pointer dog	10
German Wirehaired Pointer	GWHP	Germany	Pointer dog	2
Irish Setter	ISSET	Ireland	Setter	9
Large Munsterlander	LMUN	Germany	Pointer dog	3
Spinone Italiano	SPIN	Italy	Pointer dog	2
Vizsla	VIZS	Hungary	Pointer dog	7
Weimaraner	WEIM	Germany	Pointer dog	10
Wirehaired Pointing Griffon	WHPG	Netherlands/ Germany/France	Pointer dog	6

Note: ^a Fédération Cynologique Internationale. ^b GLPs were genotyped with either 170K or 230K canine SNP array in our previous study [10]. The genotype data of the rest dogs was collected from the study of Parker et al. 2017 [11].

5.2.2. Pedigree analysis

A pedigree consisting of 58,533 GLPs worldwide was provided by the GLP breed association. Some errors were detected in pedigree and corrections were made. Impossible birth years (e.g., year 19981, and 201) were set to unknown. Date of birth of dogs whose parents were born after their birth was also set to unknown. A pedigree loop involving a GLP was detected. To correct this, the mother of that GLP was set to unknown. The Retriever program [12] was used to perform pedigree-based analyses, such as pedigree completeness assessment (equivalent complete generations), contribution of top sires, generation interval, and litter size. The equivalent complete generations were estimated as sum of the proportions of known ancestors of an individual over all traced generations. Meanwhile, pedigree

completeness for each country, defined as the proportion of known ancestors in each generation within a country, was also assessed using optiSel package [13]. Moreover, pedigree-based inbreeding coefficient (F_{ped}) was estimated using the CFC (Coancestry, inbreeding (F) and Contribution) program [14]. Popular sires were also identified and a popular sire was defined as a male dog that sired at least 32 offspring, corresponding to 5 litters based on observed average litter size.

5.2.3. Genotype data

The overlap between the three (150, 170 and 230K) SNP genotype sets was determined based on exact location (chromosome + position) of markers and used to merge all genotype data into a single SNP genotype dataset. Plink program (v1.9) [15] was used to perform quality control with following criteria: minor allele frequency ($--maf$) > 0.05 ; missing genotype per individual ($--mind$) < 0.05 ; missing call rate per marker ($--geno$) < 0.05 , Hardy-Weinberg equilibrium exact test p-value ($--hwe$) > 0.000001 .

5.2.4. Genetic relationship between breeds

To determine the relationship between the pointer setter breeds, a principal component analysis (PCA) was performed using Plink (v1.9) and R (v4.0.3). Firstly, genetic distance between dogs was calculated using “ $--distance-matrix$ ” in Plink (v1.9). Then, classic multidimensional scaling of the distance matrix was performed using “ $cmdscale$ ” function in R and the first and second component were plotted using R. To investigate the genetic similarities between all dog breeds (GLP + 161 other dog breeds), a phylogenetic tree was constructed. Firstly, the pairwise distance between dogs was estimated using Plink with command “ $--distance 1-ibs$ ”. The ape package [16] was then used to construct a neighbor-joining phylogenetic tree from the distance matrix using “ nj ” function in default settings. The phylogenetic tree was visualized and modified using the FigTree program (v1.4.4) (<http://tree.bio.ed.ac.uk/software/figtree/>).

5.2.5. Inbreeding estimation

Besides F_{ped} , the inbreeding coefficient was also estimated based on SNP array data and compared between the pointer breeds. Homozygous genotype-based inbreeding coefficient (F_{HOM}) was estimated by Plink with command “ $--het$ ”, which is based on expected and observed autosomal homozygous genotypes. ROH across the genome were detected using Plink (v1.9) through a sliding window approach. ROH were defined according to the following criteria: (i) the minimum count of SNPs in a sliding window was 50; (ii) the minimum ROH length was set to 1 Mb; (iii) the maximum inverse density was 50 Kb per SNP; (iv) To avoid the effects of low SNP density region, the maximum gap length between consecutive SNPs was 1 Mb; (v)

The minimum hit rate of all scanning windows containing the SNP was set to 0.05; (vi) at most 1 heterozygous call allowed per scanning window; (vii) at most 5 missing calls allowed per scanning window. Total length of ROH for each dog was plotted by breeds. F_{ROH} was estimated by total length of ROH segments on auto-chromosomes divided by total length of auto-chromosomes (2,200 Mb). To eliminate the possibility that SNP chip ascertainment bias results in false positives and to validate ROH hotspots, ROH were also identified based on WGS data in Plink according to the following criteria: (i) the minimum count of SNPs in a sliding window was 50; (ii) the minimum ROH length was set to 500 Kb; (iii) the maximum inverse density was 30 Kb per SNP; (iv) To avoid the effects of low SNP density region, the maximum gap length between consecutive SNPs was 1 Mb; (v) The minimum hit rate of all scanning windows containing the SNP was set to 0.05; (vi) at most 5 heterozygous call allowed per scanning window to account for false heterozygous calls from WGS data; (vii) at most 3 missing calls allowed per scanning window.

5.2.6. Population structure

To investigate the genetic relationship between these pointer breeds, the ADMIXTURE (v1.3.0) program [17] was used to estimate the population structure base on genotype data with inferred cluster (k value) from 2 to 12. The best k value was then determined when a smallest cross-validation error was achieved from the observed data.

5.2.7. Extent of linkage disequilibrium

To eliminate the potential bias introduced by a larger sample size of GLP, we randomly sampled 10 GLPs for linkage disequilibrium (LD) decay analysis. The LD was measured using r^2 between pairs of markers. PopLDdecay [18] was used to calculate the LD decay for sub-populations identified in GLPs and each pointer breed with a maximum distance of 2000 kb between markers. Next to this, the accompanying Plot_MutiPop.pl perl script was used to plot the LD decay curve with parameters: -bin1 100; -bin2 3000; -break 2000; -method MeanBin. Due to a small sample size for GWHP, SPIN, and LMUN, these 3 breeds were not included in LD decay analysis.

5.2.8. Detection of selection signature

We identified SNPs with selection signatures in GLPs by detecting ROH islands across all autosomal chromosomes based on both SNP array and WGS data separately (see ROH analysis in inbreeding estimation section). We plotted the percentage of dogs with the SNP in a ROH against the chromosome location. The ROH island was defined to be genomic region where occurrence of ROH is in top 1%

of distribution of occurrence of ROH among all GLPs. Genes within the overlapping ROH islands identified based on SNP array and WGS data were extracted, and a gene set enrichment analysis was performed using clusterProfiler package [19].

5.3. Results

5.3.1. Pedigree based analysis

Pedigree information from 58,533 GLPs worldwide was available where the ancestor can be traced back to the year 1876. Yearly average equivalent complete generation of dogs born after year 1990 was more than 10, while most dogs have equivalent complete generations of less than 5, especially dogs born before 2013. Meanwhile, average equivalent complete generation has not gone up since 1960s (Figure 5.1A). These suggest incompleteness of the pedigree. The average yearly number of puppies born worldwide between 1990 and 2019 was 945 (ranging from 609 to 1,420) (Figure 5.1B). Average litter size was 6.1 puppies across cohorts between 1990 and 2019. Average generation interval was 4.93 years.

Before year 1938, Fped increased steadily (Figure 5.1B). From 1943 to 1986, Fped slightly and continuously decreased from 0.196 to 0.139. In 1987, there was a sharp decrease of Fped and equivalent complete generations, implying that probably some new dogs from other countries were used in the breeding in that year. After that, overall, Fped kept decreasing slowly (Figure 5.1C). Most GLPs were born in Germany (18,077 dogs) and the Netherlands (16,669 dogs). GLPs born in the Netherlands had higher inbreeding levels compared to GLPs in Germany ($0.137 > 0.048$, $p\text{-value} < 2.2e-16$, Wilcoxon rank sum test). However, the pedigree completeness of GLPs born in Germany is lower than that of GLPs born in the Netherlands (supplementary Figure S5.1). This might result in lower estimated inbreeding of GLPs born in Germany than in the Netherlands. Moreover, relationship between GLPs across countries (Germany and the Netherlands, 0.024) is more distant than that within country (within the Netherlands: 0.057, within Germany: 0.048) (supplementary Figure S5.2).

There were 4,177 sires in the GLP pedigree, of which 471 males sired more than 32 puppies each. These 471 popular sires produced 29,519 offspring in total, which sum up to 50% of the GLPs included in the pedigree. Between 1990 and 2019, the contribution of top 10 popular sires ranges from 18.1% to 47.9% per year, with an average of 28.9%. This suggests a strong sire effect among GLPs. Of the male pups 9.7% produced offspring later in life whereas 17.9% of the female pups produced offspring. The most popular sire was born in 1984 and produced 292 puppies. The most popular dam was born in 1964 and gave birth to 67 puppies.

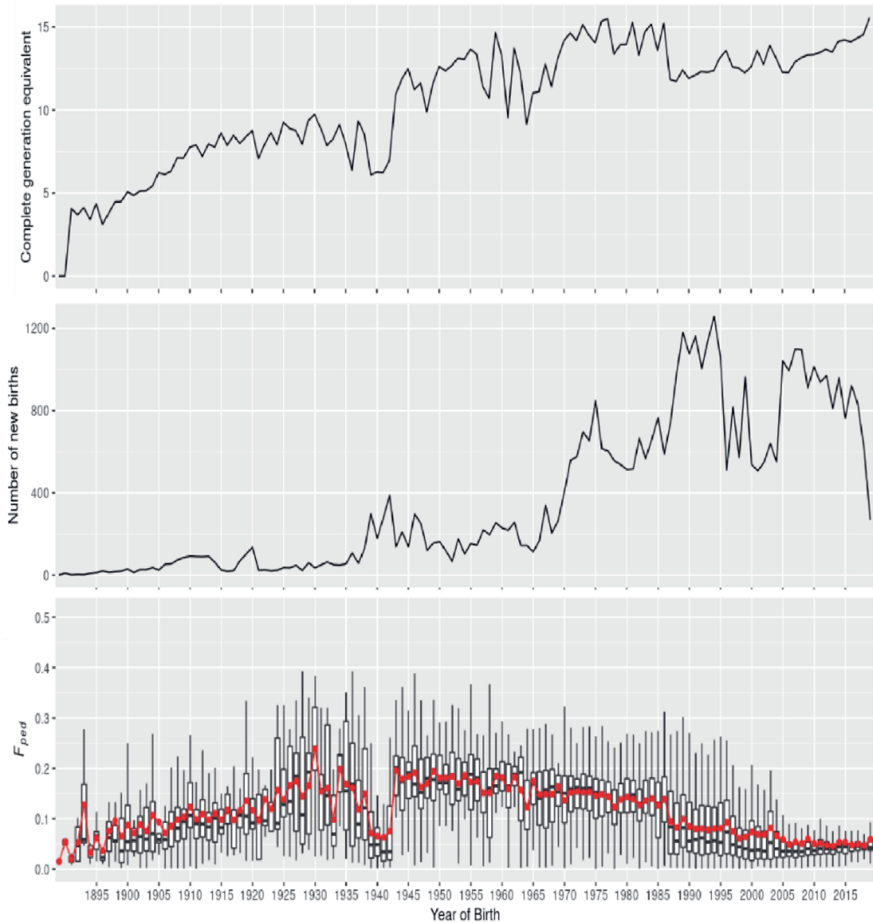


Figure 5.1. A. Average generation equivalent of GLPs born each year between 1882 and 2019. B. Number of newborn puppies per year between 1882 and 2019. C. Pedigree-based inbreeding (F_{ped}) of GLPs born each year from 1882 to 2019. Red dots indicate the average values of F_{ped} .

5.3.2. Genetic distance between pointer dog breeds

From the three SNP sets (150, 170 and 230K) 146,324 common SNP markers were selected and finally 126,144 SNPs remained after quality control. To determine the phylogenetic relationship between GLP and 161 other dog breeds, a neighbor-joining tree was constructed (supplementary Figure S5.3). The GLPs, as expected, are located on the clade of pointer setters together with the other pointer setter breeds (highlighted in red in supplementary Figure S5.3). Moreover, all

German pointer breeds were within a clade, although Vizsla (VIZS) from Hungary also appeared within this clade. VIZS is on the same sub-clade as German Shorthaired Pointer (GSHP), German Wirehaired Pointer (GWHP) and Wirehaired Pointing Griffon (WHPG) (Figure 5.2A). To investigate the genetic distance between the 12 pointer setter dog breeds, a PCA analysis was performed (Figure 5.2B). The PCA result shows that GLPs are well separated from other pointer setter breeds, and intra-population difference among the GLPs was also seen. The first component could differentiate GLP and Large Munsterlander (LMUN) from the other pointer setter breeds. The LMUN breed is the closest breed to the GLP in the PCA plot and also in the phylogenetic tree, which is in agreement with our knowledge about the breeding history of these two breeds. The second component differentiates the Weimaraner (WEIM) breed from the other pointer breeds.

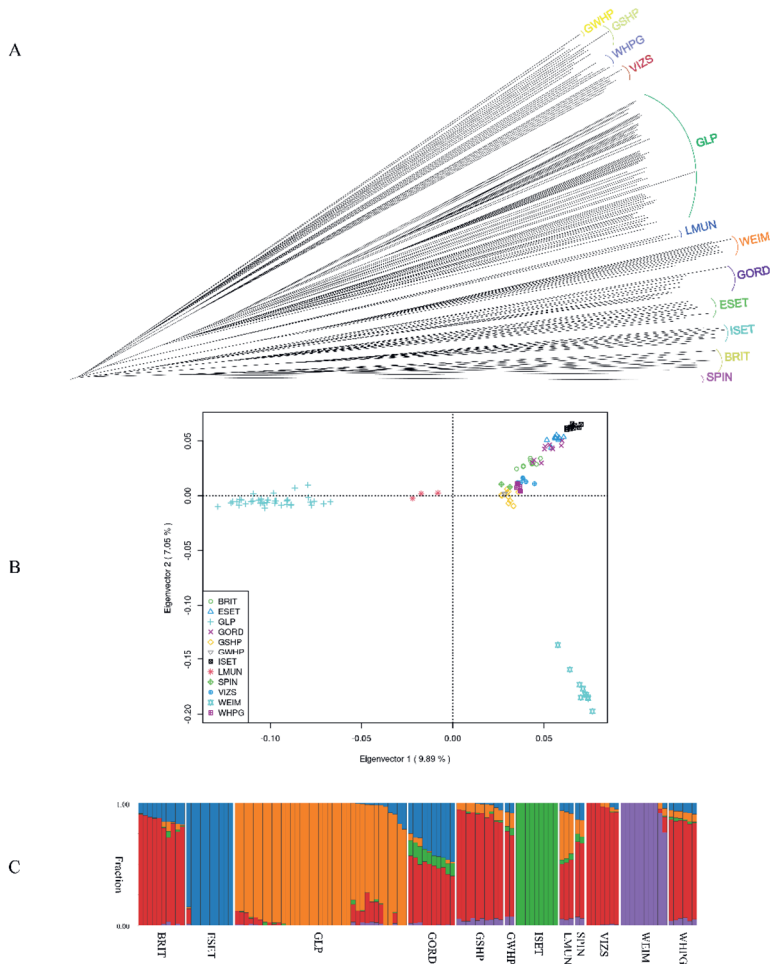


Figure 5.2. A. Subclade of 12 pointer setter breeds on the Neighbor-joining phylogenetic tree constructed based on genotype data of 1386 dogs from 162 breeds. B. Principal component analysis plot of the 12 pointer setter breeds. C. Population structure between the 12 pointer setter breeds estimated by ADMIXTURE with inferred cluster of 5. Abbreviations: BRIT – Brittany, ESET – English Setter, GLP – German Longhaired Pointer, GORD – Gordon Setter, GSHP – German Shorthaired Pointer, GWHP – German Wirehaired Pointer, ISET – Irish Setter, LMUN – Large Munsterlander, SPIN – Spinone Italiano, VIZS – Vizsla, WEIM – Weimaraner, WHPG – Wirehaired Pointing Griffon.

5.3.3. Population structure of pointer setter dogs

To characterize the population structure and admixture patterns among these pointer setter breeds, ADMIXTURE was run from $k = 2$ to $k = 12$ (supplementary Figure S5.4). According to the cross validation, the cluster number that describes the study population the best was $k = 5$ (supplementary Figure S5.5). At $k = 2$, the first breed to differentiate from the others are the GLPs. At $k = 3$, both WEIM and ISET are separated from other dogs. At $k = 5$, five distinct breeds are identified: ESET, GLP, ISET, VIZS, WEIM (Figure 5.2C). Genetic components of these five breeds are mixed in other breeds.

The genetic components identified mainly in GLPs are also detected in some of the other breeds (BRIT, GORD, GSHP, GWHP, LMUN, SPIN, VIZS, WEIM, and WHPG), but are not present in English setters and Irish setters. This suggests that those 10 breeds (except for ESET and ISET) may have shared ancestry and GLP may resemble most the ancestral breed. Among all these breeds, LMUN is genetically sharing most with GLPs, which was consistent with the PCA result. According to the breeding history, LMUN was separated in 1909 from GLP according to a difference in coat color (black color). From $k=2$ to $k=12$, we did not identify a significant proportion of genetic component coming from another breed admixed into GLPs, thus these 11 pointer setter breeds are unlikely to have served as an ancestral breed used in the development of the GLP breed.

5.3.4. Relatively low inbreeding level of GLP among pointers

To assess the inbreeding level of GLPs, we estimated the F_{ROH} and F_{HOM} of each dog and compared the results between the pointer setter breeds. Among the 12 pointer setter breeds, GLPs have relatively low average inbreeding level based on both F_{ROH} and F_{HOM} (Figure 5.3). The average F_{ROH} of GLPs is 0.16. Compared to its cousin breeds, GSHP and GWHP, GLP has a slightly higher inbreeding level. Among all these pointer setter breeds, ISET and WEIM have the highest inbreeding level. Moreover, three inbreeding parameters, F_{ped} , F_{HOM} and F_{ROH} of the 37 GLPs, have high concordance between them (supplementary Figure S5.6), and F_{ROH} and F_{HOM} estimated from genotype data had the highest correlation coefficient (0.99).

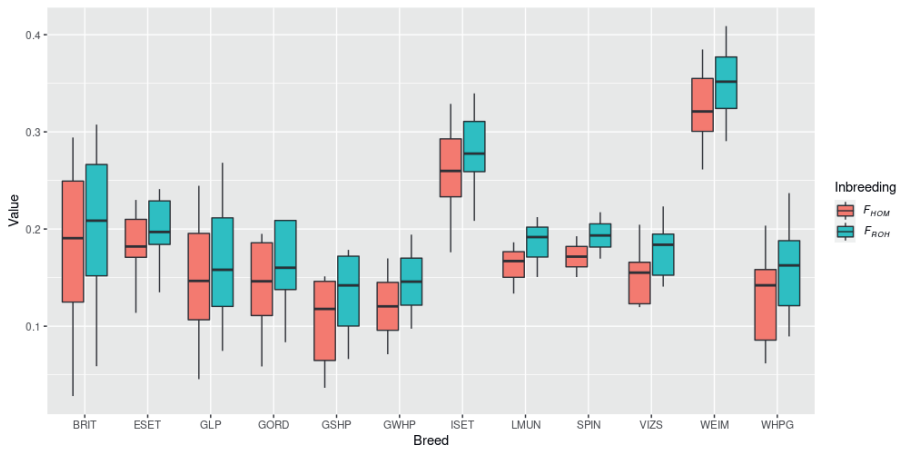


Figure 5.3. Boxplot of F_{ROH} and F_{HOM} for all the dogs grouped by breed. Abbreviations: BRIT – Brittany, ESET – English Setter, GLP – German Longhaired Pointer, GORD – Gordon Setter, GSHP – German Shorthaired Pointer, GWHP – German Wirehaired Pointer, ISET – Irish Setter, LMUN – Large Munsterlander, SPIN – Spinone Italiano, VIZS – Vizsla, WEIM – Weimaraner, WHPG – Wirehaired Pointing Griffon.

5.3.5. LD decay

To reduce the bias on estimates of LD decay because of unbalanced sample size, we randomly sampled 10 GLPs for LD decay analysis. Meanwhile, SPIN, LMUN and GWHP were not included in this study because of a small sample size (2 or 3 dogs per breed). The extent of LD, in general, corresponds well to the inbreeding level of the breed (Figure 5.4). ISET and WEIM have longest extent of LD where they also have highest inbreeding. GSHP, GORD and GLP have lower inbreeding level, where they also have shorter extent of LD.

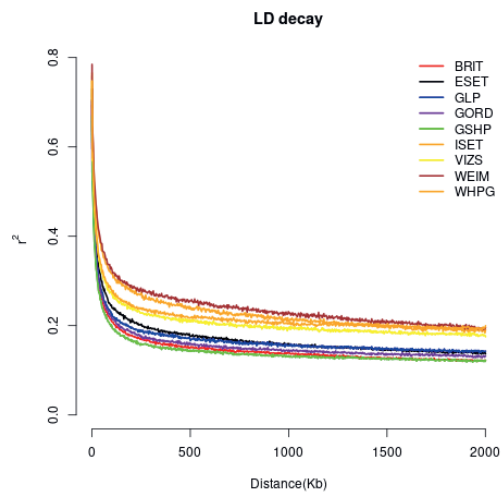


Figure 5.4. LD decay estimated for 9 pointer breeds. Abbreviations: BRIT – Brittany, ESET – English Setter, GLP – German Longhaired Pointer, GORD – Gordon Setter, GSHP – German Shorthaired Pointer, ISET – Irish Setter, VIZS – Vizsla, WEIM – Weimaraner, WHPG – Wirehaired Pointing Griffon.

5.3.6.Genomic distribution of ROH

In total, 2040 ROH with an average length of 6.57 Mb (ranging from 1.29 Mb to 64.72 Mb) were identified in the 37 GLPs based on the SNP array data. ROH are divided into 6 groups according to size (Figure 5.5), where 18% of them are longer than 10 Mb. Short ROH reflect ancestral inbreeding, while longer ROH reflect more recent inbreeding. Length distribution of ROH indicates that both ancient and recent inbreeding events have affected genomic diversity in current GLPs.

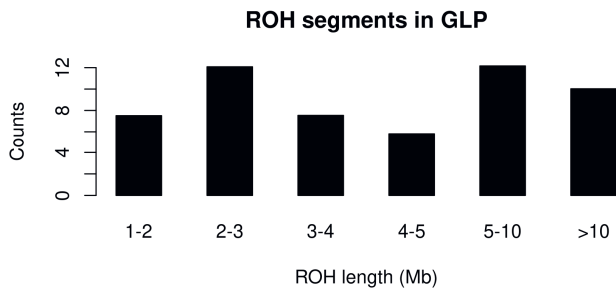


Figure 5.5. Length distribution of ROH per GLP.

ROH islands among these GLPs are identified based on either SNP array data (Figure 5.6A) or WGS data (Figure 5.6B) separately. Overlapping ROH islands between datasets were identified on chromosomes 8, 14, 22 and 30. These common ROH with low genetic diversity imply genomic signatures of selection. Genes in those overlapping ROH islands data were extracted and a gene set enrichment analysis was performed. However, no enriched GO term or KEGG pathway was identified. The 1.6Mb ROH island on chr30 indicates homozygosity for nearly all GLPs, and contains 7 genes, of which 3 are very interesting based on their function (Figure 5.7). The *RYR3* (ryanodine receptor type 3) gene is involved in skeletal muscle contraction by releasing calcium from the sarcoplasmic reticulum followed by depolarization of T-tubules [20], the *FMNI* (Formin 1) gene was reported to associate with limb deformity in mouse [21] and the *GREM1* (Gremlin 1) gene is also known to play a role in limb outgrowth and development in mammals [22].

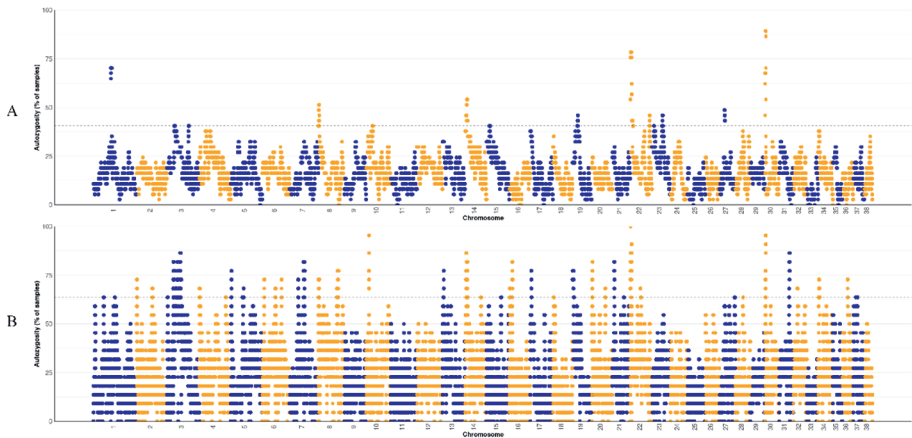


Figure 5.6. ROH islands identified on autosomes based on SNP array (A) and WGS (B) data of German Longhaired Pointers. Dashed line indicated the threshold of top 1% of empirical distribution of autozygosity.

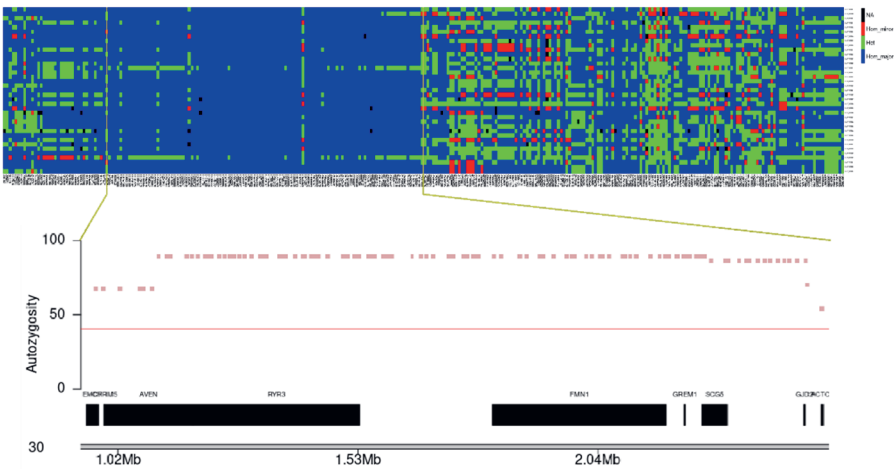


Figure 5.7. Genotypes of the variants between 0 and 5 Mb on chromosome 30 in German Longhaired Pointers. Colors denote homozygous for major allele (blue), heterozygous (green), homozygous for minor allele (red), and missing genotype (black). The plot below the genotype panel shows the genes within the zoomed region. The x-axis shows the coordinates on chromosome 30 and y-axis shows the percentage of dogs with the SNP in the ROH.

Moreover, a ROH island on chr22 was also identified and the corresponding region is an outlier not only in GLPs, but also in many other dog breeds (Figure 5.8). Interestingly, an alternative haplotype also occurs in high frequency in this region, hinting at low recombination. Five genes are located within this region: *KPNA3* (karyopherin subunit alpha 3), *EBPL* (Emopamil Binding Protein Like), *RCBTB1* (RCC1 and BTB Domain-Containing Protein), *SETDB2* (SET Domain Bifurcated Histone Lysine Methyltransferase 2), and *CAB39L* (Calcium Binding Protein 39 Like).

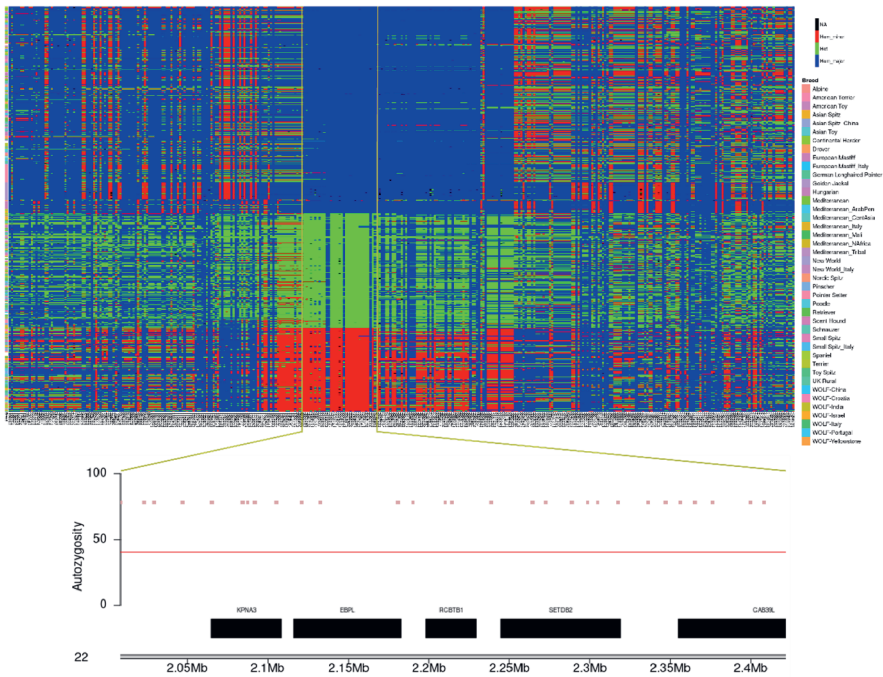


Figure 5.8. Upper panel shows Genotypes of the variants between 0 and 5 Mb on chromosome 22 in 37 GLPs and 1346 other dogs from 161 different breeds. Colors denote homozygous for major allele (blue), heterozygous (green), homozygous for minor allele (red), and missing genotype (black). Panel below shows the genes in the zoomed ROH region. The x-axis shows the coordinates on chromosome 22 and y-axis shows the percentage of dogs with the SNP in the ROH.

We also investigated ROH islands identified on other chromosomes (chromosome 8, 14, 27). These regions are less homogeneous among GLPs (supplementary Figure S5.7-S5.9), implying that these regions are not completely fixed in these GLPs and therefore not further investigated.

5.4. Discussion

In this study, we investigated the phylogenetic relationship and population structure of GLP and 11 other pointer setters. The GLP belongs to the pointer setter cluster on the phylogenetic tree as expected. The phylogenetic tree and PCA analysis show that LMUN is the breed closest to GLP, followed by the GWHP, GSHP, WHPG and VIZS. This corresponds well to the recorded breeding history of these dog breeds.

We estimated F_{ROH} and F_{HOM} and compared them between 12 pointer setter breeds included in this study. The average inbreeding of three breeds, GLP, GSHP and

GWHP, was lower compared to the other pointer breeds, while a big variation within the breeds was also observed. One thing to note is that the relatively low inbreeding of GLP, GSHP and GWHP among 12 pointer setter breeds included in the current study does not indicate that these 3 breeds have absolutely low inbreeding and do not suffer from adverse effects of inbreeding. It is known that pedigree dogs have high inbreeding rates in general due to two bottle necks in the history [5, 6, 23]. A subpopulation or country effect on the inbreeding estimation cannot be excluded. For instance, we observed a higher inbreeding level in GLPs born in the Netherlands than those in Germany, although this could also be due to unreliable pedigree. Lower inbreeding would indicate less inbreeding related health issues as we know inbreeding can cause inbreeding depression [24]. Dutch GLPs were found to be predisposed to follicular cell thyroid carcinoma and inbreeding contributed to the occurrence [9, 10]. WEIM and ISET have the highest inbreeding level among these pointer setter dogs. These two breeds have been reported to be susceptible to certain diseases, such as canine hypertrophic osteodystrophy [25]. Inbreeding based on pedigree data was determined for the GLPs. There are 4 GLPs for which estimated F_{ROH} and F_{HOM} are much higher than F_{ped} (supplementary Figure S5.6). This is likely caused by incomplete pedigree of these dogs. Pedigree based inbreeding has a lower correlation with F_{ROH} and F_{HOM} than the correlation between F_{ROH} and F_{HOM} . Pedigree based inbreeding was less accurate because of incorrect or incomplete records in the pedigree. Also, pedigree based inbreeding estimation is not able to take into account the various stochastic recombination events that occurred during meiosis.

Ancestral breeds used in the development of GLP were not detected through the population structure analysis. In addition, the breeding history of GLP is not completely clear. We thus are not able to disentangle the proportion of genetic component of ancestry in current GLPs. According to the ADMIXTURE result, with k values from 2 to 12, GLP represents a unique breed, without or with very little genetic component from other pointer breeds included in the study. This suggests that the GLP breed might resemble the shared breed ancestry of the continental pointer breeds. While other pointer breeds are more mixed with some other pointer breeds. At $k = 5$, the genetic component of VIZS is largely seen in many other pointer setter breeds, including GSHP, WHPG, SPIN, GORD, LMNU, and BRIT. It is also seen in some GLPs. It is known that VIZS was used in the development of other pointer breeds, most notably the WEIM and GSHP [26]. Our result confirmed this.

The selection signatures in GLPs were identified by exploring the ROH islands based on SNP array data. The selected genomic regions cover several genes that may underlie characteristics that are specific or important to GLP dogs. With the

ROH islands detection method, we identified several regions in the genome of GLP characterized by a loss of genetic diversity, implying selection signatures, based on both SNP array and WGS data. The ROH island identified on chromosome 30 contains 3 genes of interest. One of these, *RYR3*, plays a role in skeletal muscle contraction, which mediates the mobilization of stored Ca^{+2} in cardiac and skeletal muscle to initiate muscle contraction [20]. The *RYR3* gene was identified as a candidate gene in selection sweep analysis of hunting dogs [27]. This gene not only was under selection in GLPs, but also in many other hunting dog breeds. The other two genes in this region on chromosome 30 are *FMN1* and *GREM1*. *GREM1* is involved in limb development and growth [22]. *FMN1* and *GREM1* share the same cis-regulatory landscape and deletion of a ~180kb genomic region overlapping the *greml-fmn1* TAD disrupts *GREM1* expression in limb buds [28]. Pointer dogs run faster and are more athletic than many other dogs. The selection on these 3 genes can be linked to breeding practice of the GLPs. An especially muscular loin was included in GLP's breed standard. The selection on *RYR3* may contribute the muscular characteristic of the GLPs. Good legs are important for a hunting dog to support them searching and tracking in the field, water and forest for long periods of time. GLP breeders have stringent breed standards about limbs written in the breed standard of GLP, such as shoulder, elbow, carpus, feet and general appearance on forequarters and hindquarters. The selection on *FMN1* and *GREM1* may contribute to those as muscle and legs are important for hunting performance of a hunting dog.

The selection signature identified on chromosome 22 (from positions 2,008,422 till 2,421,940) was also reported in previous studies. Denis Akkad *et al.* [29] indicated that one particular gene in the region, *SETDB2*, could contribute to the pointing behavior, because this selection signal was detected in both pointer breeds (LMUN and WEIM), and not in herding dogs (Berger des Pyrenees and Schapendoes). However, our analysis does not support their conclusion. According to our analysis, this selection signal is present not only in pointer breeds, but also in many other breeds. Approximately 51% of all dogs (GLPs and 1346 other dogs) capture the same haplotype detected in GLPs. Especially, we found that some terrier and retriever dog breeds, Bedlington Terrier, Border Terrier, Newfoundland, Irish Water Spaniel, Scottish Terrier, Soft Coated Wheaten Terrier, Norfolk Terrier, carry the same haplotype as GLP (supplementary Figure S5.10). Therefore, we conclude that the selection signal is not related to specific pointing behavior. This ROH region characterizes with a low recombination and therefore is more sensitive to drift effects, leading to the potential of fixation of haplotypes in these inbred breeds without the need for strong selection for a gene in that region. In humans, *SETDB2* associates with left-right axis differentiation [30]. It is unknown if this association with this gene is also the case in dogs. Moreover, there are other genes located

within this ROH region on chromosome 22, such as *KPNA3*, *EBPL*, *RCBTB1*, *CAB39L*. These genes have diverse functions and are involved in different pathways (e.g. NLS-bearing protein import into nucleus, sterol metabolic process, cell cycle, intracellular signal transduction). *KPNA3* mediates nuclear import [31], *EBPL* still has unclear function [32], *RCBTB1* may play a role in angiogenesis [33] and *CAB39L* plays a role in the regulation food intake in chicken [34]. Unfortunately, without any phenotypic data, we are not able to identify specific traits potentially underlying this selected region. Further studies are needed to answer this question.

Acknowledgement

We thank the “langhaar” breeder association for providing pedigree of GLPs. We thank “Nederlands Kankerfonds voor Dieren” for financial support for genotyping with SNP array and whole genome sequencing.

Funding: This research was funded by “Nederlands Kankerfonds voor Dieren”.

Conflict of interest: The authors declare no conflict of interest.

Availability of Data: Sequencing data presented in this study are openly available at EMBL-EBI ENA database with reference number PRJEB43017. SNP array genotype data are available through ArrayExpress (accession number E-MTAB-10241).

5.5. Reference

1. Parra D, Méndez S, Cañón J, Dunner S. Genetic differentiation in pointing dog breeds inferred from microsatellites and mitochondrial DNA sequence. *Animal Genetics*. 2008;39(1):1-7.
2. Merx H, Merx A. *Der Deutsch-Langhaar-Jagdgebrauchshund*: H. Merx; 1997.
3. Kern H, Tobolik E. *Der Deutsch-Langhaar*: Neumann-Neudamm Verlag; 1996.
4. Schmutz SM, Berryere TG, Goldfinch AD. TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mammalian Genome*. 2002;13(7):380-7.
5. Wijnrocx K, François L, Stinckens A, Janssens S, Buys N. Half of 23 Belgian dog breeds has a compromised genetic diversity, as revealed by genealogical and molecular data analysis. *Journal of Animal Breeding and Genetics*. 2016;133(5):375-83.
6. Lewis TW, Abhayaratne BM, Blott SC. Trends in genetic diversity for all Kennel Club registered pedigree dog breeds. *Canine Genetics and Epidemiology*. 2015;2(1):13.
7. Leroy G, Baumung R. Mating practices and the dissemination of genetic disorders in domestic animals, based on the example of dog breeding. *Animal Genetics*. 2011;42(1):66-74.
8. Doekes HP, Bijma P, Windig JJ. How Depressing Is Inbreeding? A Meta-Analysis of 30 Years of Research on the Effects of Inbreeding in Livestock. *Genes*. 2021;12(6).
9. Yu Y, Krupa A, Keesler RI, Grinwis GCM, de Ruijscher M, de Vos J, et al. Familial follicular cell thyroid carcinomas in a large number of Dutch German longhaired pointers. *Vet Comp Oncol*. 2022;20(1):227-34.
10. Yu Y, Bovenhuis H, Wu Z, Laport K, Groenen MAM, Crooijmans RPMA. Deleterious Mutations in the TPO Gene Associated with Familial Thyroid Follicular Cell Carcinoma in Dutch German Longhaired Pointers. *Genes*. 2021;12(7):997.

11. Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, et al. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. *Cell Reports*. 2017;19(4):697-708.
12. Windig JJ, Hulsege I. Retriever and Pointer: Software to Evaluate Inbreeding and Genetic Management in Captive Populations. *Animals (Basel)*. 2021;11(5).
13. Wellmann R. Optimum contribution selection for animal breeding and conservation: the R package optiSel. *BMC Bioinformatics*. 2019;20(1):25.
14. Sargolzaei M, Iwaisaki H, Colleau JJ. CFC: A tool for monitoring genetic diversity. *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*. 2006:27-8.
15. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007;81(3):559-75.
16. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20(2):289-90.
17. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-64.
18. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*. 2019;35(10):1786-8.
19. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*. 2012;16(5):284-7.
20. Lynch AI, Irvin MR, Boerwinkle E, Davis BR, Vaughan LK, Ford CE, et al. RYR3 gene polymorphisms and cardiovascular disease outcomes in the context of antihypertensive treatment. *The Pharmacogenomics Journal*. 2013;13(4):330-4.
21. Zhou F, Leder P, Zuniga A, Dettenhofer M. Formin1 disruption confers oligodactylism and alters Bmp signaling. *Human Molecular Genetics*. 2009;18(13):2472-82.
22. Khokha MK, Hsu D, Brunet LJ, Dionne MS, Harland RM. Gremlin is the BMP antagonist required for maintenance of Shh and Fgf signals during limb patterning. *Nature Genetics*. 2003;34(3):303-7.
23. Zhang Z, Khederzadeh S, Li Y. Deciphering the puzzles of dog domestication. *Zool Res*. 2020;41(2):97-104.
24. Ujvari B, Klaassen M, Raven N, Russell T, Vittecoq M, Hamede R, et al. Genetic diversity, inbreeding and cancer. *Proceedings of the Royal Society B: Biological Sciences*. 2018;285(1875):20172589.
25. Ferguson PJ, Sandu M. Current understanding of the pathogenesis and management of chronic recurrent multifocal osteomyelitis. *Curr Rheumatol Rep*. 2012;14(2):130-41.
26. Boggs BC, Boggs SP. *The Vizsla*: Behi Publishing Company; 2000.
27. Kim J, Williams FJ, Dreger DL, Plassais J, Davis BW, Parker HG, et al. Genetic selection of athletic success in sport-hunting dogs. *Proceedings of the National Academy of Sciences*. 2018;115(30):E7212.
28. Malkmus J, Ramos Martins L, Jhanwar S, Kircher B, Palacio V, Sheth R, et al. Spatial regulation by multiple Gremlin1 enhancers provides digit development with cis-regulatory robustness and evolutionary plasticity. *Nature Communications*. 2021;12(1):5557.
29. Akkad DA, Gerding WM, Gasser RB, Epplen JT. Homozygosity mapping and sequencing identify two genes that might contribute to pointing behavior in hunting dogs. *Canine Genetics and Epidemiology*. 2015;2(1):5.
30. Ocklenburg S, Arning L, Gerding WM, Hengstler JG, Epplen JT, Güntürkün O, et al. Left-Right Axis Differentiation and Functional Lateralization: a Haplotype in the Methyltransferase Encoding Gene SETDB2 Might Mediate Handedness in Healthy Adults. *Molecular Neurobiology*. 2016;53(9):6355-61.
31. Hu B, Cheng J-W, Hu J-W, Li H, Ma X-L, Tang W-G, et al. KPNA3 Confers Sorafenib Resistance to Advanced Hepatocellular Carcinoma via TWIST Regulated Epithelial-Mesenchymal Transition. *J Cancer*. 2019;10(17):3914-25.

32. Moebius FF, Fitzky BU, Wietzorrek G, Haidekker A, Eder A, Glossmann H. Cloning of an emopamil-binding protein (EBP)-like protein that lacks sterol delta8-delta7 isomerase activity. *Biochem J.* 2003;374(Pt 1):229-37.
33. Wu JH, Liu JH, Ko YC, Wang CT, Chung YC, Chu KC, et al. Haploinsufficiency of RCBTB1 is associated with Coats disease and familial exudative vitreoretinopathy. *Hum Mol Genet.* 2016;25(8):1637-47.
34. Yuan J, Wang K, Yi G, Ma M, Dou T, Sun C, et al. Genome-wide association studies for feed intake and efficiency in two laying periods of chickens. *Genetics Selection Evolution.* 2015;47(1):82.

5.6. Supplementary materials

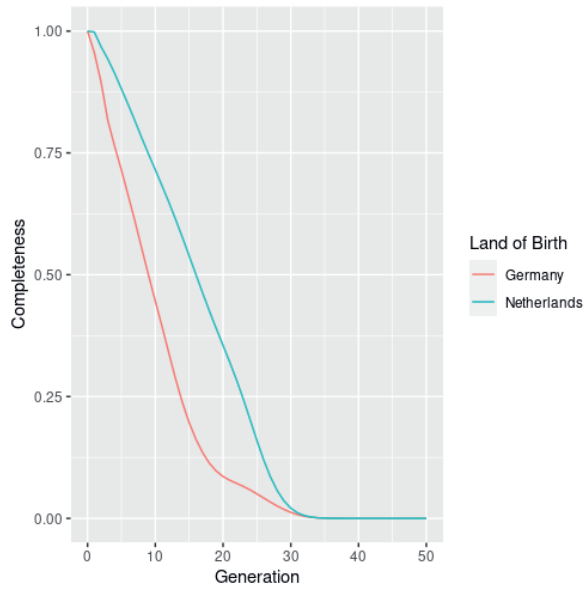


Figure S5.1. Pedigree completeness of GLPs born in Germany and the Netherlands.

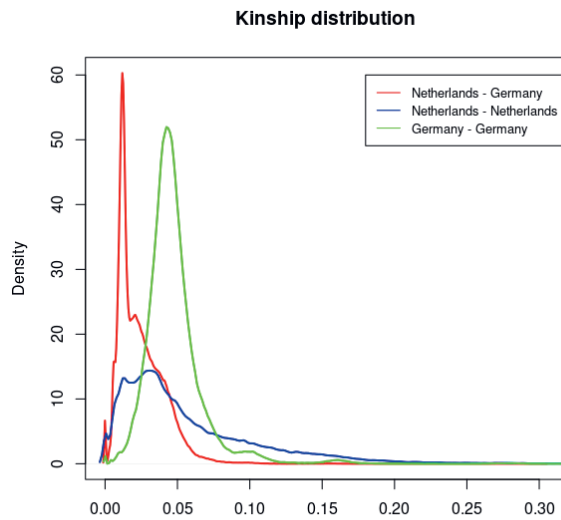


Figure S5.2. Kinship between GLPs born in the Netherlands and Germany after year 2000.

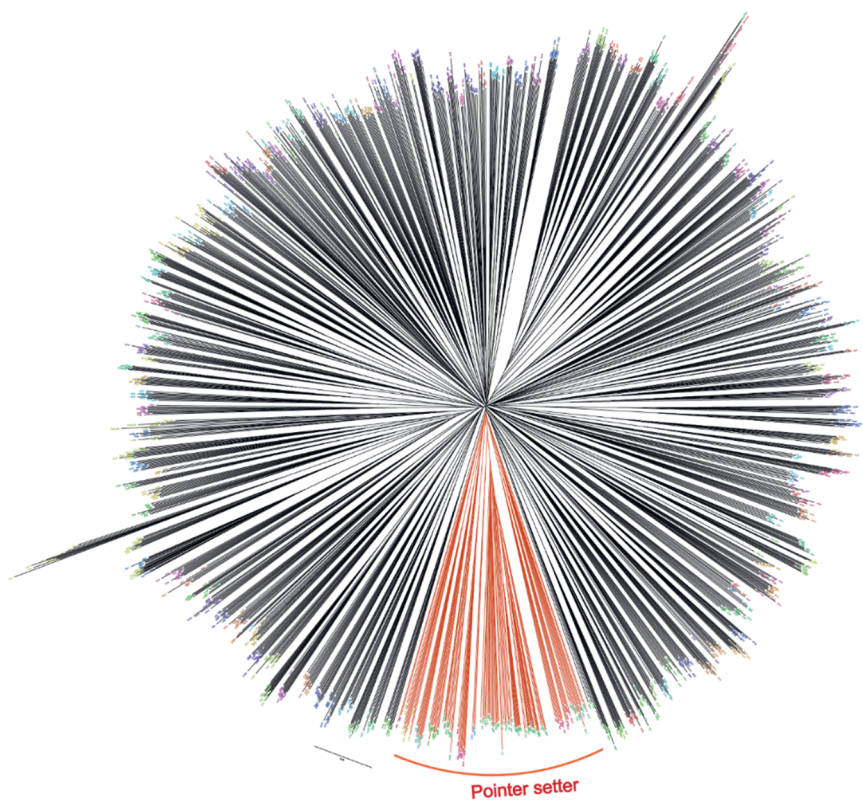


Figure S5.3. Neighbor-joining phylogenetic tree constructed based on genotype data of 1386 dog from 162 breeds. Highlighted subclade is the pointer setter subclade where German Longhaired Pointers locate on.

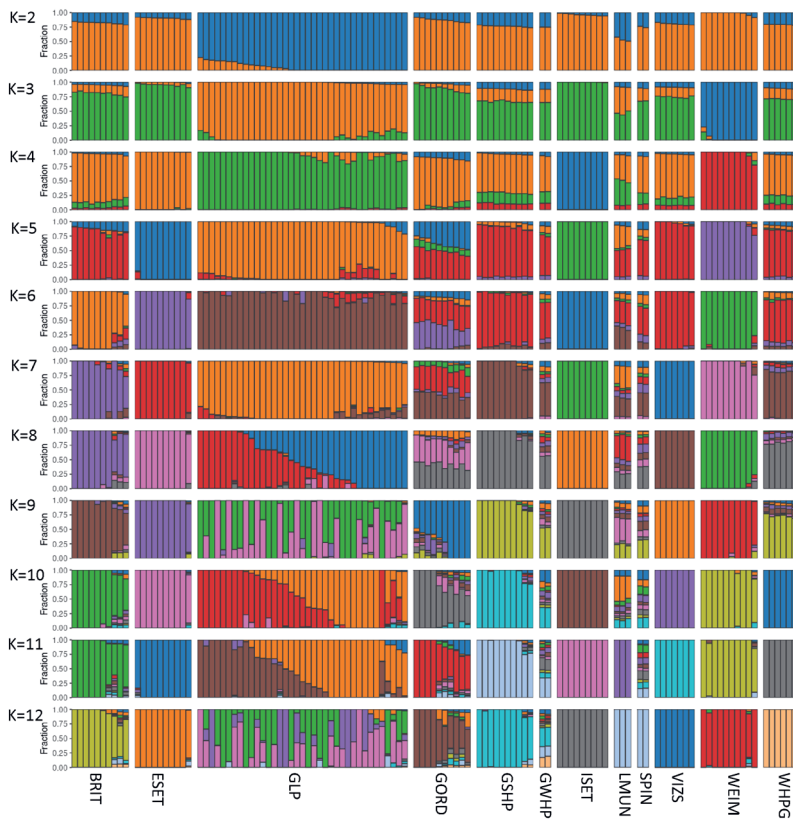


Figure S5.4. Population structure between the 12 pointer setter breeds estimated by ADMIXTURE with inferred cluster from 2 to 12.

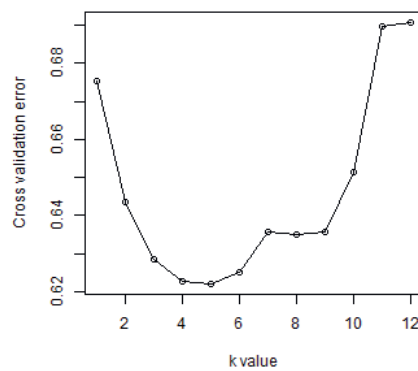


Figure S5.5. Cross-validation errors for different K values in the ADMIXTURE analysis.

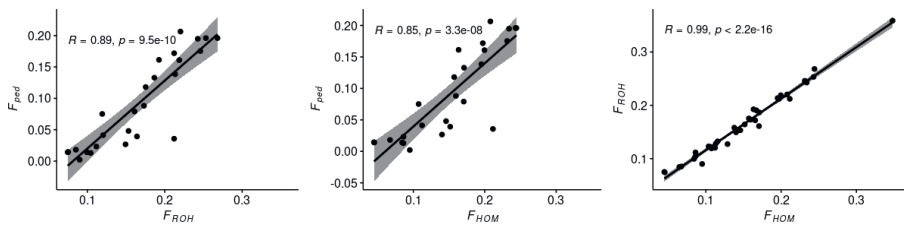


Figure S5.6. Correlation between F_{ped} , F_{ROH} , and F_{HOM} of GLPs.

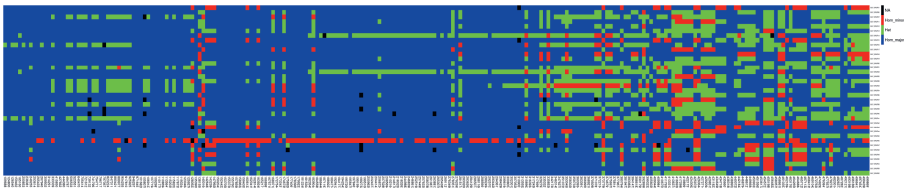


Figure S5.7. Genotypes of the variants between 0 and 5 Mb on chromosome 8 in 37 GLPs. Colors denote homozygous for major allele (blue), heterozygous (green), homozygous for minor allele (red), and missing genotype (black).

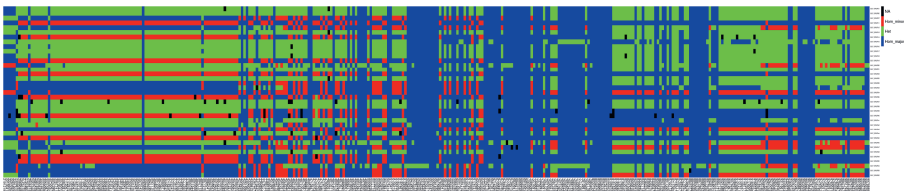


Figure S5.8. Genotypes of the variants between 2 and 10 Mb on chromosome 14 in 37 GLPs. Colors denote homozygous for major allele (blue), heterozygous (green), homozygous for minor allele (red), and missing genotype (black).

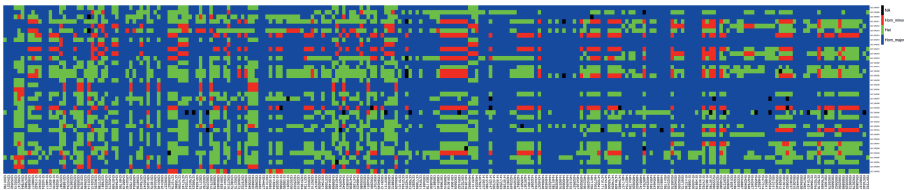


Figure S5.9. Genotypes of the variants between 12 and 16 Mb on chromosome 27 in 37 GLPs. Colors denote homozygous for major allele (blue), heterozygous (green), homozygous for minor allele (red), and missing genotype (black).

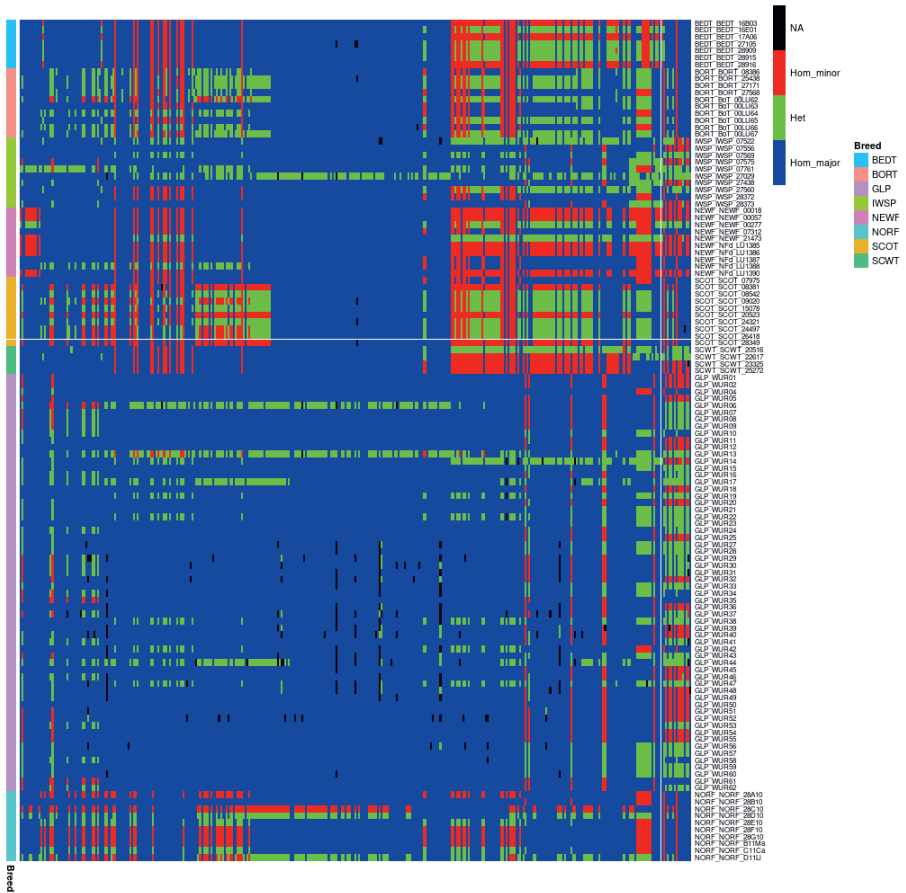


Figure S5.10. Genotypes of the variants between 0 and 5 Mb on chromosome 22 in German Longhaired Pointer (GLP), Bedlington Terrier (BEDT), Border Terrier (BORT), Irish Water Spaniel (IWSP), Newfoundland (NEWF), Norfolk Terrier (NORF), Scottish Terrier (SCOT) and Soft Coated Wheaten Terrier (SCWT). Colors denote homozygous for major allele (blue), heterozygous (green), homozygous for minor allele (red), and missing genotype (black).

6

A cancer gene score based on pathways and its application in driver mutation prediction using machine learning approach

Yun Yu¹, Pascal Duenk¹, Martien A.M. Groenen¹, Richard P.M.A. Crooijmans¹

¹ Wageningen University & Research, Animal Breeding and Genomics,
Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

Manuscript in preparation

Abstract

Identification of cancer driver mutations is one of the major challenges in oncogenic research, which helps to unveil the molecular mechanism of tumor initiation and development and enlighten targeted treatment development. Although many approaches have been developed for this purpose, it is still a challenge to identify driver mutations at high sensitivity and specificity. To achieve a higher prediction accuracy, relevant characteristics of driver mutations must be identified and taken into account as much as possible. With decades of study, many signaling pathways involved in cancer development and progression have been identified. The chance of tumor development due to mutations in the genes in these pathways varies considerably. In the current study, we proposed a new approach to incorporate abundant signaling pathway information in the driver mutation prediction. The new approach was also tested (trained) with a human dataset. We computed a cancer pathway score for each signaling pathway based on the fraction of driver genes among all genes identified in the pathway. Based on this score, we then computed a cancer gene score for each gene which is the sum of all the pathway scores of pathways that the gene is involved in. We observed higher scores for driver genes than for passenger genes, implying that this score is useful in distinguishing driver and passenger genes. We trained Random Forest Classifier models to predict driver mutations from missense mutations identified in tumors using the cancer gene score as a feature, together with 3 other features, namely, SIFT score, PolyPhen2 score, and recurrence of the mutation. On average, we observed a prediction accuracy of those trained Random Forest Classifiers, measured by F1 score (harmonic mean of precision and recall), of 0.90 (ranging between 0.85 - 0.94), demonstrating that these features, including the cancer gene score, could contribute to driver mutation prediction. Our findings enlighten a new way to incorporate in and use signaling pathway information in driver mutation prediction in the future.

6.1. Introduction

Somatic mutations in tumor cells can be classified into driver mutations and passenger mutations according to their roles in tumorigenesis. Driver mutations provide a selective advantage for the tumor cell and trigger/drive the tumor initiation and development, while functionally neutral mutations do not contribute to the tumor formation and growth and therefore are called passenger mutations [1]. The genes with driver mutations are called driver genes. Identification of driver mutations is one of the major tasks of oncogenic research, which can help to understand the molecular mechanism underlying tumorigenesis. In addition, it also contributes to the development of specific or personalized tumor treatments.

Many classifier tools have already been developed for the prediction of cancer drivers at the gene or mutation level. Mutation specific characteristics, such as conservation of the nucleotides, allele frequency, mutation ratio, and the surrounding sequence context of the mutation, contain useful information for the prediction of driver events and used by these classifier tools. For example, SIFT and PolyPhen-2 scores are in silico pathogenicity scores for missense mutations estimated based on conservation of the mutation position and physical-chemical properties of the different amino acids [2, 3]. Mutation ratio of the gene is used by the MuSiC to identify significantly mutated genes (SMG) that have a significantly higher mutation rate than the background mutation rate [4]. Likewise, signals of positive selection are used by OncodriveFML to identify driver mutations in not only coding regions but also non-coding genomic regions [5]. Some tools can integrate multiple characteristics to predict driver genes. For instance, MutSig2CV consists of three independent statistical tests based on abundance (mutation rate), clustering (mutational hotspots) and conservation (evolutionary conservation) to predict driver genes [6]. Likewise, CompositeDriver combines mutation recurrence and functional impact to identify coding and non-coding cancer drivers [7]. Meanwhile, protein domain information can also be used to predict driver genes. For instance, OncodriveCLUST identifies driver genes whose mutations are biased towards a large spatial clustering within the protein sequence [8]. e-Driver exploits the internal distribution of somatic missense mutations between the protein's functional regions (domains or intrinsically disordered regions) to find regions that show a bias in their mutation rate as compared with other regions of the same protein, providing evidence of positive selection, and suggesting that these proteins may be actual cancer drivers [9].

Besides, prior knowledge, mainly related to gene/protein interaction networks, has also been used in driver prediction. These interaction networks are constructed from prior knowledge of signaling pathways. For instance, SCS integrates expression data

and gene interaction networks to predict driver genes [10]. HotNet2 finds significantly mutated subnetworks according to both the frequency of somatic mutations in individual genes and the topology of the interactions between the corresponding proteins [11]. 20/20+ integrates features capturing mutational clustering, conservation, in silico pathogenicity scores, mutation types, protein interaction network connectivity, and other covariates to find driver genes by a machine-learning based ratiometric approach [12].

A signaling pathway describes a series of chemical reactions in which a group of molecules work together to control a cell function/activity. On the one hand, there are several signaling pathways that are associated with cancer formation and development, e.g., PI3K/Akt signaling pathways, RTK/RAS/MAP-Kinase pathway [13]. These pathways are often involved in cell proliferation, cell cycle, DNA damage repair and developmental activities. Mutations activating or deactivating these pathways can lead to uncontrolled cell growth and immortality of the cells. On the other hand, there are many signaling pathways that have an important function but that are not associated with cancer development. Examples are pathways associated with diseases instead of cancers such as the REACTOME glycosylation precursor biosynthesis pathway where mutations in genes involved in it may cause congenital disorder of glycosylation. Alterations in these kinds of pathways may influence the relevant phenotypes or cause certain diseases/disorders but would not trigger tumorigenesis. We noticed that the proportion of driver gene may be massively diverse between different pathways, for instance, between the PI3K/Akt and KEGG_PROTEASOME pathways. In the current study, we tested a new approach that incorporates abundant signaling pathway information, along with three features of the mutations themselves (i.e. SIFT, PolyPhen-2, recurrence), in the driver mutation prediction.

6.2. Materials and Methods

6.2.1. Model training data

The data used to train our machine learning model is derived from somatic mutations in coding (CDS) regions identified in 9,423 tumor exomes [14]. In total, the data consists of 751,876 somatic mutations, of which 3,437 are classified as driver mutations. To achieve a balanced dataset, we randomly sampled 3,437 passenger mutations, which resulted in 6,874 somatic mutations used for training. To avoid sampling error, the random sampling and model training were repeated 50 times. Features attached to each mutation include the SIFT score, PolyPhen-2 score, recurrence number of the mutation, and cancer gene score (see section Cancer pathway score and cancer gene score).

6.2.2. Cancer pathway score and cancer gene score

To calculate the cancer pathway and cancer gene score, we first constructed a Boolean matrix in which rows are genes and columns are signaling pathways to reflect the relationship between genes and signaling pathways. If the gene is involved in a pathway, the cell corresponding the gene and pathway is encoded as 1. If not, then the cell is encoded as 0. Signaling pathways used in this study are canonical pathways (v7.3) in curated gene sets (C2) of The Molecular Signatures Database (MSigDB) downloaded from the GSEA (Gene Set Enrichment Analysis) website [15]. In total, there are 1,569 REACTOME pathways, 615 WP (WikiPathways) pathways, 292 BioCarta pathways, 196 PID (the Pathway Interaction Database) pathways, 186 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, 10 NABA (Alexandra Naba) pathways, 10 SA (SigmaAldrich) pathways, 8 SIG (Signaling Gateway) pathways, 1 WNT pathway, respectively. This Boolean matrix was made in R. R package org.Hs.eg.db was used to convert the gene symbol to entrez ID. Package fgsea [16] was used to download and process the signaling pathways. This Boolean matrix was then used to calculate cancer pathway scores and cancer gene scores.

In theory, the more driver genes a pathway has, the higher the chance that a mutation occurring in a gene involved in that pathway leads to tumorigenesis. The percentage of driver genes in a pathway is called the cancer pathway score in this study. From the cancer pathway scores, we calculated a cancer gene score for each gene included in this study, as the sum of cancer pathway scores of all the pathways that the gene is involved in.

In the calculation of cancer pathway score, gene was removed from the constructed Boolean matrix if it has a mutation(s) included in the dataset for model training. Therefore, the target information (either driver or passenger) was not used in the cancer pathway and cancer gene score calculation. Duplicated data was removed and data with missing features was also removed from the datasets before model training.

6.2.3. Driver and passenger genes

The driver genes used in the cancer pathway score calculation are obtained from 2 resources, the Cancer Gene Census (CGC) v92 obtained from the COSMIC database (<https://cancer.sanger.ac.uk/census>) and 299 driver genes reported by the PanCancer Atlas workings group from The Cancer Genome Atlas (TCGA) [17]. The CGC is the biggest driver gene database consisting of a list of well-studied cancer genes. The CGC has two tiers of driver genes. Driver genes in tier 1 are more reliable, supported by more and stronger evidence than the genes in tier 2. In this study, only

the 570 genes in tier1 were included as driver genes. In addition, 299 driver genes identified from 33 types of cancer using 26 computational tools by the PanCancer Atlas group were also used in the cancer pathway score computation [14]. These two driver gene datasets have 180 genes in common, leading to 689 unique driver genes. Passenger genes were also obtained from the CGC. However, some of these genes were identified as driver genes in the study of the PanCancer Atlas group [17]. We therefore removed these common genes and finally retained 18,090 passenger genes.

6.2.4.Classifier model

The machine learning model trained in this study was a random forest classifier (RFC) with default hyperparameters in a supervised manner using scikit-learn package (v1.0.2) in python v3.8.5. Stratified 10-fold cross validation was applied to train the model and evaluate the performance of the model. The data was randomly divided into 10 folds of data with roughly equal size and each fold of data had approximately the same percentage of driver mutations as the complete data. Nine folds of data were used for model training and the remaining fold was used as a holdout set for validation each time. The training and validation were repeated 10 times, with each of the 10 folds of data used exactly once as the validation data. The advantage of this method is that all data is used for both training and validation, and each data is used for validation exactly once. Before training, data standardization was performed using the StandardScaler tool. To illustrate the contribution of cancer gene scores to the improvement of prediction accuracy of the machine learning model, we also trained the RFC model with the same hyperparameters and datasets but removing the cancer gene scores from the features.

6.2.5.Accuracy metrics

We evaluated the performance of our trained model using the F1 score because the driver mutations and passenger mutations were not entirely balanced. To calculate those metrics, TP (true positive), FP (false positive), TN (true negative), FN (false negative) adopted from the confusion matrix (Table 6.1) derived from the validation dataset were used.

Table 6.1. Confusion matrix.

	Predicted Passenger	Predicted Driver
Labeled Passenger	TN	FP
Labeled Driver	FssssN	TP

Recall: proportion of true driver mutations that are correctly assigned by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision: proportion of true driver mutations within predicted driver mutations.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

F1 score: the harmonic mean of the precision and recall.

$$\text{F1} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

6.3. Results

6.3.1.Cancer pathway score

The cancer pathway score has a range between 0 and 1 (supplementary Table S6.1). A score of 1 indicates that all genes in the pathway are driver genes. Only two pathways, `BIOCARTA_BARD1_PATHWAY` and `REACTOME_DISEASES_OF_MISMATCH_REPAIR_MMR`, have a cancer pathway score of 1. In contrast, a score of 0 indicates that all genes in the pathway are passenger genes. From the distribution of the cancer pathway score (Figure 6.1A), most pathways of the in total 2887 pathways, have a low cancer pathway score. Fifty percent of the pathways have a cancer pathway score of less than 0.125, of which 681 pathways (23.6%) have a cancer pathway score of zero (supplementary Table S6.1). None of these pathways with a cancer pathway score of zero is related to activities that might be involved in tumorigenesis, such as cell cycle and DNA damage repair activity. In contrast, 73 of the signaling pathways (2.5%) (supplementary table S6.1) did have a cancer pathway score higher than 0.6. Most of them are well-known cancer pathways (such as e.g. the p53, CTCF, PI3K-AKT, and

MAPK1-ERK2 pathways) and curated pathways for specific cancers (e.g., KEGG THYROID CANCER, BIOCARTA_MELANOCYTE_PATHWAY).

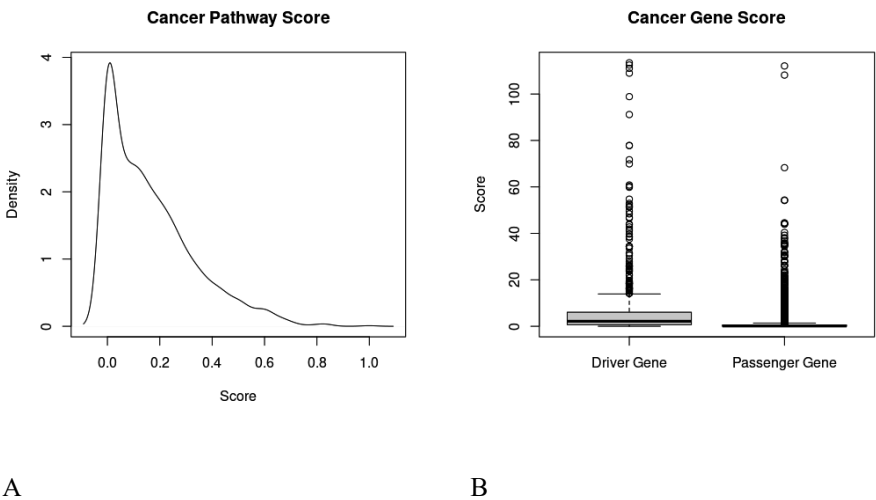


Figure 6.1. A. Density plot of the cancer pathway scores for all pathways. B. Boxplot for the cancer gene scores of driver and passenger genes.

6.3.2.Cancer gene score

The cancer gene scores range between 0.00 and 113.43 for the genes included in the current study. The gene with the highest cancer gene score is the driver gene *PIK3RI*. The average cancer gene score of driver genes is 7.08, for passenger genes it is 0.77. On average, driver genes have higher cancer gene scores than passenger genes (Figure 6.1B) (Wilcoxon rank sum test, p-value < 2.2e-16).

6.3.3.The training dataset and model training

In total, 50 training datasets were created by random sampling passenger mutations. We removed duplicated data (on average 1,504, ranging from 1,466 to 1,538) with exactly the same features and those with at least one missing value for one of the features (on average 529, ranging between 490 and 560). On average across datasets, there were 2,025 unique driver mutations and 2,818 unique passenger mutations (between 2,776 – 2,873) left for the model training.

6.3.4.Model prediction accuracy

We trained the RFC model 50 times using the 50 datasets generated. Because a stratified 10-fold cross validation strategy was used in the training, we ended up

with 500 F1 scores in total, where the average F1 score was 0.90 (ranging between 0.85 and 0.94) (Figure 6.2A).

The most important feature used by the RFC models was the cancer gene score, which has a feature importance of 0.66 on average across 50 datasets (Figure 6.2B), implicating the significant contribution of this score to the prediction of driver mutations. We also trained RFC models using the same hyperparameters and dataset but removing the cancer gene score from the feature, which resulted in an average F1 score of 0.71 (ranging between 0.63 and 0.76), a significant decrease comparing to the RFC models with the cancer gene score as a feature. This result also demonstrates that our cancer gene score contains valuable information that can be used by the RFC model to predict driver mutation more robustly. Besides the cancer gene score, the PolyPhen-2 score ranks second and the SIFT score ranks third in terms of the importance among features. The recurrence of the somatic mutation is the least important, with an average value of 0.03.

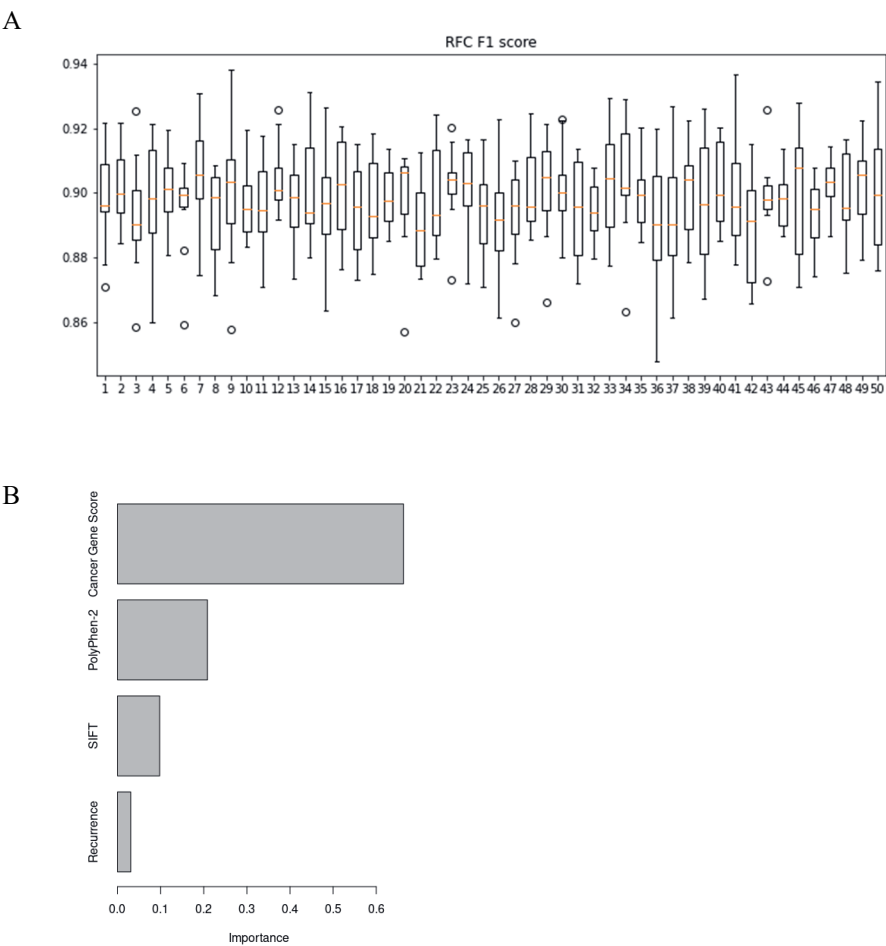


Figure 6.2. A. Boxplot of F1 scores derived from 50 Random Forest Classifiers trained with stratified 10-fold cross validation. The orange bar indicates the average value of F1 score. B. Average feature importance of 50 Random Forest Classifiers trained.

6.4. Discussion

A vast amount of signaling pathways have been identified and constructed over decades of research [18, 19]. These signaling pathways show how a group of molecules in a cell work together to regulate cell functions, such as cell division and hormone production. This information could help to identify potential driver genes

or mutations, which still is a major challenge in oncogenic research. In previous studies, protein-protein interaction networks were used to predict driver genes as well, while the consistence of prediction between them is still lacking when they are applied to the same independent dataset [10-12]. In the current study, we proposed another way to use this abundant signaling pathway information to predict driver mutations.

Driven by a simple idea that signaling pathways may have varied potential to drive tumorigenesis because of the different amount of driver genes involved in them, we computed a cancer pathway score to evaluate the potential to drive tumorigenesis for each defined signaling pathway from several resources. A pathway score of 0 means no driver genes were identified in the pathway, while a score of 1 means that all the genes in that pathway were identified as driver genes. In total, there are 681 pathways with a score of 0. Theoretically, mutations in the genes involving these pathways do not influence cell division, but of course may influence other activities of a cell, for instance hormone production. For pathways with a score of 1, a mutation in those genes resulting in changed expression or an altered protein could potentially increase the cell proliferation, and thus drive tumorigenesis.

Based on the cancer pathway score for each pathway, we calculated a cancer gene score for each gene, as the sum of score of all the pathways that the gene is involving in. The gene score is influenced by the number and score of pathways it is involved in. If a gene is only involved in pathways that have a pathway score of 0, it has a gene score of 0, implying that it is likely that mutations in that gene do not lead to tumorigenesis. If a gene is involved in many pathways with a relatively high pathway score, it has a higher gene score, implying that it is likely that mutations in that gene lead to tumorigenesis.

Some driver genes have very low scores, sometimes even zero. This is because these genes are missing in our signaling pathways. Even though we collected signaling pathways from different sources, not all the genes were represented. Additional data on signaling pathways in the future will likely yield a more accurate cancer gene score.

In the current study, the driver gene list is a combination of driver genes in the tier 1 list of the CGC database and 299 driver genes identified in the study of the PanCancer Atlas group. The passenger genes are all from the CGC database. Most of these driver genes have been identified and supported by experimental data. However, it is likely that some of the passenger genes, are misclassified due to limited number of tumor samples that have been sequenced and studied or the limited number of driver mutations identified. Likewise, we also noticed that some genes labeled as passenger gene in the CGC database have a high cancer gene score

(370 passenger genes have a cancer gene score higher than the average cancer gene score of driver genes). For instance, two genes labeled as passenger gene in the CGC database, *MAPK3* (mitogen-activated protein kinase 3) and *GRB2* (Growth factor receptor-bound protein 2), have a cancer gene score above 100. Although both genes are labeled as a passenger gene in the CGC database, they are very likely true driver genes according to their functions and both were reported to be associated with cancers [20, 21]. The role of these genes in driving tumorigenesis needs to be investigated further.

Some genes have been well studied, but some are not, simply because of difference in scientific interest, how long ago they were discovered, or how much effort has been invested in their research. The cancer gene score for genes with more information on the signaling pathways they are involved in may be skewed. Because the cancer pathway score and cancer gene score calculations completely depend on our prior knowledge on the genes and pathways.

The quality and quantity of the somatic mutation training data can be further improved. Up till now, there is no perfect benchmark driver and passenger mutation dataset available yet. The mutation data used in model training in the current study are somatic mutations in the CDS region from the report of the PanCancer Atlas group. According to their experimental validation, approximate 60%-85% of them are likely drivers. The 3,437 driver mutations in the report were identified in 5,782 out of 9,423 tumors. Still some tumors have unclear driver mutations. This implies that some of the passenger mutations could possibly be false negatives. Therefore, still some false-positive and false-negative driver mutations exist in this dataset. These potentially mislabeled mutations might limit the performance of our supervised machine learning model.

In the current study, we investigated if our gene score could be used to predict driver mutations as a feature in a machine learning model. We observed a significant improvement in F1 score after adding the gene score as a feature. Our study could enlighten driver mutation prediction algorithms in the future. To improve the performance of our model further, besides the SIFT score, PolyPhen2 score and recurrence of the mutation, other mutation related features could also be included. This could improve the prediction of, for instance, the frequency of the mutation in the population, the variant allele frequency, as well as the mutation type. Moreover, we haven't optimized the hyperparameters of the RFC model yet. A higher prediction accuracy may be achieved after hyperparameter tuning. Besides, more sophisticated machine learning models, such as a neural network machine learning model, might also yield better prediction performance.

The driver and passenger mutations used in the RFC model training are the missense mutations identified in the CDS regions of the genes. The prediction of driver mutations from synonymous mutations in the CDS region, mutations in intergenic and intronic regions of genes, or potential epigenetic drivers were not investigated in the current study. How to use cancer gene scores in the prediction of driver mutations in such non-CDS regions needs to be investigated further.

Conflict of interest: The authors have no conflict of interest to declare.

Supplementary materials: Supplementary **Table S6.1** can be found through this link

https://github.com/YunYu93/Data-depository/blob/main/Supplementary_Table_S6.1.csv

6.5. Reference

1. Greenman C, Stephens P, Smith R, Dalgleish GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446(7132):153-8.
2. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-4.
3. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-9.
4. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome research*. 2012;22(8):1589-98.
5. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*. 2016;17(1):1-13.
6. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495-501.
7. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371-85.e18.
8. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*. 2013;29(18):2238-44.
9. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics*. 2014;30(21):3109-14.
10. Guo W-F, Zhang S-W, Liu L-L, Liu F, Shi Q-Q, Zhang L, et al. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics*. 2018;34(11):1893-903.
11. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*. 2015;47(2):106-14.
12. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*. 2016;113(50):14330-5.
13. Sever R, Brugge JS. Signal transduction in cancer. *Cold Spring Harb Perspect Med*. 2015;5(4):a006098.
14. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371-85.e18.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 2005;102(43):15545.
16. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. *bioRxiv*. 2021:060012.
17. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*. 2018;18(11):696-705.
18. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30.
19. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res*. 2022;50(D1):D687-D92.
20. Zhang Y, Huang X, Wang J, Wang X, Liu X, Chen Y, et al. Nitration-induced ubiquitination and degradation control quality of ERK1. *Biochemical Journal*. 2019;476(13):1911-26.
21. Yang L, Wang X, Xu J, Wen Y, Zhang M, Lu J, et al. Integrated transcriptomic and proteomic analyses reveal α -lipoic acid-regulated cell proliferation via Grb2-mediated signalling in hepatic cancer cells. *J Cell Mol Med*. 2018;22(6):2981-92.

7

General discussion

7.1. Genetic testing

One of the major ultimate goals of the project described in this thesis was to help breeders to eradicate the familial FCC from GLPs, thus breeding healthy dogs. A genetic test can be used to achieve this goal with the premise of identification of a germline risk factor. In chapter 3, I describe two deleterious mutations (chr17:800788G>A (686F>V) and chr17:805276C>T (845T>M)) in the *TPO* gene to be germline risk mutations for the FCC. The former marker confers a slightly higher relative risk in the status of homozygous recessive genotype compared to the latter one (16.94 > 16.64), likely having a higher prediction accuracy for the risk of getting the familial FCC. We therefore developed a genetic PCR-RFLP test that determines the genotype of this marker (chr17:800788G>A (686F>V)) in dogs (chapter 3). The possible genotype derived from our genetic test can be one of three types for each tested dog, which is AA, AG, or GG. The AA genotype (homozygous recessive genotype) of the marker gives a relative risk of 16.94 compared to homozygous genotype for the reference allele (GG). According to our study, eighty-three percent of the dogs (45 out of 54) with AA genotype developed FCC, which fits well the autosomal recessive inheritance pattern as assumed according to pedigree.

A disease, including cancer, with an autosomal recessive inheritance pattern, is difficult to be eradicated from the animal population by a conventional selective breeding strategy based on phenotypes once that phenotype has spread in the population for over several generations. Because it is impossible to differentiate the heterozygous animals from homozygous wild-type animals, thus not able to remove completely the unwanted recessive allele from the population. A genetic test can be valuable in breeding with the premise of identification of the germline risk factor because the genetic test directly yields the genotype for the animals, thus making it possible to directly select on genotype, instead of phenotype. In practice, to be able to eradicate the harmful mutation that is associated with increased incidence of a recessive disease, it is necessary to remove not only animals with the homozygous recessive genotype but also those with the heterozygous genotype from the breeding program. Removing only animals homozygous for the recessive allele can prevent the disease in the next generation, but the recessive allele will still be segregating in the population through using heterozygous animals in the breeding program. Therefore, to completely remove the unwanted allele, both homozygous recessive and heterozygous animals must be removed from the breeding program in order to eradicate the recessive disease. However, a risk of removing all the dogs carrying the risk allele is that it may drastically decrease population size and restrain genetic diversity within the population by increasing inbreeding. Another breeding strategy for leveraging disease and genetic diversity is to remove only homozygous recessive

animals from the breeding program, in addition to prohibiting mating between heterozygous animals. This breeding strategy will not produce homozygous recessive animals that predispose to the disease either. Meanwhile, fewer animals will be discarded from breeding, therefore, more genetic diversity can be preserved in the population. However, the shortcoming along with this strategy is that the risk allele will stay in the population and a genetic test always has to be performed for all breeding animals.

For a disease with a dominant inheritance pattern, theoretically, directly selecting animals based on the phenotype is already able to eradicate it. For instance, for an inherited disease with an autosomal dominant inheritance pattern, the unaffected animals are usually the homozygotes for the non-dominant allele. Removing affected animals and using only unaffected animals in the breeding program will eradicate the dominant allele. Nonetheless, A genetic test is still recommended because this can speed up breeding for healthy animals. All the animals with the dominant allele can be identified in one round of genetic testing. Furthermore, some disease may only manifest signs at quite late age and selection on such a phenotype can take many generations. In this case, most importantly, a genetic test can prevent producing animals susceptible to the disease, thus improving animal welfare. Because without a genetic test, carriers of the disease might not be recognized in time and then used to produce offspring. Moreover, a dominant disease can have incomplete penetrance which makes selection based on phenotype even less efficient. When a dominant allele has a low penetrance, for instance 50%, selection on phenotype can only identify approximately 50% of carriers of the disease.

In livestock industry, a genetic test can be used to identify the animals that are susceptible to corresponding diseases, therefore, making eradication of a disease from the population possible. This strategy is also known as marker-assisted selection. A good example is the *RYRI* gene testing to identify pigs with malignant hyperthermia syndrome (also known as porcine stress syndrome) (Houde et al. 1993; Ferreira de Camargo 2019). In humans, a genetic test enables early interference and prevention of inherited cancers. A good example is the *BRCA1/2* testing for breast cancer risk (Padamsee et al. 2017). According to a survey, up till August 1, 2017, there were approximately 75,000 genetic tests available for humans, representing around 10,000 unique test types. Most of these are single-gene tests, but some are panel tests, whole-exome sequencing tests, and whole-genome sequencing tests (Phillips et al. 2018).

A genetic test is highly valuable, but also has its' limitations. One of the biggest limitations is probably that a genetic test can only be developed and used for an inherited trait/disease with a simple inheritance pattern. For diseases with complex

inheritance, only testing for one or a few markers will not allow to accurately determine the phenotype because many more mutations are involved in the phenotype formation. In this case, a polygenic risk score (PRS) can be used to assess disease risk. However, PRS is currently only applied in research studies, and not yet used in clinical diagnoses (Lewis and Vassos 2020). Meanwhile, a genetic test is also not able to predict the chance of occurrence of a sporadic cancer that accounts for the majority of cancers.

The genetic test developed based on the variant chr17:800788G>A (686F>V) identified in the *TPO* gene has been made commercially available for GLP breeders and owners through the Van Haeringen Laboratories after testing 142 GLPs at the Animal Breeding and Genomics laboratory. Among those tested GLPs, 5 are homozygous for the germline risk variant (AA genotype) while 67 are heterozygotes (GA genotype), yielding a frequency of 25.4% for the germline risk variant. In 182 samples that were genotyped for the segregation analysis (chapter 3), 54 dogs are homozygous recessive and 67 are heterozygotes, with a frequency of 48.1% for the germline risk variant. Because sampling in both rounds is biased towards affected dogs the frequencies of the germline risk variant is expected to be higher than that in Dutch GLPs as a whole. The FCC cases were over-represented in the 182 GLPs used for segregation analysis. Likewise, the 142 GLPs were tested because, according to the pedigree, they are closely related to known affected dogs.

Breeders highly appreciate this genetic test and use it to assist their breeding program to breed healthier dogs. Before a GLP will be part of the breeding program, the genetic test has to be performed for that dog. Knowing the genotype of those GLPs, the breed association allows only crosses between GLPs who are homozygous for the wild type allele (G allele) and are considering heterozygous dogs into the breeding program only in a cross with a homozygous wild type (G allele) animal. Ideally, by applying this strategy of breeding after testing, it is expected that the familial FCC associated with the homozygous status of that allele will not be present in future offspring.

The future affection status of the GLPs that underwent the genetic test will be traced in collaboration with the breeder association. This information can be valuable to evaluate the prediction performance of the genetic test and provide some clues if the variant is the actual causal mutation or just a mutation in a high linkage disequilibrium with the unknown causal mutation. A significant reduction in the incidence of thyroid cancer in the offspring of GLPs with GG genotype will demonstrate the value of this genetic test in saving the GLPs from thyroid cancer by breeding.

7.2. Follicular cell carcinoma formation

7.2.1. Genomic evidence of potential causative role of the germline risk mutation

To determine the causative role of a germline risk factor is challenging. MacArthur et al. (2014) recommended five key areas to be considered in identification of causal variants: study design; gene-level implication; variant-level implication; publication and databases; and implication for clinical diagnosis. My study to identify the potential germline causal variant for the familial FCC adhered to these recommendations well. Particularly, regarding study design, I used both SNP array data and whole-genome sequence data. The method used to locate the candidate target region in the study included a GWAS and a homozygosity mapping analysis. Homozygosity mapping was also used because the inheritance pattern of the familial FCC in these GLPs was most likely autosomal recessive according to my investigation on pedigree. In the GWAS analysis, population stratification was adjusted by incorporating the genomic relationship matrix calculated from genotype data as a random effect. Meanwhile, to exclude potential sporadic cases due to other causes, such as aging and to maximize homogeneity of the phenotype, only cases with an age at diagnosis less than 10 years were included in the GWAS and homozygosity mapping analyses. Next to the target region identification, I took into account the function of genes identified in the target region during fine-mapping. Tumor formation occurs only in the thyroid gland suggesting that the causal gene might specifically be expressed and function in the thyroid gland. The *TPO* gene was the only gene that has a pivotal role in thyroid function among all genes identified in the target region. In mice and rats, inhibition of TPO activity has been found to possibly lead to enlargement of the thyroid gland and thyroid tumors (Hoshi et al. 2009). To identify the candidate causal variant, a series of variant-level characteristics was considered, including the pathogenicity score, evolutionary conservation score, and segregation analysis. Likewise, a public database that contains SNPs derived from WGS of 722 dogs from over 144 modern breeds, 54 wild canids and 100 village dogs was also used to investigate the prevalence of the mutations identified in the target region. Taken all above evidence together, the two deleterious mutations identified in the *TPO* gene are the most likely mutations responsible for the FCC. To claim the causative role of the identified risk mutations in the *TPO* gene, however, experimental evidence is still needed beyond those statistical evidences. Meanwhile, these two deleterious mutations in the *TPO* gene are almost in complete linkage disequilibrium. It is not clear which of them is the true causal mutation or whether both are required. Further studies are needed to answer this question.

As introduced in chapter 1, due to two bottleneck events in breed dog history and continuous human selection, pedigree dogs generally have a low genetic diversity within breeds. This facilitates identification of target regions using conventional association studies with a smaller number of both markers and samples compared to many other species (Machiela and Chanock 2014). In our case, a GWAS analysis with only 64 dogs (28 cases and 36 controls) yielded the significant signal for the FCC in the GWAS analysis using 170K SNP array data. Even a smaller sample size allows identifying the target region through a GWAS analysis. For instance, for retinitis pigmentosa in the Miniature Schnauzer, only 10 cases and 33 controls yielded a significant single target region (Kaukonen et al. 2020). This again illustrates dogs' advantage in a GWAS regarding simple diseases.

7.2.2. A hypothesis on mutagenesis mechanism of the germline risk mutation

Given all the evidence from our genomic analyses, the deleterious mutations in the *TPO* gene are most likely the causal mutations. Or at least, the role of mutant *TPO* in FCC tumorigenesis is highly suspected. However, the possible molecular mechanism leading to the familial FCC is not clear yet. Nonetheless, the alteration in metabolism of hydrogen peroxide (H_2O_2) in the thyroid gland is highly suspected (Bann et al. 2019). The role of H_2O_2 in tumorigenesis is well recognized as a type of reactive oxygen species (ROS) that can lead to more oxidative DNA damages. H_2O_2 produced by *DUOX2* in follicular cells of thyroid gland, is proposed to be a very likely cause of frequent mutagenesis in the thyroid gland (Krohn et al. 2007; Ameziane El Hassani et al. 2019). It has been reported that activating *DUOX2* mutations may lead to thyroid cancer through regulating the production of H_2O_2 in thyroid follicular cells (Bann et al. 2019). H_2O_2 is involved in three steps catalyzed by *TPO* enzyme in the process of thyroid hormone production as introduced in the chapter 1. Alteration in consumption of H_2O_2 due to the deleterious mutations identified in the *TPO* gene may result in excess accumulation of H_2O_2 in the microenvironment of the thyroid gland. Particularly, those two deleterious mutations in the *TPO* gene likely impair the function of *TPO* enzyme, which might result in reduced affinity to H_2O_2 , or decreased catalytic activity of *TPO* enzyme. Anyway, the consumption of H_2O_2 might decrease due to identified mutations, thus leading to elevated accumulation of H_2O_2 in the thyroid follicular cells. Excess H_2O_2 results in higher oxidative stress and ultimately more oxidative DNA damage. Currently, to the best of my knowledge, there is no study that investigated the association between cellular H_2O_2 level and mutations in the *TPO* gene, whereas the association between the level of H_2O_2 and mutations in the *DUOX2* gene has been confirmed (Bann et al. 2019). *DUOX2* generates H_2O_2 while, *TPO* consumes H_2O_2 . It is therefore expected

that TPO has an opposite impact on cellular H_2O_2 level to DUOX2 in the case of similar type of mutations, loss-of-function or gain-of-function, in these two genes. Activating mutations in the *DUOX2* gene potentially contribute to increased tumorigenesis because they increase H_2O_2 production. It is then possible that a hypomorphic mutation (a mutation leading to decreased activity of the product) in the *TPO* gene also contributes to increase tumorigenesis because it may reduce H_2O_2 consumption thus resulting also an increased H_2O_2 level.

Almost every gene involved in thyroid hormone production has been associated with thyroid cancer. Loss-of-function mutation in the *TG* gene can lead to dysmorphogenetic goiter and further progression into thyroid carcinoma (Alzahrani et al. 2006). The *FOXO1* gene may function as a tumor suppressor in the early stage of PTC in humans. The silencing of the *FOXO1* gene significantly promotes PTC cell proliferation, migration, and invasion in vitro (Ding et al. 2019). Association between H_2O_2 change in thyroid gland and mutations in those genes has not been investigated yet to the best of my knowledge. The potential central role of H_2O_2 in thyroid tumorigenesis has been aware of but is still not sufficiently studied yet. In theory, loss-of-function mutations in these genes might lead to insufficient amounts of key substances (TG or iodide) for TPO catalyzed reactions, thus leading to excess accumulation of H_2O_2 .

It has been well established that loss-of-function mutations in the *TPO* gene and other genes involved in thyroid hormone synthesis can result in congenital hypothyroidism, (Penna et al. 2021). Congenital hypothyroidism can progress into TC in humans, even though it is not common (Penna et al. 2021). Likewise, higher concentration of H_2O_2 is also one of the proposed mechanisms of that progression.

7.2.3. Possible validation method

Comparing H_2O_2 levels in thyroid glands between normal dogs and dogs susceptible to the FCC can elucidate if H_2O_2 is potentially involved in the tumorigenesis. If there is a higher H_2O_2 level in thyroid gland tissue from dogs at a higher risk of TC than that from other dogs, the potential role of H_2O_2 in tumorigenesis is then suggested. The method to measure extracellular and intracellular H_2O_2 level has been developed (Bann et al. 2019). I recommend investigating the H_2O_2 level in the thyroid gland in different conditions, such as hypothyroidism and different types of thyroid tumors, to uncover the role of H_2O_2 in different thyroid diseases. Moreover, the association between the risk mutations identified in aforementioned genes that involved in thyroid hormone synthesis and H_2O_2 level is also worth an investigation.

Guanine is particularly susceptible to singlet oxygen. Reactive oxygen species can induce 8-oxo-7,8-dihydroguanine (8-oxoG) which can mis-pair with adenine,

leading to C>A/G>T transversion (Poetsch 2020b). To repair induced 8-oxoG, the site is excised by 8-oxoguanine DNA glycosylase (OGG1) leaving an apurinic site (AP site). Measuring the DNA damage (8-oxoG) or repair intermediates (such as AP sites) can illustrate the oxidative stress within the tissue (Poetsch 2020b). By comparing the oxidative stress between normal thyroid glands and thyroid glands at a higher risk of tumor development can also shed lights on if oxidative stress plays a role in tumorigenesis.

7.2.4. Expected serum thyroid hormone level change

If the enzymatic activity of TPO is deficient due to the mutations, the production of thyroid hormone is expected to decrease because TPO plays a pivotal role in thyroid hormone synthesis, whereas serum TSH level is expected to increase through a negative feedback loop mechanism. We had serum T4 level measured for 20 cases who are also the homozygous for the mutations identified in the *TPO* gene. Among them, eight dogs had a low level of serum T4, and one dog had a high level. The rest of the dogs had a normal range of serum T4 level. Regarding the serum TSH, only 4 homozygous cases had it measured, of which one had a normal level, one had a low level, and two had a high level. It is noteworthy that the T4 and TSH in serum were compared with reference intervals derived from general populations of dogs. However, it was warranted that serum T4 and TSH levels vary significantly between breeds (Hegstad-Davies et al. 2015). Hunting dogs are in general more athletic and active than other dogs, probably having a higher interval of thyroid hormone level in serum. Therefore, a GLP specific reference interval is needed to determine the change of serum T4 and TSH levels. Meanwhile, T4 and TSH levels were measured at the diagnosis of the FCC, which might be different from its level before tumor formation because besides the *TPO* mutation, tumor formation and growth may also influence the generation of thyroid hormones.

7.2.5. Evidence for an identified driver mutation

The *TPO* is involved in key steps in thyroid hormone synthesis but seems not to be an oncogene or a tumor suppressor gene nor is directly involved in cell proliferation or mutagenesis. It is therefore reasonable to suspect that a key somatic mutation, called driver mutation, is still required for tumorigenesis. I therefore profiled the somatic mutation landscape of several FCCs and identified a promising driver mutation chr24:43657087C>A (*GNAS* A204D) in the FCCs.

Cancer is an evolutionary process and driver mutations that drive tumorigenesis are favored by positive selection during tumorigenesis as introduced in the chapter 1. Recurrence of a specific mutation or multiple mutations in one specific gene reflects positive selection favoring the expansion of cells with the mutation(s). Therefore,

recurrence is an important indicator of a driver role of somatic mutations. The *GNAS* A204D somatic mutation was identified in our GLP FCCs with a high recurrence where 4 out of 7 whole-genome sequenced FCCs and 20 out of 32 affected GLPs harbor the somatic mutation. In addition, the *GNAS* gene was identified to be a significantly mutated gene by two prediction tools, MuSiC2 and dNdScv, used in the study. Meanwhile, given the pathogenicity prediction of damaging for the mutation and high conservation in species evolution for the locus, *GNAS* A204D is a promising driver mutation.

GNAS somatic mutations have been identified in other canine tumors as well. Kool et al. (2013) identified a few somatic mutations in the *GNAS* gene in approximately one third of 44 canine cortisol-secreting adrenocortical tumors. Somatic mutations identified in the *GNAS* gene in their study are different from the *GNAS* A204D somatic mutation identified in our canine FCCs, but analogous to somatic mutations identified in human tumors.

7.2.6. Validation of effect of *GNAS* mutation

The *GNAS* gene encodes alpha-subunit of stimulatory G protein (*Gas*), which integrates multiple extracellular factors and intracellular responses through G-protein-coupled receptor (GPCR) signaling pathways (Patra et al. 2018). *Gas* regulates a series of downstream cascades of signaling pathways, such as MAPK signaling pathway, PI3K/AKT signaling pathway, and PKA signaling pathway, through regulation of cAMP level within cells (Turan and Bastepe 2015). To determine whether the *GNAS* A204D mutation has a functional impact, a measure of cellular cyclic AMP level through an *in vitro* experiment is recommended. The method to measure cellular cAMP has been well established already (Bohnekamp and Schöneberg 2011). An increase in cAMP level in cultured cells with the mutant *GNAS* gene in comparison to cells with the wild type *GNAS* gene indicates activating role of the mutation and a decrease indicates a deactivating impact.

7.2.7. Interaction between germline risk and driver mutations

The germline risk factor and somatic driver mutation were investigated independently (chapters 3 and 4). However, is the occurrence of the *GNAS* A204D somatic mutation a stochastic event or the consequence of a specific factor? The former scenario assumes that somatic mutations accumulate randomly in the genome and highlights the role of positive selection that favors mutation expansion through subclone expansion. In the latter scenario, it is assumed that there is a common mutational process that results in the specific *GNAS* A204D somatic mutation in the thyroid follicular cells. In the studied dogs, this specific induction of *GNAS* A204D somatic mutation is very likely for two reasons: 1) a high prevalence of *GNAS*

A204D in familial FCCs and 2) the concurrence in 6 pairs of bilateral FCCs. The germline risk mutation in the *TPO* gene might be the common factor that specifically induces the *GNAS* mutation. A common environmental factor resulting in carcinogenesis is unlikely in these dogs because affected dogs were identified from different breeders and owners across the Netherlands. A germline variant by somatic mutation (G×M) association has been proposed (Ramroop et al. 2019). In human head and neck cancers, the association between germline mutation and somatic mutation was identified (Feng et al. 2022). It has also been uncovered that germline variants can associate with different types of somatic mutations, such as copy number alternations (CNAs) and single-nucleotide variants (SNVs) (Ramroop et al. 2019). A robust association between somatic mutation profiles and polygenic risk score and an inverse association between germline risk factor and total somatic mutation counts in human cancers have been detected (Zhu et al. 2016; Liu et al. 2022). In dogs that were studied in this thesis, there seems to be an interaction between germline risk mutation in the *TPO* gene and driver mutation in the *GNAS* gene as well, but it is weak as no significant interaction was detected by a Chi-squared test. However, statistical power is probably lacking in this test because I had limited sporadic cases (8 cases) according to genotypes of the germline risk mutations in the *TPO* gene.

Independent tumors are expected to have different somatic mutation landscapes. Different driver mutations between a pair of bilateral tumors might suggest independent tumorigenesis mediated by the same germline predisposition. It's reported that bilateral neuroblastoma in humans having independent lesions is mediated by a germline predisposition with evidence of no shared somatic variants between bilateral tumors (Coorens et al. 2020). Concurrent driver mutations in bilateral tumors can have a few explanations including 1) the tumor in another paired organ is a metastasis; 2) the mutation occurred during early embryogenesis before lateralization of two thyroid glands; 3) interaction between the germline predisposition and the occurrence of the driver mutation. Among our 54 histologically examined FCC cases, 37 had bilateral tumors, of which 15 were genotyped for the *GNAS* somatic mutation while 2 failed in genotyping. Among those 13 pairs of bilateral tumors, concurrent *GNAS* A204D mutation was identified in 6 dogs, single *GNAS* A204D mutations was identified in 5 dogs, and two dogs had no *GNAS* mutation. According to the presence of *GNAS* A204D in those bilateral tumors, it seems that both independent tumorigenesis and interaction between germline and somatic mutation are suggested. However, a single mutation only provides limited information on relationship between the pair of tumors. A pairwise comparison of genome-wide somatic mutations between bilateral tumors is recommended to disentangle the relationship between a pair of bilateral tumors.

7.2.8. Hypothesis of a possible bridge by H₂O₂

How could the mutant TPO induce the *GNAS* A204D mutation in thyroid follicular cells? My hypothesis is that H₂O₂ might play a bridging role between these two types of mutations. Particularly, *TPO* 686F>V results in increased accumulation of H₂O₂ in thyroid gland microenvironment, thus increases oxidative stress, which results in increased accumulation of somatic mutations in the genome and at a certain point, the occurrence of somatic mutation *GNAS* A204D (Figure 7.1). Subsequently, positive selection takes over. Follicular cells with the *GNAS* A204D mutation have faster divisions than other cells, therefore, leading to the expansion of its progeny. In this hypothesis, the *TPO* mutations do not directly increase follicular cell division thus not result in cancer directly. The somatic driver mutation increases the cell proliferation and finally leads to FCC formation and development. Both germline and somatic mutations are involved in tumorigenesis of this particular familial FCC. However, the mutational process in this case is different from that described for a two-hit model. The two-hit model describes the inactivation of a tumor suppressor gene that leads to tumorigenesis. One mutation is inherited from one of parents and another mutation of the same gene is somatically acquired (Foulkes and Polak 2020). To validate if the *GNAS* A204 was derived from oxidative damage, detection of 8-oxoG or AP site at the mutation locus in the thyroid gland of susceptible dogs may yield some clues.

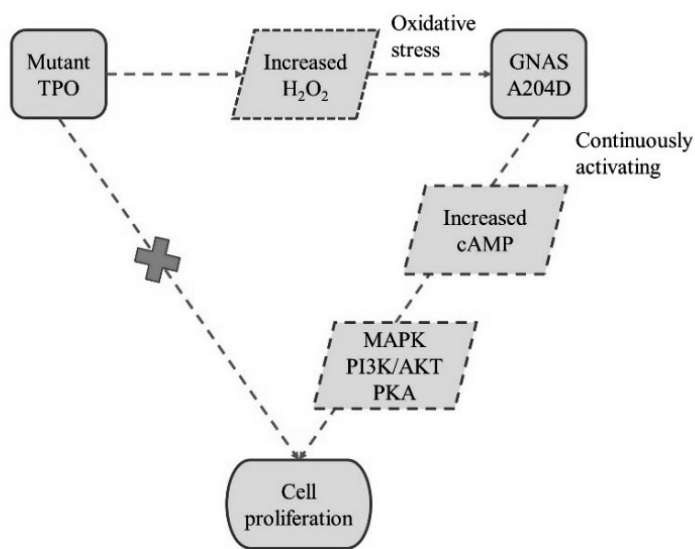


Figure 7.1. A scheme showing the hypothesis of the molecular mechanism underlying canine FCC. Trapezoids in dashed outline represent assumed and unvalidated effects.

It seems that this hypothesis can also explain several typical characteristics of this familial FCC well, such as absence of a sex difference, varied onset ages, high but incomplete penetrance, and also carcinogenesis of only follicular cells but not para-follicular cells, by emphasizing the role of H_2O_2 .

7.2.9. somatic mutations of a familial cancer

Somatic mutation profiling in chapter 4 is quite different from most studies conducted in human or canine cancers. Most studies in humans and dogs investigated sporadic cases, whereas the studied canine FCC in this thesis is a familial cancer with the same genetic cause for all affected dogs. A familial cancer has (an) inherited germline causal mutation(s) leading to an increased risk for it. A second mutation, i.e. driver mutation, is either essential (second-hit theory) or not, depending on what the germline causal mutation is. Germline causal mutations in some genes (oncogene, cell cycle gene etc.) can result in tumor initiation and development solely. For instance, germline activating mutations in the *RET* gene in humans can result in multiple endocrine neoplasia type2 (MEN) (Castellone and Melillo 2018). A somatic driver mutation is not essential in this case. However, germline causal mutations can also be located in genes that have no direct involvement in proliferation activity, such as the *DUOX2* gene (Bann et al. 2019). These genes have important functions to normal functioning of organs/tissues. For

these types of germline causal mutations, a somatic driver mutation in genes involved in proliferation is usually needed. Familial FCC belongs to the latter scenario that besides a germline risk mutation, also requires a somatic driver mutation.

7.2.10. Mutational signature

Mutational signatures can help to identify the potential mutational process that results in somatic mutations in the tumor. We fitted the somatic mutations identified in the FCCs to known human mutational signatures and identified SBS5 and SBS40 signatures in our canine FCCs. The mutational signatures with etiology of reactive oxygen species, such as SBS18 and SBS36 (Poetsch 2020a), were not identified in investigated FCCs. However, this does not necessarily reject the potential role of H_2O_2 in the FCC development. On the contrary, identification of flat mutational signatures SBS5 and SBS40 indicates an endogenous source of somatic mutations. Mutational signature can differ in frequencies of different trinucleotide substitutions between species (Riva et al. 2020). Construction of species-specific mutational signatures can improve accuracy and reliability of mutational signature identification in animal tumors, thus better inferring mutational processes. However, according to Kucab et al. (2019), H_2O_2 does not yield clear mutational patterns in human cells, although anticipated to create ROS. Thus, it is currently not possible to investigate the proportion of contribution of H_2O_2 to somatic mutations through fitting mutational signatures.

7.2.11. Passenger mutation

A somatic mutation that does not confer any growth and survival advantage to clonal expansion is called a passenger mutation. In evolution, it is a neutral mutation. The majority of somatic mutations (~99%) is classified into this type. Currently, the majority of studies were driver-centric. Recently, the role of passenger mutations has become controversial (Kumar et al. 2020). It was proposed that passenger mutations can act as a “mini driver” (Castro-Giner et al. 2015) or a “deleterious passenger” (McFarland Christopher et al. 2013). However, identifying these mini drivers is challenging. Currently available algorithms are weak to identify those “mini drivers”. Meanwhile cohort size also limits the statistical power of identification of driver events. Existing cohort sizes only allow the identification of strong, positively selected driver events, common within a cohort, but are underpowered to detect many weaker drivers and even rare strong drivers (Kumar et al. 2020).

Only 7 FCC samples were used in my genome-wide somatic mutation analysis. This probably allows me to detect only strong driver mutations. The canine FCC investigated in this study is a rather unique familial cancer. High homogeneity in

driver events was expected and a strongly selected driver was also assumed. However, the power is indeed lacking to detect those potential “mini drivers” in these canine FCCs. According to The Pan-Cancer Analysis of Whole Genomes (PCAWG) group, most human tumors have ~5 driver mutations in total (Consortium 2020). We identified only one driver mutation in four FCCs and no driver in the remaining three FCCs. This suggests that there are probably some mini drivers that have not been identified even in the 4 FCCs harboring the *GNAS* A204D mutation due to for example limitations of the used detection tools or the small cohort size.

7.2.12. The possibility of a germline non-coding causal variant

Most somatic mutations in human cancers are identified in non-coding regions. The same is seen in our canine FCCs where only 114 out of 11,250 somatic mutations (SNV + InDel) occurred in coding regions. However, somatic mutations in non-coding regions can also drive tumorigenesis. For instance, Rheinbay et al. (2017) showed that promoter regions harbor recurrent mutations in human breast cancers with functional consequences. Currently, identification of drivers in non-coding regions is still challenging due to difficulties in precisely locating non-coding elements that might contain drivers. Better functional annotation of the genome will improve driver mutation identification in non-coding regions.

Currently, large-scale annotation of functional elements is still lacking for the dog genome. The depth of annotation of the dog genome is much lower than that of the human and mouse genomes (Megquier et al. 2019). Good annotation of the genome, including not only protein coding genes, but also many kinds of functional elements (e.g., regulatory elements, non-coding RNA genes), can facilitate and improve studies involving genomic analyses such as causal variant identification for disease and elucidating the genetic basis of general phenotypes. ENCODE (Consortium 2004), Mouse ENCODE (Stamatoyannopoulos et al. 2012), DANIO-CODE (Tan et al. 2016), and GENE-SWitCH (GENE-SWitCH 2019-2023) are projects that comprehensively annotate functional elements in genomes of human, mouse, danio, and livestock respectively. Similarly, the currently ongoing Dog Genome Annotation (DoGA) project aims to improve annotation of the dog genome (DoGA 2017). Knowledge derived from this project will improve our understanding about the genetics of diseases in dogs.

7.2.13. Other types of variants

In addition to conventional genetic variants, epigenetic changes can also lead to tumorigenesis (Darwiche 2020), such as a switch-off of a tumor suppressor gene due to hypermethylation and repressive histone modification, and oncogene switch-on due to hypomethylation and activating histone modification. Besides the methylation

level changes in specific genes, variation of DNA methylation is likely to be a genome-wide regulatory pattern (Lu et al. 2020). In human TC, DNA methylation alterations and microRNAs have been revealed to be involved in pathogenesis (Zafon et al. 2019; Ghafouri-Fard et al. 2020). However, other types of epigenetics (e.g. histone modifications) in human TCs are still poorly studied (Zarkesh et al. 2018). In our canine FCCs, epigenetic changes have not been investigated though possibly they also drive FCC progression. Therefore, it is recommended to investigate the epigenome, such as methylation and histone modifications, of those canine FCCs as well.

7.2.14. New reference genome assembly

Recently, a few new canine reference assemblies have been released based on a combination of Illumina short-read sequencing and PacBio or Oxford Nanopore Technology (ONT) long-read sequencing: CanFam_GSD1.0 (Field et al. 2020), CanFam_Bas (Edwards et al. 2021), UU_CFam_GSD_1.0 (Wang et al. 2021), and UMICH_Zoey_3.1 (Halo et al. 2021). Compared to the historic CanFam3.1 reference assembly, these new assemblies in general achieved considerably higher completeness of the genome and genes, much less gaps, and better functional annotation of the sequence. In addition to these new reference genome assemblies, a pan-genome constructed from multiple individuals representing the genetic diversity of the species will also improve completeness of the reference genome (Computational Pan-Genomics 2018). This will improve downstream WGS mapping and variants detection, thus can benefit all the analyses based on detected variants, such as fine-mapping, driver mutation identification, selective sweep analysis etc.

7.2.15. Challenge in somatic InDel, CNA and SV identification

Somatic InDel identification in canine tumor bulk tissue is still challenging according to my experience. In chapter 4, three callers were used to identify consensus somatic InDels (captured by ≥ 2 callers). Of the final consensus somatic InDels identified, according to our visual inspection in a genome browser, only 57% are true InDels. Somatic SVs and CNAs are even notoriously more difficult to detect than small variants (SNV and InDel) using short-read next-generation sequencing data. This is because of several challenges, such as low purity of tumor samples, intratumor heterogeneity, limitations of short-read sequencing data, and sequence alignment ambiguities (Zare et al. 2017; Zaccaria and Raphael 2020; Gong et al. 2021). Many somatic SV callers have been developed over the past several years. These tools use one or more of four methods: 1) split reads; 2) paired reads; 3) local assembly; 4) read depth (Lin et al. 2014). However, there is not a perfect one overwhelming others. Each caller has inherent advantages and limitations. For

instance, varied sensitivity to detect SVs of different sizes, and varied sensitivity to detect different types of SVs between callers. Both specificity and sensitivity of these tools are currently still limited. Therefore, it is recommended to use a combination of a few somatic SV callers to identify consensus somatic SVs (Lin et al. 2014). Likewise, somatic CNA detection through short-read WGS in bulk tumor tissue is complicated and challenging due to potential tumor purity and subclone structure (Zaccaria and Raphael 2020).

Characterizing intra-tumor heterogeneity and reconstruction of phylogeny of subclones within a tumor is important for understanding how a tumor evolves as cancer progresses and responds to treatment. However, this is also challenging, and different tools have been developed although giving quite conflicting results (Tanner et al. 2021). Subclone reconstruction is highly sensitive to both SNV and CNA detection (Salcedo et al. 2020). Accurately identifying somatic mutations is critical to reconstruct subclone structure.

Nowadays, long-read sequencing from PacBio or Oxford Nanopore sequencing platform has proven its superiority in SV identification compared to short-read sequencing. Whereas lower sequencing accuracy with Oxford Nanopore (~95%) is insufficient for precise identification of somatic point mutations (Sakamoto et al. 2020), PacBio HiFi sequencing can yield long reads with accuracies greater than 99.5% (Hon et al. 2020), which seems to enable precise identification of SVs and point mutations. Somatic point mutation identification requires deep sequencing-depth while long-read sequencing usually has low yields which may not satisfy somatic point mutation identification. Compared to long-read sequencing, short-read sequencing still has its advantages in many aspects, such as lower cost, high yield, and higher sequencing accuracy. Therefore, currently, a combination of short-read sequencing and Nanopore long-read sequencing might be the ideal strategy to accurately identify all types of somatic genetic mutations. However, this strategy is costly as somatic mutation identification requires deep coverage of the genome.

Likewise, single-cell sequencing may also solve problems in somatic mutation identification led by low purity and clone structure in a tumor bulk. Currently, this technology is rapidly evolving but still facing some technological challenges, such as amplification bias, genome loss, mutations and chimaeras arising during whole-genome amplification, uneven coverage across the genome and severe allelic drop events (Gawad et al. 2016; Huang and Lee 2022).

These emerging technologies are not completely ready to solve current challenges in somatic SV and CNA identification and subclone reconstruction within a tumor yet. Nevertheless, they have promising potential to boost somatic mutation research in

cancers with their inherent advantages compared to commonly applied methods currently.

7.3. Selection and incidence of FCC

The target region identified by GWAS and homozygosity mapping analyses shows evident genomic footprint of selection in affected dogs, while not in controls (Figure 3.1d). Affected GLPs have higher inbreeding than unaffected GLPs resulting from selection by humans. These imply that specific selection occurred in the subpopulation of those GLPs favoring the spread of the FCC in the population. From the pedigree, it is obvious that breeding favored spreading of the FCC. According to the pedigree, dam GLP52 and sire GLP905 contributed most to the number of affected GLPs. These two dogs seem to be cherished by breeders and used for breeding frequently. GLP905 had 140 offspring and GLP52 had 47 offspring up to the moment of pedigree collection. Unfortunately, frozen sperm of GLP905 is still being used even though this sire has passed away.

Since thyroid cancer cannot be an advantage for fitness or a desirable phenotype by the breeders, it is likely that thyroid cancer is linked to a desirable phenotype that was preferred by the breeders. In chapter 3, we identified a ROH segment of approximately 5 Mb which harbors the germline risk mutations in the *TPO* gene (chapter 3, Figure 3.1d). Germline risk mutations in the *TPO* gene may hitchhike an adjacent variant that is positively selected because of the high linkage between SNPs in that ROH segment. This selection resulted in increasing frequency of the risk mutations. In human genomes, deleterious mutations can hitchhike to a higher frequency due to linkage to sites that have been under positive selection (Chun and Fay 2011). For instance, it was found that causal variants that are responsible for Crohn's disease in humans hitchhiked to a relatively high frequency due to linkage disequilibrium with a positively selected variant, *OCTN1* 503F (Huff et al. 2012). However, unfortunately, no other phenotypes are available for these dogs, which makes the identification of the selected phenotype(s) impossible.

7.4. Difference between FCC in dogs and human TC

7.4.1 Difference in genetics of TC between dogs and humans

The thyroid glands in dogs and humans share similar functions. The morphology and histology of FCC in dogs is also highly similar to its corresponding type of TC in humans. It is therefore logical to assume that the genetics of the FCC is also comparable between the two species. This, however, seems not to be the case. To the best of my knowledge, reports of germline risk factors in the *TPO* gene in human TCs are limited (Cipollini et al. 2013; Zhu et al. 2015). Nevertheless, germline risk factors in the genes involved in the same signaling pathways, namely the thyroid

hormone synthesis pathway, have been reported a lot as mentioned before. Therefore, although different susceptible genes are identified between studied dogs and humans, the underlying molecular mechanism might be the same. Besides difference in germline risk factors, the driver mutation identified in canine FCCs in my study is also different from that in human TCs. *GNAS* somatic mutations were only identified in 2 out of 496 human thyroid tumors investigated in the TCGA project, and of different location from the *GNAS* A204 identified in our study. Besides our study, Campos et al. (2014) investigated the somatic mutation landscape of driver genes that were identified in human TCs and identified only two analogous mutations in 59 canine thyroid tumors, suggesting a difference in driver mutation landscape between canine and human TCs. In combination with our results in GLP, canine TCs seem to have distinct driver mutations from human TCs. I suggest adding the *GNAS* A204D mutation into the list of to-be-test driver gene that can lead to a canine thyroid tumor.

7.4.2 Differences in somatic mutations between different types of canine FCC

In humans, the number and type of somatic mutations was found to be similar between sporadic and familial TC (Moses et al. 2011). Although only familial cases were included in a whole-genome profiling of somatic mutations in this thesis. *GNAS* A204D was genotyped in all canine FCC samples we collected, including familial type, and assumed sporadic type according to the genetic test. It turned out that the *GNAS* A204D mutation occurred in both sporadic and familial FCC, while the prevalence of the *GNAS* A204D mutation in sporadic FCC (3/8) is lower than the familial FCC (16/23). However, the prevalence of *GNAS* A204D in assumed sporadic FCCs is considerably higher than the frequency of somatic mutations in the *GNAS* gene in human TCs included in the TCGA project (2/496).

Driver mutations are consistently shared across TCs in different histological types, PTC, PDTC, and ATC (Agrawal et al. 2014; Landa et al. 2016; Yoo et al. 2019), while this seems not to be the case with FTC in humans (Swierniak et al. 2016). FTCs have been revealed to have a different somatic mutation landscape from other types of TC, where in FTCs *RAS* gene mutations are more frequent while *BRAF* mutations are less frequent (Jung et al. 2016; Swierniak et al. 2016; Nicolson et al. 2018; Prete et al. 2020). In dogs studied in this thesis, *GNAS* A204D was detected in all FCC subtypes (FTC, FCTC, CTC, PTC, and carcinosarcoma) (Table 7.1). Compared to FCTC, the prevalence of *GNAS* A204D in FTC is lower. Because a limited number of cases are available for other types of FCC, there is limited power to detect differences in prevalence compared to FTC. More samples are needed to achieve a higher statistic power.

Table 7.1. *GNAS* somatic mutation landscape in different subtypes of FCC.

		<i>GNAS</i> genotype		
		CA	CC	NA ^c
FCC Subtype	FTC	8	8	0
	CTC	1	3	1
	FCTC	10	0	0
	PTC	3	1	2
	Carcinosarcoma	2	0	0
	Suspected ^a	1	1	0
	Unknown ^b	1	5	1
	Adenoma	0	1	0

^a Dogs showed typical signs of thyroid cancer but have not gone through a sample removal and histology examination. ^b Histology type is unknown due to other reasons like unknown side of origin. ^c Unknown genotype due to failure in typing.

7.5. Animal model

7.5.1 Canine model of human diseases

Dogs are the most popular pet globally, owned by one third (33%) of households according to a survey of the Growth from Knowledge (GfK) (Knowledge 2016). Dogs are second only to humans in medical surveillance and preventative health care. Dogs can get cancer spontaneously or due to inherited mutations as humans. Due to heavy inbreeding in pure breed dogs, the incidence of cancers in pedigreed dogs is high. All these result in a high number of dogs affected by different types of cancers. Human cancers are still difficult to cure and in dogs, this is even more difficult. Cancer is a leading cause of death in dogs. One in three dogs will be diagnosed with certain cancer at some point of their lives and 50% of dogs older than 10 years develop the disease and eventually one forth dying from it (Davis and Ostrander 2014).

Dogs are gaining increasing attention of scientists to serve as disease models. Compared to the mostly used rodent disease models, dogs have many important advantages. Contrary to rodents that are produced by genetic engineering techniques

and reared in a protected lab environment, dogs share with their human owners the same environment, therefore are affected by the same environmental factors. Also, based on physiology, dogs are closer to humans than mice and their genomes are more similar to humans than the genomes of mice (Lindblad-Toh et al. 2005). Furthermore, dogs' immune system works much like that in humans (Felsburg 2002). Very often, dogs show great similarity to humans in clinical signs, histological pattern, and progression of tumors (Gardner et al. 2016). Furthermore, many dog tumors have similar genetic alterations that are responsible for the disease, including both germline and somatic alterations (Decker et al. 2015; Elvers et al. 2015; Gardner et al. 2016; Sakthikumar et al. 2018; Wong et al. 2019; Amin et al. 2020; Alsaihati et al. 2021). Currently, in preclinical drug discovery, mouse models or human cell lines are still mostly used. However, those really don't match the experience of these drugs in humans very well (Mak et al. 2014). Even though such drugs work well in a preclinical research trial, they fail once used in humans. Dog cancer models can possibly fill the gap between rodent cancer models and human cancers.

Dogs can contribute to beat cancer in both humans and dogs. On the one hand, dogs can play a role in cancer research and therapy development as a disease model, or at least as a partner to understand the etiology of the disease. On the other hand, therapies developed for humans can also benefit the cure of disease in dogs. Therapies to cure cancer applied in dogs are mainly adopted from human cancer treatments. Development of novel therapies or medicines involves considerable investments. Such big investment impedes development of novel and better treatments for dogs. A collaboration across species in cancer research and treatment development is an efficient way to improve our understanding about cancers and to innovate effective treatments to cure cancer in all species. Therefore, a win-win situation between species can be achieved.

7.5.2. Value of FCC dogs to be used as a disease model

To breeders, my research provided a genetic test that can help them to eradicate FCC from their GLP population. To cancer treatment research, these dogs might be used as a valuable disease model. More than half of FCCs identified in the studied GLPs capture the driver mutation *GNAS* A204D. This high homogeneity of driver mutations makes them unique models for a variety of human cancers with similar driver mutations in the *GNAS* gene. In addition to aforementioned advantages of dogs with a sporadic cancer, another advantage of these dogs is that these affected dogs are flexibly reproduceable. Production of affected dogs can be planned as needed according to the requirement of a trial. Of course, animal welfare of these

dogs should also be carefully ensured, and guidelines must be strictly adhered to while raising these dogs and during the preclinical trial.

7.5.3. *GNAS* somatic mutation -- potential drug target

One of the ultimate aims of somatic mutation studies in human tumors is driven by the theory of target therapy that cures cancer by targeting proteins that control how cancer cells grow, divide, and spread. Targeted therapy research for human tumors with the *GNAS* R201C mutation has been initiated (Haridas et al. 2021). Human *GNAS* R201C is a well-recognized driver mutation that can continuously activate downstream signaling pathways thus lead to a higher cell proliferation rate. Although *GNAS* A204D identified in canine FCC in our study is different from that mutation, possibly they have the same/similar molecular mechanism underlying tumorigenesis. If that is the case, these dogs might be a valuable animal model for preclinical trials of those target therapies after tests in a mouse or rat model.

7.5.4. Animal cancer genetics investigation

Somatic mutation landscape of tumors in mammals other than humans, dogs, mice, and rats, have been poorly investigated at a genome-wide scale. Besides dogs, pigs have also been proposed to be good cancer models to fill the gap between rodents and primates (Kalla et al. 2020). Uncovering driver mutations is critical for understanding the molecular mechanism underlying specific tumor development and potential therapy development. Thus, not only germline risk mutation identification, but also somatic mutation profiling of tumors is important in other animals that can potentially be used as disease models. Somatic mutation landscapes of several canine tumors have been profiled and revealed to be similar to corresponding human cancers (Decker et al. 2015; Elvers et al. 2015; Sakthikumar et al. 2018; Wong et al. 2019; Amin et al. 2020; Alsaihati et al. 2021). For tumors in other species (except for mice and rats), it is unclear if the same driver mutations/genes are shared with humans. Many genetically modified pigs were created to model human tumors, such as osteosarcoma, pancreatic cancer, colorectal cancer, and breast cancer (Luo et al. 2011; Sieren et al. 2014; Li et al. 2015; Callesen et al. 2017; Kalla et al. 2020). In addition to genetically modified pigs, pigs with naturally occurring cancers can also be valuable models for scientific research and used for novel therapy development and trial before a clinical use. However, the incidence of cancer in pigs is low, approximately 40 cases per 1 million slaughtered pigs (Jagdale et al. 2019). Likely, this is because most pigs are slaughtered before the age of 6 months. Nevertheless, some natural pig models for spontaneous cancers have been established, such as MeLiM for melanoma (Bourneuf 2017), although their somatic mutation landscapes have not been profiled yet. It is therefore recommended to unravel the driver mutations of those tumors to make better use of these models.

7.6. Reference

- Agrawal N Akbani R Aksoy BA Ally A Arachchi H Asa Sylvia L Auman JT Balasundaram M Balu S Baylin Stephen B et al. 2014. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell* **159**: 676-690.
- Alsaihati BA, Ho K-L, Watson J, Feng Y, Wang T, Dobbin KK, Zhao S. 2021. Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. *Nature Communications* **12**: 4670.
- Alzahrani AS, Baitei EY, Zou M, Shi Y. 2006. Clinical case seminar: metastatic follicular thyroid carcinoma arising from congenital goiter as a result of a novel splice donor site mutation in the thyroglobulin gene. *J Clin Endocrinol Metab* **91**: 740-746.
- Ameziane El Hassani R, Buffet C, Leboulleux S, Dupuy C. 2019. Oxidative stress in thyroid carcinomas: biological and clinical significance. *Endocrine-Related Cancer* **26**: R131-R143.
- Amin SB, Anderson KJ, Boudreau CE, Martinez-Ledesma E, Kocakavuk E, Johnson KC, Barthel FP, Varn FS, Kassab C, Ling X et al. 2020. Comparative Molecular Life History of Spontaneous Canine and Human Gliomas. *Cancer Cell* **37**: 243-257.e247.
- Bann DV, Jin Q, Sheldon KE, Houser KR, Nguyen L, Warrick JI, Baker MJ, Broach JR, Gerhard GS, Goldenberg D. 2019. Genetic Variants Implicate Dual Oxidase-2 in Familial and Sporadic Nonmedullary Thyroid Cancer. *Cancer Research* **79**: 5490-5499.
- Bohnekamp J, Schöneberg T. 2011. Cell Adhesion Receptor GPR133 Couples to G_s Protein *. *Journal of Biological Chemistry* **286**: 41912-41916.
- Bourneuf E. 2017. The MeLiM Minipig: An Original Spontaneous Model to Explore Cutaneous Melanoma Genetic Basis. *Frontiers in Genetics* **8**.
- Callesen MM, Árnadóttir SS, Lyskjær I, Ørntoft M-BW, Høyer S, Dagnæs-Hansen F, Liu Y, Li R, Callesen H, Rasmussen MH et al. 2017. A genetically inducible porcine model of intestinal cancer. *Molecular Oncology* **11**: 1616-1629.
- Castellone MD, Melillo RM. 2018. RET-mediated modulation of tumor microenvironment and immune response in multiple endocrine neoplasia type 2 (MEN2). *Endocrine-Related Cancer* **25**: T105-T119.
- Castro-Giner F, Ratcliffe P, Tomlinson I. 2015. The mini-driver model of polygenic cancer evolution. *Nature Reviews Cancer* **15**: 680-685.
- Chun S, Fay JC. 2011. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS genetics* **7**: e1002240-e1002240.
- Cipollini M, Pastor S, Gemignani F, Castell J, Garritano S, Bonotti A, Biarnés J, Figlioli G, Romei C, Marcos R et al. 2013. TPO genetic variants and risk of differentiated thyroid carcinoma in two European populations. *International Journal of Cancer* **133**: 2843-2851.
- Computational Pan-Genomics C. 2018. Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics* **19**: 118-135.
- Consortium EP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636-640.
- Consortium ITP-CAoWG. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**: 82-93.
- Coorens THH, Farndon SJ, Mitchell TJ, Jain N, Lee S, Hubank M, Sebire N, Anderson J, Behjati S. 2020. Lineage-Independent Tumors in Bilateral Neuroblastoma. *New England Journal of Medicine* **383**: 1860-1865.
- Darwiche N. 2020. Epigenetic mechanisms and the hallmarks of cancer: an intimate affair. *Am J Cancer Res* **10**: 1954-1978.
- Davis BW, Ostrander EA. 2014. Domestic dogs and cancer research: a breed-based genomics approach. *ILAR J* **55**: 59-68.
- Decker B, Parker HG, Dhawan D, Kwon EM, Karlins E, Davis BW, Ramos-Vara JA, Bonney PL, McNeil EA, Knapp DW et al. 2015. Homologous Mutation to Human BRAF V600E Is

- Common in Naturally Occurring Canine Bladder Cancer--Evidence for a Relevant Model System and Urine-Based Diagnostic Test. *Mol Cancer Res* **13**: 993-1002.
- Ding Z, Ke R, Zhang Y, Fan Y, Fan J. 2019. FOXE1 inhibits cell proliferation, migration and invasion of papillary thyroid cancer by regulating PDGFA. *Molecular and Cellular Endocrinology* **493**: 110420.
- DoGA. 2017. THE DOG GENOME ANNOTATION (DoGA) PROJECT. Vol 2022.
- Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, Johnson GS, Rice ES, Hillier LD, Hammond JM et al. 2021. Chromosome-length genome assembly and structural variations of the primal Basenji dog (*Canis lupus familiaris*) genome. *BMC Genomics* **22**: 188.
- Elvers I, Turner-Maier J, Swofford R, Koltoonian M, Johnson J, Stewart C, Zhang CZ, Schumacher SE, Beroukhir R, Rosenberg M et al. 2015. Exome sequencing of lymphomas from three dog breeds reveals somatic mutation patterns reflecting genetic background. *Genome Res* **25**: 1634-1645.
- Felsburg PJ. 2002. Overview of immune system development in the dog: comparison with humans. *Human & Experimental Toxicology* **21**: 487-492.
- Feng G, Feng H, Qi Y, Wang T, Ni N, Wu J, Yuan H. 2022. Interaction analysis between germline genetic variants and somatic mutations in head and neck cancer. *Oral Oncology* **128**: 105859.
- Ferreira de Camargo GM. 2019. The role of molecular genetics in livestock production. *Animal Production Science* **59**: 201-206.
- Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Edwards RJ, Barton K, Lyons RJ, Tuipulotu DE, Hayes VM et al. 2020. Canfam_GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (*Canis lupus familiaris*) using a combination of long reads, optical mapping, and Hi-C. *GigaScience* **9**.
- Foulkes WD, Polak P. 2020. Bilateral Tumors — Inherited or Acquired? *New England Journal of Medicine* **383**: 280-282.
- Gardner HL, Fenger JM, London CA. 2016. Dogs as a Model for Cancer. *Annual review of animal biosciences* **4**: 199-222.
- Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* **17**: 175-188.
- GENE-SWitCH. 2019-2023. The regulatory GENomE of SWine and CHicken: functional annotation during development. Vol 2022.
- Ghafouri-Fard S, Shirvani-Farsani Z, Taheri M. 2020. The role of microRNAs in the pathogenesis of thyroid cancer. *Non-coding RNA Research* **5**: 88-98.
- Gong T, Hayes VM, Chan EKF. 2021. Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics* **22**.
- Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, Kirby LE, Myers B, Sliwerska E, Emery S et al. 2021. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proceedings of the National Academy of Sciences* **118**: e2016274118.
- Haridas V, Ito I, Shen JPYC. 2021. Targeting GNAS mutant appendiceal adenocarcinoma. *Journal of Clinical Oncology* **39**: 470-470.
- Hegstad-Davies RL, Torres SMF, Sharkey LC, Gresch SC, Muñoz-Zanzi CA, Davies PR. 2015. Breed-specific reference intervals for assessing thyroid function in seven dog breeds. *Journal of Veterinary Diagnostic Investigation* **27**: 716-727.
- Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data* **7**: 399.
- Hoshi S, Hoshi N, Okamoto M, Paiz J, Kusakabe T, Ward JM, Kimura S. 2009. Role of NKX2-1 in N-bis(2-hydroxypropyl)-nitrosamine-induced thyroid adenoma in mice. *Carcinogenesis* **30**: 1614-1619.

- Houde A, Pommier SA, Roy R. 1993. Detection of the ryanodine receptor mutation associated with malignant hyperthermia in purebred swine populations1. *Journal of Animal Science* **71**: 1414-1418.
- Huang AY, Lee EA. 2022. Identification of Somatic Mutations From Bulk and Single-Cell Sequencing Data. *Frontiers in Aging* **2**.
- Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, Hakonarson H, Whiting A, Davis CT, Wu W et al. 2012. Crohn's disease and genetic hitchhiking at IBD5. *Molecular biology and evolution* **29**: 101-111.
- Jagdale A, Iwase H, Klein EC, Cooper DK. 2019. Incidence of Neoplasia in Pigs and Its Relevance to Clinical Organ Xenotransplantation. *Comp Med* **69**: 86-94.
- Jung SH, Kim MS, Jung CK, Park HC, Kim SY, Liu J, Bae JS, Lee SH, Kim TM, Lee SH et al. 2016. Mutational burdens and evolutionary ages of thyroid follicular adenoma are comparable to those of follicular carcinoma. *Oncotarget* **7**: 69638-69648.
- Kalla D, Kind A, Schnieke A. 2020. Genetically Engineered Pigs to Study Cancer. *International Journal of Molecular Sciences* **21**: 488.
- Kaukonen M, Quintero IB, Mukarram AK, Hytönen MK, Holopainen S, Wickström K, Kyöstiä K, Arumilli M, Jalomäki S, Daub CO et al. 2020. A putative silencer variant in a spontaneous canine model of retinitis pigmentosa. *PLOS Genetics* **16**: e1008659.
- Knowledge Gf. 2016. Man's best friend: global pet ownership and feeding trends. Vol 2022.
- Kool MMJ, Galac S, Spandauw CG, Kooistra HS, Mol JA. 2013. Activating Mutations of GNAS in Canine Cortisol-Secreting Adrenocortical Tumors. *Journal of Veterinary Internal Medicine* **27**: 1486-1492.
- Krohn K, Maier J, Paschke R. 2007. Mechanisms of disease: hydrogen peroxide, DNA damage and mutagenesis in the development of thyroid tumors. *Nat Clin Pract Endocrinol Metab* **3**: 713-720.
- Kucab JE, Zou X, Morganello S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP et al. 2019. A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**: 821-836.e816.
- Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, Harmanci A, Martinez-Fundichely A, Chan CWY, Nielsen MM et al. 2020. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell* **180**: 915-927.e916.
- Landa I, Ibrahimspic T, Boucai L, Sinha R, Knauf JA, Shah RH, Dogan S, Ricarte-Filho JC, Krishnamoorthy GP, Xu B et al. 2016. Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. *The Journal of Clinical Investigation* **126**: 1052-1066.
- Lewis CM, Vassos E. 2020. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**: 44.
- Li S, Edlinger M, Saalfrank A, Flisikowski K, Tschukes A, Kurome M, Zakhartchenko V, Kessler B, Saur D, Kind A et al. 2015. Viable pigs with a conditionally-activated oncogenic KRAS mutation. *Transgenic Research* **24**: 509-517.
- Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D. 2014. Making the difference: integrating structural variation detection tools. *Briefings in Bioinformatics* **16**: 852-864.
- Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803-819.
- Liu Y, Gusev A, Heng YJ, Alexandrov LB, Kraft P. 2022. Somatic mutational profiles and germline polygenic risk scores in human cancer. *Genome Medicine* **14**: 14.
- Lu Y, Chan Y-T, Tan H-Y, Li S, Wang N, Feng Y. 2020. Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy. *Molecular Cancer* **19**: 79.

- Luo Y, Li J, Liu Y, Lin L, Du Y, Li S, Yang H, Vajta G, Callesen H, Bolund L et al. 2011. High efficiency of BRCA1 knockout using rAAV-mediated gene targeting: developing a pig model for breast cancer. *Transgenic Research* **20**: 975-988.
- MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**: 469-476.
- Machiela MJ, Chanock SJ. 2014. GWAS is going to the dogs. *Genome Biology* **15**: 105.
- Mak IW, Evaniew N, Ghert M. 2014. Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res* **6**: 114-118.
- McFarland Christopher D, Korolev Kirill S, Kryukov Gregory V, Sunyaev Shamil R, Mirny Leonid A. 2013. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences* **110**: 2910-2915.
- Megquier K, Genereux DP, Hekman J, Swofford R, Turner-Maier J, Johnson J, Alonso J, Li X, Morrill K, Anguish LJ et al. 2019. BarkBase: Epigenomic Annotation of Canine Genomes. *Genes* **10**: 433.
- Moses W, Weng J, Kebebew E. 2011. Prevalence, clinicopathologic features, and somatic genetic mutation profile in familial versus sporadic nonmedullary thyroid cancer. *Thyroid : official journal of the American Thyroid Association* **21**: 367-371.
- Nicolson NG, Murtha TD, Dong W, Paulsson JO, Choi J, Barbieri AL, Brown TC, Kunstman JW, Larsson C, Prasad ML et al. 2018. Comprehensive Genetic Analysis of Follicular Thyroid Carcinoma Predicts Prognosis Independent of Histology. *The Journal of Clinical Endocrinology & Metabolism* **103**: 2640-2650.
- Padamsee TJ, Wills CE, Yee LD, Paskett ED. 2017. Decision making for breast cancer prevention among women at elevated risk. *Breast Cancer Research* **19**: 34.
- Patra KC, Kato Y, Mizukami Y, Widholz S, Boukhali M, Revenco I, Grossman EA, Ji F, Sadreyev RI, Liss AS et al. 2018. Mutant GNAS drives pancreatic tumorigenesis by inducing PKA-mediated SIK suppression and reprogramming lipid metabolism. *Nature Cell Biology* **20**: 811-822.
- Penna G, Rubio IGS, Brust ES, Cazarin J, Hecht F, Alkmim NR, Rajão KMAB, Ramos HE. 2021. Congenital hypothyroidism and thyroid cancer. *Endocrine-Related Cancer* **28**: R217-R230.
- Phillips KA, Deverka PA, Hooker GW, Douglas MP. 2018. Genetic Test Availability And Spending: Where Are We Now? Where Are We Going? *Health Affairs* **37**: 710-716.
- Poetsch AR. 2020a. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Computational and structural biotechnology journal* **18**: 207-219.
- Poetsch AR. 2020b. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput Struct Biotechnol J* **18**: 207-219.
- Prete A, Borges de Souza P, Censi S, Muzza M, Nucci N, Sponziello M. 2020. Update on Fundamental Mechanisms of Thyroid Cancer. *Frontiers in Endocrinology* **11**.
- Ramroop JR, Gerber MM, Toland AE. 2019. Germline Variants Impact Somatic Events during Tumorigenesis. *Trends in Genetics* **35**: 515-526.
- Rheinbay E, Parasuraman P, Grimsby J, Tiao G, Engreitz JM, Kim J, Lawrence MS, Taylor-Weiner A, Rodriguez-Cuevas S, Rosenberg M et al. 2017. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**: 55-60.
- Riva L, Pandiri AR, Li YR, Droop A, Hewinson J, Quail MA, Iyer V, Shepherd R, Herbert RA, Campbell PJ et al. 2020. The mutational signature profile of known and suspected human carcinogens in mice. *Nature Genetics* **52**: 1189-1197.
- Sakamoto Y, Sereewattanawoot S, Suzuki A. 2020. A new era of long-read sequencing for cancer genomics. *Journal of Human Genetics* **65**: 3-10.

- Sakthikumar S, Elvers I, Kim J, Arendt ML, Thomas R, Turner-Maier J, Swofford R, Johnson J, Schumacher SE, Alföldi J et al. 2018. SETD2 Is Recurrently Mutated in Whole-Exome Sequenced Canine Osteosarcoma. *Cancer Res* **78**: 3421-3431.
- Salcedo A, Tarabichi M, Espiritu SMG, Deshwar AG, David M, Wilson NM, Dentre S, Wintersinger JA, Liu LY, Ko M et al. 2020. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nature Biotechnology* **38**: 97-107.
- Sieren JC, Meyerholz DK, Wang X-J, Davis BT, Newell JD, Jr., Hammond E, Rohret JA, Rohret FA, Struzynski JT, Goeken JA et al. 2014. Development and translational imaging of a TP53 porcine tumorigenesis model. *The Journal of Clinical Investigation* **124**: 4052-4066.
- Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13**: 418.
- Swierniak M, Pfeifer A, Stokowy T, Rusinek D, Chekan M, Lange D, Krajewska J, Oczko-Wojciechowska M, Czarniecka A, Jarzab M et al. 2016. Somatic mutation profiling of follicular thyroid cancer by next generation sequencing. *Molecular and Cellular Endocrinology* **433**: 130-137.
- Tan H, Onichtchouk D, Winata C. 2016. DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish. *Zebrafish* **13**: 54-60.
- Tanner G, Westhead DR, Droop A, Stead LF. 2021. Benchmarking pipelines for subclonal deconvolution of bulk tumour sequencing data. *Nature Communications* **12**: 6396.
- Turan S, Bastepe M. 2015. GNAS Spectrum of Disorders. *Current Osteoporosis Reports* **13**: 146-158.
- Wang C, Wallerman O, Arendt M-L, Sundström E, Karlsson Å, Nordin J, Mäkeläinen S, Pielberg GR, Hanson J, Ohlsson Å et al. 2021. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Communications Biology* **4**: 185.
- Wong K, van der Weyden L, Schott CR, Foote A, Constantino-Casas F, Smith S, Dobson JM, Murchison EP, Wu H, Yeh I et al. 2019. Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. *Nature Communications* **10**: 353.
- Yoo S-K, Song YS, Lee EK, Hwang J, Kim HH, Jung G, Kim YA, Kim S-j, Cho SW, Won J-K et al. 2019. Integrative analysis of genomic and transcriptomic characteristics associated with progression of aggressive thyroid cancer. *Nature Communications* **10**: 2764.
- Zaccaria S, Raphael BJ. 2020. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature Communications* **11**: 4301.
- Zafon C, Gil J, Pérez-González B, Jordà M. 2019. DNA methylation in thyroid cancer. *Endocrine-Related Cancer* **26**: R415-R439.
- Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. 2017. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics* **18**: 286.
- Zarkesh M, Zadeh-Vakili A, Azizi F, Foroughi F, Akhavan MM, Hedayati M. 2018. Altered Epigenetic Mechanisms in Thyroid Cancer Subtypes. *Molecular Diagnosis & Therapy* **22**: 41-56.
- Zhu B, Mukherjee A, Machiela MJ, Song L, Hua X, Shi J, Garcia-Closas M, Chanock SJ, Chatterjee N. 2016. An investigation of the association of genetic susceptibility risk with somatic mutation burden in breast cancer. *British Journal of Cancer* **115**: 752-760.
- Zhu H, Peng YG, Ma SG, Liu H. 2015. TPO gene mutations associated with thyroid carcinoma: Case report and literature review. *Cancer Biomark* **15**: 909-913.

Summary

A familial cancer is like a curse to individuals in the family. In Dutch German Longhaired Pointer (GLP) dogs, a familial thyroid cancer (TC) was identified. Many GLPs were affected by, and died of, this familial TC. The aim of the research project described in this thesis was to eradicate this familial TC from the GLP population, and also to further our knowledge on causes of familial cancers in dogs.

In **Chapter 2**, we described the clinical details and histological diagnoses of all the affected GLPs diagnosed in the past ~20 years. The TCs identified in those 54 histologically diagnosed GLPs belong to thyroid follicular cell carcinoma (FCC) and manifested five sub-types: 1) follicular thyroid carcinoma (FTC), 2) compact thyroid carcinoma (CTC), 3) follicular-compact thyroid carcinoma (FCTC), 4) papillary thyroid carcinoma (PTC), and 5) carcinosarcoma. Most of the affected GLPs are closely related where 45 of them can be traced to a pair of first-half cousins. With the pedigree we estimated the heritability to be 0.62. According to the pedigree, this familial FCC has most likely an autosomal recessive inheritance pattern. Pedigree-based inbreeding was estimated for each dog, and it turned out that affected GLPs had higher inbreeding levels, suggesting that inbreeding contributed to incidence of the familial FCC.

In **Chapter 3**, I identified the germline risk mutations of the familial FCC in the Dutch GLPs. I performed a genome-wide association study and homozygosity mapping based on a combination of SNP array genotype and whole-genome sequencing (WGS) data to identify the target genomic region. Subsequently, I used WGS data from 11 affected and 11 unaffected GLPs to fine-map the potential causal variant(s) for the FCC. This yielded two deleterious mutations (chr17:800788G>A (686F>V) and chr17:805276C>T (845T>M)) in the *TPO* gene to be the germline risk mutations. We genotyped these two variants in 186 GLPs (59 affected and 127 unaffected) and revealed that homozygous recessive genotypes of these two SNPs confer a relative risk of 16.94 and 16.64 respectively, in comparison to homozygous wild-type genotypes.

In **Chapter 4**, I investigated the somatic mutations of the familial FCCs with the aim to identify driver mutations. I identified somatic single nucleotide variants (SNVs), insertions and deletions (InDels), structural variants (SVs), and copy number alternations (CNAs) using WGS data from FCC tissue and matched blood samples from 7 affected GLPs. Among somatic SNVs, I identified a recurrent deleterious mutation in the *GNAS* gene (chr24:43657087C>A, *GNAS* A204D) in 4 of 7 sequenced FCC tissues. Through Sanger sequencing, this somatic mutation was

Summary

further identified in FCC tissues of 20 out of 32 GLPs. The high prevalence of the *GNAS* A204D mutation indicates that it is a promising driver mutation. I also found that this *GNAS* A204D somatic mutation is associated with lower somatic mutation burden. Meanwhile, to reveal potential mutational processes during tumorigenesis, I constructed mutational signatures of these FCCs and identified signatures that are similar to human SBS5 and SBS40, suggesting an endogenous mutagenesis factor in tumorigenesis.

In **Chapter 5**, I investigated the genetic diversity of Dutch GLPs and compared inbreeding levels between GLP and 11 other pointer setter breeds. I revealed that Dutch GLPs have relatively low inbreeding in comparison to the 11 other pointer setter breeds. Furthermore, I investigated the genetic relationship between GLP and those 11 pointer setter breeds and revealed good consistence between identified genetic relationship and breeding history of these breeds. Lastly, I identified the genomic selection signatures in GLPs using a runs of homozygosity (ROH) islands approach. I showed that a ROH segment identified on chromosome 30, harboring the *RYR3*, *FMNI*, and *GREM1* genes, might be selected for athletic performance.

In **Chapter 6**, I tested a new approach to use prior knowledge on signaling pathways to predict driver mutations. I calculated a cancer pathway score for each signaling pathway and then computed a cancer gene score for each gene. I showed that driver genes have higher cancer gene scores than passenger genes, implying that this cancer gene score is useful in distinguishing driver and passenger genes. I then trained Random Forest Classifier models using the cancer gene score as a feature, along with SIFT score, PolyPhen2 score, and recurrence of the mutation. On average, I observed a prediction accuracy of those trained Random Forest Classifiers, measured by F1 score (harmonic mean of precision and recall), of 0.90 (ranging between 0.85 - 0.94), demonstrating that these features, including the cancer gene score, can contribute to driver mutation prediction.

Finally, in **Chapter 7**, I brought major findings in chapter 2-6 together and discussed them in context. I discussed the utility of genetic test based on the PCR-RFLP experiment described in chapter 2 for the identification of GLPs predisposed to the FCC and how to use this genetic test to assist selective breeding for healthy GLPs. I connected the germline risks (chapter 3) and somatic driver mutation (chapter 4) and discussed the mechanisms underlying FCC development. Lastly, I discussed the advantage of using dogs as disease models and the specific value of GLPs studied in this thesis as cancer models.

Appendices

Curriculum vitae

About the author

Yun Yu was born on 06-04-1993 in Hubei, China. He was fascinated by Biology when he took his first lesson of it in middle school. He studied Animal Sciences in Huazhong Agricultural University between 2011 – 2015 and obtained his Bachelor's degree with a thesis entitled "The expression and localization of Sororin in mouse granulosa cells" under the supervision of Prof. Dr. Lijun Huo. He finished his Master study at Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education, in Huazhong Agricultural University. During his Master study, he investigated the association between single nucleotide polymorphisms of the *MC4R* gene and local temperature adaption of sheep and finished his Master thesis entitled "The study on variations of *MC4R* genes and their effects in sheep and goat" under the supervision of Prof. Dr. Xunping Jiang. In October 2018, he started his PhD study at Animal Breeding and Genomics group, Wageningen University. The outcomes of his work during the past four years are presented in this thesis entitled "Genomics underlying a canine hereditary thyroid follicular cell carcinoma" under the supervision of Prof. Dr. Martien A.M. Groenen and Dr. Richard P.M.A. Crooijmans.

Publications

Yu Y, Bovenhuis H, Wu Z, Laport K, Groenen MAM, Crooijmans RPMA. Deleterious Mutations in the TPO Gene Associated with Familial Thyroid Follicular Cell Carcinoma in Dutch German Longhaired Pointers. *Genes*. 2021; 12(7):997. <https://doi.org/10.3390/genes12070997>

Yu Y, Krupa A, Keesler, RI, Grinwis GCM, de Ruijscher M, de Vos J, Groenen MAM, Crooijmans RPMA. Familial follicular cell thyroid carcinomas in a large number of Dutch German longhaired pointers. *Veterinary and Comparative Oncology*. 2022; 20(1): 227- 234. doi:[10.1111/vco.12769](https://doi.org/10.1111/vco.12769)

Acknowledgements

A saying that is often spoken and heard during my PhD journey is that it is your PhD. What it conveys is that you are responsible for the progress, outcome, plan, and many other things of your PhD project. In this sense, you independently do your PhD study. While many people actually played indispensable roles in the whole or part of the PhD journey. Some of them gave big supports and insights to your research work and some brought flowers and drinks to your boring/leisure days. A big thank I would like to express to all of them here.

Martien, you are the first Dutchman who I contacted and met online and in person. You opened the open Dutch world to me. Pearls are everywhere but not the same as the eyes. Thank you for offering me this opportunity to do a PhD study at ABG group, even though I was nearly blank in genetics and genomics studies. Your sharp insight, kindness, inclusivity enlightened and encouraged me now and then.

Richard, I appreciate having you as my daily supervisor. You are the one standing closest to me during my whole PhD journey. You witnessed my ups and downs and always encouraged me with your bright eyes and minds. You prepared everything needed for this project, funding, samples, and so on. You also lead me into the dog cancer field. You care not only about my work and but also my life. Thank you for your trust and encouragement even though when I was doing something completely out of the plan sometimes.

Guy, thank you for performing all those histology diagnoses of samples and showing me how cancer cells look like under the microscope. **Rebekah**, thank you for your work in diagnosis of those samples as well. **Ada**, thank you and also other people in the AniCura for clinical examination and organizing all clinical results. Especially thank you **Johan** for identifying thyroid cancer issue in GLPs and all preparation work. Your work really laid the foundation of this research. Nederlands Kankerfonds voor Dieren provided funding for genotyping and sequencing. Thanks for your kind financial support. **Annie**, thank you for providing me the pedigree of GLPs. Especially, thank you for the special and unforgettable moments with your lovely GLP dog.

Henk, thank you for patiently explaining many concepts related to GWAS and statistics to me. **Jack**, you are the only one at ABG who dedicates the research to dogs, the lovely creatures. Thank you for lessons about inbreeding of dogs. **Mirte**, I have been enlightened many times by your comments in the weekly Genomics meeting. Also thank you for your valuable help and insights in GLP selection signature study. **Ole**, you are charming as a scientist and a colleague. Although you are not directly involved in my PhD research, but I have been inspired uncountable times by your visions and comments in the Genomics meeting. **Hendrik-jan**, I was inspired by your good taste of and enthusiasm to genomic science now and then. **Pascal**, thank you for many nice discussions about machine learning and

contributions to the driver mutation prediction study. **Kimberley**, thank you for preparing all samples for sequencing and performing typing experiments, which were really great helps. **Alex** and **Bart**, thank you for helping me out with ASReml. Weekly Genomics meeting really opens my eyes in genomics research with such diverse studies. Thank all the people who participated in that meeting, especially **Martijn, Rayner, Henri, Chiara, Vinicius, Lim, Jani, Gibbs, Annemiek, Marta, Carolina, Jeroen, Pedro**. Meanwhile, I must say thank you to all colleagues at whole ABG, especially **Harmen, Benan, Renzo, Dries, Farid, Priadi, Lisette, Fatma, Ibrahim**, for those delightful chats we had.

Zhou (舟哥) and **Langqing (刘郎)**, it is one of my biggest luck to know both of you. You taught me many about genomics and bioinformatics. I also appreciate your friendships in life. I would award you my best ABG buddy, gym buddy, and Haarweg buddy. **Lu Cao (曹露)**, it is a special predestination to have you as a friend. I appreciate the interesting moments with you, especially the Christmas in Denmark. You three really have a soul full of fun and love. **Shuwen (包姐)**, I appreciate your help in many analyses (especially ASReml) and your insights about science, life, and all other things, **You (常优)**, it was such a big surprise that we are from the same county. **Haibo (海博)**, I miss our badminton match, also drinks and meals afterwards. **Xue (学哥)**, you are a nice house mate, badminton team mate, swimming mate, travel mate etc. It is my luck to develop a friendship with you. **Xiaofei (飞哥)**, we started our PhD the same day. Thanks for many drinks, foods, trips, talks about all different things we had together. We witnessed each other's growth in science. **Ruimin (乔老师)**, it is a great luck to meet you and develop friendship and collaborations with you! **Haniel**, it was nice to have beers with you and miss the trips we had together.

Zhuoshi (卓识), **Siyuan (思远)**, **Wenye (文晔)**, **Aixin (爱心)**, **Rui (瑞)**, **Ziwei (紫薇)**, **Qitong (其通)**, **Liyan (立言)**, **Yebo (业博)**, **Qiuyu (秋雨)**, **Wenzhe (文哲)**, thank you all for fill in blank moments with laughs, bright colors, and fragrances. All annoyances and stresses go away immediately after a wonderful badminton match. Thank all the badminton friends, old and new, here and there, for every great match we had.


I thank the China Scholarship Council (CSC) for providing the scholarship to finish my PhD study. Thank you Prof. Dr. **Zhao Shuhong** for your selfless help in applying my PhD position at ABG group.

To finish this PhD journey, supports from my family, of course, is critically important. **Papa** and **Mom**, thank you for your love and the best supports all the way around. Thank you for supporting my decisions on studies. Thank all my relatives for your encouragement all the way here. Also thank my friends back in

Appendices

China. Although we are thousands of kilometers away from each other, messages and calls from you warmed my heart and encouraged my study.

This journey comes to its end. Doing a PhD study is probably the coolest decision in my life. In this journey, there were struggling, proud, self-doubt, and joy. Looking back, I would like to say to coming journeys, be fun, be inclusive, stay hungry, stay foolish.

Training and Supervision Plan (TSP)		Graduate School WIAS	
			
A. The Basic Package	year	credits *	
WIAS Introduction Day (mandatory)	2018	0.3	
Course on philosophy of science and/or ethics (mandatory)	2019	1.5	
Course on essential skills (Frank Little) <i>(recommended)</i>	2019	1.2	
Subtotal Basic Package		3	
B. Disciplinary Competences	year	credits	
ABG-30306 Genomics	2018	6.0	
Writing a research proposal	2019	6.0	
Advanced statistics course Design of Experiments	2018	0.8	
Getting started in ASReml	2019	0.3	
summer course on ChIP-seq (wet-lab) and basic functional animal genome analysis	2019	1.5	
Genome Maintenance & Cancer	2021	0.8	
Interactive post-graduate course on characterization, management and exploitation of genomic diversity in animals	2019	1.5	
Subtotal Disciplinary Competences		17	
C. Professional Competences	year	credits	
Project and Time Management	2019	1.5	
Scientific Writing	2019	1.8	
Research data management	2019	0.5	
Presenting with Impact	2019	1.0	
D. Presentation Skills <i>(maximum 4 credits)</i>	year	credits	
WIAS science day poster presentation	2019	1.0	
WIAS annual conference oral presentation	2021	1.0	
ISAG 2021, poster online	2021	1.0	
WIAS annual conference oral presentation	2022	1.0	
E. Teaching competences <i>(max 6 credits)</i>	year	credits	
Supervising MSc major theses	2020-2021	2.0	
Assisting Genomics course	2020 P5	0.4	
Assisting Genomics course	2021 P2	0.4	
Assisting Genomics course	2021 P5	0.2	
Subtotal Teaching competences		3	
Education and Training Total (minimum 30 credits)*		32	

Colophon

This research was financially supported by Wageningen University and Nederlands Kankerfonds voor Dieren.

Yun Yu was sponsored by the Chinese Scholarship Council (CSC).

