Genetic Predisposition to Gastric Cancer: *CDH1* and Beyond

Christopher Hakkaart

A thesis submitted for the degree of

Doctor of Philosophy

at the

Centre for Translational Cancer Research,

University of Otago, Dunedin,

New Zealand

July 2017

Abstract

Gastric cancer is a complex disease influenced by strong genetic and environmental factors. Hereditary gastric cancer syndromes are thought to account for between 1-3% of all cases. The most common hereditary gastric cancer syndrome is Hereditary Diffuse Gastric Cancer (HDGC), an autosomal dominant cancer syndrome that is primarily characterised by an extreme risk of developing diffuse-type gastric cancer. Approximately 40% of families that fit the clinical criteria for HDGC carry a pathogenic variant in germline *CDH1*. An explanation for the remaining 60% of cases remains largely elusive.

While New Zealand as a whole is a country with a low-incidence of gastric cancer, Māori have an age-standardised incidence of gastric cancer more than three times that of non-Māori. Additionally, Māori have an average age of diagnosis approximately 10 years younger than non-Māori, and are one of the few populations worldwide with a higher incidence of the diffuse-type disease. To assess the contribution of HDGC to the high-incidence of diffuse gastric cancer for Māori, we analysed the *CDH1* sequence from an unselected cohort of 94 Māori gastric cancer patients and 200 healthy matched controls using next-generation amplicon sequencing, multiplex ligation-dependent probe amplification, and Sanger sequencing. Pathogenic *CDH1* variants were identified in 18% of all cases, 34% of diffuse gastric cancers, and 66% of early-onset cases (< 45 years of age). After adjusting for the effect of clinical genetic testing for known Māori HDGC families, we estimate 6% of all Māori gastric cancer patients and 13% of diffuse gastric cancer patients carry pathogenic germline *CDH1* variants.

Chile is a country with a high-incidence of gastric cancer and no formal genetic screening programme for gastric cancer patients. To explore pathogenic

germline *CDH1* variants as a cause of gastric cancer in Chile, next-generation amplicon sequencing and Sanger sequencing were used to screen a cohort of 51 Chilean gastric cancer patients that presented with a striking family history or early-onset disease. Overall, one clear pathogenic *CDH1* variant was identified, representing 2.0% of all probands and 3.6% of probands who met the clinical criteria for HDGC. Although pathogenic *CDH1* variants were rare in this Chilean cohort, we were able to screen the extended family of the one proband with a confirmed mutation and identify five further carriers. These carriers will now benefit from surveillance and early intervention.

Finally, whole-exome sequencing was used to examine 14 diffuse gastric cancer patients that fit the clinical criteria for HDGC and did not carry a pathogenic variant in their germline *CDH1*. Variants in these patients were filtered and prioritised for further evaluation and validation using Sanger sequencing. Single probands were found to carry pathogenic variants in *ATM* and *TP53*, genes which are not associated with HDGC, but are known to increase gastric cancer risk. Additional mutations of interest were identified in *FARP2*, *FGD4*, and *LMO7*, genes that are important in the coordination of the actin cytoskeleton and/or cell adhesion, pathways which are dysregulated in diffuse-type gastric tumours. Until now, *FARP2*, *FGD4*, and *LMO7* were not linked with diffuse gastric cancer risk. It is clear from the current study and other HDGC studies, that there is no other common gene for HDGC, however families may carry private variants in genes rarely associated with disease.

Taken together, these studies demonstrate the variable frequency of pathogenic variants in germline *CDH1* in different populations, the absence of other commonly mutated genes in familial diffuse gastric cancers, and the importance of genetic screening and targeted interventions for those that carry pathogenic variants.

Acknowledgements

First and foremost, I would like to thank my supervisor Parry Guilford for his patience, encouragement, and guidance throughout the course of this PhD. The support you have given me has allowed me to grow as a scientific and see the big picture.

I am also grateful to Rob Day, Anita Dunbier, Tony Reeve, and Stephen Robertson for their expert help as my PhD advisory committee.

Thank you to all the members of the Cancer Genetics Laboratory for sharing their knowledge and extending their friendship during the course of my PhD. Special thanks to Rob Day and Les McNoe for being my source of sequencing knowledge and troubleshooting advice, and fellow PhD students Aziz, Henry, Tyler, and Andrew for taking this journey alongside me.

I extend my sincere thanks to the collaborators for projects described in this thesis. In particular thank you Dr. Lis Ellison-Loschmann, Soo Cheng, and Andrew Sporle for answering my continuous questions about the Māori case-control cohort, and Dr. Alejandro Corvalan, Dr. Enrique Norero Munoz, Paulina Gonzalez, and Alejandra Alarcon for answering my questions about the Chilean study cohort.

To my parents Heather and Charles, thank you for supporting my education and given me the opportunity to challenge myself.

Finally, thank you to my partner Millie, who has supported me through the ups and downs of my studies. Thank you for keeping me going, I could not have completed this without you.

Contents

List of Figures

List of Tables

List of Abbreviations

°C	Degrees Celsius
μg	Microgram
μL	Microlitre
μM	Micromole
Alt	Alternate allele
A-T	Ataxia-Telangiectasia
BLAST	Basic Local Alignment Search Tool
BMI	Body Mass Index
bp	Base pair
BQSR	Base quality score recalibration
BWA	Burrows Wheeler Aligner
Chr	Chromosome
CSI	Chromosomal instable
csv	Comma-separated values
Del	Deletion
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
EB	Elution buffer
EBV	Epstein-Barr virus
ENCODE	Encyclopedia of DNA Elements
ESP6500	Exome Sequencing Project
ExAC	Exome Aggregation Consortium
FAP	Familial Adenomatous Polyposis
g	Gram
GAPPS	Gastric Adenocarcinoma and Proximal Polyposis of the Stomach
GATK	Genome Analysis Toolkit
gDNA	Genomic DNA
GS	Genomically stable
pM	Picomole
h	Hour
H_20	Water
HDGC	Hereditary Diffuse Gastric Cancer
Het	Heterozygote
Hom	Homozygote
HR	Homologous recombination
ID	Identified

Indel	Insertion/deletion
Ins	Insertion
IGCLC	International Gastric Cancer Linkage Consortium
kg	Kilogram
L	Litre
M	Molar
MA	Mutation assessor
MAF	Minor allele frequency
MAP	MUTYH-associated polyposis
MLPA	Multiplex Ligation-dependant Probe Amplification
min	Minute
mL	Millilitre
mM	Millimolar
MSI	Microsatellite instable
n	Number
ng	Nanogram
NHI	National Health Index
nM	Nanomole
NGS	Next-generation Sequencing
NZCR	New Zealand Cancer Registry
NZGL	New Zealand Genomics Limited
PCR	Polymerase chain reaction
PJS	Peutz-Jeghers Syndrome
PPH2	Polyphen 2
Ref	Reference allele
sec	Second
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SRCC	Signet ring cell carcinoma
TAE	Tris acetic acid EDTA
TCGA	The Cancer Genome Atlas
TE	Tris EDTA
V	Volt
VCF	Variant call format
VQSR	Variant quality score recalibration
VUS	Variant of uncertain significance

Chapter 1

Introduction

1.1 Introduction

Gastric cancer is a complex disease influenced by strong genetic and environmental factors. Hereditary gastric cancer syndromes, in which inherited deleterious variants are predisposing to a heightened risk of developing gastric cancer, are thought to account for between 1-3% of all gastric cancer cases. The most frequent hereditary gastric cancer syndrome is Hereditary Diffuse Gastric Cancer (HDGC), an autosomal dominant syndrome characterised by an extreme risk of developing diffuse gastric cancer and lobular breast cancer. Germline *CDH1* explain approximately 40% of families meeting the clinical criteria for HDGC. An explanation for the other 60% of HDGC cases remains largely elusive. Here, I have used next-generation sequencing (NGS) and bioinformatic tools to explore how pathogenic germline *CDH1* variants contribute to the diagnosis of gastric cancer in the Māori and Chilean populations, and search for predisposing variants in HDGC patients whom do not carry pathogenic germline *CDH1* variants.

1.1.1 Gastric cancer

Gastric cancer is the fifth leading cause of cancer globally and is ranked third for cancer related mortality (Ferlay et al., 2015). Globally, gastric cancer affects nearly one million individuals each year, of whom 70-85% die within 5-years of diagnosis (Ferlay et al., 2015). The high mortality associated with the disease is the main result of typically late diagnoses and limited therapeutic options. The incidence of gastric cancer is twice as high in men as in women and varies greatly across countries and ethnicities. Geographic variation, in part, reflects differences in exposure to environmental risk factors. The incidence of gastric

cancer is highest in Eastern Asia, Central Europe, Eastern Europe, and South America; and lowest in Northern America, most parts of Africa, Australia, and New Zealand (Oliveira, Senz, et al., 2009; Ferlay et al., 2015). Due to a reduction in the prevalence of risk factors, the incidence of gastric cancer has been declining in most parts of the world (Oliveira, Senz, et al., 2009).

1.1.1.1 Classification and staging

Histopathologically, approximately 90% of gastric cancer diagnoses are classified as adenocarcinomas and 10% are classified as mucosa-associated lymphoid tissue (MALT) lymphomas or carcinoid tumours (Bosman, Carneiro, Hruban, & Theise, 2010). According to the Lauren classification system, gastric adenocarcinomas can be divided into two main histological subtypes, intestinal and diffuse (Lauren, 1965). The relative frequency of the intestinal and diffuse gastric cancers are approximately 54% and 32%, respectively. The remaining 15% of gastric adenocarcinomas are characterised as intermediate-type and are histologically made up of a mixture of the two main subtypes (Polkowski et al., 1999). It is widely accepted that the intestinal and diffuse subtypes represent distinct disease entities with marked pathology, epidemiology, and etiology (Hu et al., 2012). Although the Lauren classification system dates back to 1965, it is still widely accepted and employed by pathologists and physicians as a simple but robust classification approach.

Intestinal gastric cancer is defined by the presence of tumour cells with glandular, tubular, or papillary growth patterns, with various degrees of differentiation. These tumours typically grow in a unifocal, expanding fashion, and are commonly a response to chronic inflammation (Lauren, 1965).

Intestinal gastric cancer typically presents in older patients and is often associated with environmental factors, such as *Helicobacter pylori* (*H. pylori*) infection, diet, and life style (Kaneko & Yoshimura, 2001). The incidence of intestinal gastric cancer is declining in most parts of the world.

In contrast, diffuse gastric cancer is defined by solitary or small clusters of poorly cohesive cells that frequently infiltrate the stomach wall in a diffuse pattern, with or without a small component of gland formation (Lauren, 1965). Diffuse gastric cancer frequently presents with signet ring cells and is sometimes referred to as signet ring cell carcinoma (SRCC) (Lauren, 1965). SRCC is a rare form of highly malignant adenocarcinoma that is characterised by mucin filled vacuoles and present with an appearance similar to a signet ring. Family history is the strongest risk factor for diffuse gastric cancer, and unlike intestinal-type gastric cancer, is not strongly associated with environmental factors (Caldas et al., 1999). Additionally, diffuse gastric cancer typically presents at an earlier age compared to intestinal-type disease (Lauren, 1965).

Another set of criteria commonly used to classify gastric tumours is the World Health Organisation (WHO) classification (Bosman et al., 2010). The WHO classification is based on the predominant histological pattern of each tumour and recognises four major subtypes of gastric adenocarcinomas: tubular, papillary, mucinous, and poorly cohesive (including signet ring cell carcinoma) (Bosman et al., 2010). When compared to the Lauren classification: papillary, tubular, and mucinous adenocarcinomas are all categorised as intestinal-type gastric cancer; signet ring cell carcinomas and other poorly cohesive carcinomas are categorised as diffuse-type gastric cancer, and mixed carcinomas are categorised as intermediate-type gastric cancer (Berlth, Bollschweiler, Drebber, Hoelscher, & Moenig, 2014).

Recent advances in sequencing technologies have made it possible to classify tumours by their genetic profiles, rather than by their histology. The Cancer Genome Atlas (TCGA), in the most comprehensive study to date, classified gastric adenocarcinomas into four molecular subtypes, Epstein–Barr-virus-associated (EBV), microsatellite instability (MSI), genomically stable (GS); and chromosomal instability (CIN) (Bass et al., 2014). Importantly, classification of these molecular subtypes has provided valuable insight into some of the molecular mechanisms underling the different histological subtypes. TCGA showed, GS tumours were enriched for diffuse-type tumours and frequently contain pathogenic variants in *RHOA* and *CDH1*, or harboured a *CLDN18-ARHGAP* translocations, all of which impact cell adhesion (Bass et al., 2014). Conversely, the EBV subtype, characterised by Epstein–Barr virus infection, pathogenic *PIK3CA* variants, DNA hypermethylation, and amplification of the genes *JAK2*, *CD274*, and *PDCD1LG2*; the MSI subtype, characterised by elevated genome wide mutation rates; and the CIN subtype, characterised by marked aneuploidy and focal amplification of receptor tyrosine kinases, were all enriched for the intestinal subtype (Bass et al., 2014).

The extent to which a cancer has spread is known as its stage. The earliest stage stomach cancers is called stage 0 (carcinoma in situ), and then range from stages I (1) through IV (4). Generally, the lower the number, the less the cancer has spread. Although each person's cancer is unique, cancers with a similar stage tend to have a similar prognosis and are often treated in the same way.

The staging system most often used for gastric cancer is the American Joint Committee on Cancer (AJCC) TNM system (Washington, 2010), which is based on the size of the primary tumour (T), spread to nearby lymph nodes (N), and spread (metastasis) to distant sites (M). The "T" plus a letter or number (0 to 4)

is used to describe the how far the tumour has grown into the stomach wall. Some tumour stages are divided into even smaller groups that help describe in more detail. The "N" uses a system of numbers and letters as an indicator of if the cancer has spread to regional lymph nodes (defined as lymph nodes within the abdomen), and if it has, how widespread the cancer is. The "M" is described using a 0 or 1, indicating whether the cancer has spread to other parts of the body, called distant metastasis.

1.1.2 Risk factors

Gastric cancer risk is complex and is modified by a number of environmental exposures and genetic factors. Individually, these factors may only contribute a minimal amount towards an individuals risk of disease risk, but together they may explain larger population trends. To improve the management of gastric cancer patients, we must identify and understand both the environmental and genetic factors that influence the risk of this disease. Known risk factors for gastric cancer include diet, obesity, low socioeconomic status, *H. pylori* infection, and a family history of gastric cancer (Alemán et al., 2014; Kelley & Duggan, 2003).

1.1.2.1 Diet

In 2007, salt and salty/salted foods were classified as a 'probable' risk factor for gastric cancer (Marmot et al., 2007). Diets with a high intake of salt and various traditional salt preserved foods such as salted fish, cured meat, and salted vegetables have been linked to a 22% increased risk of gastric cancer (Ge et al., 2012; Peleteiro, Lopes, Figueiredo, & Lunet, 2011; Tsugane & Sasazuki, 2007). Additionally, in rodents, high salt intake was shown to damage the

stomach mucosa by inducing proliferative change, and increase susceptibility to gastric carcinogens (Takahashi, Kokubo, Furukawa, Kurokawa, & Hayashi, 1984; Tatematsu, Takahashi, Fukushima, Hananouchi, & Shirai, 1975). The global decline in the incidence of gastric cancer over the last 50 years has, in part, been attributed to the increased use of refrigeration, decreasing the need for salting and other salt-based food preservation methods (Karimi, Islami, Anandasabapathy, Freedman, & Kamangar, 2014; B. Park et al., 2011). Additionally, with increased awareness of hypertension there has been a shift towards lowering salt intake.

Low consumption of fruit and vegetables has also been associated with an increased risk of gastric cancer. Fruit and vegetables are a rich source of vitamin C, folate, and carotenoids, all of which are suggested to inhibit carcinogenesis by regulating metabolism of xenobiotic-enzymes (Wiseman, 2008). A 100 gram increase in vegetable and fruit intake was shown to reduce the relative risk of gastric cancer to 0.81 (95% CI 0.75-0.87) and 0.74 (95% CI 0.69-0.81), respectively. However, support for this association remains inconsistent, with some reports showing no association between fruit and vegetable consumption and gastric cancer risk (Freedman et al., 2008; Gonzalez et al., 2012).

1.1.2.2 Obesity

Obesity is a growing problem in modern societies and is a well-established risk factor for several diseases, including gastric cancer (M. Ng et al., 2014; Yang et al., 2009). Individuals with a body mass index (BMI) of 30-35 and > 40, have a 2-fold and 3-fold increased risk of cancer of the gastroesophageal junction, respectively, when compared to individuals with a BMI of < 25

(Hoyo et al., 2012). The biochemical mechanisms linking obesity to cancer include: hyperinsulinemia, increased levels of insulin-like growth factors (IGF), and altered IGF/IGF-binding protein ratios that promote cell division and inhibit cell death (Alemán et al., 2014). Additionally, excess body weight may directly cause gastroesophageal reflux disease, another known risk factor for gastric cancer (Alemán et al., 2014).

1.1.2.3 Low socioeconomic status

Low socioeconomic status is a well established risk factor for most cancer types (Ward et al., 2004). Markers of low socioeconomic status, including low income and low education, are associated with gastric cancer and its precursor lesions (Eusebi, Zagari, & Bazzoli, 2014). Low socioeconomic groups tend to have higher rates of *H. pylori* infection, obesity, and high salt diets, which may also mediate this relationship (Eusebi et al., 2014; Guggenheim & Shah, 2013).

1.1.2.4 *Helicobacter pylori*

H. pylori is the strongest known risk factor for gastric cancer and the most common agent for infection related cancers (Helicobacter and Cancer Collaborative Group, 2001). In 1994, the International Agency for Research on Cancer (IARC) and the WHO classified *H. pylori* as a type-I carcinogen (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2010a). *H. pylori* is normally acquired in childhood, and without antimicrobial intervention, can persist for a lifetime. It is estimated that half of the worlds population is infected with *H. pylori* , of which 1-3% will consequently develop gastric cancer (Everhart, 2000; Peek & Crabtree, 2006). It is estimated that

65-80% of all gastric cancer cases are caused by a *H. pylori* infection (Helicobacter and Cancer Collaborative Group, 2001).

The association between *H. pylori* infection and gastric carcinoma is not fully understood. It is thought *H. pylori* induces chronic gastric inflammation, which can lead to mucosal atrophy, metaplasia, dysplasia, and in some cases, carcinoma (Correa, 1992). *H. pylori* infection significantly increases the risk of developing both diffuse-type and intestinal-type gastric adenocarcinoma; however, only the intestinal-type is linked to chronic inflammation, suggesting that a different mechanism is required for the development of diffuse-type disease (Polk & Peek, 2010).

Certain strains of *H. pylori* positive for specific virulence factors are more likely to cause gastric cancer. One virulence factor that clearly influences gastric cancer risk is the cytotoxin-associated gene A (CagA) pathogenicity island (Parsonnet et al., 1991). When compared to an uninfected person, CagA-positive *H. pylori* are associated with a 5.8-fold increased risk of developing gastric cancer. In comparison, CagA-negative stains have a reduced 2.2-fold increased risk (Parsonnet, Friedman, Orentreich, & Vogelman, 1997). Another virulence factor strongly linked to an increased gastric cancer risk is the VacA toxin, encoded by the *vacA* gene (Boquet & Ricci, 2012). All *H.pylori* possess *vacA*; however, there is considerable variation in the 5' of the gene, for which only certain alleles are associated with a higher gastric cancer risk (Parsonnet et al., 1997). *H. pylori* strains containing the s1 and m1 alleles are associated with a particularly high risk of cancer. Notably, *H. pylori* strains with multiple virulence factors are at a greater risk compared to strains lacking these factors (Cover & Peek, 2013).

The risk of gastric cancer is not only influenced by *H. pylori* strain specific virulence factors, but also by host and environmental factors. As previously

described, diets high in salt increase the risk of developing gastric cancer (Section 1.1.2.1). Markedly, *H. pylori* infections appear to synergise with high-salt diets to further increases gastric cancer risk, relative to those who consumed less salt (Lee et al., 2003; Tsugane & Sasazuki, 2007). The association between *H. pylori* and gastric cancer risk is supported by animal models, for which both Mongolian gerbils and IL-10 deficient mice infected with *H. pylori* and fed a high salt diet have shown increased incidence of gastric cancer (Gamboa-Dominguez et al., 2007; J. Park et al., 2014). Although the mechanism causing this effect is not fully understood, recent work has shown that both *cagA* and *vacA* are upregulated when *H. pylori* is cultured in a media with a high salt content (Gancz, Jones, & Merrell, 2008; Loh, Torres, & Cover, 2007), suggesting these virulence factors are important. Additionally, host genetic factors can influence the immune and inflammatory response to *H. pylori* infection and contribute to carcinogenesis. For example, polymorphisms in the human interleukin-1 gene cluster enhance production of interleukin-1-beta (IL-1beta) in response to *H. pylori* infection and create favourable conditions for carcinogenesis (El-Omar, Carrington, Chow, & McColl, 2000). It is likely that both host and environmental factors explain why some individuals infected with *H. pylori* develop gastric cancer, while others do not.

1.1.2.5 Family history

Familial aggregation occurs in approximately 10% of gastric cancers and is significantly associated with an increased risk of disease (Yaghoobi, Bijarchi, & Narod, 2010). In most populations, having a first-degree relative with gastric cancer increases the risk of developing gastric cancer by between 1.5 and 3.5-fold (Yaghoobi et al., 2010). Shared environmental exposures, such as *H. pylori* infections or a common diet could also all explain this increased familial

risk. Alternatively, low penetrance genetic variants or a pathogenic variant in a gene associated with a highly penetrant cancer susceptibility syndrome could explain the aggregation of cancers within these families (Lynch, Grady, Suriano, & Huntsman, 2005).

1.1.2.6 Alcohol consumption

A relation between alcohol consumption and gastric cancer risk is biologically plausible. Heavy alcohol consumption could cause gastric inflammation and chronic gastritis, which could increase gastric cancer risk (Franceschi & La Vecchia, 1994). In 1988, the IARC concluded that there was inadequate evidence for alcohol to be classified as a risk factor for gastric cancer (International Agency for Research on Cancer and World Health Organization, 1988). In 2007, the IARC reviewed the evidence and concluded that there were suggestions that alcohol consumption might be associated with an increased risk (Tramacere et al., 2011). Recently, in a meta-analysis of 44 case-control and 15 cohort studies, there was a risk associated with moderate alcohol drinking ($\leq$ 4 drinks per day; RR 1.07, 95% CI 1.01-13), and a stronger association with heavy alcohol consumption ($>$ 4 drinks per day; RR 1.20, 95% CI 1.01–1.44) (Tramacere et al., 2011). The study by Tramacere et al. (2011) suggests that alcohol is a modest risk factor for gastric cancer, however, confounding effects due to dietary habits and *H.pylori* infections could not be ruled out (Tramacere et al., 2011).

1.1.2.7 Smoking

Smoking is a well established risk factor for multiple cancer types. Despite this, it wasn't until 2002 that the IARC classified smoking as a risk factor for

gastric cancer (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2010b). The reason for the delay, in part, is due to inconstancies in the relationship between smoking and gastric cancer risk between studies. While most studies have reported a small risk (less than two-fold increased risk), some studies had reported no risk (Kelley & Duggan, 2003). A recent meta-analysis of cohort studies concluded that when compared to 'never smokers,' the relative risk for male and female smokers was 1.62 (95% CI 1.50–1.75) and 1.20 (95% CI 1.01–1.43), respectively (Ladeiras-Lopes et al., 2008). Trend estimation analysis by Ladeiras-Lopez et al, (2008) found that the relative risk of gastric cancer increased from 1.3 to 1.7 for low and high consumption of cigarettes, respectively (Ladeiras-Lopes et al., 2008).

1.1.2.8 Other Risk Factors

Numerous other factors have been proposed to increase gastric cancer risk. However, evidence for these risk factors is limited and sometimes contradictory. Other risk factors include: exposure to radiation (D. E. Thompson et al., 1994), blood-type A (Aird, Bentall, & Roberts, 1953), prior gastric surgery (Stalnikowicz & Benbassat, 1990), and infection with Epstein-Barr virus (Shibata & Weiss, 1992). Further research will be required before these are considered putative gastric cancer risk factors.

1.1.3 Familial Gastric Cancer

Hereditary gastric cancer syndromes, where inherited pathogenic variants are predisposing to a heightened risk of developing gastric cancer, are thought to account for between 1-3% of all gastric cancers. These syndromes predispose affected individuals to an extreme risk of developing gastric cancer, and may

also cause the early-onset of the disease. There are three main gastric cancer syndromes: hereditary diffuse gastric cancer (HDGC), familial intestinal gastric cancer (FIGC), and gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS). Furthermore, gastric cancers have been identified as a part of the tumour spectrum in several other hereditary cancer syndromes. The considerable uncertainty associated with hereditary cancer syndromes makes the identification of genetic factors and families who are at risk of great importance.

1.1.3.1 Hereditary Diffuse Gastric Cancer (HDGC)

HDGC is an autosomal dominant cancer syndrome primarily characterised by an extreme risk of developing diffuse gastric cancer. Women with HDGC syndrome have an additional risk of developing lobular breast cancer (van der Post et al., 2015). Familial clustering of diffuse gastric cancer was first reported in 1964 (Jones, 1964), however it wasn't until 1998 when a genetic basis for the disease was identified for the syndrome to be recognised (Guilford et al., 1998). In 1999 the International Gastric Cancer Linkage Consortium (IGCLC) proposed the first clinical criteria for the recommended screening of *CDH1* (Caldas et al., 1999). In 2010, and again in 2015, as more about this syndrome was discovered and the cost of sequencing declined, the criteria were relaxed to include more families that may be at risk (Fitzgerald et al., 2010; van der Post et al., 2015). The current 2015 criteria for HDGC include: (1) families with two or more patients with gastric cancer at any age, one confirmed diffuse gastric cancer; (2) individuals with diffuse gastric cancer before the age of 40; and (3) families with diagnoses of both diffuse gastric cancer and lobular breast cancer (one diagnosis before the age of 50) (van der Post et al., 2015). Additionally, the 2015 criteria identified families whom may benefit from

testing as: (1) patients with bilateral or familial lobular breast cancer before the age of 50; (2) patients with diffuse gastric cancer and cleft lip/palate; and (3) those with precursor lesions for signet ring cell carcinoma (van der Post et al., 2015).

Pathogenic germline variants in the E-cadherin gene *CDH1* are responsible for approximately 40% of families that meet the clinical criteria for HDGC (Hansford et al., 2015). More than 120 HDGC families with pathogenic *CDH1* variants have been described in published literature (Hansford et al., 2015), although there are more than 500 known HDGC families worldwide (P. Guilford, personal communication). An explanation of the remaining 60% of HDGC cases remains largely elusive. Germline variants in the α-catenin gene *CTNNA1* have been identified in a small number of HDGC families, but do not appear to be common (Hansford et al., 2015; Majewski et al., 2013; Schuetz et al., 2012). A small number of additional candidate genes have been identified (described in detail in Section 1.1.4.3), but still require further validation before they can be considered putative HDGC genes.

1.1.3.2 Familial Intestinal Gastric Cancer (FIGC)

Guidelines for the classification of FIGC include: (1) intestinal gastric cancer in two or more first or second degree relatives, with at least one confirmed case of intestinal pathology diagnosed before age 50; and (2), intestinal gastric cancer in three or more first or second degree relatives, independent of age (Kluijt et al., 2012). The few cases of FIGC that have been documented have shown an autosomal dominant inheritance pattern (Caldas et al., 1999; Kluijt et al., 2012). Aggregates of FIGC are believed to be a combination of both genetic and environmental factors, with any genetic cause yet to be identified

(Kluijt et al., 2012). Currently no clinically relevant recommendations are available for the management of FIGC families (Kluijt et al., 2012).

1.1.3.3 Gastric Adenocarcinoma and Proximal Polyps of the Stomach

GAPPS is characterised by the autosomal dominant transmission of fundic gland polyposis (including dysplastic lesions and intestinal-type gastric adenocarcinoma, or both) that are restricted to the proximal stomach, with no evidence of duodenal or colorectal polyposis (Worthley et al., 2012). Clinical criteria from GAPPS include: (1) gastric polyps restricted to the body of the fundus with no evidence of colorectal or duodenal polyposis; (2) more than 100 polyps in the index case or more than 30 polyps in a first degree relative of another case; (3) mainly fundic gastric polyps, some with regions of dysplasia; (4) an autosomal dominant pattern of inheritance; and (5), exclusion of other heritable gastric polyposis syndromes. Recently, specific variants in the *APC* promoter 1B were found to be a cause of GAPPS in six families (J. Li et al., 2016).

1.1.3.4 Other hereditary cancer syndromes

Hereditary cancer syndromes can predispose to cancer at a large set of different body sites. Gastric cancer is recognised as a part of some specific cancer syndromes. These syndromes are detailed below:

Lynch syndrome is a highly penetrant colorectal cancer syndrome. Mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*) are frequently mutated, causing microsatellite instability and mismatch repair deficiency (Rahner et al., 2010). While Lynch syndrome predominantly predisposes to

colorectal and endometrial cancers, it is estimated that 1.6% of all Lynch syndrome patients will develop gastric cancer, typically of the intestinal-type (Capelle et al., 2010). Strikingly, patients whom carry pathogenic variants in specific Lynch syndrome associated genes have a much greater risk of the developing gastric cancer. For example, individuals with a germline defect in *MLH1* or *MSH2* have a 4.8% and 9.0% risk of developing gastric cancer, respectively (Capelle et al., 2010). Gastric surveillance is recommended for patients diagnosed with Lynch syndrome (Capelle et al., 2010).

Li-Fraumeni syndrome is an autosomal dominant inherited cancer predisposition syndrome that predisposes to a wide range of tumour types. Cancers associated with Li-Fraumeni syndrome typically present with an early-onset (< 45 years of age) (Malkin et al., 1990). Pathogenic germline variants in the DNA repair gene *TP53* have been identified in approximately 70% of cases whom fit the clinical criteria for Li-Fraumeni syndrome (Malkin et al., 1990). While sarcomas, breast carcinomas, brain tumours, and leukemia are most common, gastric cancer represents 1.8-4.9% of cancers in Li-Fraumeni syndrome families (Masciari et al., 2011). Furthermore, 40% of families with a pathogenic *TP53* variant will have at least one family member diagnosed with gastric cancer (Malkin et al., 1990; Masciari et al., 2011). Currently, both intestinal- and diffuse-type gastric cancers appear to be associated with Li-Fraumeni syndrome (Masciari et al., 2011).

Familial adenomatous polyposis (FAP) is caused by pathogenic germline *APC* variants and is characterised by early-onset colorectal cancer with colonic and rectal adenomas (Lipton & Tomlinson, 2006). Adenomas can also develop in the upper gastrointestinal tract, and if left untreated, can progress to malignant disease. Gastric fundic gland polyps and adenomas in the atrum of the stomach can also occur (Lipton & Tomlinson, 2006). Gastric

adenocarcinomas in FAP are generally thought to arise from fundic gland polyps (Spigelman, Talbot, Williams, Domizio, & Phillips, 1989; Vasen et al., 2008). Notably, the risk of gastric cancer in these families is not much higher than the general population (Vasen et al., 2008).

Peutz-Jeghers syndrome (PJS) is primarily caused by pathogenic variants in the gene *STK11*. PJS is characterised by the development of hamartomatous gastrointestinal polyps with mucocutaneous pigmentation, and an increased risk of gastrointestinal and breast cancers at a young age (Utsunomiya, Gocho, Miyanaga, Hamaguchi, & Kashimure, 1975). After small intestine and colorectal carcinoma, gastric cancer is the third most common tumour in PJS (Chun & Ford, 2012). Meta-analysis suggests a cumulative risk of gastric cancer for pathogenic variant carriers of 29% by the age of 65 years. Increased surveillance is recommended for these families (Van Lier et al., 2010).

Juvenile polyposis syndrome (JPS) is characterised by numerous juvenile polyps developing in the colon and/or stomach. It is an autosomal dominant syndrome that is caused by pathogenic variants in several different genes, most commonly *SMAD4* and *BMPR1A* (Allen & Terdiman, 2003). Gastric cancer develops in 21% of patents who are affected by gastric polyps (Howe et al., 2004). Regular surveillance and screening is recommended for patients from an early age or when symptoms of polyps present (Allen & Terdiman, 2003; Howe et al., 2004).

Familial breast and ovarian cancer syndrome is predominantly caused by pathogenic germline *BRCA1* and *BRCA2* variants and predisposes women to both breast and ovarian cancer (Breast Cancer Linkage Consortium, 1999). Pathogenic variants in *BRCA1* and *BRCA2* also predispose both women and men to an increased risk of gastric cancer (Friedenson, 2004). A large meta-analysis study has shown the average relative risk of gastric cancer is

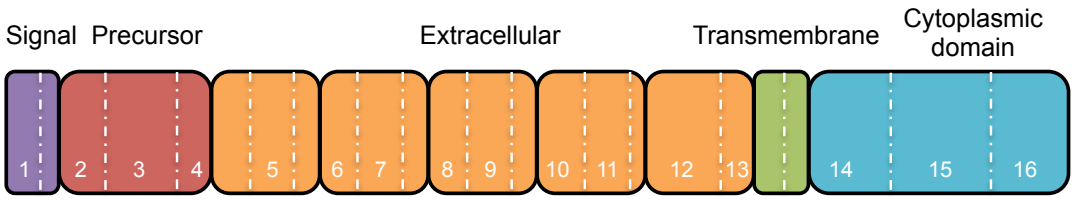


Figure 1.1: The E-cadherin protein. The *CDH1* gene maps to chromosome 16q22.1 and consists of 16 exons. The *CDH1* gene encodes the 120-kDa protein E-cadherin. This protein has three major components: signal peptide consisting of 27 amino acids encoded by exons 1 and 2 (purple), precursor peptide consisting of 154 amino acids encoded by exons 2 to 4 (red), and mature protein containing 728 amino acids encoded by exons 4 to 16. The mature protein segment has an extracellular domain that includes exons 4–13 (orange), a smaller transmembrane domain that includes exons 13 and 14 (green), and a cytoplasmic domain that comprises exons 14–16 (blue).

collectively higher for *BRCA1* and *BRCA2* pathogenic variant carriers, first degree relatives of pathogenic variant carriers, women that have had a breast cancer diagnosis, and woman that meet the clinical criteria for *BRCA1* and *BRCA2* sequencing (RR = 1.69; 95% CI 1.21-2.38) (Friedenson, 2004). Currently, it is not known if a particular histological subtype is associated with familial breast and ovarian cancer syndrome (Jakubowska et al., 2003).

1.1.4 The cell adhesion molecule E-cadherin (*CDH1)*

The human E-cadherin gene *CDH1* is located on chromosome 16q22.1 and belongs to a family of genes encoding calcium-dependent cell-adhesion molecules. The *CDH1* gene comprises 16 exons and spans a region of approximately 100 kb (Figure 1.1). *CDH1* is tightly regulated and has a high density CpG island regulatory region in intron 1 (Berx et al., 1995).

The E-cadherin protein is composed of three main parts: a large extracellular domain, comprised of five tandem repeats that contains sites with adhesive activity for forming bonds with E-cadherin on opposing cells; a small

transmembrane segment, that binds p120-catenin and supports cadherin clustering and adhesive strength between cells; and a short cytoplasmic tail, that binds beta-catenin and interacts with the actin-cytoskeleton via a protein complex with α-catenin (Weber, Bjerke, & DeSimone, 2011) (Figure 1.2).

E-cadherin is expressed very early in development and is important for the differentiation and polarisation of cells (Papusheva & Heisenberg, 2010). Animal models heterozygous for mutated E-cadherin show normal development, while animals homozygous for mutated E-cadherin show severe developmental abnormalities that are lethal during embryonic development (Larue, Ohsugi, Hirchenhain, & Kemler, 1994). E-cadherin is one of the most important molecules for cell-cell adhesion in normal epithelial tissue, where it is predominantly expressed at the basolateral membrane of epithelial cells and is an essential part of the adherens junction (Vleminckx, Vakaet, Mareel, Fiers, & Van Roy, 1991). In addition to its primary roles in the structure and maintenance of adherens junctions, there is strong evidence that E-cadherin mediates complex interactions between signalling pathways that establish and maintain cell-polarity, migration, proliferation, differentiation, and apoptosis (Van Roy & Berx, 2008).

1.1.4.1 E-cadherin and cancer

The E-cadherin gene *CDH1* is a well established tumour suppressor (Vleminckx et al., 1991). Loss of E-cadherin is associated with malignant transformation and tumour progression (Jeanes, Gottardi, & Yap, 2008). E-cadherin negative tumours are often characterised by major changes in the organisation of the cytoskeleton, loss of adhesion, and abnormal adhesion-mediated signalling (Paredes et al., 2012). Disruptive genetic

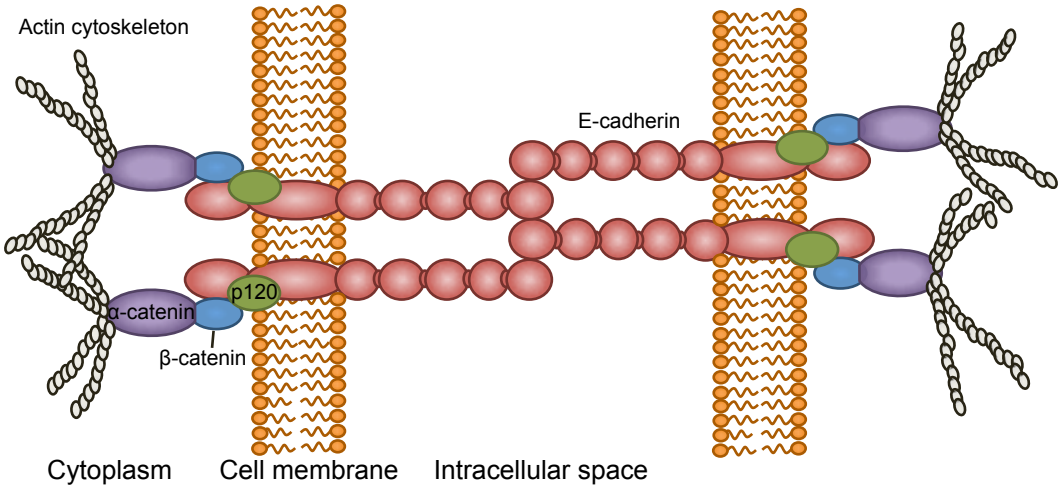


Figure 1.2: A schematic diagram of the E-cadherin, beta-catenin and α-catenin complex. Opposing E-cadherin extracellular domains bind to one another via calcium-dependent dimerisation. Intracellular E-cadherin domain binds to beta-catenin which complexes with α-catenin and the actin cytoskeleton.

aberrations in the *CDH1* gene and subsequent loss of functional E-cadherin is a feature of many epithelial tumours, including prostate, ovarian, and lung carcinomas; and is the hallmark of both the sporadic and familial forms of diffuse gastric cancer and lobular breast cancer (Van Roy & Berx, 2008). Consistent with classical tumour suppressor activity, a two-hit mechanism often occurs in tumours, mutating or down-regulating both copies of *CDH1*, for a complete loss of functional E-cadherin (Grady et al., 2000; Oliveira, de Bruin, et al., 2004). Tumours with loss of E-cadherin sometimes undergo an epithelial-mesenchymal transition (EMT), characterised by the altered expression of transcription factors, cell surface receptors, and cytoskeletal proteins. An EMT causes epithelial cells to take on a mesenchymal phenotype, that enhances their migration, invasion, and resistance to apoptosis (Onder et al., 2008; Tiwari, Gheldof, Tatari, & Christofori, 2012).

1.1.4.2 Germline *CDH1* and HDGC

As previously described, pathogenic variants in germline *CDH1* are associated with an extreme risk of developing diffuse gastric cancer and explain approximately 40% of families that meet the clinical criteria for HDGC (Hansford et al., 2015). More than 120 pathogenic *CDH1* variants have been identified across all coding regions of *CDH1* and do not appear to be restricted to any particular functional domain of the E-cadherin protein (Hansford et al., 2015). Strikingly, three pathogenic variants (c.1137G>A, c.1565+1G>A, and c.1792C>T) have been found in multiple families from seemingly unrelated backgrounds, raising the possibility of variant hotspots (Hansford et al., 2015).

Geographically, the frequency of reported pathogenic variants in germline *CDH1* varies substantially between low-incidence and middle/high-incidence areas. In a recent meta-analysis describing 122 pathogenic germline *CDH1* variants, 94 (77.0%) were described in low-incidence areas, 16 (13.1%) were described in middle/high-risk incidence areas, and 9.9% were without ethnicity information (Corso, Marrelli, Pascale, Vindigni, & Roviello, 2012). The large enrichment of pathogenic variants from low-incidence areas suggests that pathogenic germline *CDH1* variants are rarely identified in middle/high-incidence areas (Corso et al., 2012). It is likely that in these middle/high-incidence areas, true HDGC families are lost in a background of family clusters caused by a shared environmental factors.

The stomachs of *CDH1* germline variant carriers contain multifocal, stage T1a signet-ring cell carcinomas, that are generally indolent but can show an unpredictable tendency towards progression (Blair, 2012). Nearly all stomachs from pathogenic *CDH1* variant carriers who have undergone total gastrectomies have multiple foci of stage T1a signet-ring cell carcinomas (Blair,

2012). The number of foci observed per patient is highly variable. The average number of foci ranged between 2 and 487 per family when multiple members of different HDGC families have had their foci number determined (Blair, 2012; Charlton et al., 2004). Foci are typically less than 1 mm in diameter but can be as large as 10 mm (Charlton et al., 2004). In some families, the transition zone of the stomach have the greatest number of signet-ring cell carcinomas (Charlton et al., 2004), however this is not always the case (Rogers et al., 2008). The great variability in the number, size, and location of foci suggests that background genetics or environmental exposures are important in the progression of these tumours.

The trigger for the initial development of these cancer foci is downregulation of the second *CDH1* allele (Humar & Guilford, 2009). It is hypothesised that the loss of E-cadherin expression disrupts the normal orientation of the mitotic spindle, allowing a proportion of epithelial cells to divide out of the normal epithelial cell plane of division and invade the lamina propria (Humar & Guilford, 2009). It has been proposed a proportion of the foci of signet-ring cells that make it to the the lamina propria are transient, while others undergo an EMT and progress (P. Guilford, personal communication). The EMT in early HDGC is often incomplete, with some cells retaining expression of some epithelial cell markers (Humar et al., 2007). Mesenchymal-like morphology can be seen in some of the larger early lesions and is a dominant feature of all tumour stages beyond T1a (Humar et al., 2007). The shift to the mesenchymal cell type correlates with the activation of the EMT-inducer c-SRC and its downstream targets (Humar et al., 2007).

The average age of onset for HDGC is approximately 38 years, but has affected some as young as 14 years of age (Guilford et al., 1998; Hansford et al., 2015). Recent penetrance analysis of HDGC families with pathogenic germline *CDH1*

variants have estimated the cumulative risk of developing diffuse gastric cancer by the age of 80 years is 70% for men (95% CI 59%-80%) and 56% for women (95% CI 44%-69%). In addition, women carry a 42% (95% CI 23%-68%) cumulative risk of developing lobular breast cancer by the age of 80 years (Hansford et al., 2015). The variable penetrance is not well understood and it is likely that both environmental and genetic factors act as modifiers of risk in these families.

Families found to meet the clinical criteria for HDGC are recommended to undergo genetic screening for *CDH1* variants (van der Post et al., 2015). If a pathogenic *CDH1* variant is found, unaffected relatives may be offered genetic counselling services and subsequent testing for risk stratification. Genetic counselling for pathogenic variant carriers is extensive and tailored to the individuals age, sex, and nutritional issues (Kluijt et al., 2012). Those found to carry a pathogenic *CDH1* variants are offered a total prophylactic gastrectomy, a radical procedure with high morbidity that removes the risk of disease. Alternatively, pathogenic variant carriers can elect for regular endoscpoic screening to survey the stomach for early stage foci. Currently, it is still very difficult to detect diffuse gastric cancer at its earliest stages using endoscopic screening making prophylactic gastrectomy the recommended procedure for individuals with a strong family history. The recommended age for prophylactic gastrectomy surgery for *CDH1* variant carriers is greater than 20 years (van der Post et al., 2015).

1.1.4.3 Beyond *CDH1* variation as a cause of HDGC

For families that meet HDGC criteria but don't carry a pathogenic variant in *CDH1*, there is major uncertainty in how to best manage risk and identify

those who would benefit from intervention. In a recent study of *CDH1* variant-negative families, analysis of 55 genes associated with HDGC and other gastrointestinal cancer syndromes revealed candidate variants in 11% of probands. Pathogenic variants were found in genes of high and moderate penetrance including *ATM*, *BRCA2*, *CTNNA1*, *MSR1*, *PALB2*, *PRSS1*, *SDHB*, and *STK11* (Hansford et al., 2015).

Of particular interest were the variants in α-catenin (*CTNNA1*), that have now been observed in multiple families that meet HDGC criteria (Hansford et al., 2015; Majewski et al., 2013). α-catenin is also a part of the cadherin-catenin complex at the cell membrane, suggesting that a similar mechanism may be responsible for the progression of disease in these families. Early studies that have screened α-catenin in HDGC families did not discover a large number of pathogenic variant carriers, suggesting *CTNNA1* is not a common HDGC gene (Schuetz et al., 2012). Unfortunately, there is currently insufficient data to determine the penetrance of *CTNNA1* variants. The most recent clinical guidelines for HDGC families now include sequencing of *CTNNA1* as an option for *CDH1* variant-negative families (van der Post et al., 2015).

Additional studies have implicated a small number of genes as directly predisposing to HDGC. A variant in the mitogen-activated protein kinase kinase kinase 6, encoded by the gene *MAP3K6*, was identified in one large family with a history of diffuse gastric cancer (Gaston et al., 2014). In analysis of further *CDH1* variant-negative HDGC families, an additional five variants in six unrelated families were identified (one nonsense and four missense). In another family with a strong history of diffuse gastric cancer, rare variants in the genes *DOT1L*, *FBXO24*, and *INSR* were identified. While these gene variants may be of importance to these families, further research will be required before they are able to guide clinical intervention. These studies and

genes are discussed in greater detail in Chapter 5 Section 5.1.2.

Regulation of E-cadherin by unspecific defects at the *CDH1*-locus may also explain HDGC in some families. Under normal conditions *CDH1* is biallelically expressed (Gimelbrant, Hutchinson, Thompson, & Chess, 2007), however in tumours occurring in patients with a pathogenic germline *CDH1* variant, *CDH1* expression is monoallelicly expressed or completely absent (Pinheiro et al., 2010). Interestingly, in a study of germline RNA extracted from peripheral blood lymphocytes from cancer-free, HDGC *CDH1* variant-positive, and HDGC variant-negative probands, *CDH1* showed biallelic expression in all cancer free samples, and irregular mono-allelic expression in 80% and 71% of variant positive and negative probands, respectively (Pinheiro et al., 2010). The study by Pinheiro et al. suggests that some sort of non-coding variant or other yet to be identified regulatory element could be impairing the regulation of E-cadherin in some *CDH1* variant-negative families.

1.1.5 Next-Generation Sequencing (NGS)

Prior to the advent of NGS, human genomics was primarily concerned with the characterisation of single genes and relied largely on linkage-analysis with subsequent candidate gene selection. With NGS, it became possible to sequence entire genomes and survey for genetic aberrations that may be responsible for inherited syndromes, disease, and cancer. NGS has revolutionised the field of genetics, creating a paradigm of using genetic sequencing to address biological questions at a genome-wide scale.

1.1.5.1 Advances in sequencing technologies

In 1977, Frederick Sanger developed the first rapid DNA sequencing technique (Sanger, Nicklen, & Coulson, 1977). Known as the Sanger's dideoxy chain-termination sequencing method, this technique incorporates fluorescently labelled chain-terminating dideoxynucleotides into DNA extension reactions, that when separated by size, can be used to determine a genetic sequence (Sanger et al., 1977). The dideoxy chain-termination method, hereinafter referred to as Sanger sequencing, revolutionised genetics and was the most widely used sequencing method for nearly 40 years.

In the mid 2000s, Sanger sequencing was exceeded by next-generation sequencing technology (Mardis, 2008). Instead of inferring nucleotide identity by fluorescent labels, NGS utilises pyrophosphate synthesis to detect nucleotides as they are incorporated in real time (Metzker, 2010). In the first NGS machines, libraries of DNA molecules were attached to beads that underwent a water-in-oil PCR reaction. These DNA coated beads were washed over a plate of microscopic wells only large enough to hold a single bead, and were sequenced in parallel (Metzker, 2010). Later, several new parallel sequencing techniques were developed. Of the new technologies, bridge amplification excelled. For this, adapter bracketed DNA libraries are passed over complementary oligonucleotides attached to a flow cell (Metzker, 2010). The single-stranded sequences are amplified in a solid phase PCR to produce millions of dense double-stranded clusters that can be detected using sensitive cameras (Metzker, 2010). NGS revolutionised the sequencing process by using array-based systems that allow millions of small DNA fragments to be sequenced in parallel, greatly increasing the yield of sequencing efforts.

NGS technology was initially used to study whole genomes, including both

coding and non-coding regions. However, the amount of data generated by whole-genome sequencing can be difficult to manage and is frequently superfluous to requirements. Because of this, a variety of targeted sequencing approaches that address defined regions of the genome have been developed. Currently, there are two main techniques used to target specific regions of the genome (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013). The first is by PCR, typically involving multiplexed primer pairs that are designed to target specific regions of DNA. Following amplification of targeted regions, platform-specific adapters are added to each end of the PCR products to form a library suitable for sequencing. The second approach involves hybrid capture, for which probes are designed to bind selected regions of DNA and are captured using magnetic beads. Hybrid capture is commonly used to capture all known coding regions of the genome in an approach referred to as 'exome capture' (Koboldt et al., 2013). Alternatively, a panel of probes can be synthesised to target specific regions of DNA in a 'targeted panel.' Targeting NGS is an efficient strategy for uncovering pathogenic variants that are attributable to rare hereditary syndromes, such as hereditary cancers, but are limited to genes and variants that have previously been associated with risk.

Along with the advances in the sequencing technology and target enrichment strategies, multiplexing has revolutionised the efficiency of sequencing. Multiplexing allows for the pooling of multiple samples into a single sequencing reaction, cutting the costs and time associated with separate sequencing reactions (Elshire et al., 2011; Smith et al., 2010). For multiplexing, unique sample-specific sequence indices are included in adapters used during library preparation, making the origin of each strand of DNA identifiable. Being able to trace DNA to a sample of origin enables the sequencing of accumulated samples in a single sequencing reaction. Multiplexing is proving to be the most effective way to classify recurrent variants in large cohorts,

when large numbers of samples and regions of interest can be sequenced in a single reaction.

1.1.5.2 Utilising NGS for HDGC gene discovery

A genomic era of research is developing rapidly. In particular, the field of cancer genomics has been impacted profoundly by the application of NGS technology (McKenna et al., 2010). NGS has accelerated the pace of gene discovery while dramatically reducing the cost of data production, enabling remarkable advances in our understanding of the somatic and germline variant profiles of cancer genomes. Hence, there has been a rapid progression from using traditional Sanger sequencing, to either targeted, or whole-genome sequencing using these massively parallel sequencing platforms. There still remains significant challenges in understanding HDGC, but our fundamental understanding of which genes are frequently mutated in cancer cells, the pathways that are impacted by these variants, and how these contribute to cancer biology, will undoubtedly mature as a result.

Pathogenic germline variants in *CDH1* are a well established cause of HDGC. However, for approximately 60% of families that fit the clinical criteria for HDGC, a genetic cause remains elusive (Hansford et al., 2015). NGS has assisted in the identification of a small number of genes that may be predisposing in some families (Section 1.1.4.3), however the majority of *CDH1* variant-negative families are still unexplained. As the cost of sequencing continues to drop, it is increasingly feasible to address some of the questions surrounding HDGC. Families and populations who appear to be at a high-risk of HDGC can be examined for genetic variants that may be predisposing to disease.

1.2 Thesis aims and outline

The principal aim of this thesis was to develop a greater understanding of the genetic factors that increase diffuse gastric cancer risk. To address this, I have used NGS and bioinformatic tools to identify predisposing germline variants that are carried by gastric cancer patients. This thesis includes three results chapters, each identifying and assessing predisposing germline variants in different study groups, and providing distinctive insights into different aspects of hereditary gastric cancer risk and management.

Chapter 3 describes the investigation of pathogenic germline *CDH1* variants as a cause of the high incidence of diffuse gastric cancer in New Zealand Māori. This chapter examines both rare variants and common polymorphisms in germline *CDH1*. In **Chapter 3**, I also discuss the impact genetic screening for *CDH1* variants has had on the New Zealand Māori gastric cancer population.

Chapter 4 explores pathogenic *CDH1* variants as a cause of cancer in a cohort of Chilean gastric cancer patients that presented with a striking family history or early-onset disease. This chapter provides a direct insight into the importance of screening for germline *CDH1* variants, as well as the application of HDGC testing criteria in identifying gastric cancer patients in a high-incidence population such as Chile.

Chapter 5 expands the search for predisposing germline variants beyond *CDH1*. In this chapter, whole-exome sequencing is used to identify rare germline variants carried by *CDH1* variant-negative diffuse gastric cancer patients that fit the clinical criteria for HDGC. This chapter identifies variants in genes that are associated with increased gastric cancer risk, as well as genes that have not previously been linked to diffuse gastric cancer risk. The

findings from Chapter 5 have implications for future HDGC studies and the clinical screening and management of gastric cancer patients.

Finally, **Chapter 6** discusses the results from the preceding chapters in relation to the biggest challenges for improving the screening and management of those who are/or may be at a risk of diffuse gastric cancer.

Chapter 2

Methods and Materials

2.0.1 Reagents

Agarose - AppliChem, USA

Agencourt AMPure XP beads - Beckman Coulter, USA

dsDNA High Sensitivity Assay Kit - Thermo Fisher, USA

Ethidium Bromide - Merck, USA

GF–1 Nucleic Acid Extraction Kit - Vivantis, Malaysia

Kapa Hotstart Readymix with dye - Kapa Biosystems, South Africa

Kapa HiFi Hotstart Readymix with dye - Kapa Biosystems, South Africa

Kapa Robust Hotstart Readymix with dye - Kapa Biosystems, South Africa

Kapa Universal Ladder - Kapa Biosystems, South Africa

MiSeq Reagent Nano Kit, v2 (300 cycles) - Illumina Inc., USA

MiSeq Reagent Nano Kit, v2 (500 cycles) - Illumina Inc., USA

MiSeq Reagent Kit v2 (500cycle) - Illumina Inc., USA

MiSeq Reagent Kit v3 (150cycle) - Illumina Inc., USA

MiSeq Reagent Kit v3 (600cycle) - Illumina Inc., USA

Nextera Expanded Exome Library Preparation Kit - Illumina Inc., USA

SALSA MLPA P083 *CDH1* probemix (vC1) - MRC-Holland, Netherlands

Tris-ultrapure - Applichem, USA

TruSeq Exome Enrichment Library Preparation Kit - Illumina Inc., USA

Tween-20 - Sigma-Aldrich, USA

2.0.2 Equipment

ABI 3730xl DNA Analyser - Thermo Fisher, USA

Agilent 2100 Bioanalyzer - Agilent Technologies, USA

GelDoc Transilluminator - BioRad, USA

Illumina HiSeq 2000 - Illumina Inc., USA

Illumina MiSeq - Illumina Inc., USA

Milli-Q Ultrapure Water Purification System - Millipore, USA

Nanodrop ND-1000 Spectrophotometer - Nanodrop Technologies, USA

QubitTM 2.0 Fluorometer - Invitrogen, USA

2.0.3 Software

4peaks (v.1.7.2) - http://nucleobytes.com/4peaks/

Bioanalyzer 2100 (v1.0) Expert Software - Agilent Technologies, USA

Coffalyser - MRC-Holland, Netherlands

Integrative Genomics Viewer (IGV),

http://software.broadinstitute.org/software/igv/

QuantityOne (v.4.6.5), BioRad, USA

R (v3.3.2), https://www.r-project.org/

2.0.4 Online tools

BLAST 2 Sequences - http://www.ncbi.nlm.nih.gov/tools/primer-blast/

CBioPortal - Gao et al. (2013), http://www.cbioportal.org/

PrimerBlast - NCBI, http://www.ncbi.nlm.nih.gov/

Primer3 - Untergasser et al. (2012), http://www.simgene.com/Primer3/

KEGG - Kanehisa Laboratories, Japan, http://www.kegg.jp/

Reactome - Croft et al. (2013), http://www.reactome.org/

2.0.5 Bioinformatic tools

AnnoVar (v.2015Dec14) - http://annovar.openbioinformatics.org/

Burrows-Wheeler Aligner (v0.7.7) - http://bio-bwa.sourceforge.net/

FastQC (v0.10.1) - http://bioinformatics.babraham.ac.uk/projects/fastqc/

Genome Analysis Toolkit (v3.2-2) - https://software.broadinstitute.org/gatk/

Picard (v1.109) - https://broadinstitute.github.io/picard/

SNPEff (v4.1) - http://snpeff.sourceforge.net/SnpEff.html

SNPSift (v4.1) - http://snpeff.sourceforge.net/SnpSift.html

Trimmomatic (v0.33) - https://github.com/timflutre/trimmomatic

2.0.6 *In silico* prediction tools

Condel - http://bg.upf.edu/fannsdb/

SIFT - http://sift.jcvi.org/

Polymorphism Phenotyping v2 (PolyPhen2) -

http://genetics.bwh.harvard.edu/pph2/

MutationAssessor (Release 3) - http://mutationassessor.org/r3/

Functional Analysis through Hidden Markov Models (v2.3) (FatHMM) -

http://fathmm.biocompute.org.uk/

MaxEntScan -

http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html

NNSplice - http://www.fruitfly.org/seq_tools/splice.html

Human Splicing Finder (v3.0) - http://www.umd.be/HSF3/

2.0.7 R packages

ggplot2 (v2.2.1) - (Wickham, 2016)

SNPassoc (v1.9.2) - (González et al., 2007)

2.1 General laboratory protocols

2.1.1 DNA quantification

DNA and PCR products were quantified using either the NanoDrop ND-1000 Spectrophotometer (Nanodrop Technologies) or Qubit 2.0 Fluorometer (Thermo Fisher). The Nanodrop utilises UV absorbance and compares the ratio of absorbance between 260 nm and 280 nm (260/280) to quantify DNA. In contrast, the Qubit uses fluorescence-based dyes to bind and quantify DNA.

2.1.1.1 NanoDrop

Tris-EDTA (TE) or mqH_20 was used as a blank reading and to calibrate the instrument before each use. Samples were quantified by loading 2 μL of sample onto the NanoDrop pedestal and reading the absorbance. The purity of the DNA was assessed by the A260/280 ratio.

2.1.1.2 Qubit

Samples were quantified using a Qubit dsDNA High Sensitivity Assay Kit and assay tubes (Thermo Fisher). Qubit working solution was prepared by diluting Qubit reagent 1:200 in Qubit buffer. Qubit standards were prepared by mixing 190 μL of working solution with 10 μL of each standard in separate assay tubes. 2 μL of the sequencing library was mixed with 198 μL of working solution. All tubes were left to incubate at room temperature for 2 min. The Qubit Fluorometer was calibrated using the Qubit standards when samples were read.

2.1.2 Gel electrophoresis

2% agarose gels were used for resolution of PCR products. The gels were prepared by dissolving agarose powder (HydraGene) in 1 x Tris acetic acid EDTA (TAE) buffer containing ethidium bromide (10 mg/mL). Gels were submerged in 1 x TAE buffer containing ethidium bromide (10 mg/mL). 4 μL of PCR products were mixed with 1 μL of 6 x DNA loading dye on a piece of parafilm and loaded into the gel. 4 μL Kapa Universal Ladder (Kapa Biosystems) was loaded as a size marker in each gel. Gels were run for up to 40 min at 90 V. The products were visualised and photographed under UV light using the GelDoc Transilluminator (BioRad) with QuantityOne software (v.4.6.5).

2.1.3 Sample purification

2.1.3.1 Ethanol precipitation

PCR products were resuspended in 5 x their total volume of 100% ethanol with 0.3 M sodium acetate, and transferred to a spin column and collection tube. Samples were centrifuged for 1 min at 5000 x g and the supernate discarded. The DNA was cleaned in two wash steps by adding 200 μL 70% ethanol and centrifuging at 5000 x g for 1 min. Columns were then centrifuged at 14000 x g for 3 min to dry the column, and transferred to a new collection tube. 50 μL of warmed elution buffer (EB) was added to the column and left at room temperature for 5 min. Finally, the column was centrifuged at 5000 x g for 1 min. The eluted DNA was stored at 4°C until used.

2.1.3.2 Bead clean-up

DNA samples and sequencing libraries were cleaned with Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol. Briefly, an aliquot of beads was taken and left to warm to room temperature for 30 min. Samples were resuspended in 1.8 x volume of Ampure beads, incubated at room temperature for 15 min, and placed upon a magnetic stand until all the beads had aggregated to one point. The supernatant was carefully removed and discarded. The beads were washed twice in 200 μL of 70% ethanol and left to air dry for 15-20 min. The samples were taken from the magnetic stand and the DNA was resuspended in 40 μL of EB and transferred to a new tube.

2.1.4 Primer design

Primer sets were designed using the open source primer design software Primer3 (http://simgene.com/Primer3) and checked for off target binding using the online tool BLAST 2 Sequences (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). When designing primers, consideration was given to the size, melting temperature, and formation of hairpins. All primers were ordered from Integrated DNA Technologies (IDT. Ltd., Singapore) and reconstituted in 1 x TE to a stock concentration of 100 μM. Aliquots of primer stocks were further diluted with dd$H_2$0 to a working solution of 10 μM. Primers were kept at -20 °C until use.

2.1.5 Sanger sequencing

Purified PCR products were sent for sequencing at the Genetic Analysis Services (GAS; Department of Anatomy, University of Otago). Products were

sequenced using a capillary ABI 3730xl DNA Analyser (Thermo Fisher). Samples were prepared for sequencing by combining 1 ng/100 bp/5 μL of PCR product with 3.2 pM/5 μL in a total volume of 5 μL.

2.2 Māori case-control study

2.2.1 Cases

All New Zealand Māori diagnosed with gastric cancer (ICD10 C16), based on histology reports sent to the New Zealand cancer registry (NZCR) between 1 February 2009 and 31 October 2013, were eligible for inclusion in this study. As the time of registration for gastric cancer can be up to one year post diagnosis, an additional method for identifying cases was used. For the identification of additional controls, the pathology reports for all cancer diagnosis that were sent to the New Zealand Cancer Registry (NZCR) for confirmation, were also sent to the study centre, where gastric cancer cases were identified. Ethnicity of each case was identified from the NZCR data that assigned Māori ethnicity if the person had self-identified as Māori on a previous health record. The clinician named on the pathology report was contacted via letter or phone call for permission to contact the patient. Follow up reminder letters were faxed to the clinician within a two week period. When consent and contact details were provided, each case was contacted by post. If no reply was received, two reminder letters were sent.

2.2.2 Controls

Controls were block samples from the New Zealand electoral roll based on the age of cases in 5-year age bands. In New Zealand, registration on the electoral roll is compulsory for all people age 18 years and over. Māori can choose to enrol on either the 'General' or 'Māori' electoral roll. All those who enrol on the general electoral role are asked to self-identify whether they are Māori of Māori descent. Māori controls were randomly chosen in equal numbers from both rolls using the most recent 2008 and 2010 electoral rolls. Due to a large number of non-responders, an additional database of Māori population-based controls was used. A control cohort which had been established using the same methods, whose recruitment period had ended approximately one year before that of the current study, and whom more recent contact details were available, were contacted. Controls were sampled from this additional database, including both those who had, and had not, consented to take part in the earlier study.

2.2.3 Exposure information

Consenting study participants were given the option of completing a questionnaire with a trained interviewer face-to-face, returning it by mail, or completing it over the telephone. The questionnaire comprised of sections on childhood socioeconomic demographics, household crowding, parent and sibling cancer history, occupational history, and life course exposures to health behaviours, such as smoking and exposure to second hand smoke. Age was defined as the age of diagnosis for cases and the time of interview for controls.

For current exposures, both cases and controls were asked to report their

lifestyles for the past year including exercise frequency, alcohol intake, as well as dietary intakes of red meat, white meat, fish and dried/salty food, and servings of fruit and vegetables per week. Participants were also asked their height and weight from which body mass index (BMI) was calculated. Questions regarding diagnosis of diabetes, having ever been tested for *H. pylori*, and diagnosed dyspepsia were also included.

2.2.4 Case pathology

Cases were matched to their National Health Index (NHI) number, corresponding pathology reports, and information reported to the cancer registry. Pathology reports were reviewed for information regarding tumour pathology and details related to previous variant screening, prophylactic gastroectomies and endoscopic screening. Tumour histology classified using the WHO classification (Bosman et al., 2010) was aligned with the Lauren classification (Lauren, 1965).

2.2.5 Collection of blood samples

For those completing the interview face-to-face, participants were given the opportunity to accompany the interviewer to the nearest medical laboratory for collection of a blood sample. For those unable to do this at the time of interview, the laboratory staff came to the participants home to collect a sample, or the collection kit was left with the participant to organise collection of a sample at their convenience. Once blood samples were taken, they were couriered directly to Canterbury Health Laboratories (Christchurch, New Zealand). At the conclusion of the study collection period, blood samples were sent to relevant institutions and testing facilities for assessment. Blood

samples were used to assess a variety of gastric cancer associated risk factors, including *H. pylori* serology, heritable genetic alterations, antioxidant levels, and trace elements of heavy metals. The latter information is provided for completeness; however, only the heritable genetic alterations are being analysed as a part of this thesis.

2.2.6 Ethical approval

The study was granted ethics approval by the Multi-region Ethics Committee (ref: MEC/08/08/102/AM03).

2.2.7 DNA extractions

Blood samples were kept frozen at -20°C for short-term storage and -80°C for long-term storage. When a sample was required, the sample was thawed at room temperature and an aliquot was taken. The aliquot of blood was kept on ice until used.

Germline DNA was extracted from blood samples using a GF–1 Nucleic Acid Extraction Kit (Vivantis) according to the manufacturer's protocol. Briefly, 200 μL of whole blood was mixed with 200 μL of buffer and 20 μL of proteinase K, and incubated at 65°C for 15 min. Following this, 200 μL of absolute ethanol was added and mixed, then the entirety transferred to a GF-1 spin column and collection tube. The spin column and collection tube were centrifuged at 5,000 x g for 1 min, capturing DNA in the spin column. The DNA was washed by adding 500 μL of wash buffer 1 to the spin column and centrifuging at 5000 x g for 1 min. The DNA was then washed a second time by adding 500 μL of wash buffer 2 to the spin column, and centrifuging at 5000 x g for 1 min. The column

was washed and dried in a final wash step by adding 500 μL of wash buffer 2 to the spin column and centrifuging at 15,000 x g for 3 min. Finally, the spin column was placed into a new collection tube and the DNA was eluted in 100 μL of elution buffer (EB) and centrifuged at 5,000 x g for 2 min. Eluted DNA was stored at 4°C for short-term storage and -20°C for long-term storage.

2.3 *CDH1* amplicon sequencing library design

To sequence all coding regions and the proximal promoter of the *CDH1* gene, I adapted a two-step PCR strategy used by New Zealand Genomics Limited (NZGL; Dunedin, New Zealand) and Illumina, that uses two PCR reactions to create a DNA library suited for sequencing on an Illumina MiSeq.

Targeted PCR reactions amplify the regions of interest in separate reactions. These PCR reactions are then pooled and amplified in a second reaction using primers designed to substitute for Illumina sequencing adapters. Both the forward and reverse primers used in the second PCR reaction were designed to contain unique indices, the sequences required for binding to an Illumina flow cell, and the sequences required for the binding of Illumina MiSeq extension primers. Each sample was amplified using a unique combination of indices. These steps are summarised in Figure 2.1.

2.3.1 PCR 1: Amplicon specific primer design

Primers were designed to amplify each of the 16 exons and the proximal promoter of *CDH1*. Each primer was designed with a targeted sequence and 18 bp of non-specific sequence. The targeted sequence was designed to be complementary to the target DNA, while the non-specific sequence was

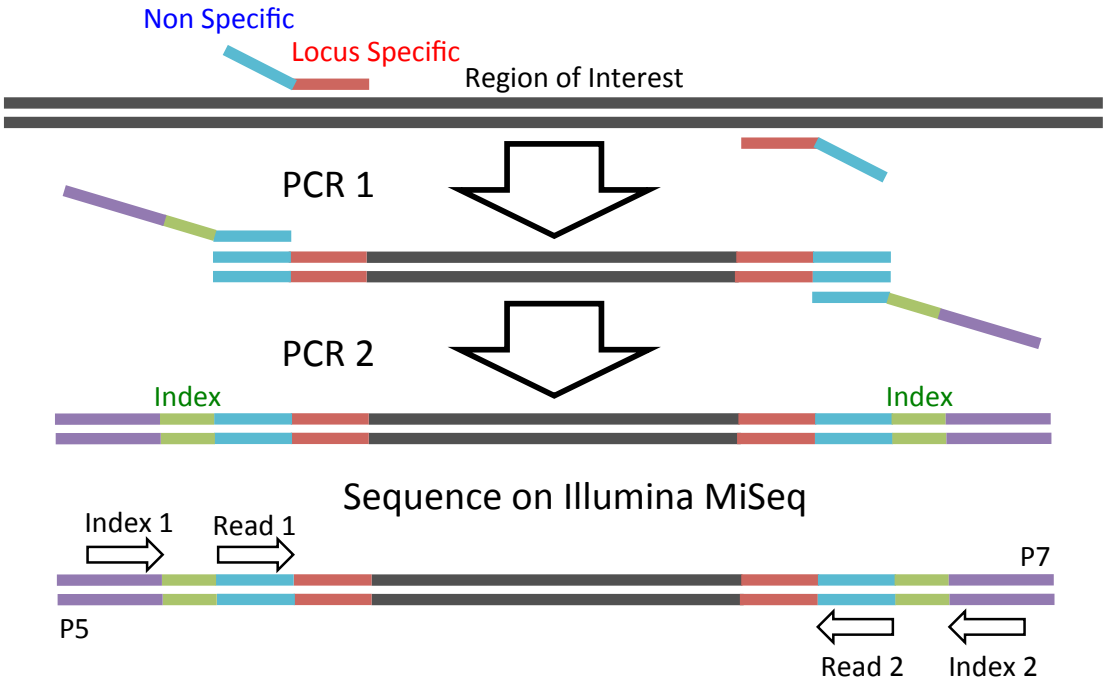


Figure 2.1: Two-step PCR strategy for creating NGS amplicon sequencing libraries. A schematic diagram of the two-step PCR strategy for generating sequencing libraries. Locus specific primers amplify region of interest in the first PCR reaction. PCR products are amplified in a second PCR reaction using primers which substitute for Illumina sequencing adapters. The primers used in the second PCR reaction contain unique indices and binding sites for sequencing primers used by the Illumina MiSeq.

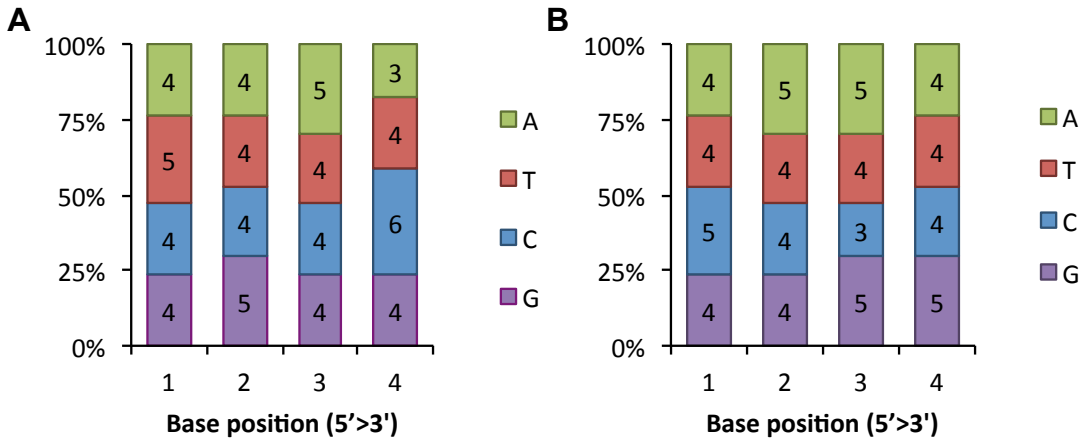


Figure 2.2: Base diversity of first four sequenced bases. Diversity of the first four bases sequenced for custom *CDH1* amplicon sequencing libraries for the (A) forward sequencing primers, and (B) reverse sequencing primers.

designed to act as the binding site for primers used in the second PCR reaction. Each primer pair was designed to produce an amplicon less than 500 bp in length to fit within the capabilities of 500-cycle MiSeq reagent kits.

Fluorescence calibration for base calling occurs during the first four bases of sequence during a MiSeq run. Consideration was given to the distribution of bases in these positions to enable the best sequence quality possible. There are two channels that need to fluoresce at each of these sites: G/T and A/C.

In total, 17 primer pairs were required to cover the coding regions and proximal promoter of the *CDH1* gene. Primers were designed to provide the best possible distribution of bases in the first four read positions. Amplicon specific primers are displayed in Appendix A.1. Figure 2.2 shows the distribution of bases in the forward and reverse primers in the first four read positions in each direction.

2.3.2 PCR 2 : Adapter primer design

The primers used in the second reaction were designed as a substitute for the adapters used in standard TruSeq library preparations. These primers were based on the standard sequences used for Illumina TruSeq custom amplicon sequencing, and contain the necessary primer binding sites for pair-end dual-index sequencing.

The 3′ end of each adapter primer was complementary to the lagging strand of the non-specific overhang sequence introduced to the products during the first round of PCR amplification. This meant that any residual primers left after the first PCR reactions would not bind and inhibit the primers used in the second PCR reaction. Primers contain the sequences required for sequencing primers to bind during the sequencing reaction and for flow cell binding.

The same 6 bp indices used in the TruSeq and short RNA sequencing kit designed by Illumina were used for both the i7 (reverse) and i5 (forward) adapters. These indices are designed to allow for two miss-incorporations per index and still have enough redundancy to be unique compared to other indices. In total, 20 different forward primers and 30 reverse primers were designed and are listed in Appendix Table A.2 and A.3, respectively.

2.4 Amplicon sequencing and analysis

2.4.1 Amplicon specific PCR

Exonic regions were amplified in 10 μL PCR reactions. As there were large numbers of reactions required to amplify all 17 amplicons, a pre-made Kapa

Hotstart Ready Mix with Dye (Kapa Biosystems) was used to reduce the time required for master mix preparation and handling of samples when visualising products on an agarose gel. For negative controls mqH_20 was substituted for DNA. Amplicon specific PCR reactions were made as specified in Table 2.1.

Reagent	**Volume**
mqH_20	1.8 μL
10 μM forward primer	0.6 μL
10 μM reverse primer	0.6 μL
Kapa Hotstart readymix	5 μL
DNA (20 ng total)	2 μL
Total	10 μL

Table 2.1: Amplicon specific PCR mastermix.

PCRs reactions were amplified on a DNA Engine Thermal Cycler (BioRad). The standard cycling conditions for PCR were are outlined in Table 2.2. Temperature gradient PCRs for each of the primer pair were carried out using control DNA. Annealing temperatures with the strongest product band and absence of multiple banding were selected. Variations in annealing temperature for each primer pair are stated in Appendix Table A.1.

	Time	**Temperature**	**Cycles**
Initial denaturation	3 min	95°C	1 x
Denaturation	14 sec	98°C	35 x
Annealing	14 sec	64-66°C	
Extension	14 sec	72°C	
Final extension	1 min	72°C	1 x

Table 2.2: PCR conditions for amplicon specific PCRs.

After PCR, products were run on a 2% agarose gels to verify successful amplification and absence of contamination. Gels were prepared and run as described in Section 2.1.2. As PCR products from the first reactions had the

same 18 bp sequences on each end, they could be combined and amplified in a second PCR reaction using primers targeting these sequences as priming sites. In a UV treated hood, 2 μL of each amplicon product was combined into sample-specific pools. To remove any residual primers in each reaction, 10 μL of each sample specific pool was purified using Ampure XP beads (as previously described in Section 2.1.3.2).

2.4.2 Adapter PCR

Unique combinations of forward and reverse adapter primers were used in the second round of PCR reactions. Adapter combinations were cycled through between runs to reduce the chance of contamination between sequential runs. PCR reactions were made as described in Table 2.3.

Reagent	**Volume**
mqH_20	1.8 μL
10 μM forward primer	0.6 μL
10 μM reverse primer	0.6 μL
Kapa HiFi hotstart readymix	5 μL
DNA (2 ng total)	2 μL
Total	10 μL

Table 2.3: Amplicon specific PCR mastermix.

PCR conditions were optimised for complete amplification of all 17 amplicon specific PCR products while limiting the number of cycles and required DNA input. The standard cycling conditions for these PCR reactions are shown in Table 2.4.

Post PCR, all amplicon libraries contained a unique index combination, making each sample identifiable. Each reaction was visualised on a 2% agarose gel to confirm adapter PCR amplification (Section 2.1.2). 5 μL of each

	Time	Temperature	Cycles
Initial denaturation	2 min	95°C	1 x
Denaturation	14 sec	98°C	10 x
Annealing	14 sec	68°C	
Extension	14 sec	72°C	
Final extension	1 min	72°C	1 x

Table 2.4: Thermocycler conditions for adapter PCR of amplicon products.

post adapter PCR reaction was pooled into a single tube. 50 μL of the combined sample libraries was purified using Ampure XP beads and the same protocol described in Section 2.1.3.2.

2.4.3 Library preparation

Purified sequencing libraries were quantified using the Qubit. An Agilent 2100 Bioanalyzer (Agilent Technologies) was used to determine the average size of the combined libraries. 500 pg of combined library was loaded into a Agilent high sensitivity DNA chip (Agilent Technologies) and read using the bioanalyzer. The Bioanalyzer 2100 Expert Software (v1.0; Agilent Technologies) was used to determine the average size of each library.

Before sequencing, DNA libraries were denatured and diluted to an appropriate concentration for loading. The library concentration was converted to a nM amount using equation 2.1 and diluted to 4 nM with EB.

$$concentration\ (nM) = \frac{DNA\ conc.\ (ng/ul)\ \times 1 \times 10^6}{average\ size\ (bp)\ \times\ 656.4\ (g/M)} \tag{2.1}$$

Frozen aliquots of 200 μL 1.0 M NaOH were thawed at room temperature. Once thawed, aliquots were diluted in 800 μL of mqH_20 to make 1 mL of 0.2 M NaOH. A fresh dilution of NaOH was made each time a library was made.

DNA was denatured into single strand libraries by combing 5 μL of 4 nM DNA library with 5 μL of freshly diluted 0.2 NaOH. This sample solution was vortexed briefly and left to incubate at room temperature for 5 min. 10 μL of denatured DNA was mixed with 990 μL of chilled HT1 buffer, resulting in a 20 pM denatured library at 1 mM NaOH.

For the best sequence quality, the MiSeq machine requires diversity within the sequencing library, so no single base is over-represented, causing overexposure and making cluster definition difficult. PhiX is a control virus genome added to sequencing libraries to add diversity to sequencing libraries. A 4 nM PhiX library was prepared by combining 2 μL of 10 nM stock PhiX library with 3 μL 10 mM Tris-Cl with 0.1% Tween20. 5 μL of this library was combined with 5 μL of 0.2 M NaOH and vortexed briefly. The template solution was left to denature at room temperature for 5 min. A 20 pM library was made by diluting 10 μL into 990 μL of pre chilled HT1 buffer.

Denatured 20 pM PhiX was added to sequencing runs at various concentrations. In low diversity runs, PhiX made up 20% of the total run. For libraries with high diversity, the proportion of PhiX was reduced to 5% of the total run. In later runs, sequencing runs were shared with sequencing libraries from other projects using the same library preparation strategy. The proportion of PhiX was adjusted to compensate for the diversity of the additional libraries added to the MiSeq run.

Sequencing libraries were loaded at various concentrations depending on the version of reagents being used and complexity of the library. Loading varied between 6-16 pM. Low diversity runs were loaded at a lower concentration with more PhiX, and high diversity runs were loaded at a higher concentration with less PhiX.

2.4.4 MiSeq workflow

2.4.4.1 Reagents

Reagent cartridges were stored at the recommended conditions of -20°C. Before being used, the reagent cartridge was removed from storage and thawed in a room temperature water bath for approximately 2 hours. Once thawed the cartridge was removed from the water bath, mixed by inverting the cartridge multiple times, and visually inspected to make sure all reagents were appropriately thawed. The cartridge was tapped on a bench top to reduce air bubbles in the reagents and stored on ice until used.

2.4.4.2 Sample sheet

The MiSeq system required a sample sheet to control the reaction cycles during sequencing and demultiplex the sequencing libraries post run. Due to the indices in the custom primer adapters used in our library preparations, the standard Illumina Experimental Manager (IEM) used for amplicon sequencing was not suitable to make a sample sheet for our custom sequencing libraries. Instead, a comma-separated values (CSV) formatted sample sheet was made. The CSV sample sheet was designed to imitate sample sheets created by the standard IEM software, and was loaded into the appropriate file location on the MiSeq machine.

2.4.4.3 Sequencing

Prior to the sequencing run, the MiSeq system was rebooted to restore the machine to its full memory. Post reboot, the MiSeq was cleaned with three

cycles of a pre-run standby wash using 0.5% Tween. The flow cell was removed from its storage buffer solution and rinsed rigorously with mqH_20 and 100% ethanol, and then wiped clean. Once dry, the flow cell was loaded into the MiSeq machine.

600 μL of library was loaded into the sample reservoir in the MiSeq reagent cartridge and visually inspected to make sure it had moved to the bottom of the well. The reagent cartridge and incorporation buffer were loaded into the MiSeq machine and the sequencing run was started. Post run the MiSeq was washed with 0.5% tween using the post run wash tray and wash bottle. The wash tray, wash bottle, and flow cell were left in place until the MiSeq was next used.

2.4.5 Quality checks

FastQC is a bioinformatic tool used for checking the quality of raw sequence data generated by NGS. FastQC (version 0.11.2) was used for assessing the quality of the fastq sequencing files. Fastq files were assessed for quality as well as other parameters, including the GC content, N content, sequence length distribution, duplication levels, overrepresented sequences, and Kmer content.

2.4.6 Data processing and analysis

Processing of high throughput NGS data can be split into three major steps: pre-processing, variant calling, and variant evaluation. Each step is necessary for the identification of true genetic variants from a background of sequencing artefacts and errors. A shell script was written in a text editor and was used

for the step-wise processing of the raw fastq sequencing reads into annotated variant call formatted (VCF) files. A schematic diagram of the sequencing pipeline is shown in Figure 2.3 and is described in detail below.

2.4.6.1 Pre-processing

Data produced by high throughput sequencing technology produces raw sequence data that is unable to be processed by evaluation tools immediately. First, the data needs to be processed to prepare it for the variant calling. There are three main steps for this pre-processing; trimming, mapping, and base quality score recalibration (BQSR).

Demultiplexed fastq files were downloaded from the MiSeq and loaded onto the local machine for processing. Fastq files were trimmed using Trimmomatic (Bolger, Lohse, & Usadel, 2014). For trimming, adapters, amplicon specific primers, and leading and trailing low quality bases were removed. A sliding window of 4 bases was used to cut reads when average quality dropped below Q20. Only reads with matching pair end reads were retained for further processing.

Trimmed sequence reads were mapped to the reference human genome (Hg19) using the Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2009). BWA relies on the Burrows-Wheeler transform algorithm for compression of text rich sequences to increase alignment efficiency and reduce memory usage. The BWA software comes with three algorithms: *BWA-backtrack*, *BWA-sw*, and *BWA-mem*. The *BWA-mem* algorithm was selected as the most recent and best aligner for long reads, and was used to produce a SAM formatted file.

Next, Picard tools (v1.109; http://broadinstitute.github.io/picard/) was used

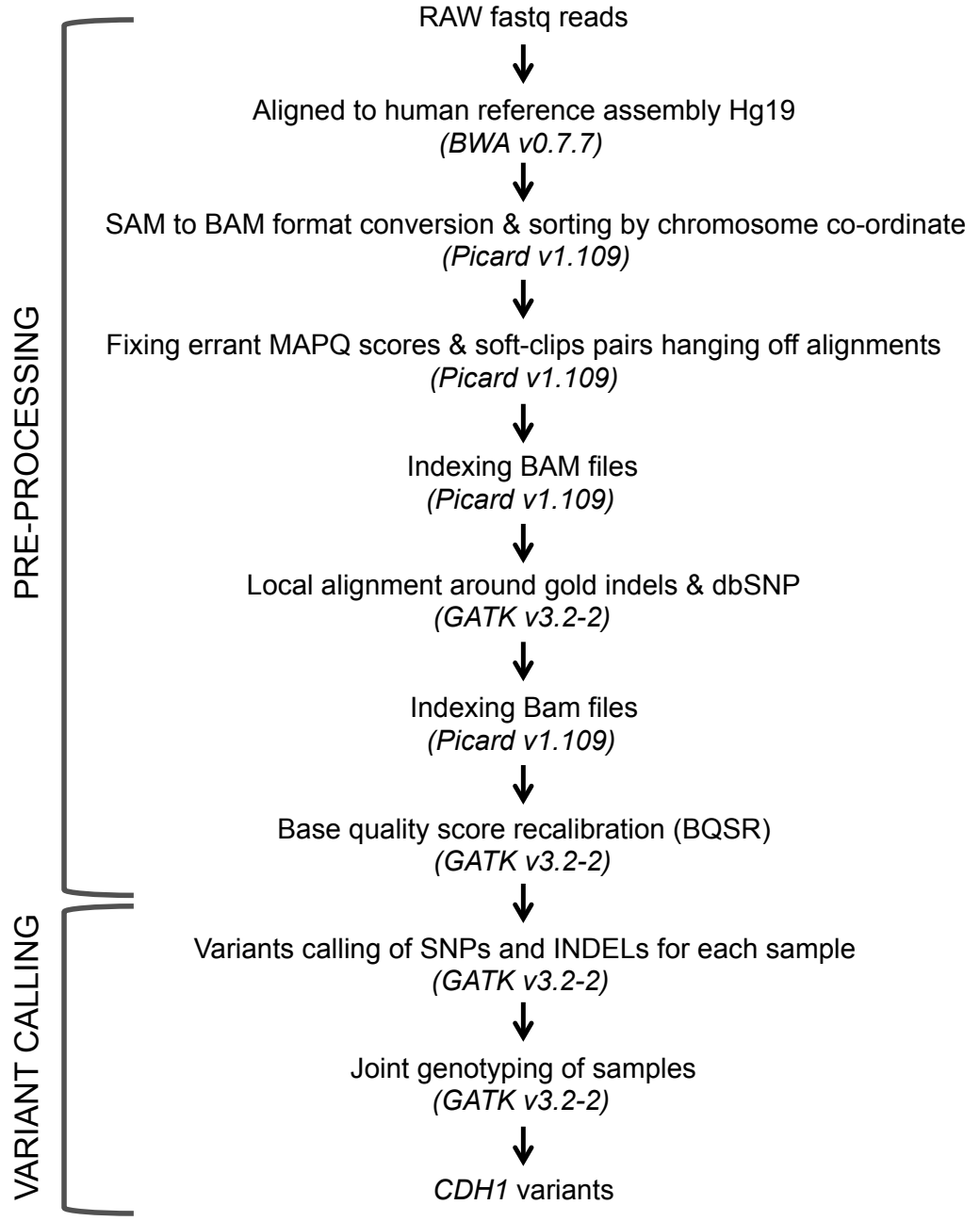


Figure 2.3: Summary of pre-processing and variant calling analysis of *CDH1* amplicon sequencing data workflow. Fastq files for *CDH1* amplicon sequencing libraries were processed to produce analysis ready variant files for annotation and filtering. BQSR, base quality score recalibration; INDELS, insertions/deletions; SNPs, single nucleotide polymorphisms.

to convert the SAM file to a BAM file and sort the reads by genomic coordinates, soft-clip reads beyond the end of reference alignments, verify mate-pair information, and set MAPQ scores for unmapped reads to 0.

Finally, the Genome Analysis Toolkit (GATK) *BaseRecalibrator* was used to apply base quality score recalibration (BQSR) and adjust the quality scores of of each base. BQSR applies machine learning to model per-base estimates of error emitted by the sequencing machine and adjust the quality scores accordingly. A machine learning model was applied to mapped reads based on the data and known variants in the dbSNP and goldindel files to analyse the patterns of covariation in the sequence dataset. The covariation was used in a second pass of the data using the same known sites. Plots were generated to evaluate the recalibration. The recalibration was then applied to the BAM file to recalibrate base quality scores.

2.4.6.2 Variant calling

The GATK *HaplotypeCaller* was run for each BAM file to generate VCF files. A bed file was used to restrict variant calling to the coding exons and splice regions of *CDH1*. GATK *GenotypeGVCFs* was used to merge the VCF files and annotate them with dbSNP (build 137).

2.4.6.3 Variant evaluation

The effect of variants were predicted using SnpEFF (Cingolani et al., 2012). Annotations included information about how gene transcripts were affected and the variant type. Variant calls were further annotated with population frequencies from the Exome Aggregation Consortium (ExAc; (Lek et al.,

2016)), NHLBI Exome Sequencing Porject (ESP6500; http://evs.gs.washington.edu), and 1000 Genomes Project (1000 Genomes, (1000 Genomes Project Consortium, 2010)). Sequence alignment files and variant call files were visually inspected in Integrative Genomics Viewer (IGV, (Robinson et al., 2011)).

FannsDB 1.0 Condel (v2.0) (González-Pérez & López-Bigas, 2011) was used to predict the effect of missense variants. Condel incorporates predictions from MutationAssessor (Reva, Antipin, & Sander, 2011) and FatHMM (Shihab et al., 2013) to produce a single effect score and prediction. Notably, previous versions of Condel incorporated predictions scores from five bioinformatic tools, namely: SIFT (P. C. Ng & Henikoff, 2003), PolyPhen2 (Adzhubei et al., 2010), MutationAssessor (Reva et al., 2011), Ensembl-variation (McLaren et al., 2010), Multivariate Analysis of Protein Polymorphism (MAPP) (Stone & Sidow, 2005) and LogR Pfam E-value (Clifford, Edmonson, Nguyen, & Buetow, 2004). Although the latter tools are not incorporated into Condel v2.0, they are still reported in result outputs and are included in these analyses.

The effect of splice site mutations were predicted using MaxEntScan (Yeo & Burge, 2004), NNSplice (Reese, Eeckman, Kulp, & Haussler, 1997), and Human Splicing Finder 3.0 (HSF) (Desmet et al., 2009).

Average coverage was used as check for successful sequencing of targeted regions. Amplicon coverage was determined from sequence alignment files using GATK's *DepthofCoverage* tool. Coverage plots were generated and visualised in the statistical platform R (v3.3.2; https://www.r-project.org/). Amplicons with a coverage of less than 40 reads were identified and were re-sequenced by NGS or Sanger sequencing.

2.4.7 Validation of variants

DNA was re-extracted from blood samples using the protocol described in Section 2.2.7. New primers were designed and checked for off target binding. Primers used for sequencing validation are listed in Appendix Table A.4. Pathogenic variants were validated in forward and reverse direction using Sanger sequencing.

2.5 Multiplex ligation-dependent probe amplification

Multiplex ligation-dependent probe amplification (MLPA) is a multiplex PCR method used for the detection of abnormal copy numbers of DNA or RNA sequences. MLPA uses a series of paired oligonucleotides, each pair specific for one target. The pair of probes are hybridised against two target sequences immediately adjacent to each other, and ligated together to form a single product. Primer pairs are fluorescently labelled and contain a 'stuffer' sequence of variable length. Ligated products are amplified in a subsequent PCR reaction and separated using capillary electrophoresis. The amount of fluorescence from PCR products is proportional to the amount of target DNA present in the sample, making this technique suitable for quantitative measurements.

2.5.1 MLPA methodology

Structural rearrangement were were analysed in cases without a clear pathogenic *CDH1* variants using a SALSA MLPA P083 CDH1 probemix (vC1; MRC-Holland). These reactions were performed on a BioRad thermocycler

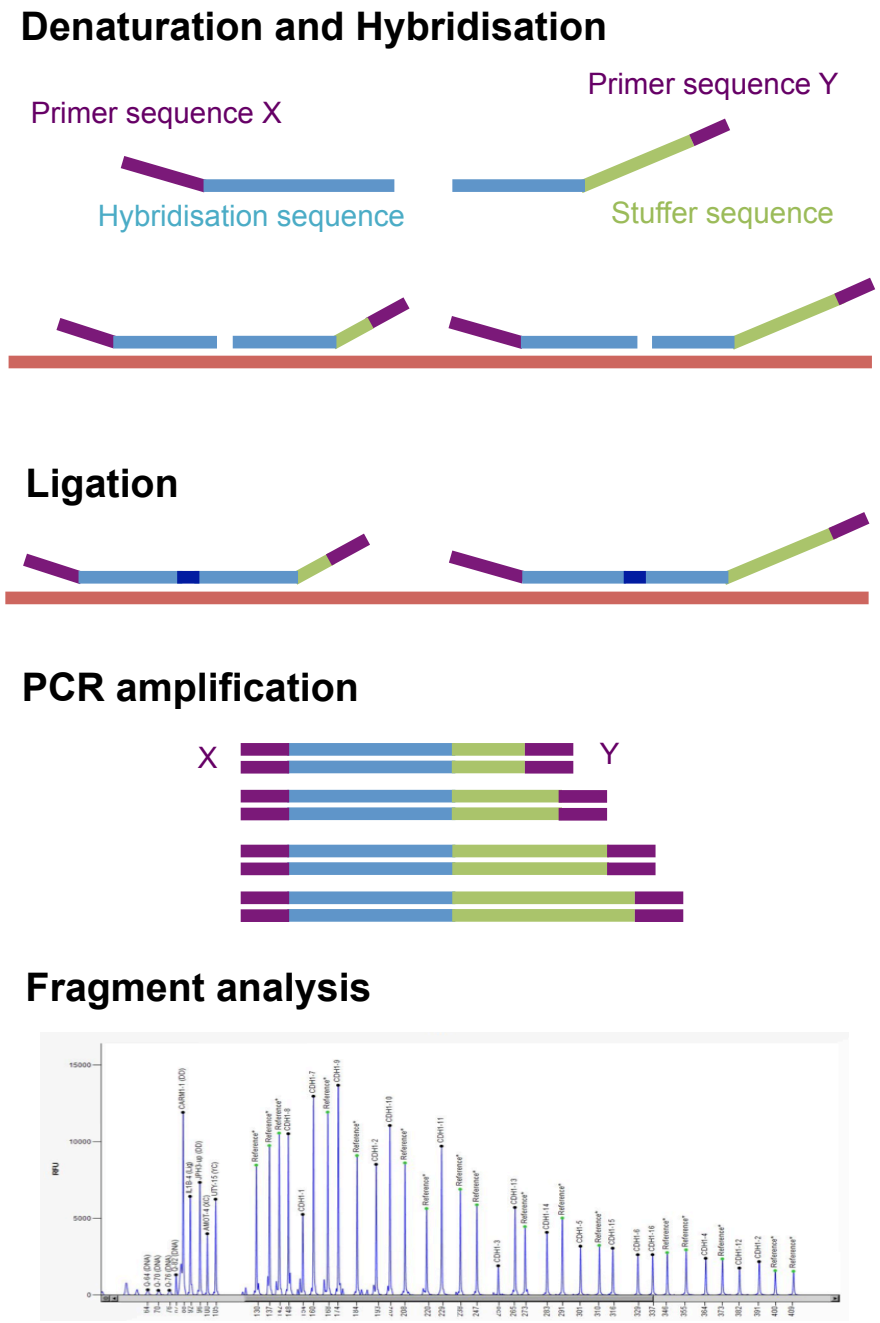


Figure 2.4: Schmatic diagram of MLPA reaction. MLPA reaction showing the hybridisation of probes to denatured DNA, ligation of probes to form single strands, PCR amplification of ligated strands, and fragment analysis of amplified products.

Time	Temperature	Cycles	
Step 1: DNA denaturation	5 min	98°C	
	Pause	25°C	
Step 2: Hybridisation reaction	1 min	95°C	
	18 hours	60°C	
Step 3: Ligation reaction	Pause	54°C	
	15 min	54°C	
	5 min	98°C	
	Pause	20°C	
Step 4: PCR reaction	30 sec	95°C	35 cycles
	30 sec	60°C	
	60 sec	72°C	
	20 min	72°C	
	Pause	15°C	

Table 2.5: Thermocycler programme for MLPA reactions.

with heated lid (105°C) in thin-walled PCR tubes. Samples were left on the thermocycler to maintain the required temperature when reagents were being added. Thermocycler conditions are displayed in Table 2.5.

80 ng of DNA in 5 μL was denatured at 95°C for 5 min, then cooled to 25°C. 1.5 μL MLPA buffer and 1.5 μL probemix was added to each denatured DNA sample and mixed gently. Samples were heated to 95°C for one min, incubated at 60°C for 18 hours to allow probes to hybridise, and cooled to 54°C. 3 μL ligase buffer B, 25 μL of mqH_2O and 1 μL of ligase were added to each sample and mixed gently. Samples were incubated for 15 min at 54°C , then 98°C for 5 min. After ligation, 1 μL SALSA PCR-primers, 2 μL SALSA enzyme dilution buffer, 15.75 μL mqH_2O, 0.25 μL SALSA Polymerase and 5 μL of Polymerase were added to each sample and the PCR reaction was started (PCR conditions, 35 cycles: 30 sec 95°C; 30 sec 60°C; 60 sec 72°C). The reaction ended with a 20 min incubation at 72°C. Post PCR, amplified products were stored in the dark at 4°C until fragment analysis.

2.5.2 Analysis

10 μL of post-PCR samples were sent to the Genetic Analysis Service for genotyping. Samples were genotyped on an ABI3730XL using LIZ-600 size standard. The output for each MLPA reaction displayed a series of peaks which corresponded to the amount of amplified ligated probe present for each target region. Additionally, MLPA probe sets contain various control fragments used to detect problems in the MLPA reactions. Internal quality control fragments include: one nucleotide benchmark probe, used to compare other control fragments to; four quantify fragments (Q-fragments), used to measure whether sufficient DNA was added and ligation was successful; two Denaturation fragments (D-fragments), used to measure whether DNA fragments denature properly; and gender specific X and Y fragments, used to identify sample gender to control for sample swapping.

Peak heights were analysed using the MRC-Holland software Coffalyser (MRC-Holland). Standard presets for the P083 *CDH1* (vC1) probemix were used to identify peaks. Within each sample, targeted probe peaks were compared to reference probe peaks to determine a relative probe ratio. This relative probe ratio was compared to the average relative probe ratio in the reference samples to determine a final ratio, which is known as the Dosage Quotient (DQ). Copy number status was considered normal for probes with a DQ greater than 0.7 and less than 1.3.

2.6 Chilean gastric cancer cohort

2.6.1 Study cohort

Cases were identified at various medical institutions in Chile between 2006 and 2017. Patients were consented and blood samples were taken at referring institutions. Germline DNA was extracted and sent to our laboratory for sequencing. The probands clinical details and family history were collected and provided by the referring institution.

2.6.2 Amplicon sequencing and validation of rare variants

Germline *CDH1* was sequenced by a combination of NGS and Sanger sequencing. NGS and subsequent analysis was performed as previously described in Section 2.4 and 2.4.6, respectively. Sanger sequencing was performed as previously described in Section 2.1.5. Deleterious and rare missense variants and were validated by Sanger sequencing as previously described in Section 2.4.7.

2.7 Whole-exome sequencing

2.7.1 Case selection

Gastric cancer patients who had previously tested negative for pathogenic *CDH1* variants were considered for this study. Cases were evaluated for their tumour histology, age at the time of diagnosis, and family history of cancer.

Cases with the most striking family history of cancer, or an unusual early-onset of diffuse gastric cancer, were selected for further sequencing.

2.7.2 Library preparation and sequencing

DNA samples were checked for integrity by agarose gel electrophoresis and the nanodrop. Samples were quantified using the Qubit. Whole-exome capture and sequencing was contracted to NZGL. Libraries were prepared and sequenced in two separate batches: eleven samples were prepared with the TruSeq Exome Enrichment library preparation kit (Illumina Inc.) and 2 x 100 pair-end sequenced on one lane of HiSeq 2000; and three samples were prepared with the Nextera Expanded Exome library preparation kit (Illumina Inc.) and 2 x 100 pair-end sequenced on one lane of HiSeq 2000. Sequence reads were returned as FASTQ formated files for processing and analysis.

2.7.3 Data processing and analysis

Similar to the amplicon sequencing data produced by the Illumina MiSeq, the data produced by the Illumina HiSeq requires bioinformatic processing before variants can be identified and evaluated. The processing of whole-exome sequencing data is more complex than amplicon sequencing data and requires more computational power and processing steps to recalibrate the quality of variant calls.

Pre-processing and variant calling of whole-exome sequencing data was performed with the assistance of Dr. David Markie (University of Otago, New Zealand). The pipeline showing the pre-processing and variant calling is summarised in Figure 2.5. Notably, many of these steps are the same as those

used in the processing of the amplicon sequencing data (Section 2.4.6). These steps are briefly outlined, while the steps which have been added or changed to account for the complexity of whole-exome data are described in greater detail.

2.7.3.1 Pre-processing

Fastq files were trimmed using Trimmomatic (Bolger et al., 2014). Adapters, amplicon specific primers, and leading and trailing low quality bases were removed. A sliding window of 4 bases was used to cut reads when average quality dropped below Q20 and short reads (< 40 bases) were removed. Only reads with matching pair end reads were retained for further processing.

BWA-mem and was used to map reads to the reference human genome and produce SAM formatted files. Picard tools was used to convert the SAM file to a BAM file and sort the reads by genomic coordinates, soft-clip reads beyond the end of reference alignments, verify mate-pair information, and set MAPQ scores for unmapped reads to 0. In an additional step compared to the processing of the targeted amplicon sequencing data, duplicate reads were removed using Picard tools. BQSR was applied to adjust the quality scores of of each base and plots were generated to evaluate the recalibration.

2.7.3.2 Variant calling

The GATK *HaplotypeCaller* was used to generate VCF files from each BAM file. Joint genotyping was done using GATK *GenotypeGVCF* in conjunction with an additional 394 unrelated exome samples, which leveraged information from all samples when calling the genotypes for the individual samples. These

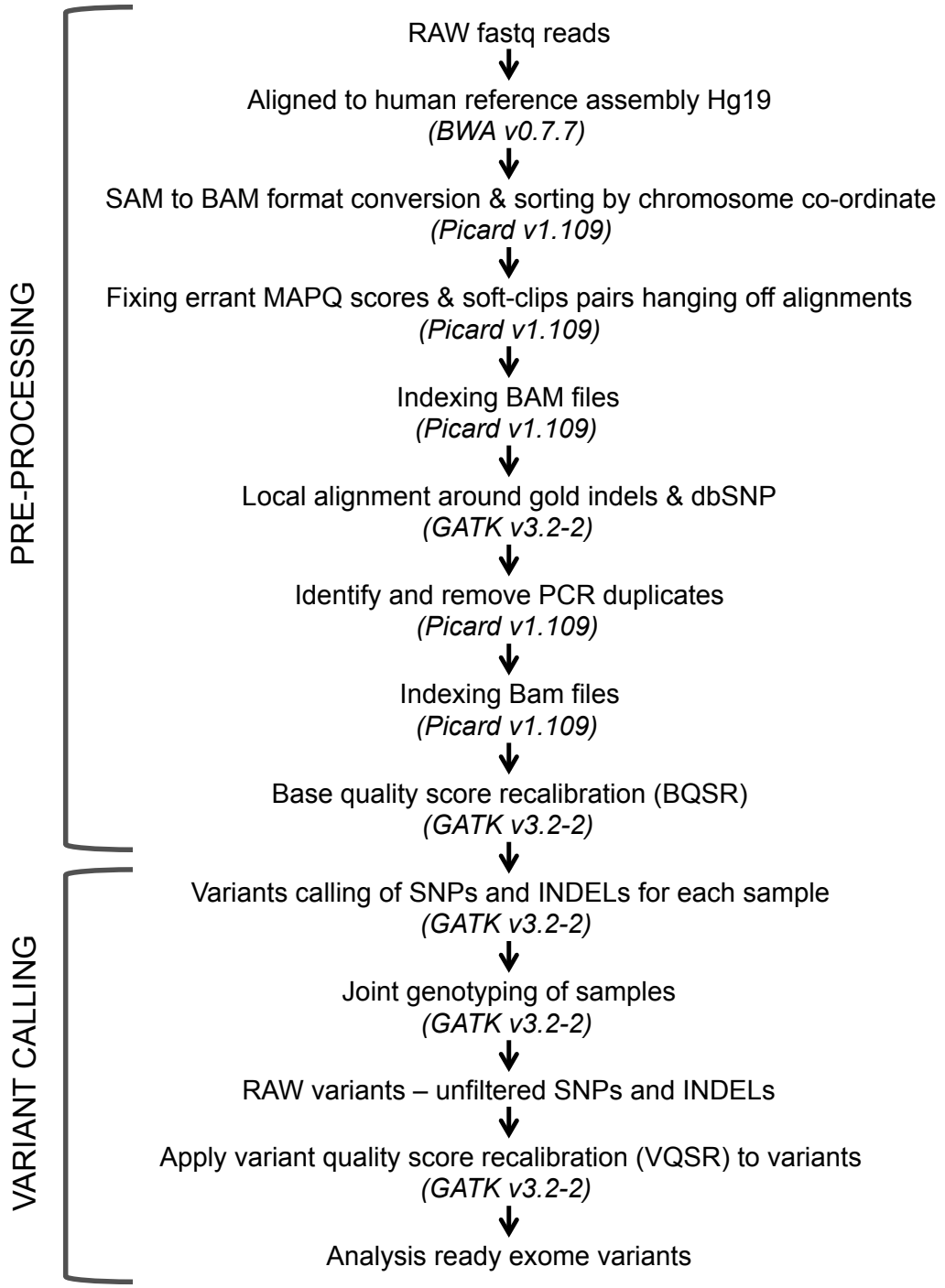


Figure 2.5: Summary of pre-processing and variant calling analysis for whole-exome sequencing data. Fastq files were processed to produce analysis ready variants for annotation and filtering.

additional exomes were produced using the same sequencing and alignment procedures.

Variant quality score recalibration (VQSR) was performed on the 14 exomes in this study in combination with 394 exomes from unrelated subjects using GATK *VariantRecalibrator* and *ApplyRecalibration*. VQSR is a statistical soft filtering that is applied to the raw variant calls to filter false positives and improve the accuracy of base calling by determining the probability that a variant is real. An adaptive error model is built using a training set of variants verified in the 1000 Genomes database to discover the properties of these variants, including things such as strand bias, read position, mapping quality, and haplotype score. This model is then applied to known and novel variation discovered in the call set of interest to estimate the probability that each variant in the call set is a true genetic variant or a machine/alignment artefact. The filtering criteria are derived from the data itself and therefore it is recommended that at least 30 exomes are used to empower VQSR. GATK *ApplyRecalibration* uses a sensitivity thresholds to bin variants into tranches, that establish thresholds indicating different levels of sensitivity relative to the training sets. The higher tranches specify a more accurate call set than the lower tranches, which are more sensitive and therefore contain more true variants but also including more false positives. The values assigned to each variant by VQSR acts as an estimate of the accuracy of that call by assigning a relative ranking to help identify true variants.

2.7.3.3 Variant evaluation

ANNOVAR (v.2015Dec14; http://annovar.openbioinformatics.org/; Wang et al., 2010) was used for the annotation of VCF files. Annotated variant files

were then loaded into R and filtered successively for quality, location, and frequency in population databases. *In silico* predictions for the missense variants were performed using FannsDB 1.0 Condel (v2.0) (González-Pérez & López-Bigas, 2011).

2.7.4 Validation of variants identified by whole exome sequencing

Prioritised variants were validated using Sanger sequencing. Primers were designed as previously described in Section 2.1.4 and are shown in Appendix Table A.5.

2.8 Declaration

The following people contributed to work described within this thesis.

Chapter 3

Blood samples, survey information, and clinical data was collected by research nurse and clinical staff. All experimental work and analysis was completed by the candidate.

Chapter 4

Clinical information and DNA samples were collected by numerous research nurse and clinical staff from referring institutions. All experimental work and analysis was completed by the candidate.

Chapter 5

Library preparation and whole-exome sequencing was contracted to NZGL. Processing of raw reads was completed with the assistance of Dr. David

Markie (University of Otago, Dunedin). All other analysis and experimental work was completed by the candidate.

Chapter 3

Germline *CDH1* Variants as a Cause of Gastric Cancer in Māori

3.1 Introduction

While New Zealand as a whole is considered to be a country with a low incidence of gastric cancer, New Zealand Māori have a notably elevated incidence of gastric cancer compared to non-Māori, and are one of the few populations worldwide with a higher incidence of the diffuse form of the disease. Multiple Māori families have been diagnosed with pathogenic germline *CDH1* variants and the cancer syndrome HDGC, which predisposes to early-onset diffuse gastric cancer. As part of a study examining the known risk factors of gastric cancer in Māori, next-generation amplicon sequencing, MLPA, and Sanger sequencing were used to search for variants in germline *CDH1*. Our aim was to determine the prevalence of pathogenic *CDH1* variants in the Māori gastric cancer population, and understand what impact this is having on the incidence of gastric cancer in New Zealand.

3.1.1 The Māori population

Originating from eastern Polynesia, Māori settlers migrated to New Zealand in several waves of canoe voyage between 1200 and 1300 CE (Oliver, 1981). For many centuries the Māori population lived as an isolated group, developing characteristic cultures, language, and mythology. Māori were the only ethnic population in New Zealand until Europeans first visited New Zealand in 1642, with subsequent settlements beginning in the early 1800's (Oliver, 1981).

Modern day New Zealand has a population of more than 4.5 million people from diverse ethnic backgrounds. Although people can identify as multiple ethnicities, 14.9% of the population (approximately 600,000 people) identified as Māori in the most recent census (Statistics New Zealand, 2013a). Other

major ethnic groups were European (74.0%), Asian (11.1%), and Pasifika (7.4%) (Statistics New Zealand, 2013a). Māori are a comparatively young population with a median age of just 23.9 years, an age much lower than that of the non-Māori population (encompassing all other ethnic groups) of 38.0 years. Geographically, 86.0% of Māori live in the North Island of New Zealand, with 23.8% of all Māori concentrated in New Zealand's largest urban centre, Auckland (Statistics New Zealand, 2013b).

Historically, the Māori population has experienced poorer health outcomes than the general population (R. Harris et al., 2006a). Māori are overrepresented in almost all poor health indicators, including high rates of cot death, glue ear, teenage pregnancy, youth suicide, self injury, diabetes, stroke, pneumonia, influenza, and mental health (R. Harris et al., 2006b; Ministry of Health, 2015). The reasons for the disparities in these indicators and the overall poor health of Māori involve a complex mix of socioeconomic, environmental, and lifestyle factors. Understanding these complexities and identifying ways in which they can be addressed is a primary concern for health research in New Zealand.

3.1.2 Cancer in New Zealand

Cancer is a substantial burden on the New Zealand health care system. In 2013, the New Zealand Cancer Registry received 22,166 new cancer registrations for nearly 90 different cancer sites, an incidence of 335.5 new registrations per 100,000 population (Ministry of Health, 2016b). Prostate cancer was the most common registration type, making up 14.1% of the total cancer burden. This was followed closely by colorectal cancer (13.9%), breast cancer (13.7%), and melanoma (10.7%) (Ministry of Health, 2016b). In 2013 9,063 people died as the result of cancer, accounting for approximately one

third of all deaths in New Zealand that year (Ministry of Health, 2016b). Cancer related deaths were most commonly a result of lung cancer (18.3%), colorectal cancer (13.8%), and prostate cancer (7.1%) (Ministry of Health, 2016b). The age-standardised mortality rate was 122.8 deaths per 100,000 population (Ministry of Health, 2016b).

When considering Māori and non-Māori as two separate groups, Māori experience disproportionately high cancer registration and mortality rates. In 2013, the cancer registration rate was 1.3 times higher for Māori compared to non-Māori (418.9 versus 328.2 per 100,000 population) (Ministry of Health, 2016b). Similar differences were seen in terms of mortality, for which the age-standardised mortality rate was 1.7 times higher for Māori compared to non-Māori (197.9 versus 116.0 per 100,000 population) (Ministry of Health, 2016b).

3.1.2.1 Gastric cancer

In 2013, the world age-standardised gastric cancer incidence in developed countries was 14.4 per 100,000 population, almost three times the New Zealand incidence of 5.4 per 100,000 population (Ferlay et al., 2015; Ministry of Health, 2016b). In line with international trends, the overall incidence of gastric cancer in New Zealand has fallen in recent decades (Ferlay et al., 2015; Ministry of Health, 2016b). Since the 1980's, the age-standardised incidence of gastric cancer in New Zealand has approximately halved (Figure 3.1). A similar trend is seen for the age-standardised mortality rate, which has also halved since the 1980's (Ministry of Health, 2016b). However, despite this drop in the mortality rate, the survival outcomes for gastric cancer patients in New Zealand are still very poor. The 5-year survival for gastric cancer patients

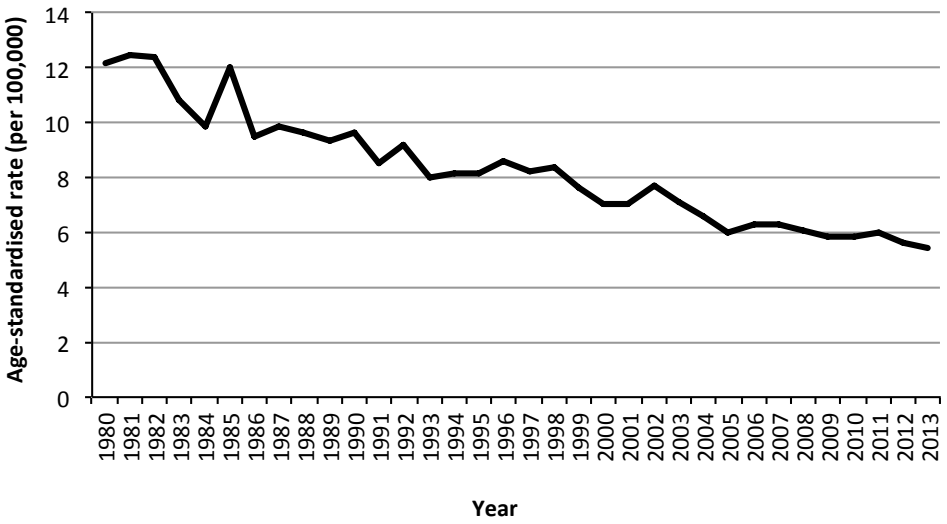


Figure 3.1: Age-standardised gastric cancer registration rate in New Zealand, 1980-2013. Age-standardised rate per 100,000 population; standardised to the WHO world standard population. Data source: New Zealand Cancer Registry, Cancer: New Registrations and Deaths.

in New Zealand is just 20%, much lower than the 60% 5-year survival for all cancer sites combined (Service, 2006), but comparable to international rates of gastric cancer (Fock, 2014).

While gastric cancer registration and mortality rates for both Māori and non-Māori have declined for the last two decades, there are still clear disparities between the two groups. The incidence of Māori gastric cancer has consistently been 2-3 times higher than that of non-Māori (Figure 3.2). Similarly, gastric cancer mortality rates mirror this pattern of inequity, and have been up to three times higher for Māori compared to non-Māori (Ministry of Health, 2016b). These trends are exemplified in the most recent 2013 NZCR data, for which the age-standardised gastric cancer incidence and mortality rates were 3.3 (14.6 versus 4.4 per 100,000 population) and 3.5 (11.4 versus 3.3 per 100,000 population) times higher for Māori compared to non-Māori, respectively (Ministry of Health, 2016b).

The clinical features of gastric cancer also differ between Māori and

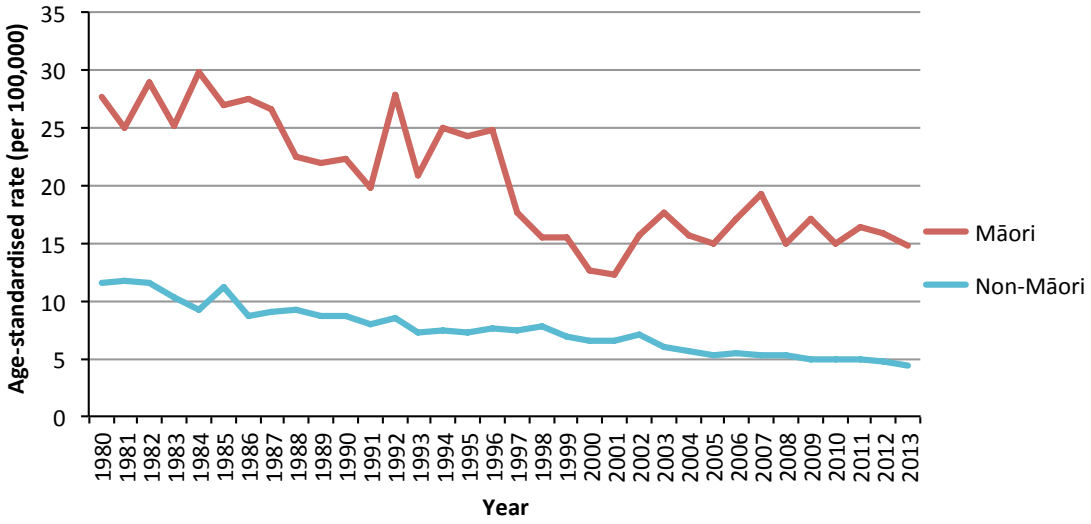


Figure 3.2: Age-standardised incidence of gastric cancer in New Zealand, Māori and non-Māori, 1980-2013. Age-standardised incidence per 100,000 population; standardised to the WHO world standard population. Data source: New Zealand Cancer Registry, Cancer: New Registrations and Deaths

non-Māori. Internationally it is widely accepted that intestinal-type gastric cancer is more common than diffuse-type disease (histological subtypes previously described in section 1.1.1.1). Although New Zealand Europeans align with this trend and are predominately diagnosed with intestinal-type gastric cancer, Māori do not, and are more commonly diagnosed with diffuse-type disease (Biggar et al., 2011). Additionally, Māori are more likely to be diagnosed with tumours in the distal stomach (43% Māori versus 26 % non-Māori), and on average, develop gastric cancer approximately 10 years younger that non-Māori (Signal et al., 2015). Together, these finding suggest that there may be differing aetiological factors driving the high-incidence of gastric cancer observed for Māori.

3.1.3 Disparities in exposure to environmental risk

As described in Section 1.1.2, there are a number of risk factors that influence gastric cancer risk. It is presumed that Māori experience a higher exposure to a number of these risk factors and that these are contributing to their high incidence of gastric cancer relative to non-Māori.

One of these risk factors is socioeconomic status, a well established determinant of health. In New Zealand, socioeconomic status can be measured by a series of indicators from census information, including: education, income, employment status, occupation, living standards, wealth, deprivation, and poverty (Robson, Purdie, & Cormack, 2010). In general, a lower socioeconomic status is associated with an increased risk of cancer. Although both Māori and non-Māori with low socioeconomic status are at a heightened risk, Māori are highly over-represented in the most deprived deciles of all socioeconomic indicators (Robson et al., 2010). In the 2002–2006 period, it is estimated that a higher exposure to socioeconomic deprivation accounted for 27% of the disparity in cancer incidence between Māori and non-Māori (Robson et al., 2010).

Another risk factor disproportionately experienced by Māori is *H. pylori* infection. *H. pylori* is a well established gastric cancer risk factor which has been shown to greatly increase risk. Although *H. pylori* infection rates have declined across all ethnic groups in New Zealand, differences in infection prevalence between different ethnic groups have increased. Currently the prevalence of *H. pylori* infection is nearly double for Māori (35%) compared to Europeans (18%) (McDonald, Sarfati, Baker, & Blakely, 2015). It is thought that this difference may partially be explained by household overcrowding, which is five times more common for Māori, and has been associated with increased

rates of *H. pylori* infection (Brown, 2000; Baker, Goodyear, Telfar Barnard, & Howden-Chapman, 2006). *H. pylori* infection has also been linked to the development of distal stomach cancer (Crew & Neugut, 2006; Forman & Burley, 2006), which may explain the high rates of distally located tumours for Māori.

The rates of smoking for both Māori and non-Māori have dropped in recent years, however 2015/2016 New Zealand health survey data showed that the proportion of Māori 15 years and over who were regular smokers was more than double that of New Zealand Europeans (32.7% versus 14.1%, respectively) (Ministry of Health, 2016a). While the direct risk of gastric cancer caused by smoking has been debated, it is still considered a risk factor for gastric cancer and has a series of co-morbidities. Of note, smoking is thought to interact with *H. pylori* and increase the risk of gastric cancer more than what would be expected for each risk factor alone (Forman & Burley, 2006). Smoking is one of the leading modifiable risks to health.

3.1.3.1 Gastric cancer risk factors for Māori

The research in this thesis is an adjunct to a large population based case-control study investigating the relationships between established risk factors and gastric cancer for New Zealand Māori. Recently, a paper presenting the methods and findings in relation to known gastric cancer risk factors was published (Ellison-Loschmann et al., 2017). The full study includes 165 Māori gastric cancer cases, identified from the New Zealand Cancer Registry between 1 February 2009 and 31 October 2013, and 480 controls identified from the New Zealand electoral roll and matched to cases by 5-year age band (Ellison-Loschmann et al., 2017).

Results from this study confirmed the importance of a number of risk factors identified in previous studies. Of note, being tested for *H. pylori* (OR 12.17, 95% CI 6.15-24.08), having diabetes (OR 2.03, 95% CI 1.32-3.14) or dyspepsia (OR 2.61, 95% CI 1.70-4.01), sharing a bedroom (>2 people) in childhood (OR 3.30, 95% CI 1.95-5.59), and having a parent diagnosed with gastric cancer (OR 4.54, 95% CI 2.45-8.40) were all significantly associated with an increased risk of gastric cancer. Being an ex-smoker (OR 2.26, 95% CI 1.44-3.54) and being exposed to environmental tobacco smoke (OR 3.29, 95% CI 1.94-5.59) were both significantly associated with gastric cancer risk, however being a current smoker was not (OR 1.46, 95% CI 0.86-2.48). Apart for the highest quantile, deprivation was not associated with an elevated risk of gastric cancer. There were no significant associations between alcohol intake, red and white meat consumption, or obesity and increased gastric cancer risk. Markedly, this study noted nearly half of all cases in the cohort (49.7%) were of the diffuse-type, almost double that of the intestinal-type.

3.1.4 HDGC in New Zealand

Familial gastric cancer in a kindred of Māori ethnicity was first reported in 1964 (Jones, 1964). In this pedigree of 98 family members (known as family A), 28 (28.6%) were affected by primary gastric carcinoma, with many at an early age. While the clustering of gastric cancer in this family was consistent with the dominant inheritance of a susceptibility gene with incomplete penetrance, genetic linkage and sequencing technology was limited, making a search for a genetic cause of the disease impossible.

Almost 30 years later, the genetic basis underlying the familial clustering of gastric cancer in family A was discovered (Guilford et al., 1998). In a

partnership between the family and scientists at the University of Otago, genetic linkage analysis and meticulous genealogy records assembled by family A were used to identify a conserved haplotype containing the E-cadherin gene *CDH1*. Subsequent sequencing of the *CDH1* gene identified a heterozygous missense c.1008G>T variant in the last base of exon 7. RT-PCR experiments revealed that the variant was deleterious and resulted in a 7-bp insertion and a premature stop codon derived from an intronic sequence between the normal splice donor site and an adjacent cryptic splice site (Guilford et al., 1998). In this family, the proportion of individuals with the c.1008G>T variant who were affected with disease provided a lifetime penetrance estimate of approximately 70% (Guilford et al., 1998).

In addition to the c.1008G>T variant identified in family A, deleterious variants in germline *CDH1* were identified in two additional Māori families with early-onset diffuse gastric cancer (also known as family B and C) (Guilford et al., 1998). In family B, two nonsynonymous variants were identified. The first was an insertion of a cytosine residue in a run of five cytosines in exon 15 (c.2382_ 2386insC), that resulted in an E-cadherin molecule lacking a section of its cytoplasmic domain (Guilford et al., 1998). An additional missense variant (c.1409C>T) was identified in exon 10 of *CDH1* in the proband from this family. However, this variant did not segregate with the disease phenotype and was not thought to be causative in this family (Guilford et al., 1998). In family C, a c.2095C>T variant was identified in exon 13 of *CDH1*. This variant is an inactivating nonsense variant that results in an E-cadherin peptide lacking its transmembrane and cytoplasmic domains (Guilford et al., 1998). As a result of the identification of germline *CDH1* variants in these families, further pathogenic *CDH1* variants were identified in similar families around the world, culminating in the description of a new familial cancer syndrome - hereditary diffuse gastric cancer.

In New Zealand genetic testing has been offered to families with history of diffuse gastric cancer since the time variants were first discovered in 1997. For those found to carry a predisposing variant, regular endoscopic screening and total prophylactic gastrectomy have been offered since the year 1998. Between the years of 1998 and 2008, 68 Māori carrying deleterious *CDH1* variants were identified in New Zealand (P. Guilford, personal communication). In more recent years, testing for *CDH1* variants has moved away from research labs into clinical laboratories, making the exact number of Māori whom carry these variants unknown.

3.1.5 Clustering of sporadic gastric cancer

Clustering of sporadic gastric cancers, for which no major high-penetrance variants have been identified, appear to be due to a complex combination of environmental and genetic factors. Environmental risk factors for gastric cancer have been well established and when shared by a family group, community, or ethnicity, may explain high rates of the disease. However, common low-penetrance genetic variants could also be contributing to the clustering of gastric cancer in these groups. Single nucleotide polymorphisms (SNPs) are one type of common genetic variant, which through either a direct genetic mechanism or by being in linkage disequilibrium with another variant, have proven useful as markers for genetic risk. Accordingly, associations between SNPs and gastric cancer risk have been established in various variants and populations world-wide (Gonzalez, Sala, & Capellá, 2002; Skierucha et al., 2016).

Genetic variants located in genes relating to several pathways critical for gastric cancer progression have been identified through a series of candidate

gene and genome wide association studies. SNPs have been identified in genes related to: (1) cell-to-cell adhesion (*CDH1*); (2) protection against invading pathogens (*MUC1*); (3) the inflammatory response (*IL-1*, *IL-17*, and toll-like receptors); (4) the repair of DNA damage related to *H. pylori* (*ERCC2*, *XPA*, *XPC*); (5) the metabolism of polycyclic aromatic hydrocarbons (*EPHX1*, *GSTT1*, *NAT2*, *SULT1A1*); (6) the metabolism of xenobiotic chemicals (*Cyp2e1*); and (7) further genes whose functions are not fully understood, for example *PSCA* (Gonzalez et al., 2002; Skierucha et al., 2016).

Of particular interest are SNPs in close proximity to the *CDH1* gene. Loss of E-cadherin expression has been observed in diverse types of sporadic human cancer, in particular, diffuse-type gastric cancer and lobular breast cancer (Van Roy & Berx, 2008). Accordingly, SNPs in and in close proximity to *CDH1*, may promote to the high incidence of sporadic gastric cancer and sporadic diffuse gastric cancer in some populations. The most widely studied polymorphism is *CDH1* -160C>A (rs16260), where the A allele was found to have reduced transcription factor binding strength, and reduced transcriptional activity of the *CDH1* gene (L. Li et al., 2000).

Associations between the rs16260 SNP and gastric cancer risk have been inconsistent and sometimes contradictory (Jiang et al., 2015). Medina-Franco, Ramos-De la Medina, Vizcaino, and Medina-Franco (2007) found that the homozygous AA genotype had a significantly elevated risk for gastric cancer in the Mexican population (OR = 6.5, 95% CI = 2.1–19.6). Al-Moundhri et al. (2010) found a similar result in the Omani population (OR = 3.6, 95% CI = 1.1–11.8). Conversely, Wu et al. (2002) observed that the frequency of the AA genotype was significantly lower in gastric cancer cases than controls in the Taiwanese population (OR = 0.20, 95% CI 0.06–0.56). In a recent meta-analysis of the rs16260 SNP, including 4218 cases and 5461 controls from 22 case-control

studies, no association between the *CDH1* -160C>A polymorphism and risk of gastric cancer was found (Jiang et al., 2015).

When exclusively considering sporadic diffuse-type gastric cancer, there is some evidence that the A allele increases disease risk. In an Italian case-control study, the odds ratio associated with the A-allele was 2.27 (95% CI 1.16-4.44) for CA heterozygotes and 7.84 (95% CI 2.89-21.24) for AA homozygotes (Humar, Graziano, et al., 2002). Furthermore, this study noted that the age at diagnosis was highest in CC homozygotes and lowest in AA homozygotes (mean ages: CC=59, CA=52.6, AA=52.3 years), and that a three marker haplotype containing the -160C>A polymorphism was significantly associated with disease risk (Humar, Graziano, et al., 2002). As of yet, no studies have examined the associations between SNPs at the *CDH1* locus and gastric cancer risk for the Māori population.

3.2 Results

HDGC is well documented in Māori families from New Zealand. However, the contribution of germline *CDH1* variants to overall high incidence of diffuse gastric cancer in the Māori population is unknown. In this chapter, the most recent data from the New Zealand Cancer Registry was examined to identify differences in the presentation of gastric cancer between Māori and non-Māori. Next, I used next-generation amplicon sequencing, MLPA, and Sanger sequencing to identify germline *CDH1* variants in a case-control cohort of Māori gastric cancer patients and matched healthy controls. Variants were evaluated and rare and potentially pathogenic *CDH1* variants were validated using Sanger sequencing. Pathology reports from cases were reviewed to identify those who were likely to be members of HDGC families. Using this

information, the contribution of HDGC families to the high incidence of gastric cancer in the Māori population was estimated. Finally, associations between common variants at the *CDH1* locus and the development of sporadic gastric cancer and sporadic diffuse gastric cancer in the Māori population were analysed.

3.2.1 Characterisation of New Zealand cancer registry data

Data from the New Zealand Cancer Registry was acquired to examine the trends for gastric cancer registrations for the years 2009-2013. During this period there were 1,882 gastric cancer registrations, of which 379 (20.1%) were Māori and 1,503 (79.9%) were non-Māori. For the early-onset 5-year age bands 15-19, 20-24, 25-29, and 35-39, there were markedly more gastric cancer registrations for Māori than non-Māori (Figure 3.3a). For all age bands greater than 45 years, there were more non-Māori gastric cancer registrations (Figure 3.3b).

The crude age-standardised registration rate of gastric cancer increased with age for both Māori and non-Māori. However, in the crude number of registrations there was a striking over-representation of early-onset gastric cancer registrations which were absent from the non-Māori population (Figure 3.3c). For the early-onset age-bands, less than 45 years of age, the rate of registration was 3.9 times higher for Māori compared to non-Māori. The age-standardised registration rate peaked at the 80-84 age band for both Māori (181.1 per 100,000 population) and non-Māori (56.6 per 100,000 population) (Figure 3.3d). For this period, the relative risk for Māori was 3.29 (95% CI 1.19-9.06) time higher relative to non-Māori (15.7 versus 4.8 per 100,000 population, respectively).

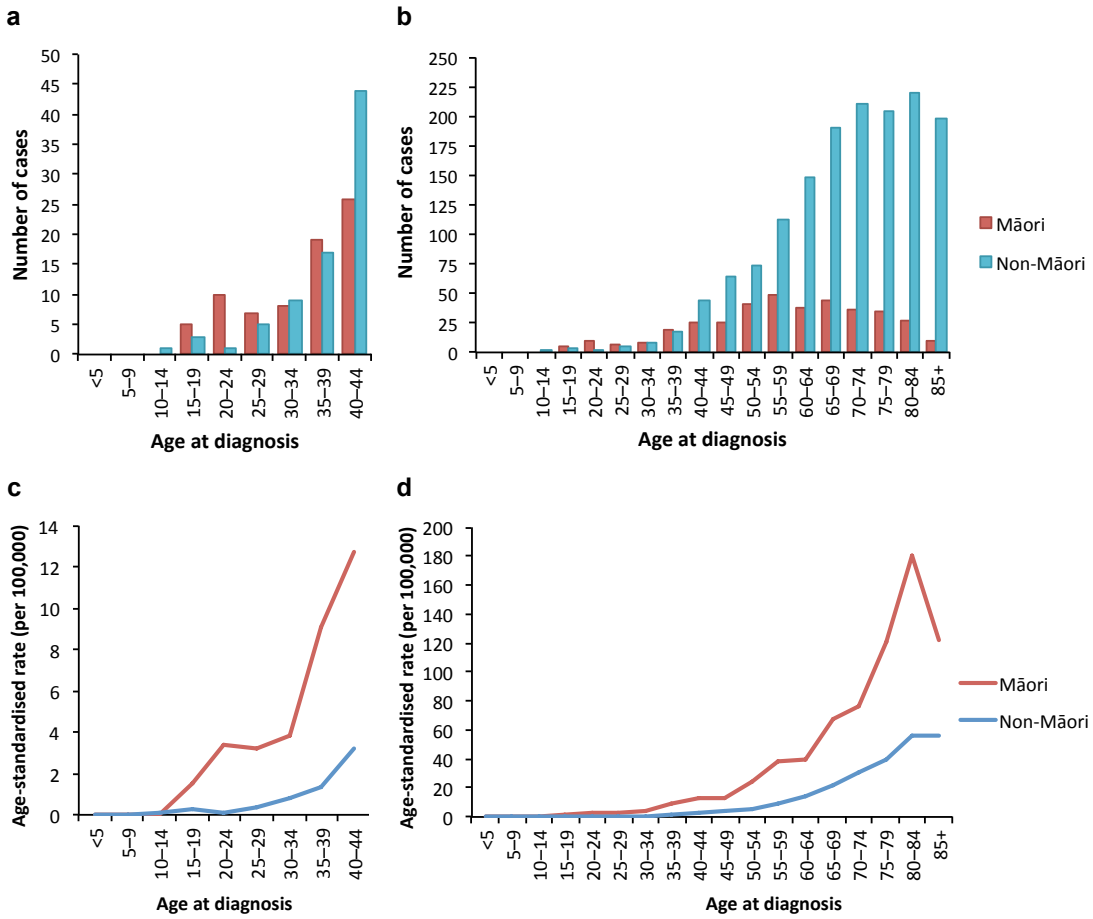


Figure 3.3: Crude gastric cancer registrations and age-adjusted incidence by 5-year age band, Māori and non-Māori, 2009-2013. Distribution of crude gastric cancer registrations by 5-year age band for (a) early onset cases and (b) all cases. Age-adjusted incidence by 5-year age band for (c) early onset cases and (d) all cases. Data source: New Zealand Cancer Registry, Cancer: New Registrations and Deaths.

3.2.2 Cohort characteristics

Germline *CDH1* was sequenced for 94 cases and 200 healthy controls matched for 5-year age band and gender. The case cohort comprised 50 (53%) males and 44 (47%) females with an average age at diagnosis of 55.5 years (range 17–81 years). More than half of cases (53%) were classified as diffuse-type gastric cancers, nearly double the number of cases classified as intestinal-type gastric cancer (27%). Reflecting the large number of diffuse-type tumours, exactly half of tumours (47 tumours; 50%) were poorly differentiated. In contrast to earlier studies (Biggar et al., 2011) there were slightly more tumours located in the proximal stomach (32%) than the distal stomach (21%). Patients were most commonly diagnosed at an early stage when tumours were still localised (28%). Full clinical details for cases are presented in Table 3.1.

Controls were cancer free at the time of blood draw and interview. The control cohort comprised 104 (52%) males and 96 (48%) females with an average age of 57.6 years (range 19–84 years).

3.2.2.1 Clinical features of early-onset gastric cancers

Frequency of clinical features distributed by age at diagnosis are shown in Figure 3.4. Of particular interest is the proportion of early-onset gastric cancers, defined as cases younger than 45 years of age at the time of diagnosis. Of the 94 cases in the cohort, 21 (22.3%) were in this group. Twenty (95.2%) of the early-onset cases were diffuse-type, while one (4.8%) was unspecified (Fig 3.4a). The earliest intestinal-type tumour was diagnosed in a patient aged 49 years. Tumour differentiation was most frequently unspecified (61.9%), while just eight cases were classified as poorly differentiated (38.1%) (Fig 3.4 b).

	Total 94 (n)	%
Gender		
Male	50	53
Female	44	47
Age at diagnosis (years)		
<45	21	22
45-59	32	34
60-74	30	32
$\geq$ 75	11	12
Tumour subtype		
Diffuse	50	53
Intestinal	22	23
Other	9	10
NOS	13	14
Tumour grade		
Well differentiated	4	4
Moderately differentiated	10	11
Poorly differentiated	47	50
NOS	33	35
Tumour site		
Proximal	30	32
Distal	20	21
Mixed	4	4
Oesophageal junction	5	5
NOS	35	37
Extent		
Local	26	28
Lymph node involvement	20	21
Regional spread	7	7
Metastatic spread	13	14
NOS	28	30

Table 3.1: Characteristics of Māori case cohort. *Other subtype includes intermediate type tumours described using the World Health Organization classification. NOS, Not otherwise specified.

	Total 200 (n)	%
Gender		
Male	104	52
Female	96	48
Age (years)		
<45	39	20
45-59	67	33
60-74	70	35
>75	20	12

Table 3.2: Characteristics of control cohort.

There did not appear to be any trend in the site of the early-onset cancers, of which four (19.0%) were distal, six (28.6%) were proximal, and 12 (52.4%) were unspecified (Fig 3.4c). Nearly half (47.6%) of these early-onset cancers were still localised, although two (9.5%) had lymph node involvement, and five (23.8%) had metastatic spread (Fig 3.4d).

3.2.2.2 Review of pathology reports

Pathology reports from all 94 patients were reviewed for information regarding diagnosis and prior variant screening. Reports from 15 of the 94 cases described prophylactic gastrectomies or endoscopic screening as a part of the clinical pathway, and/or noted the patient was a *CDH1* carrier of a pathogenic variant. As prophylactic gastrectomies and endoscopic screening procedures are offered to *CDH1* pathogenic variant carriers, these cases were likely to be *CDH1* pathogenic variant carriers who had elected for prophylactic surgery or who had had foci of gastric cancer identified during endoscopic screening.

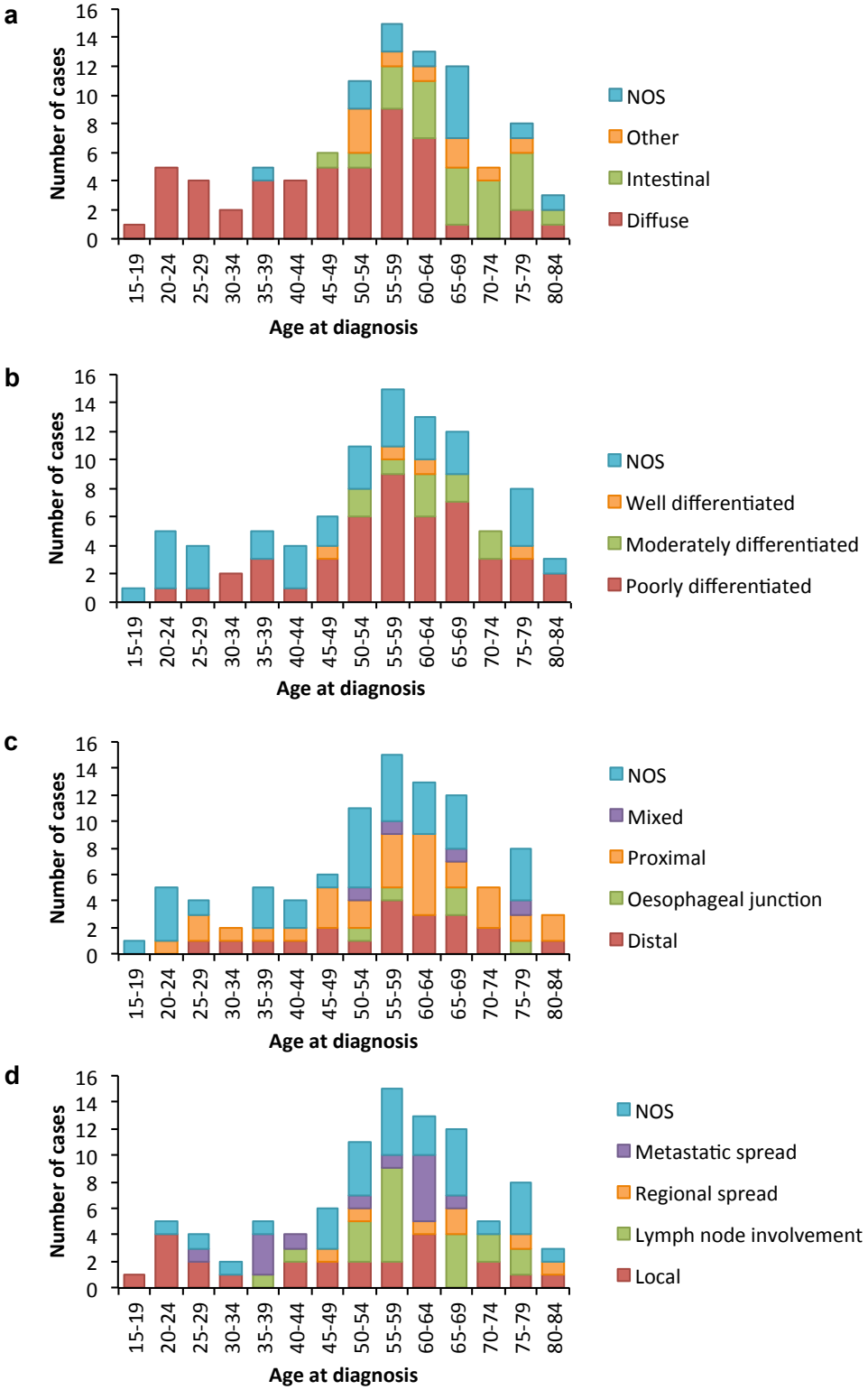


Figure 3.4: Frequency of clinical features by age at diagnosis of gastric cancer in the Māori population. Distribution of (a) histological subtypes, (b) tumour grade, (c) tumour site, and (d) tumour extent. NOS; Not Otherwise Specified.

3.2.3 Sequencing libraries and data

Amplicon libraries were prepared for the 94 cases and 200 matched controls in this study. All libraries were prepared the same way. First, 17 target sites were amplified in separate PCR reactions and run on 2% agarose gels to check for successful amplification (Figure 3.5a). Second, diluted pools of amplicon products were amplified using a pair of primers that were designed to add indices and sequences necessary for sequencing on an Illumina MiSeq (Figure 3.5b). Amplicon products were designed to fit within the capabilities of a 500 cycle MiSeq reagent kit. For each sequencing run, indexed libraries were pooled and checked on a bioanalyzer to determine the average size of each library (Figure 3.5c).

As a part of this PhD, multiple MiSeq sequencing runs were used to sequence *CDH1* amplicon libraries for different study cohorts. Data was generated from a total of seven MiSeq runs. Basic details including the reagents and flow cell used, loading density, number of reads obtained, percentage of reads which passed filter, percentage of reads which aligned to PhiX, and percentage of reads with a quality greater than Q30, are displayed in Table 3.3.

Five different sequencing runs were used to generate sequence data for the Māori case and control samples relevant to this chapter. Two of these runs (run one and run four) were used to generate short pair-end reads which were used to optimise the library loading density and representation of each sample for the long read runs. For samples with a low number of reads or low representation of target amplicons, extra sample specific library was added. Three full length runs (run two, three, and five) were used to generate pair-end reads that were long enough to cover the full target regions: In run two, 87 case samples and eight control samples were sequenced; in run three, eight case

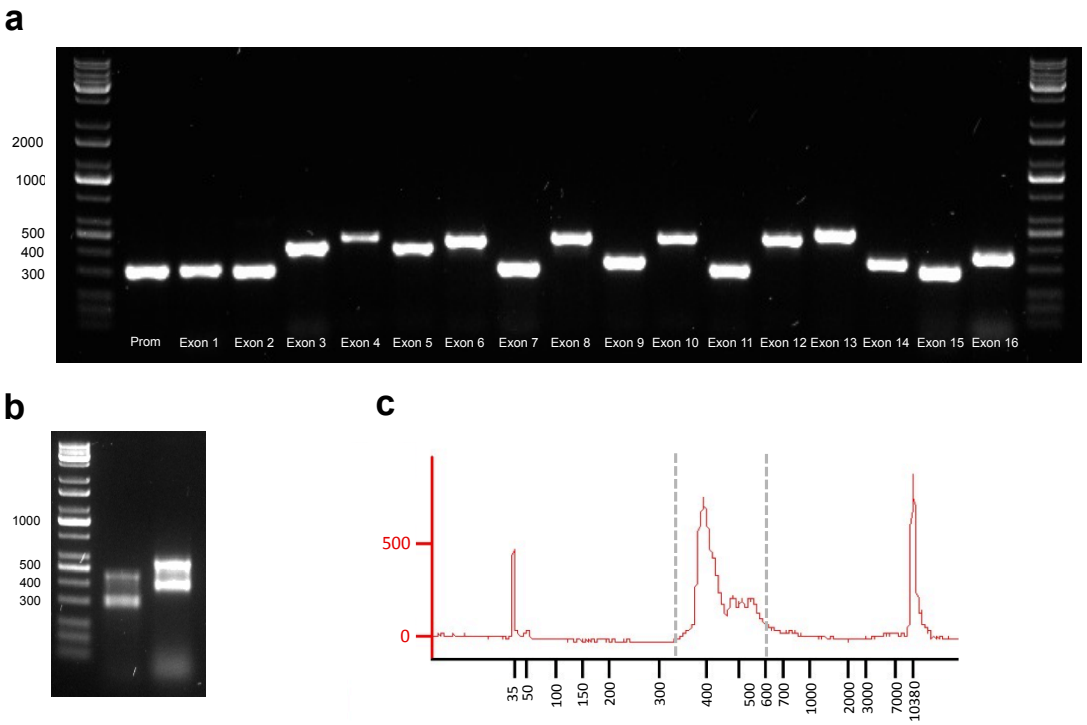


Figure 3.5: Preparation of *CDH1* MiSeq Libraries. Exemplar library preparation showing: (a) amplification of targeted *CDH1* exon target sites; (b) adapter PCR used for adding sequences necessary for sequencing on an Illumina MiSeq; and (c) bioanalyzer trace showing the average size of an amplified library.

MiSeq Sequencing Run	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8
Reagents used	V3-150	V3-600	V3-600	V2-300	V3-600	V2-500	V2-500	V2-500
Flow cell used	Full	Full	Full	Nano	Full	Nano	Full	Nano
Loading density	334 +/-8	571 +/-14	360 +/-10	615+/-10	635+/-10	1025+/-25	695+/-26	1020+/-2
Total reads	8,402,476	13,912,235	8,876,392	856,500	15,733,683	1,359,181	13,391,679	1,248,637
Passing filter (%)	98.8	94.4	97.2	96.9	97.4	87 .3	94.5	86.5
Aligned (%)	21.6	20.2	17.9	4.6	4.7	8.6	11.24	10.7
Reads >Q30 (%)	98.1	91.8	92.1	96.7	91.3	87.8	83.62	84.31
Cycles (read 1 | read 4)	68 | 68	241 | 241	241 | 241	150 | 150	241 | 241	241 | 241	241 | 241	241 | 241

Table 3.3: Characteristics of MiSeq sequencing runs. Basic details of MiSeq runs used to sequence *CDH1* libraries during this PhD. Note: Runs three, five, six, and seven contained libraries from sequencing projects not a part of this PhD.

samples were sequenced; and in run five, 192 control samples were sequenced. Run five, six, seven, and eight were used for sequencing of samples in Chapter 4. Notably, in runs three, five, six, and seven, additional samples from other projects within the lab were included in the sequencing run.

A negative control library was prepared and sequenced along side each set of samples as a measure of contamination and erroneous reads. mqH_20 was used as the negative control and underwent the same library preparation method as the DNA sample. The number of aligned reads identified in the negative control library was compared to the lowest number of reads for each amplicon. In all sequencing runs the number of reads identified in the negative control sample were less than 1% of the lowest read depth for each amplicon. Therefore there was a low likelihood of false positives caused by erroneous reads or contamination between samples. Additionally, the Fisher strand values and allele counts were used to assess each variant call to identify any variant calls which had an abnormal distribution of reads. All variant calls had expected allele ratios and Fisher strand values.

3.2.4 *CDH1* variants

In total 26 different variants were identified across the case and control cohorts (Table 3.4). Of these: 18 variants were located in the coding exons of the *CDH1* gene (six synonymous, eight missense, three nonsense, and one frameshift); five variants were located in the proximal promoter and 5'UTR; and three were located in intronic splice sites. Table 3.4 shows the position of each variant, allele frequencies in the case and control cohort, and allele frequencies from major population databases. All rare variants (MAF < 0.05) were critically reviewed to assess the clinical significance of variants identified in this study.

The effects of missense and splice variants were predicted using *in silico* prediction tools. The functional consequences of missense variants were predicted using Condel, a bioinformatics tool for the evaluation of nonsynonymous single nucleotide variants (SNVs) (González-Pérez & López-Bigas, 2011). Condel incorporates two prediction tools (MutationAssessor and FatHMM) to create a Condel pathogenicity score that is used for variant classification. Condel also reports pathogenicity scores from other prediction software, namely SIFT and PolyPhen2. The scores from each prediction tool and the overall Condel score and variant classification are shown in Table 3.5. Possible splice site alterations were predicted using MaxEntScan, NNSplice and Human Splicing Finder (HSF) 3.0 and are shown in Table 3.6.

Classification of variants reported in the public archive ClinVar are included in the review of variants. ClinVar reports the clinical significance using the five terms for Mendelian diseases recommended by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (Richards et al., 2015). The five clinical significance values are: (1) benign, (2) likely benign, (3) variants of uncertain significance (VUS), (4) likely pathogenic, and (5) pathogenic, respectively.

3.2.4.1 Frameshift variants

c.2381_ 2386insC

Four cases diagnosed with diffuse gastric cancer, aged 17, 20, 26, and 44 years, were found to carry a *CDH1* c.2381_ 2386insC variant. This variant is located in exon 15 of the *CDH1* gene and has been shown to cause a truncated E-cadherin protein missing a portion of its cytoplasmic domain (Guilford et

Position[a]	SNP ID	Region	Class	Variant	Protein	Cases Het/Hom	Controls Het/Hom	MAF Cases	MAF Controls	MAF ExAc	MAF 1000 Genomes	MAF ESP6500
68771034	rs16260	Promoter	.	c.-285C>A (-160C>A)	.	19/1	62/3	0.112	0.170	NA	0.2360	NA
68771043	rs34149581	Promoter	.	c.-276T>C (-152T>C)	.	0/0	1/0	0	0.002	NA	NA	NA
68771122	rs28372783	Promoter	.	c.-197A>C (-73A>C)	.	17/5	40/3	0.144	0.115	NA	0.0463	NA
68771143	rs34500817	Promoter	.	c.-176C>T (-52C>T)	.	0/0	2/0	0	0.005	NA	0.0018	NA
68771248	rs34033771	5′ UTR	.	c.-71C>G	.	0/0	3/0	0	0.007	NA	0.0056	NA
68771372	rs3743674	Intron 1	Splice	c.48+6C>T	.	10/82	36/163	0.926	0.905	0.81110	0.7634	0.8603
68772239	rs139866691	Exon 2	Missense	c.88C>A	p.Pro30Thr	0/0	1/0	0	0.002	0.00094	0.0004	0.0010
68835599	.	Exon 3	Nonsense	c.190C>T	p.Gln64*	4/0	0/0	0.021	0	NA	NA	NA
68835754	rs1801023	Exon 3	Synonymous	c.345G>A	p.Thr115Thr	0/0	1/0	0	0.002	0.00392	0.0026	0.0030
68835801	rs113055163	Intron 3	Splice	c.387+5G>A	.	1/0	3/0	0.005	0.007	0.00001	NA	NA
68845648	rs139110184	Exon 7	Synonymous	c.894C>T	p.Ala298Ala	0/0	1/0	0	0.002	0.00018	NA	0.0002
68847292	.	Exon 9	Missense	c.1214A>G	p.Asn405Ser	0/0	1/0	0	0.002	0.00003	NA	NA
68849506	.	Exon 10	Missense	c.1409C>T	p.Thr470Ile	3/0	4/0	0.016	0.010	0.00003	NA	NA
68853297	rs35741240	Exon11	Synonymous	c.1680G>C	p.Thr560Thr	0/0	2/0	0	0.005	0.00372	0.0018	0.0022
68855966	rs35187787	Exon 12	Missense	c.1774G>A	p.Ala592Thr	1/0	1/0	0.005	0.002	0.00321	0.00120	0.0045
68855984	rs121964877	Exon 12	Nonsense	c.1792C>T	p.Arg598*	4/0	0/0	0.021	0	NA	NA	NA
68856041	rs33935154	Exon 12	Missense	c.1849G>A	p.Ala617Thr	1/0	0/0	0.005	0	0.00423	0.0144	0.0160
68856088	rs33969373	Exon 12	Synonymous	c.1896C>T	p.His632His	0/0	2/0	0	0.005	0.01142	0.0288	0.0245
68857441	rs1801552	Exon 13	Synonymous	c.2076T>C	p.Ala692Ala	48/21	105/30	0.479	0.4125	0.65480	0.7187	0.7161
68862107	.	Exon 14	Missense	c.2195G>A	p.Arg732Gln	1/0	0/0	0.005	0	NA	NA	NA
68862165	rs33964119	Exon 14	Synonymous	c.2253C>T	p.Asn751Asn	5/0	18/0	0.027	0.044	0.03955	0.0545	0.0376
68862199	.	Exon 14	Nonsense	c.2287G>T	p.Glu763*	4/0	0/0	0.021	0	NA	NA	NA
68863590	rs372989292	Exon 15	Missense	c.2329G>A	p.Asp777Asn	1/0	0/0	0.005	0	0.00013	NA	0.0001
68863643	.	Exon 15	Frameshift	c.2381_2386insC	p.Arg796fs	4/0	0/0	0.021	0	NA	NA	NA
68867187	rs139757930	Intron 15	Splice	c.2440-6C>G	.	0/0	2/0	0	0.005	0.00180	0.0008	0.0021
68867309	.	Exon 16	Missense	c.2556G>T	p.Glu852Asp	0/0	2/0	0	0.005	0.00002	NA	NA

Table 3.4: Variants identified in Māori case-control cohort. Variants in the regulatory region of *CDH1* are commonly denoted based on GenBank accession number NT010498. These denotions are shown in brackets next the variant description.
[a]Reference sequence for variant position is Hg19. MAF, Minor Allele Frequency; Het, Heterozygote; Hom, Homozygote; ins, insertion; fs, frameshift; NA, Not Available.

al., 1998). *CDH1* c.2381_ 2386insC is a well established pathogenic HDGC variant in New Zealand and was first seen by Guilford et al. (1998), in one of the three Māori families germline *CDH1* variants were first identified in. ClinVar classifies *CDH1* c.2381_ 2386insC as pathogenic.

3.2.4.2 Nonsense variants

c.190C>T

The *CDH1* c.190C>T variant was identified in four cases diagnosed with diffuse gastric cancer, aged 24, 29, 48, and 61 years. This nonsense variant causes the premature termination in the E-cadherin sequence by changing a glutamine residue to a stop codon at codon 64 of the E-cadherin protein (p.Gln64*). *CDH1* c.190C>T was originally described by Guilford et al. (1999) in a Māori family with a history of gastric cancer. In this family, the proband was diagnosed with poorly differentiated, diffuse gastric cancer at age 22 years. The probands' mother was also affected by diffuse gastric cancer and had died of the disease aged 28 years. ClinVar classifies *CDH1* c.190C>T as pathogenic.

c.1792C>T

CDH1 c.1792C>T was identified in four cases diagnosed with diffuse gastric cancer, aged 29, 20, 23, and 24 years. *CDH1* c.1792C>T causes a nonsense change (p.Arg598*) in exon in 12 of *CDH1*, resulting in a truncated E-cadherin protein. The variant was first seen by Gayther et al. (1998) in a European family in which two identical twins were both diagnosed with diffuse gastric cancer. Later, Humar, Toro, et al. (2002) identified the same variant in a Māori family from New Zealand. The European and Māori families are not known to be related, and as of 2005, no further analysis of relatedness had been done

(Suriano et al., 2005). ClinVar classifies *CDH1* c.1792C>T as pathogenic.

c.2287G>T

Four cases diagnosed with diffuse gastric cancer were found to carry a *CDH1* c.2287G>T variant. These cases were aged 31, 39, 41, and 50 years at the time of diagnosis. The *CDH1* c.2287G>T variant is a nonsense change (p.Glu763*) located in exon 14 which truncates the protein in the cytoplasmic domain. *CDH1* c.2287G>T has previously been described in a HDGC family in New Zealand (Charlton et al., 2004; Guilford, Humar, & Blair, 2010). ClinVar classifies *CDH1* c.2287G>T as pathogenic.

3.2.4.3 Missense variants

c.88C>A

The germline c.88C>A *CDH1* variant was identified in one healthy control, aged 56 years. This variant causes a proline to threonine substitution at codon 30 of the E-cadherin protein (p.Pro30Thr). Between species, the proline residue is weakly conserved and there is only a small physiochemical difference between proline and threonine. Condel predicts c.88C>A is deleterious. This variant is present at low frequencies in ExAc (MAF 0.00094), 1000 Genomes (MAF 0.0004), and ESP6500 (MAF 0.0010) variant databases. ClinVar classifies *CDH1* c.88C>A variant as likely benign.

c.1214A>G

CDH1 c.1214A>G was identified in one healthy control aged 74 years. *CDH1* c.1214A>G causes an asparagine to serine substitution at codon 405 (p.Asn405Ser), located in the cadherin three domain of extracellular E-cadherin. The asparagine residue is moderately conserved across species and there is a small physiochemical difference between asparagine and serine.

Position[a]	SNP ID	Region	Variant	Protein	Cases Het/Hom	Controls Het/Hom	SIFT	PPH2	MA	FatHMM	Condel Score	Condel Label
68772239	rs139866691	Exon 2	c.88C>A	p.Pro30Thr	0/0	1/0	0.400	0.886	2.045	-1.890	0.567	Deleterious
68847292	.	Exon 9	c.1214A>G	p.Asn405Ser	0/0	1/0	0.130	0.080	1.130	0.760	0.464	Neutral
68849506	.	Exon 10	c.1409C>T	p.Thr470Ile	3/0	4/0	0.000	1.000	2.725	0.340	0.587	Deleterious
68855966	rs35187787	Exon 12	c.1774G>A	p.Ala592Thr	1/0	1/0	0.000	0.092	1.440	0.090	0.494	Neutral
68856041	rs33935154	Exon 12	c.1849G>A	p.Ala617Thr	1/0	0/0	0.190	0.035	1.545	0.530	0.475	Neutral
68862107	.	Exon 14	c.2195G>A	p.Arg732Gln	1/0	0/0	0.100	0.991	2.215	0.300	0.529	Deleterious
68863590	rs372989292	Exon 15	c.2329G>A	p.Asp777Asn	1/0	0/0	0.440	0.997	2.280	-1.060	0.606	Deleterious
68867309	.	Exon 16	c.2556G>T	p.Glu852Asp	0/0	2/0	0.010	0.020	-0.690	-1.140	0.397	Neutral

Table 3.5: Prediction of functional consequence of missense variants. Condel deleterious prediction cut off >0.522. [a]Reference sequence for variant position is Hg19. Abbreviations: Het, Heterozygote; Hom, Homozygote; MA, Mutation assessor; PPH2, Polyphen2.

Condel predicts this variant to be neutral. *CDH1* c.1214A>G has been seen at a low frequency in the ExAc variant database (MAF 0.00003), but was absent from both the 1000 Genomes and ESP6500 databases. ClinVar classifies *CDH1* c.1214A>G as a variant of unknown significance.

c.1409C>T

CDH1 c.1409C>T was identified in three cases and four controls in this study. The three cases varied in age and cancer morphology and were diagnosed with intestinal-type, unspecified-type, and diffuse-type gastric cancer, aged 71, 66, and 21 years, respectively. Of note, the case diagnosed with diffuse-type disease carried a second rare frameshift variant (c.2381_ 2386insC), a well documented HDGC variant. Controls identified to carry this variant were aged 35, 57, 59, and 74 years.

CDH1 c.1409C>T causes a threonine to isoleucine at codon 470 of the E-cadherin protein (p.Thr470Ile). The threonine residue is highly conserved between species and there is a moderate physicochemical difference between threonine and isoleucine. Condel predicts *CDH1* c.1409C>T to be damaging. This variant is rare in ExAc database (MAF 0.00003) and is absent from the 1000 Genomes and ESP6500 databases. ClinVar classifies *CDH1* c.1409C>T as likely benign.

This variant has previously been described in a Māori family affected with early-onset diffuse gastric cancer alongside a deleterious c.2381_2386insC *CDH1* variant (Guilford et al., 1998). In this family, the second variant segregated with the disease while *CDH1* c.1409C>T did not, suggesting that the c.1409C>T substitution was not the primary cause of disease in this family (Guilford et al., 1998). As homozygosity for deleterious *CDH1* variants is embryonic lethal, the co-occurrence of c.1409C>T with another pathogenic variant argues against its pathogenicity. Furthermore, as this variant was

found in an additional four currently healthy controls, it is likely this variant may be more common in the Māori population than other populations, making the identification of this variant in cases unremarkable.

c.1774G>A

The *CDH1* c.1774G>A variant was identified in one case diagnosed with diffuse gastric cancer aged 57 years and one control aged 64 years. *CDH1* c.1774G>A causes an alanine to threonine substitution at codon 592 of the E-cadherin protein (p.Ala592Thr.) The alanine in this position is not well conserved between species and there is another rare substitution (c.1774G>T; ExAc MAF 0.00042) in the same position. *CDH1* c.1774G>A was predicted to be neutral by Condel and is classified as benign and likely benign by ClinVar. *CDH1* c.1774G>A has been reported at low frequencies in the ExAc (MAF 0.0032), 1000 Genomes (MAF 0.0012), and ESP6500 (MAF 0.0045) variant databases.

CDH1 c.1774G>A has previously been identified in a Swedish cancer family (Jonsson, Bergh, Stattin, Emmanuelsson, & Grönberg, 2002). In the Swedish family, the variant was found in two male siblings, both with prostate cancer. A sister with breast cancer did not carry the variant. Structural analysis from *in vitro* (Keller et al., 2004) and *in silico* studies (Suriano, Seixas, Rocha, & Seruca, 2006) do not support a pathogenic classification for *CDH1* c.1774G>A.

c.1849G>A

CDH1 c.1849G>A was identified in one case diagnosed with intestinal-type gastric cancer, age 77 years. *CDH1* c.1849G>A causes an alanine to threonine substitution at codon 617 (p.Ala617Thr), in the fifth extracellular repeat of the E-cadherin protein. The c.1849G>A variant is present at low frequencies in the ExAc (MAF 0.00423), 1000 Genomes (MAF 0.0144) and ESP6500 (MAF 0.0160) databases, and is almost exclusive to the African population (MAF

approximately 0.05). *CDH1* c.1849G>A is predicted to be neutral by Condel and is classified as benign by ClinVar.

CDH1 c.1849G>A was first reported by Risinger, Berchuck, Kohler, and Boyd (1994) in a sample of endometrial cancer tissue. Later, Ascaño et al. (2001) reported a germline variant in a diffuse gastric cancer patient of African-American decent. Similarly, El-Husny et al. (2016) also noted a germline variant in a diffuse gastric cancer patient of African-American decent. *In vitro* assays examining the p.Ala617Thr variant in Chinese hamster ovary cells did not show increased motility or invasiveness as a result of the *CDH1* c.1849G>A variant, but did observe a significant reduction of cellular adhesion (Suriano et al., 2003).

c.2195G>A

CDH1 c.2195G>A was identified in one case diagnosed with diffuse gastric cancer aged 38 years. *CDH1* c.2195G>A causes a missense arginine to glutamine at codon 732 of the E-cadherin protein (p.Arg732Gln). *CDH1* c.2195G>A is located in a highly conserved residue, which is located at the intracellular border of the cytoplasmic domain of the E-cadherin protein. The variant was absent from population databases and was predicted to be damaging by Condel.

The c.2195G>A variant has previously been identified in two HDGC families of European decent (Kaurah et al., 2007; Brooks-Wilson et al., 2004). *In silico* analysis of p.Arg732Gln predicted this missense change to cause a new acceptor splice site in the E-cadherin protein (Kaurah et al., 2007). RT-PCR analysis of RNA extracted from white blood cells and a gastrectomy specimen of a p.Arg732Gln variant carrier showed a complex splicing and deletion of 32 base pairs at the start of exon 14 (Kaurah et al., 2007).

c.2329G>A

CDH1 c.2329G>A was identified in one case diagnosed with a gastric cancer of an unspecified type aged 81 years. *CDH1* c.2329G>A causes an aspartic acid to asparagine substitution at codon 777, in cytoplasmic tail of the E-cadherin protein (p.Asp777Asn). The aspartic acid residue is highly conserved and there is only a small physicochemical difference between aspartic acid and asparagine. *CDH1* c.2329G>A is present at low frequencies in the ExAc (MAF 0.00013) and ESP6500 (MAF 0.0001) variant databases. Condel predicts *CDH1* c.2329G>A to be deleterious, however ClinVar describes the variant as likely benign.

CDH1 c.2329G>A has been reported in a families with prostate (Jonsson et al., 2002), breast (Tung et al., 2016), and colon cancers (Kraus et al., 2015). *CDH1* c.2329G>A has also been identified in an individual from a healthy, ancestrally diverse cohort using cancer susceptibility gene panel testing (Bodian et al., 2014).

c.2556G>T

CDH1 c.2556G>T variant was identified in two controls age 62 and 68 years. *CDH1* c.2556G>T causes a glutamic acid to aspartic acid at codon 852 in the cytoplasmic domain of the E-cadherin protein (p.Glu852Asp). The glutamic acid is weakly conserved between species and there is only a small physicochemical difference between glutamic acid and aspartic acid. *CDH1* c.2556G>T is rare in variant databases and is only reported in the ExAc database (MAF 0.00002). Condel predicts this variant to be neutral. *CDH1* c.2556G>T has not been described in any published literature.

Position[a]	Variant	MaxEntScan	NNSplice	HSF 3.0
68835801	c.387+5G>A	9.80 → 7.88	0.99 → 0.64	100 → 99.0
68867187	c.2440-6C>G	10.9 → 9.85	0.98 → 0.95	84.3 → 80.8

Table 3.6: Splice site prediction scores. MaxEntScan: splice site if >= 3; NNSplice: splice site if >= 0.95; HSF 3.0: splice site if >= 65. HSF, Human Splice Finder. [a]Reference sequence for variant position is Hg19.

3.2.4.4 Intronic splice variants

c.387+5G>A

One case, diagnosed with intestinal gastric cancer aged 77 years, and three controls, aged 39, 47, and 67 years were identified to carry a *CDH1* c.387+5G>A variant. *In silico* models predict *CDH1* c.387+5G>A to weaken, but not disrupt, the nearby natural donor site of intron 3. *CDH1* c.387+5G>A was not recorded in the ESP6500 or 1000 Genomes databases but was seen at a very low frequency in the ExAc database (MAF 0.00001). The guanine nucleotide that is altered is not conserved across species. There have been no RNA or functional studies examining the effects of this variant. ClinVar classifies *CDH1* c.387+5G>A as likely benign.

c.2440-6C>G

CDH1 c.2440-6C>G was identified in two controls aged 59 and 66 years. *In silico* models predict *CDH1* c.2440-6C>G to weaken the nearby acceptor site of intron 15. *CDH1* c.2440-6C>G is present at low frequencies in the ExAc (MAF 0.00180), 1000 Genomes (MAF 0.00002), and ESP6500 (0.0021) variant databases. ClinVar classifies *CDH1* c.2440-6C>G as benign or likely benign.

The *CDH1* c.2440-6C>G variant has been controversially discussed in the literature. More et al. (2007) first described *CDH1* c.2440-6C>G in a patient with diffuse gastric cancer. The authors identified an unusual splicing pattern by gel electrophoresis of RT-PCR products and classified the *CDH1*

c.2440-6C>G variant as disease causing. Later, two groups could not confirm the unusual banding in others with the *CDH1* c.2440-6C>G variant using RT-PCR analysis and Sanger sequencing analysis (Grodecká et al., 2014; Molinaro et al., 2014). Analysis of a heterozygous SNP in mRNA from blood cells from a *CDH1* c.2440-6C>G variant carrier demonstrated that both alleles are transcribed (Grodecká et al., 2014).

3.2.4.5 5′ UTR and promoter variants

c.-276T>C (-152T>C)

CDH1 c.-276T>C was detected in one control aged 64 years. *CDH1* c.-276T>C is not described in the three variant databases used in this study but does has a SNP ID (rs34149581). This variant has previously been identified in a case-control study for which no risk of developing breast cancer was detected (Lei et al., 2002). Additionally, transient transfection experiments using reporter constructs with the *CDH1* c.-276T>C variant did not significantly decrease transcriptional activity compared to the wild-type construct (Lei et al., 2002).

c.-176C>T (-52C>T)

CDH1 c.-176C>T was identified in two controls aged 69 and 71 years. This variant is seen at a low frequency in the 1000 Genomes database (MAF 0.0018), but was not reported in the ExAc or ESP6500 databases. *CDH1* c.-176C>T has been reported in a first degree relative of a sporadic gastric cancer patient, but was not seen in the affected patient (Garziera et al., 2013).

c.-71C>G

CDH1 c.-71C>G was identified in three healthy controls aged 60, 64, and 73 years. *CDH1* c.-71C>G is reported at low frequency in the 1000 Genomes

variant database (MAF 0.0056) and is classified as benign and likely benign in ClinVar. *CDH1* c.-71C>G has previously been reported in both sporadic gastric cancer cases and healthy controls and was not thought to affect gene expression (Avizienyte, Launonen, Salovaara, Kiviluoto, & Aaltonen, 2001).

3.2.4.6 Summary of rare variant classification

Evidence supporting the classification of the variants identified in this cohort was varied and sometimes contradicting. Using all available information, the pathogenicity of variants were evaluated in accordance with ACMG/AMP guidelines.

In summary, five variants (c.190C>T, c.1792C>T, c.2195G>A, c.2287G>A, and c.2381_ 2386insC) were classified as pathogenic; two variants (c.1214A>G and c.2556G>T) were classified as variants of uncertain significance; and ten variants (c.-276T>C, c.-176C>T, c.-71C>G, c.88C>A, c.387+5G>A, c.1409C>T, c.1774G>A, c.1849G>A, c.2329G>A, and c.2440-6C>G) were classified as benign or likely benign. Notably, both variants of uncertain significance variants were predicted to be neutral by Condel and were exclusively identified in controls aged greater than 60 years.

3.2.5 Validation of variants

All rare and pathogenic variants identified using our NGS amplicon sequencing were validated using Sanger sequencing. For validation, germline DNA was re-extracted from blood samples, amplified using new primer sets, and Sanger sequenced. 100% of variants identified by NGS were validated successfully. An example of the pathogenic variants identified in this cohort

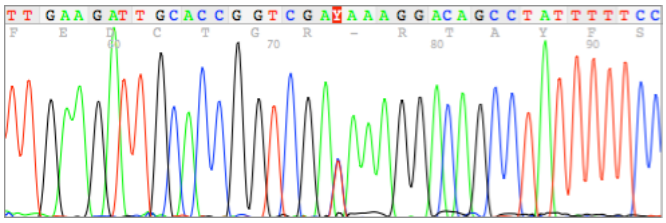
are shown in Figure 3.6. Sanger sequence of all pathogenic, rare missense, and rare splice site variants are shown in Appendix B.

3.2.6 MLPA

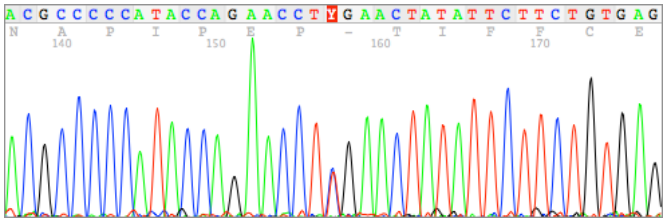
MLPA is a method used for the identification of abnormal copy numbers of DNA sequences. In this study MLPA was used to screen samples for large genomic deletions or duplications of the *CDH1* gene that may have been missed by our PCR based amplicon sequencing approach. MLPA was used to screen samples from the 77 cases who did not carry a clear pathogenic germline *CDH1* variant, and ten samples from healthy controls. Samples from healthy controls were assumed to have a normal copy number and were used as reference samples. No positive control was included as a sample from an individual with a known large genomic deletion or duplication was not available.

The dosage quotient and standard deviation of probes for test and reference sample are shown in Appendix C tables C.1 and C.2. For 8/10 reference samples (80%), the standard deviation of both reference and test probes were less than 0.10 and had DQ values ranging between 0.85-1.15 for both target and reference probes. Two reference samples (Y357 and Y638) showed considerable variation and had probes with a standard deviation larger than 0.10. Because of this, they were excluded from further analysis. For test samples the standard deviation of reference probes ranged between 0.04-0.15. DQ values ranged between 0.76-1.24 and were considered to be normal. Accordingly, no further *CDH1* variants were identified in the case cohort using MLPA.

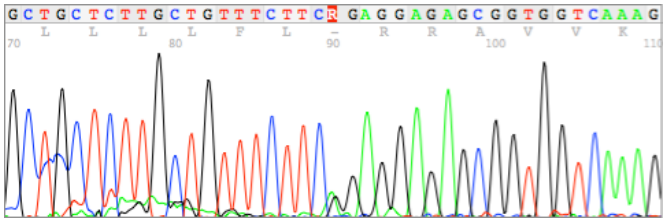
c.190C>T



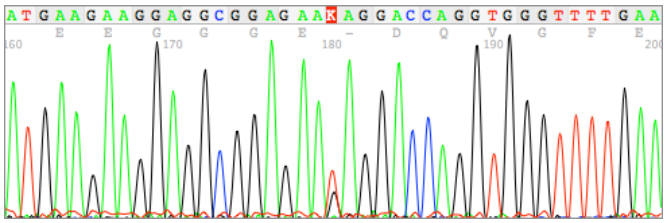
c.1792C>T



c.2195G>A



c.2287G>T



c.2381_2382insC

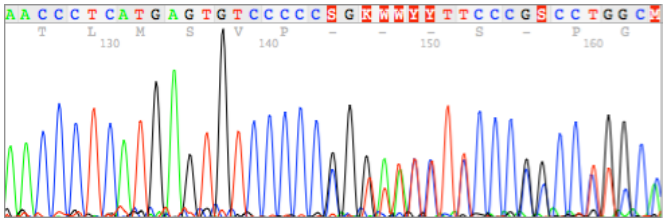


Figure 3.6: Validation of pathogenic *CDH1* variants. Forward orientated chromatograms of the five different pathogenic variants identified in the Māori gastric cancer cohort.

3.2.7 Characteristics of pathogenic *CDH1* variant carriers

Clinical characteristics of the 17 cases carrying a clear pathogenic germline *CDH1* variant were reviewed and are displayed in Table 3.7. Cases were diagnosed between the age of 17-61 years. All cases with pathogenic germline *CDH1* variants were diagnosed with diffuse-type gastric cancer. 15/17 (88.2%) cases were diagnosed with early-stage localised tumours. These 15 pathogenic variant carriers were the same cases as those who were likely diagnosed as a result of HDGC familial cancer screening (Section 3.2.2.2). The remaining two cases with pathogenic variants were diagnosed with late-stage metastatic disease. There was no indication from the pathology reports these latter two cases had *CDH1* pathogenic variants identified at the time of diagnosis. The degree of differentiation and stage of these 17 tumours were mostly unspecified, however tumours for which this information was specified were all described as poorly differentiated (3/17, 17.6%) and were identified in the proximal stomach (3/17, 17.6%).

3.2.8 Frequency of pathogenic *CDH1* variants

Overall, pathogenic *CDH1* variants were identified in 17/94 (18.1%) of the total gastric cancer case cohort and 17/50 (34.0%) of diffuse gastric cancers (Figure 3.7). The proportion of cases with a pathogenic *CDH1* variant that were aged less than 45 years at diagnosis was 14/21 (66.7%). Only 3/73 (4.1%) of cases with pathogenic variants were aged 45 years and over. The average age of diagnosis for pathogenic variant carriers was 33.2 years (range 17-61), and 60.0 years (range 28-81) for sporadic cases.

Case	Age[a]	Gender	Position[b]	Exon	Variant	Designation	Subtype	Differentiation	Grade	Site
Y240	24	Female	68835599	3	c.190C>T	p.Q64*	Diffuse	Local	NOS	NOS
Y382	29	Male	68835599	3	c.190C>T	p.Q64*	Diffuse	Local	NOS	Proximal
Y704	48	Female	68835599	3	c.190C>T	p.Q64*	Diffuse	Local	NOS	NOS
Y647	61	Female	68835599	3	c.190C>T	p.Q64*	Diffuse	Local	NOS	Distal
Y649	20	Female	68855984	12	c.1792C>T	p.R598*	Diffuse	Local	NOS	NOS
Y579	23	Female	68855984	12	c.1792C>T	p.R598*	Diffuse	Local	NOS	NOS
Y709	24	Male	68855984	12	c.1792C>T	p.R598*	Diffuse	Local	Poorly differentiated	Proximal
Y255	29	Female	68855984	12	c.1792C>T	p.R598*	Diffuse	Local	NOS	Proximal
Y616	38	Female	68862107	14	c.2195G>A	p.R732Q	Diffuse	Metastatic spread	Poorly differentiated	NOS
Y435	31	Female	68862199	14	c.2287G>T	p.E763*	Diffuse	Local	Poorly differentiated	Proximal
Y670	39	Female	68862199	14	c.2287G>T	p.E763*	Diffuse	Metastatic spread	NOS	NOS
Y706	41	Male	68862199	14	c.2287G>T	p.E763*	Diffuse	Local	NOS	NOS
Y335	50	Male	68862199	14	c.2287G>T	p.E763*	Diffuse	Local	NOS	NOS
Y638	17	Male	68863643	15	c.2381_2386insC	p.R796fs	Diffuse	Local	NOS	NOS
Y425	20	Male	68863643	15	c.2381_2386insC	p.R796fs	Diffuse	Local	NOS	NOS
Y666	26	Female	68863643	15	c.2381_2386insC	p.R796fs	Diffuse	Local	NOS	NOS
Y386	44	Male	68863643	15	c.2381_2386insC	p.R796fs	Diffuse	Local	NOS	NOS

Table 3.7: Characteristics of cases carrying pathogenic germline *CDH1* variants. [a]Age at time of diagnosis. [b]Reference sequence for variant position is Hg19. ins, insertion; fs, frameshift; NOS, Not Otherwise Specified

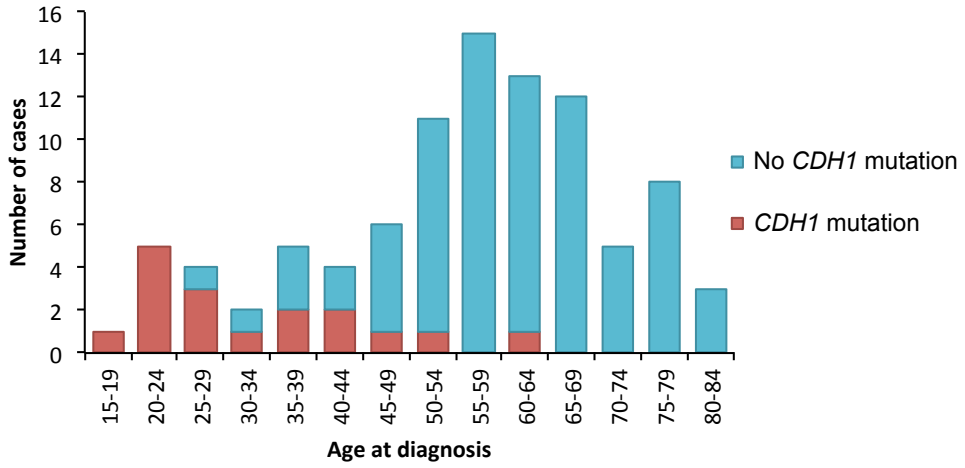


Figure 3.7: Frequency of pathogenic *CDH1* variants by age at diagnosis of gastric cancer in the Māori gastric cancer cohort. Red, *CDH1* pathogenic variant positive; blue, *CDH1* pathogenic variant negative.

3.2.9 The impact of HDGC screening

Our raw data shows that approximately 18% (17/84) and 34% (17/50) of Māori gastric cancer and diffuse gastric cancer cases, respectively, were caused by pathogneic germline *CDH1* variants. However, these figures are skewed by the impact of familial HDGC screening and the incomplete penetrance of pathogenic germline *CDH1* variants, for which without the increased surveillance via endoscopic screening and elected prophylactic surgery available to pathogenic germline *CDH1* variant carriers, it is likely the 15 cases identified as a result of prior genetic screening would have presented with late-stage disease.

Accounting for the length of time Māori HDGC families have been screened for pathogenic germline *CDH1* variants in New Zealand (since the first families were identified in 1998), the number of Māori identified with a pathogenic germline *CDH1* variant between 1998 and 2008, and a penetrance

estimate of 70% between the age of 20-70 years, we estimate that in the absence of familial HDGC screening, pathogenic germline *CDH1* variants would account for 6% (95% CI 0.9-11.3) and 13% (95% CI 2.4-23.9) of all advanced Māori gastric cancer and diffuse gastric cancer, respectively.

3.2.10 Associations between *CDH1* polymorphisms and sporadic gastric cancer

Multiple studies have examined common variants in and around the *CDH1* locus for associations with various cancer types. Notably, associations between common *CDH1* variants and risk for sporadic gastric cancer have been inconsistent. Here, variant calls generated by direct amplicon sequencing to investigate associations between common *CDH1* variants and both sporadic gastric cancer and sporadic diffuse gastric cancer in the Māori population.

Four common *CDH1* variants (MAF > 0.05) were covered by our NGS amplicon sequencing: rs16260, rs28372783, rs3743674, and rs1801552. The rs16260 (-160C>A) and rs28372783 (-74A>C) SNPs are both located in the proximal promoter of the *CDH1* gene. rs16260 has been shown to reduce its transcriptional efficiency of *CDH1 in vitro* while the effect of the rs28372783 SNP is unknown (L. Li et al., 2000). rs3743674 (c.48+6T>C) is located in the donor splice site of intron 1 and is not known to have an effect on the transcription or translation of the *CDH1* gene (Avizienyte et al., 2001). rs1801552 (c.2076T>C) is a synonymous variant (p.Ala692Ala) located in exon 13 of the of *CDH1* gene and is thought to be silent (Risinger et al., 1994).

After removing the 17 cases that carried pathogenic germline *CDH1* variants, there were 77 sporadic gastric cancer cases in this study. These sporadic cases comprised 43 males and 34 females, and had a mean age of 69.1 years (range

28-81 years.) Of the sporadic cases, 34 were diffuse-type, and had a mean age of 54.9 years (range 28-81 years). All 200 sequenced controls were included in the association analysis (previously described in Section 3.2.2). Odds ratios, 95% confidence intervals (CI), and p-values for trend associations for the case-control study were obtained by conditional logistic regression analysis and adjusted for age. Associations between cases and control samples were tested using co-dominant, dominant, recessive, and over-dominant association models.

The frequencies of all SNPs were in Hardy-Weinberg equilibrium for the case and control groups ($p > 0.134$). Overall, there were no significant associations between the four SNPs tested and sporadic gastric cancer risk (Table 3.8) or sporadic diffuse gastric cancer (Table 3.9), for any of the association models tested. The only association to approach significance was for the promoter SNP rs3743674 when tested for an association with sporadic diffuse gastric cancer using a dominant model (OR 2.25 (95% CI 0.98-5.18); $p = 0.064$). Notably, the rs16260 A allele, which has been shown to decrease the transcriptional activity of *CDH1*, was not significantly associated with risk of sporadic gastric cancer or sporadic diffuse gastric cancer in this study ($p > 0.152$).

3.3 Discussion

To our knowledge, this is the first study that has examined the frequency of gastric cancers that are attributable to germline *CDH1* variants in a specific ethnic group. In keeping with previous studies (Biggar et al., 2011; Nevalainen, Laurén, & Gavin, 1988), a high proportion of diffuse gastric cancers was observed in our Māori gastric cancer cohort. Overall, pathogenic

SNP	Model	Genotype	Controls		Cases		Odds Ratio (95% CI)		p-value	Adjusted p-value
			n	(%)	n	(%)				
rs16260	Codominant	CC	135	(67.5)	58	(75.3)	1			
-160C>A		CA	62	(31.0)	18	(23.4)	0.65	(0.35-1.20)		
		AA	3	(1.0)	1	(1.3)	0.73	(0.07-7.23)	0.376	1.000
	Dominant	CC	135	(67.5)	58	(75.3)	1			
		CA-AA	65	(32.5)	19	(24.7)	0.66	(0.36-1.20)	0.163	0.652
	Recessive	CC-CA	197	(98.5)	76	(98.7)	1			
		AA	3	(1.5)	1	(1.3)	0.83	(0.08-8.12)	0.869	1.000
	Overdominant	CC-AA	138	(69.0)	59	(76.6)	1			
		CA	62	(31.0)	18	(23.4)	0.66	(0.36-1.21)	0.170	0.680
rs28372783	Codominant	AA	157	(78.5)	59	(76.6)	1			
-73A>C		AC	40	(20.0)	15	(19.5)	1.01	(0.50-1.91)		
		CC	3	(1.5)	3	(3.9)	3.18	(0.61-16.63)	0.399	1.000
	Dominant	AA	157	(78.5)	59	(76.6)	1			
		AC-CC	43	(21.5)	18	(23.4)	1.11	(0.59-2.10)	0.747	1.000
	Recessive	AA-AC	197	(98.5)	74	(96.1)	1			
		CC	3	(1.5)	3	(3.9)	3.20	(0.62-16.63)	0.176	0.705
	Overdominant	AA-CC	160	(80.0)	62	(80.5)	1			
		AC	40	(20.0)	15	(19.5)	0.94	(0.48-1.84)	0.859	1.000
rs3743674	Codominant	TT	163	(81.5)	69	(89.6)	1			
c.48+6T>C		CT	36	(18.0)	7	(9.1)	0.48	(0.20-1.13)		
		CC	1	(0.5)	1	(1.3)	2.02	(0.12-33.32)	0.175	0.700
	Dominant	TT	163	(81.5)	69	(89.6)	1			
		CT-CC	37	(18.5)	8	(10.4)	0.53	(0.23-1.19)	0.107	0.416
	Recessive	TT-CT	199	(99.5)	76	(98.7)	1			
		CC	1	(0.5)	1	(1.3)	2.21	(0.13-36.43)	0.584	1.000
	Overdominant	TT-CC	164	(82.0)	70	(90.9)	1			
		CT	36	(18.0)	7	(9.1)	0.47	(0.20-1.12)	0.072	0.288
rs1801552	Codominant	CC	65	(32.5)	23	(29.9)	1			
c.2076T>C		TC	105	(52.5)	37	(48.1)	1.05	(0.57-1.93)		
		TT	30	(15.0)	17	(22.1)	1.66	(0.77-3.58)	0.377	1.000
	Dominant	CC	65	(32.5)	23	(29.9)	1			
		TC-TT	135	(67.5)	54	(70.1)	1.19	(0.67-2.11)	0.560	1.000
	Recessive	CC-TC	170	(85.0)	60	(77.9)	1			
		TT	30	(15.0)	17	(22.1)	1.62	(0.83-3.15)	0.165	0.660
	Overdominant	CC-TT	95	(47.5)	40	(51.9)	1			
		TC	105	(52.5)	37	(48.1)	0.87	(0.51-1.47)	0.599	1.000

Table 3.8: Associations between sporadic gastric cancer and common variants at the *CDH1* locus. Allele frequencies of four *CDH1* variants in sporadic gastric cancer cases and control subjects and the odd ratios (ORs) for co-dominant, dominant, recessive and over-dominant models. P-values were corrected using the Bonferroni correction. CI, Confidence Interval; OR, Odds Ratio; SNP, Single Nucleotide Polymorphism.

SNP	Model	Genotype	Controls		Cases		Odds Ratio (95% CI)		p-value	Adjusted p-value
			n	(%)	n	(%)				
rs16260	Codominant	CC	135	(67.5)	26	(78.8)	1			
-160C>A		CA	62	(31.0)	7	(21.2)	0.59	(0.24-1.44)		
		AA	3	(1.5)	0	(0.0)	0	(0.00-0.00)	0.294	1.000
	Dominant	CC	135	(67.5)	26	(78.8)	1			
		CA-AA	65	(32.5)	7	(21.2)	0.56	(0.23-1.37)	0.185	0.740
	Recessive	CC-CA	197	(98.5)	33	(100)	1			
		AA	3	(1.5)	0	(0.0)	0	(0.00-0.00)	0.318	1.000
	Overdominant	CC-AA	138	(69.0)	26	(78.8)	1			
		CA	62	(31.0)	7	(21.2)	0.60	(0.25-1.30)	0.152	0.608
rs28372783	Codominant	AA	157	(78.5)	22	(66.7)	1			
-73A>C		AC	40	(20.0)	9	(27.3)	2.02	(0.853-4.90)		
		CC	3	(1.5)	2	(6.1)	4.64	(0.68-31.49)	0.131	0.524
	Dominant	AA	157	(78.5)	22	(66.7)	1			
		AC-CC	43	(21.5)	11	(33.3)	2.25	(0.98-5.18)	0.064	0.256
	Recessive	AA-AC	197	(98.5)	31	(93.9)	1			
		CC	3	(1.5)	2	(6.1)	3.93	(0.60-25.96)	0.181	0.724
	Overdominant	AA-CC	160	(80.0)	62	(80.5)	1			
		AC	39	(19.5)	9	(27.3)	1.89	(0.78-4.54)	0.167	0.668
rs3743674	Codominant	TT	163	(81.5)	29	(87.9)	1			
c.48+6T>C		CT	36	(18)	4	(12.1)	0.63	(0.20-1.96)		
		CC	1	(0.5)	0	(0.0)	0	(0.00-0.00)	0.653	1.000
	Dominant	TT	163	(81.5)	29	(87.9)	1			
		CT-CC	37	(18.5)	4	(12.1)	0.62	(0.20-1.92)	0.390	1.000
	Recessive	TT-CT	199	(99.5)	33	(100)	1			
		CC	1	(0.5)	0	(0.0)	0	(0.00-0.00)	0.683	1.000
	Overdominant	TT-CC	164	(82.0)	29	(87.9)	1			
		CT	36	(18.0)	4	(12.1)	0.63	(0.21-1.96)	0.411	1.000
rs1801552	Codominant	CC	65	(32.5)	8	(24.2)	1			
c.2076T>C		TC	105	(52.5)	20	(60.6)	1.52	(0.63-3.71)		
		TT	30	(15.0)	5	(15.2)	1.16	(0.34-3.92)	0.620	1.000
	Dominant	CC	65	(32.5)	8	(24.2)	1			
		TC-TT	135	(67.5)	25	(75.8)	1.44	(0.61-3.40)	0.401	1.000
	Recessive	CC-TC	170	(85.0)	28	(84.8)	1			
		TT	30	(15.0)	5	(15.2)	0.88	(0.31-2.50)	0.805	1.000
	Overdominant	CC-TT	95	(47.5)	13	(39.4)	1			
		TC	105	(52.5)	20	(60.6)	1.44	(0.65-1.98)	0.343	1.000

Table 3.9: Associations between sporadic diffuse gastric cancer and common variants at the *CDH1* locus. Allele frequencies of four *CDH1* variants in sporadic diffuse gastric cancer cases and control subjects and the odd ratios (ORs) for co-dominant, dominant, recessive and over-dominant models. P-values were corrected using the Bonferroni correction. CI, Confidence Interval; OR, Odds Ratio; SNP, Single Nucleotide Polymorphism.

germline *CDH1* variants were identified in 17/94 (18%) of the total case cohort and 17/50 (34%) of diffuse gastric cancers. The proportion of cases with a pathogenic *CDH1* variant aged less than 45 years at the time of diagnosis was 14/21 (67%). Only 3/73 (4%) of cases with pathogenic germline *CDH1* variants were aged 45 years and over. No pathogenic germline *CDH1* variants were identified in 200 healthy controls, nor were any associations between common variants at the *CDH1* locus and sporadic gastric cancer. After adjusting for the effect of *CDH1* screening, an estimated 6% of all advanced gastric cancer and 13% of all advanced diffuse gastric cancers carry pathogenic germline *CDH1* variants . This study demonstrates that inherited genetic predisposition to gastric cancer is an important cause of cancer, and a major contributor to the high incidence of early-onset diffuse gastric cancer in the Māori population.

3.3.1 The impact of familial gastric cancer screening

In New Zealand, familial gastric cancer genetic screening has been offered since 1998 when pathogenic germline *CDH1* variants were first identified in three Māori kindred (Guilford et al., 1998). For pathogenic variant carriers, endoscopic screening and total prophylactic gastrectomy have been available since the early 2000s. Some pathogenic variant carriers elect to undergo regular endoscopic screening to identify early-stage foci before the cancer progresses. However, foci can be small and hard to identify making screening imperfect (van der Post et al., 2015). Currently, total prophylactic gastrectomy is the only option available to completely eliminate the risk associated with an inherited pathogenic *CDH1* variant. While the mortality rate associated with total prophylactic gastrectomy is approximately 1%, there are still major morbidities associated with this procedure (van der Post et al., 2015).

The impact of prior *CDH1* variant screening was apparent in the case cohort in this study. Almost all of the cases (15/17; 88.2%) with pathogenic *CDH1* variants were identified as a part of the clinical pathway for HDGC variant carriers. Interestingly, the two cases which did not appear to know about their pathogenic *CDH1* variants (Y616 and Y670), both presented with advanced metastatic disease. Y670 carried a known HDGC variant (c.2287G>T) that has previously been documented in New Zealand (Charlton et al., 2004). As *CDH1* c.2287G>T had been identified in a Māori kindred from New Zealand prior to this study, it is a tragedy that this case was not diagnosed sooner. It is unclear if case Y670 knew about their risk as a part of a known HDGC family and had declined screening, or did not know about their risk of being a pathogenic variant carrier. Y616 carried a missense c.2195G>A pathogenic variant, which to our knowledge, has not previously been documented in New Zealand, but has been documented in a HDGC families of European decent (Kaurah et al., 2007; Brooks-Wilson et al., 2004).

Our data also highlights the importance of the identification and proper clinical management of those with pathogenic *CDH1* variants. The 15 cases that were identified as a result of genetic screening were all still alive five years post diagnosis, while the two cases that were not identified, both died shortly after diagnosis (data not shown). It is almost certain that without intervention the majority of cases that were identified as a result of genetic testing would have presented with advanced diffuse gastric cancer and died. Although there is risk of missing early lesions with endoscopic screening, and there is a high level of morbidity associated with total prophylactic gastrectomy, as seen here, the outcomes for both procedures are typically better than the prognosis associated with the late detection of tumours.

3.3.2 *CDH1* c.2195G>A: a new HDGC variant in New Zealand

Of particular interest is case Y616, who was found to carry a pathogenic *CDH1* c.2195G>A variant. Due to the lack of cancer diagnoses in their first degree relatives, and no reference to any of the procedures normally available to pathogenic *CDH1* variant carriers, it appears this case did not know they were a pathogenic *CDH1* variant carrier. The apparent lack of family history in this family could be explained by two main hypotheses. The first is that the c.2195G>A variant has a low penetrance in this family. The second is that the variant recently arose *de novo* in this individual or a recent ancestor.

If the *CDH1* c.2195G>A variant was present in the family of Y616 for multiple generations and has been shared by a large number of family members, there is a significant deviation from the expected presentation of disease in this family. Current estimates of the cumulative risk of developing diffuse gastric cancer by the age of 80 years is 70% for men (95% CI 59%-80%) and 56% for women (95% CI 44%-69%) (Hansford et al., 2015). However, previous estimates of disease penetrance for variant carriers have ranged between 40% and 83%, depending on gender and ethnicity (Kaurah et al., 2007; Pharoah, Guilford, Caldas, International Gastric Cancer Linkage Consortium, & International Gastric Cancer Linkage Consortium, 2001). Evaluation of *CDH1* variant penetrance in large families from more recent studies are lower (Kaurah et al., 2007), suggesting that family ascertainment for HDGC may have skewed early estimates. The vast majority of individuals with truncating *CDH1* variants had a family history of diffuse gastric cancer, however as sequencing is becoming more freely available, families that deviate away from the expected pattern of disease are emerging. Recently, a series of studies have reported pathogenic *CDH1* variants in individuals with no family history of

diffuse gastric cancer (Huynh & Laukaitis, 2016; Lajus & Sales, 2015; Xie et al., 2011).

Alternatively, the c.2195G>A variant could have arisen *de novo* in case Y616 or a very recent ancestor, explaining why there was no history of disease in the family. Haplotype analysis of the two European families with the same c.2195G>A variant found that the three haplotypes markers that were most tightly linked to the c.2195G>A variant were identical, however, the two markers flanking each side of these markers were different (Kaurah et al., 2007). Because the variant could either have arisen independently in these two families, or as an ancient variant with differing haplotypes as a result of two separate recombination events, the results from this analysis were inconclusive (Kaurah et al., 2007).

Already, three variants (c.1137G>A, c.1565+1G>A, and c.1792C>T) have been found in multiple families from seemingly unrelated backgrounds and have been proposed as *CDH1* variation hotspots (Hansford et al., 2015). Strikingly, including the c.2195G>A variant, all four of these hotspot sites are located in cytosine-guanine (CpG) dinucleotide sites. CpG dinucleotides are known to be a hotspot for variants in the human genome. This hypermutability is related to its role as the major site of cytosine methylation, with the attendant risk of spontaneous deamination of 5-methylcytosine (5mC) to thymine (Cooper & Youssoufian, 1988). Recurrence of germline variants at CpG sites have been reported for a number of genetic syndromes. For example, germline variants in the methyl-CpG binding protein 2 gene *MECP2* is causative of Rett syndrome, a neurodevelopmental disorder characterised by loss of acquired skills after a period of normal development in infant girls, are most common in CpG hotspots (Wan et al., 1999). The identification of *CDH1* c.2195G>A in our geographically and ethnically distant cohort of New Zealand Māori,

supports the notion that *CDH1* c.2195G>A a *de novo* variant in a variant hotspot. Regardless of the history of this variant, the family of Y616 can now be screened for this pathogenic variant.

3.3.3 A high number pathogenic *CDH1* variants in the Māori gastric cancer population

Founder variants have been identified as a common cause of cancer in some populations. Of note is the Ashkenazi Jewish population, for which approximately 2% of the general population carry one of three founder variants in the tumour suppressor genes *BRCA1* and *BRCA2* (Levy-Lahad et al., 1997). Deleterious variants in *BRCA1* and *BRCA2* are associated with an increased risk of both breast and ovarian cancer (King, Marks, & Mandell, 2003). Subsequently, approximately 12% of breast cancers (Warner et al., 1999) and 29-40% (Modan et al., 2001; Moslehi et al., 2000) of ovarian cancers in the Ashkenazi Jewish population are attributable to these specific variants.

Similarly, a founder variant in *CDH1* has previously been linked to an increased incidence of gastric cancer in Newfoundland Canada (Kaurah et al., 2007). Interestingly, Newfoundland has an elevated incidence of gastric cancer compared to the Canadian average and the regions these families come from are the highest-risk areas within the province (Kaurah et al., 2007; McLaughlin et al., 2006). As of yet, the overall contribution of pathogenic *CDH1* variants to the high incidence of gastric cancer has not been determined.

It is unclear why the prevalence of pathogenic *CDH1* variants is so high in the Māori gastric cancer population. Similar to the common variant seen in Newfoundland, pathogenic *CDH1* variants could have arisen as founder variants prior to the Māori migration to New Zealand. However, the relatively

high number of distinct *CDH1* variants (5 pathogenic variants in this study alone) suggests that rather than being an illustration of a simple genetic bottleneck, it is possible *CDH1* variants may have provided a selective advantage to pathogenic variant carriers in ancestral Māori populations.

One possible explanation is that some *CDH1* variant carriers may have some innate resistance to infection with *Listeria monocytogenes* (*L. monocytogenes*), a food-born pathogen that can cause gastroenteritis, meningitis, and miscarriage in pregnant women (da Silva Tatley, Aldwell, Dunbier, & Guilford, 2003; Hamon, Bierne, & Cossart, 2006). The bacteria is normally internalised into epithelial cells by a process requiring the binding of the bacterial protein internalin-A (*InlA*) to the N-terminus of the E-cadherin protein (Hamon et al., 2006). Notably, some truncating E-cadherin variants produce short soluble N-terminal peptides containing the InlA binding site, which have been shown *in vivo* to act as decoy receptors for invading *L. monocytogenes* (da Silva Tatley et al., 2003). Additionally, previous work has shown *L. monocytogenes* takes advantage of junctional remodelling and exposed E-cadherin to adhere and invade epithelial cells (Pentecost, Otto, Theriot, & Amieva, 2006). Conceivably, deleterious *CDH1* variant carriers may have less functional E-cadherin available for bacterial adherence and subsequent invasion. Alternatively, some *CDH1* variants may interfere with the dynamics of *L. monocytogenes* internalisation by misregulating the dynamics of the actin cytoskeleton necessary for actin assembly at the bacterial entry site.

3.3.4 Conclusions

In conclusion, these results show pathogenic germline *CDH1* variants are a major contributor to the high incidence of early-onset diffuse gastric cancer

seen in the Māori population. The importance of screening for predisposition to gastric cancer was also apparent in this study group. We speculate that some *CDH1* variants could have provided an evolutionary advantage, which may explain the high number of pathogenic variants seen in this cohort. Routine sequencing of germline *CDH1* should be considered for all Māori who present with diffuse gastric cancer to enable genetic testing and early intervention in their wider family.

Chapter 4

Pathogenic variants in germline *CDH1* as a cause of gastric cancer in Chile

4.1 Introduction

Chile is a country with a high incidence of gastric cancer. In this country, a high prevalence of environmental risk factors is causing a high number of early-onset and familial gastric cancer cases in Chile. Consequently, the suggestion that some cases may be hereditary is often ignored. Currently, there is no formal clinical genetic screening available for gastric cancer patients in Chile. Despite the lack of screening, hereditary gastric cancers are expected to explain a small but important portion of cases. Our aim was to identify pathogneic variants in germline *CDH1* carried by early-onset and familial gastric cancer patients from Chile and highlight the importance of genetic screening in high-incidence countries.

4.1.1 Chile

Occupying the narrow strip of land between the Andes mountains and the Pacific Ocean, Chile is one of South America's largest countries. Chile has been populated by several tribes of Native Americans since 3000 B.C. (Collier & Sater, 2004). Chile was conquered and colonised by Spain in the 16th century, but later attained its independence in 1818 (Collier & Sater, 2004). Currently, Chile has a population of approximately 18 million people, 40% of whom are located in Chile's capital city, Santiago (de la Jara et al., 2015).

There are three main ethnic groups in Chile: European, Mestizo, and Mapuche (Fernández, 2005). European are the largest ethnic group, accounting for 59% of the total population. The European group is primarily made up of individuals with Croatian, French and Italian ancestry, but also includes those of German, English, and Polish decent. Mestizo are the second largest ethnic

group and account for 25% of the population. Mestizo includes those with mixed European and Amerindian heritage. The Mapuche are third largest ethnic group and are the main indigenous population of Chile. Approximately 9% of the population identify as Mapuche, the majority of whom are located in the south-central region of the country. Collectively, Mapuche refers to a wide-ranging ethnicity composed of various groups that share a common social, religious and economic structure (Crow, 2013). The remaining 7% of the Chilean population is made up of several small ethnic groups, namely Africans, Aymara, Rapa Nui, Likan Antai, Colla, Yagan, Kawesqar, and Quechua (Fernández, 2005).

4.1.2 Gastric cancer in Chile

In Chile, an estimated 35,000 people are diagnosed with cancer every year (Goss et al., 2013; de la Jara et al., 2015). The age-adjusted incidence rates of cancer are 226.1 and 180.0 per 100,000 population, for males and females, respectively (Vallebuona et al., 2011). Notably, the burden of cancer related mortality in Chile has grown consistently for the past several decades: 8.4% in 1960, 12.1% in 1970, 15.8% in 1980, and 18.1% in 1990 (de la Jara et al., 2015). In 2011, approximately 24% of deaths in Chile were cancer related (de la Jara et al., 2015).

Gastric cancer is of particular concern for the Chilean population. The IARC estimated incidence of gastric cancer in Chile is 15.6 per 100,000 population, much higher than most countries worldwide (Ferlay et al., 2015). For example, the IARC estimated incidence of gastric cancer in New Zealand is just 2.9 per 100,000 population (Ferlay et al., 2015). Additionally, gastric cancer is the main cause of cancer related mortality in Chile. In 2012, an estimated 3,371 people

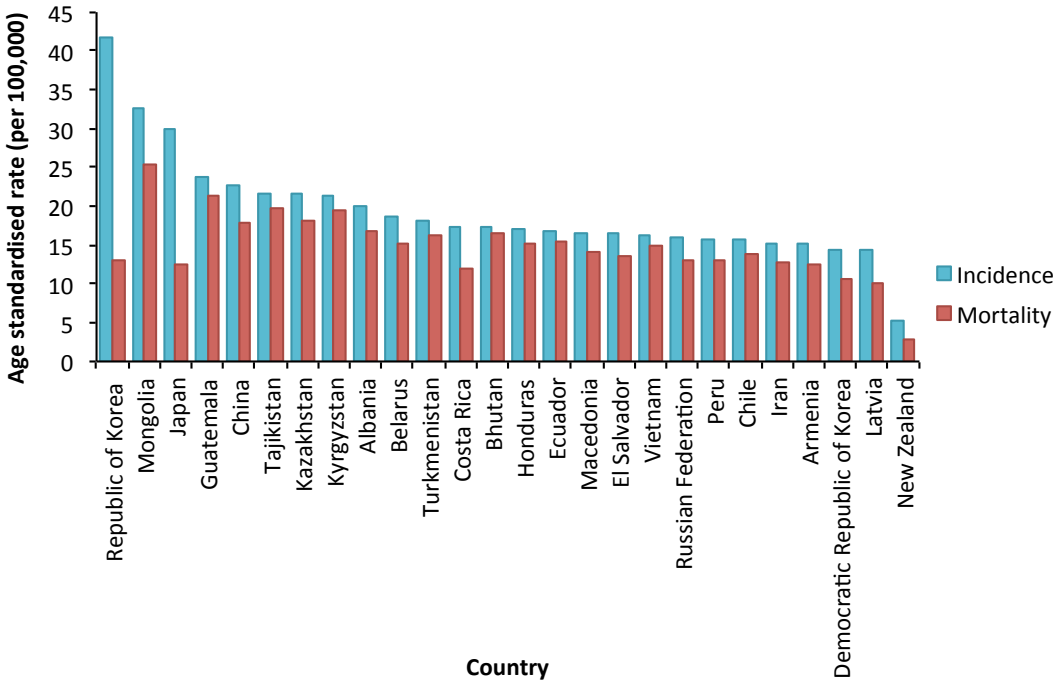


Figure 4.1: Countries with high gastric cancer incidence and mortality rates. The 25 countries with the highest IARC estimated gastric cancer incidence and mortality rates. A low incidence country, New Zealand, is shown for comparison. Data acquired from: https://www.iarc.fr/.

died as a results of this disease, a rate of 13.8 per 100,000 population (Ferlay et al., 2015). Strikingly, there was a significant increase in the number of diffuse-type tumours diagnosed in Chile between 1986-1995 and 1996-2005 (46% versus 62%, respectively, $p < 0.001$) (Tapia et al., 2010).

Similar to other indigenous populations groups, the Mapuche people are known to have a higher incidence of gastric cancer than the general population (Caglevic, Silva, Mahave, Rolfo, & Gallardo, 2016). The age-standardised incidence of gastric cancer for Mapuche has been reported to be as high as 47.2 and 28.1, per 100,000 males and females, respectively (Heise, Bertran, Andia, & Ferreccio, 2009). Higher mortality rates have been observed in the south-central region of Chile where the Mapuche people historically settled and diffuse-type tumours account for 55% of tumours (Heise et al., 2009; Tapia et al., 2010).

4.1.2.1 High exposure to environmental risk

As previously described in Section 1.1.2, there are a series of environmental risk factors that influence gastric cancer risk. High levels of a number of these risk factors are thought to be contributing to the high rates of gastric cancer in Chile (de la Jara et al., 2015).

H. pylori infection is recognised as the principal risk factor for gastric cancer in Chile (Caglevic et al., 2016). The prevalence of *H. pylori* infection is reported to be as high as 73% in adults and 18% in children (Jaime, Villagrán, Serrano, Cerda, & Harris, 2013). Diets high in salt are thought to facilitate *H. pylori* infection and contribute to the high incidence of disease. The average adult salt intake in Chile is estimated to be 10.4 grams per day, almost double the WHO recommended intake of 5.0 grams per day, and significantly higher than British adults who, on average, consume 8.1 grams per day (Crovetto & Uauy, 2013; López-Rodríguez, Galván-García, & Muzzo, 2009; Sadler et al., 2011). It is suggested *H. pylori* and a high salt intake may synergise to promote the development of gastric cancer (Caglevic et al., 2016).

Obesity is another risk factor thought to contribute to the high incidence of gastric cancer in Chile. It is estimated that up to 20% of all cancers in Chile are obesity related (de la Jara et al., 2015). Recent statistics show that 67% of the Chilean population are overweight and 25% are obese (de la Jara et al., 2015). Moreover, it is estimated that 88% of the population lead a sedentary lifestyle that contributes to obesity (de la Jara et al., 2015).

4.1.2.2 Screening for gastric cancer in Chile

Chile has taken steps to reduce the burden of gastric cancer by introducing a gastric cancer screening programme. In 1995, the Chilean Ministry of Health started an endoscopic-based screening pilot programme in Santiago. For the 1996–2006 period, the pilot screened 10,284 individuals, during which time 190 gastric cancers were identified (8.5 per 100 000 population, per year). On average, 70 gastroscopies were required to identify one gastric cancer. One third of these cancers were early stage and the average 5-year survival rate of those diagnosed with gastric cancer via screening was 40%. In 2006, the Ministry of Health initiated a nationwide gastric cancer detection programme that focused on symptomatic individuals. It guaranteed endoscopic examination for any patient older than 40 years of age that experienced epigastric pain lasting more than 15 days, heavy bleeding, anaemia, weight loss of unknown origin, general feeling of weakness, tiredness, loss of appetite, or dysphagia. History of gastrectomy and family history of gastric cancer are also taken into consideration but do not automatically lead to screening. Additionally, unless endoscopic examination is normal, *H. pylori* eradication is recommended for individuals with duodenal or stomach ulcers, atrophic gastritis, lymphoma, gastric cancer, or a family history of gastric cancer. Genetic testing is not a part of standard care for those that appear to have a strong family history or early-onset of disease.

4.1.3 Hereditary gastric cancer in Chile

During the last decade there has been increased awareness of the importance of identifying individuals with inherited cancer syndromes. Chile is a country with good health standards and reliable services, but does not have an

extensive clinical genetics programme (Castillo Taucher, 2015). In 2015, there were 30 clinical geneticists in Chile, the majority of which resided in Santiago (Castillo Taucher, 2015). Based on the WHO recommendation of having one clinical geneticist per 100,000 population, Chile should have had approximately 170 clinical geneticist's at this time (Castillo Taucher, 2015).

Pathogenic variants in *CDH1* are known to predispose to hereditary gastric cancer and may be responsible for some of early-onset and familial gastric cancer cases in Chile. To our knowledge, prior to this study, there have not been any pathogenic variants in germline *CDH1* in gastric cancer patients from Chile. Identification of pathogenic variants in *CDH1* in high-risk individuals will improve awareness of hereditary gastric cancer syndromes and highlight the importance of genetic testing HDGC families in Chile.

4.2 Results

In this chapter, next-generation amplicon sequencing was used to identify variants in germline *CDH1* in a series of Chilean gastric cancer probands with either a striking early-onset and/or family history of gastric cancer. Germline *CDH1* variants were evaluated, and rare and potentially pathogenic variants were validated using Sanger sequencing. I then screened the family of probands with pathogenic *CDH1* variants to determine carrier status.

4.2.1 Characterisation of Chilean cohort

Germline *CDH1* was sequenced in 51 Chilean probands diagnosed with gastric cancer between 2006 and 2017. Study inclusion criteria consisted of

having early-onset gastric cancer or a striking family history of gastric cancer. DNA samples and clinical details were collected by the referring institutions.

Clinical details from the proband cohort are summarised in Table 4.1. The cohort consisted of 27 males (53%) and 24 females (57%). The average age of diagnosis was 47.8 years (range 17-80 years). The majority of probands (74.5%) were diagnosed with diffuse gastric cancer, while a small proportion were diagnosed with intestinal-type cancer (21.6%), and no subtype was recorded for two probands (3.9%). Tumour stage was defined according to TNM classification (Strong, D'Amico, Kleinberg, & Ajani, 2013) and were near evenly distributed across the four stages. Two thirds of cases (66.7%) were poorly differentiated. Tumour location was not recorded.

The history of cancer in the proband's family was also collected and is summarised in Table 4.2. Family history data included the cancer type and age at diagnosis for first, second and third degree relatives. Many of the probands had a striking family history of gastric cancer and/or breast cancer. Four probands are a part of families with five or more extended family members that had been diagnosed with gastric cancer. Eight probands had one or more family members diagnosed with breast cancer, four of whom also had multiple gastric cancers diagnosed in their families. Remarkably, one proband in the Chilean study cohort had an additional eight family members who have been diagnosed with gastric cancer.

Unfortunately the subtype of gastric and breast cancers in the extended family of probands were not recorded. Because of this, I was unable to determine if many of these probands would have met the clinical criteria for HDGC testing (van der Post et al., 2015). Twenty-eight probands were diagnosed with diffuse gastric cancer less than 40 years of age, or were diagnosed with diffuse gastric cancer and had one or more family members who had been diagnosed with

	Total 51 (n)	(%)
Gender		
Male	27	52.9
Female	24	47.1
Age at diagnosis (years)		
<45	21	41.2
45-59	19	37.3
60-74	8	15.7
$\geq$ 75	3	5.9
Tumour subtype		
Diffuse	38	74.5
Intestinal	11	21.6
NOS	2	3.9
Tumour stage		
I	13	25.5
II	9	17.6
III	13	25.5
IV	16	31.4
Tumour grade		
Well differentiated	4	7.8
Moderately differentiated	13	25.5
Poorly differentiated	34	66.7

Table 4.1: Characteristics of Chilean gastric cancer probands. The stage of tumours are defined by TNM staging. Tumours graded as poorly differentiated include signet ring cell carcinomas. NOS, Not otherwise specified.

gastric cancer, fitting the most recent clinical criteria for HDGC. The other 23 probands in this study had a family history of gastric and/or breast cancer, however without the subtypes of the cancers diagnosed in the proband's extended family, the probands could not be accurately compared to the HDGC testing criteria.

4.2.2 Sequencing libraries and data

DNA samples from Chilean probands arrived in small groups over three years. Consequently, despite a relatively small number of samples, to ensure

	Total 51 (n)	(%)
Number of family members with gastric cancer		
1	20	39.2
2	10	19.6
3	15	29.4
4	2	3.9
≥ 5	4	7.8
Number of gastric cancers diagnosed < 50 years of age		
0	7	13.7
1	35	68.6
2	7	13.7
3	2	3.9
Number of family members diagnosed with breast cancer		
0	43	84.3
1	7	13.7
2	1	2.0
Proband meets clinical criteria for HDGC testing		
Yes	28	54.9
No	23	45.1

Table 4.2: Summary of family history data collected for Chilean probands. The number of family members with gastric cancer and the number of gastric cancers diagnosed under 50 years of age include the study proband. The 2015 IGCLC guidelines are were used to define HDGC testing criteria (van der Post et al., 2015).

timely testing of these samples, sequence data was generated across multiple MiSeq runs and by Sanger sequencing. Sequencing libraries for these samples were prepared and tested the same way as previously described in Section 3.2.3. Sequencing libraries were included in MiSeq runs five, six, seven, and eight (previously described in Chapter 3 Table 3.3). PCR products for Sanger sequencing were prepared using primers listed in Appendix A.4.

4.2.3 *CDH1* variants

In total, 11 different germline *CDH1* variants were identified across the 51 Chilean probands (Table 4.3). Of these, eight variants were located in the coding exons of *CDH1* (six synonymous, one missense, and one nonsense),

two variants were located in the proximal promoter, and one was located in an intronic splice site. Table 4.3 shows the position of each variant, the minor allele frequency in this series of probands, and allele frequencies from major population databases. To assess the clinical significance of variants identified in this cohort, all rare non-synonymous variants (MAF < 0.05) were reviewed. The effect of missense variants were predicted as previously described in Section 3.2.4.

4.2.3.1 Nonsense variants

c.1531C>T

Proband 6 was found to carry a *CDH1* c.1531C>T variant. This variant causes the premature termination of the E-cadherin protein by changing a glutamine residue to a stop codon at codon 511 (p.Gln511*). Proband 6 was was diagnosed with advanced diffuse gastric cancer aged 23 years. Proband 6 did not have a family history of gastric cancer or lobular breast cancer. *CDH1* c.1531C>T has reported as pathogenic variant in ClinVar.

4.2.3.2 Missense variants

c.88C>A

Proband 17 was was found to carry a *CDH1* c.88C>A variant. The proband was diagnosed with advanced diffuse gastric cancer aged 59 years. The probands father and paternal grandfather were also diagnosed with gastric cancer, aged 65 and 51 years, respectively. *CDH1* c.88C>A causes a proline to threonine substitution at codon 30 of the E-cadherin protein (p.Pro30Thr). The proline residue is weakly conserved between species and there is only a small physiochemical difference between proline and threonine. ConDel predicts

Position[a]	SNP ID	Region	Class	Variant	Protein	Probands (Het/Hom)	MAF Probands	MAF ExAc	MAF 1000 Genomes	MAF ESP6500
68771034	rs16260	Promoter	.	c.-285C>A	.	21/9	0.3824	NA	0.2356	NA
68771122	rs28372783	Promoter	.	c.-197A>C	.	11/1	0.1274	NA	0.0463	NA
68771372	rs3743674	Intron 1	Splice	c.48+6C>T	.	14/38	0.8627	0.8111	0.7634	0.8603
68772239	rs139866691	Exon 2	Missense	c.88C>A	p.Pro30Thr	1/0	0.0098	0.0009	0.0004	0.0010
68847350	rs61756284	Exon 9	Synonymous	c.1269A>T	p.Thr424Thr	1/0	0.0098	0.0008	0.0018	0.0008
68849628	.	Exon 10	Nonsense	c.1531C>T	p.Gln511*	1/0	0.0098	NA	NA	NA
68856085	.	Exon 12	Synonymous	c.1893A>T	p.Thr631Thr	1/0	0.0098	NA	NA	NA
68856088	rs33969373	Exon 12	Synonymous	c.1896C>T	p.His632His	1/0	0.0098	0.0114	0.0228	0.0245
68857417	rs764379691	Exon 13	Synonymous	c.2052C>T	p.Ser684Ser	1/0	0.0098	0.0001	NA	NA
68857441	rs1801552	Exon 13	Synonymous	c.2076T>C	p.Ala692Ala	28/17	0.6078	0.6548	0.7161	0.7161
68862165	rs33964119	Exon 14	Synonymous	c.2253C>T	p.Asn751Asn	4/0	0.0392	0.0396	0.0545	0.0376

Table 4.3: Variants identified in 51 Chilean gastric cancer probands. [a]Reference sequence for variant position is Hg19. MAF, Minor Allele Frequency; Het, Heterozygote; Hom, Homozygote.

c.88C>A is deleterious. *CDH1* c.88C>A is present at low frequencies in ExAc (MAF 0.00094), 1000 Genomes (MAF 0.0004), and ESP6500 (MAF 0.0010) variant databases. ClinVar most frequently classifies this variant as likely benign. Notably, this variant was also identified in one healthy control in the Māori cohort (Section 3.2.4.3).

4.2.4 Frequency of pathogenic *CDH1* variants in probands

Proband 6 was the only proband found to carry a pathogenic *CDH1* variant. As a patient diagnosed with diffuse gastric cancer aged 23 years, Proband 6 fits the criteria for *CDH1* screening (van der Post et al., 2015). With the possible exception of Proband 17, none of the other 28 probands fitting *CDH1* testing criteria, or the other 23 probands who lacked enough detailed family history to apply the clinical criteria for HDGC, carried a pathogenic *CDH1* variant. Overall, patogenic variants in germline *CDH1* were identified in 3.6% (1/28) of probands who met HDGC testing criteria and 2.0% (1/51) of all probands in this study.

4.2.5 Inheritance of *CDH1* c.1531C>T (p.Gln511*) in the family of Proband 6

A further eight family members of proband 6 were sequenced to determine carrier status of the pathogenic c.1531C>T (p.Gln511*) variant. Of these eight family members, five were found to carry the same pathogenic variant (Figure 4.2a). The proband's sibling (IV4; age 19 years), mother (III4; age 45 years), maternal cousin (IV1; age 20 years), maternal aunt (III2; age 44 years), and maternal grandmother (II7; age 66 years), were all found to carry the c.1512C>T variant (Figure 4.2b). The proband's maternal grandfather (II6;

aged 70 years), father (III5; age 44 years), and maternal cousin (IV2; age 23 years) were also tested, but were homozygous for the reference allele (Figure 4.2c). Chromatograms showing the variant site for all eight family members are shown in Appendix E. Variants were also independently validated by clinical services in Chile.

Further members of proband 6's extended family could also carry the pathogenic variant (Figure 4.2a). The half-sibling of the proband (IV3) has a 50% chance of carrying the pathogenic variant but is currently too young for genetic testing. If the c.1512C>T variant was inherited from either of the the proband's maternal great-grandparents (I3 and I4), the maternal grand-aunts and grand-uncles of the proband (I8-I13) may also be at risk. Both maternal great-grandparents are deceased and died of unrelated causes. I3 died as a result of respritory problems aged 75 years, while I4 died as a result of bladder intervention age unknown.

The five probands that were found to carry the pathogenic *CDH1* c.1512C>T variant are now receiving counselling.

4.3 Discussion

Chile has some of the world's highest gastric cancer incidence and mortality rates. High levels of environmental risk are largely to blame, while the possibility of genetic predisposition is generally ignored. In this study, next-generation amplicon sequencing was used to sequence germline *CDH1* in a series of 51 Chilean gastric cancer probands with striking early-onset or familial gastric cancer, 28 of whom fit the clinical criteria for HDGC. Overall, I identified one proband with a clear pathogenic *CDH1* variant. Screening of an

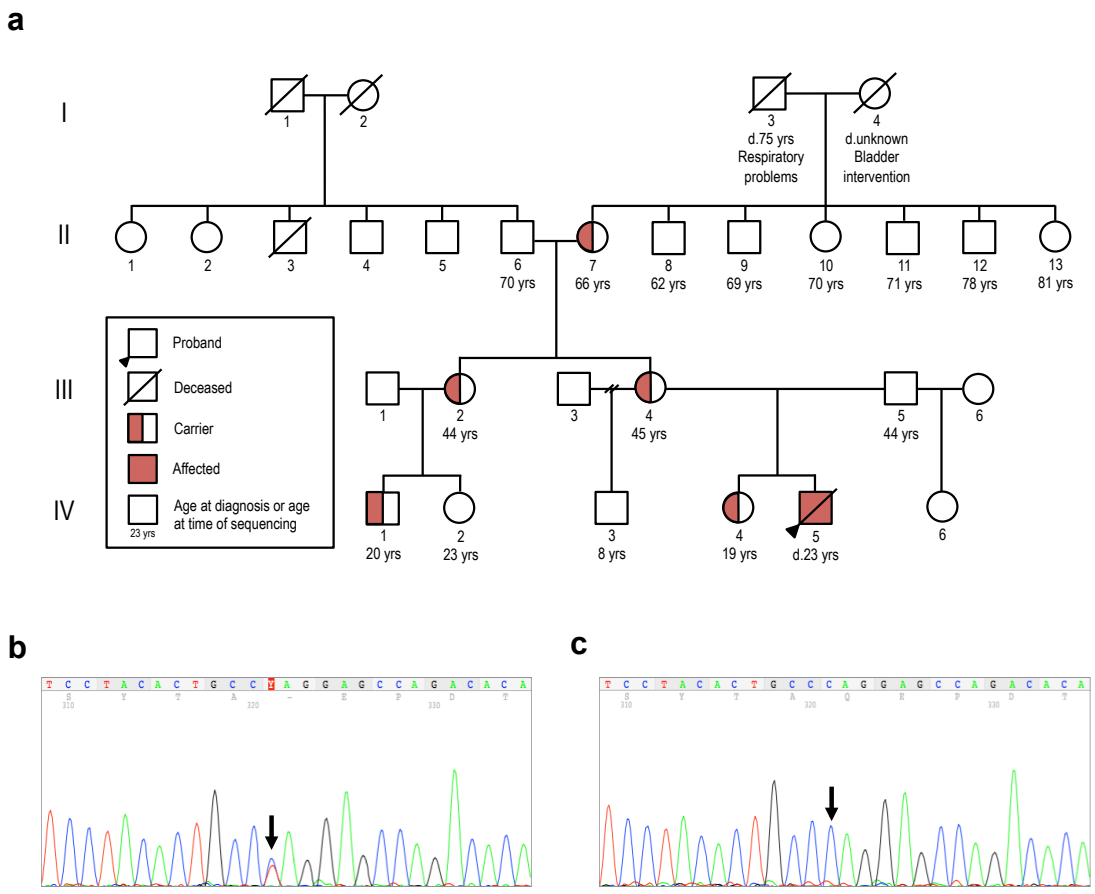


Figure 4.2: Analysis of proband 6 family members. (a) Family tree showing the extended family of Proband 6. (b) Exemplar Sanger sequence of the c.1531C>T variant. (c) Exemplar Sanger sequence of homozygous reference allele. Both Sanger sequences are shown in the forward direction.

additional eight family members of the affected proband identified a further five pathogenic variant carriers. To our knowledge, this is the first pathogenic *CDH1* variant to be reported in Chile and one of a handful of HDGC families with a confirmed pathogenic *CDH1* variant to be identified in South America.

4.3.1 A low rate of variant detection

The frequency of cases that fulfil HDGC testing criteria and carry a pathogenic variant in germline *CDH1* varies between populations with high or low gastric cancer incidence (Oliveira, Seruca, & Carneiro, 2009). To illustrate this, Oliveira, Seruca, and Carneiro (2009) analysed the frequency of pathogenic *CDH1* variants in families fulfilling the criteria for HDGC from regions with low, moderate, and high incidence of gastric cancer. Overall the frequency of germline *CDH1* alterations in countries with a low-incidence of gastric cancer was approximately 40% in HDGC families, and 20% in isolated patients with diffuse gastric cancer under the age of 35 years. In contrast, in moderate and high-incidence countries, the frequency of pathogenic *CDH1* variants was less than 20% both in HDGC families and isolated patients with diffuse gastric cancer less than 35 years of age (Oliveira, Seruca, & Carneiro, 2009). The detection rate of pathogenic variants in germline *CDH1* in HDGC and early-onset diffuse gastric cancers from the high-incidence countries of Portugal, Japan, and South Korea, range between 8% and 15% (S. Kim et al., 2013; Oliveira, Ferreira, et al., 2004; Yamada et al., 2011).

In the current study, 3.6% (1/28) of probands whom meet the 2015 clinical criteria for HDGC (van der Post et al., 2015) and 2.0% (1/51) of all probands in this study were found to carry a pathogenic pathogenic variant in germline *CDH1*. The frequency of pathogenic variants reported here are noticeably

lower than those published in the aforementioned high-incidence populations. There are a series of reasons that may explain this difference. Firstly, although large deletions in *CDH1* are rare (Oliveira, Senz, et al., 2009), I can not exclude the possibility that some of these probands carry a large deletion in *CDH1* which may have been missed by our amplicon sequencing. Secondly, the 2015 criteria used to define HDGC cases in this study are broader than those used in earlier studies. By broadening the clinical criteria for HDGC, probands were included in this analysis whom would not have been included if an older version of HDGC criteria were used. For example, if the 2010 HDGC *CDH1* screening criteria (Fitzgerald et al., 2010) were used to define those tested, 6.7% (1/15) of probands would carry a pathogenic *CDH1* variant. Thirdly, 11 probands in this study were diagnosed with intestinal-type gastric cancer. Intestinal-type gastric cancer is not a part of HDGC and is not caused by pathogenic *CDH1* variants. These intestinal-type cases do not affect the frequency of pathogenic variants in probands that fit HDGC testing criteria in this study, however if all 11 intestinal-type cases were removed from the overall cohort, 2.5% (1/40) of all probands in this study would carry a pathogenic variant.

4.3.2 The importance of genetic screening in high-incidence countries

Pathogenic *CDH1* variant carriers have an extreme risk of developing diffuse gastric cancer and lobular breast cancer. Consequently, the identification of *CDH1* variants in probands and subsequently in their relatives, is of great importance. Unfortunately, *CDH1* variants have incomplete penetrance, making it hard to distinguish hereditary cancer families from sporadic early-onset cases or familial clusters without genetic testing.

It is thought that environmental factors are driving the overall high incidence of gastric cancer in Chile. Accordingly, environmental factors are likely causing some early-onset cases and sporadic family clusters, which mask true hereditary gastric cancer families. Until the prevalence of environmental factors is reduced, the detection of true hereditary gastric cancer families will be difficult and will return high rates of negative results. Despite this, I stress the importance of screening individuals that meet HDGC testing criteria. As seen in this study, by identifying one individual with a pathogenic *CDH1* variant, I was able to screen and identify an additional five family members that are at an extreme risk of diffuse gastric cancer. These family members are now receiving additional targeted care to mitigate disease risk and will undoubtedly have a better prognosis than if they were to present with late-stage disease.

In high-incidence populations such as Chile, there will be large numbers of patients diagnosed with compelling family histories and clinical features. Twenty-three of the probands in this study were not diagnosed with diffuse gastric cancer and/or did not have enough detailed family history to show they met the clinical criteria for HDGC. It is likely that if detailed clinical records and family histories were available, some of these cases would not have met HDGC testing criteria and would not typically be offered *CDH1* screening. In particular, patients diagnosed with intestinal-type gastric cancer or those that present with a family history of intestinal-type gastric cancer are not likely to benefit from *CDH1* variant screening. To prevent resource-consuming testing of patients that are unlikely to carry pathogenic variants, detailed clinical features and family history of cancer should be collected and considered before genetic testing. For families with a variety of cancer types and/or subtypes, screening of other cancer predisposition genes may be a better alternative, especially for those with well known

co-morbidities.

4.3.3 Conclusions

Overall, our results show pathogenic germline *CDH1* variants were not a common cause of gastric cancer in a series of Chilean probands. I show the importance of screening *CDH1* in cases who meet the clinical criteria for HDGC, and if a pathogenic variant is identified, their extended family. This study is also a reminder that a detailed family history of cancer should be collected for gastric cancer patients, and that *CDH1* screening need not be offered to those diagnosed with intestinal-type gastric cancer. I stress the importance of applying the HDGC criteria to identify those that may be at risk and provide them with appropriate genetic testing. This research is relevant to the development of clinical guidelines for screening germline *CDH1* in Chile, as well as other high-incidence populations.

Chapter 5

Novel germline variants in *CDH1* variant-negative diffuse gastric cancer patients

5.1 Introduction

Pathogenic variants in germline *CDH1* explain approximately 40% of cases that fit the clinical criteria for HDGC. An explanation for the remaining 60% of HDGC cases remains largely elusive. In this chapter, whole-exome sequencing was used to examine HDGC patients that did not carry pathogenic *CDH1* variants and identify variants that may be increasing diffuse gastric cancer risk.

5.1.1 Discovery of cancer predisposing genes

More than 100 cancer predisposition genes have been identified over the last three decades (Rahman, 2014). Since the year 1990, at least one new cancer predisposition gene has been identified each year (Rahman, 2014). Genome-wide linkage analysis has been the most successful strategy for gene identification, leading to the discovery of 60 high-risk cancer predisposition genes in the 1990s (Rahman, 2014). Candidate gene studies have also been utilised for gene discovery, but with less success (Rahman, 2014). More recently, the field of cancer genomics has been impacted by the application of next-generation sequencing (NGS) technology that has accelerated the pace of gene discovery while dramatically reducing the cost of sequencing.

Although there have been many important breakthroughs in the identification of cancer predisposition genes, the missing heritability of many familial cancer syndromes suggests that more cancer predisposition genes remain to be discovered. Next-generation whole-genome and whole-exome sequencing studies are ideally suited for the identification of high-risk genes. As with other common complex conditions, identification of low- and

moderate-penetrance genes will be challenging and will require large collaborative efforts.

5.1.2 Genetic predisposition to gastric cancer

On average familial clustering of gastric cancer occurs for approximately 10% of cases, but has been reported to be as high as 39% in some populations (Yaghoobi et al., 2010). Hereditary gastric cancer syndromes, where inherited germline variants are predisposing to a heightened risk of developing gastric cancer, are thought to account for between 1-3% of all gastric cancers (Section 1.1.2).

HDGC is the most common hereditary gastric cancer syndrome. Pathogenic germline variants in the E-cadherin gene *CDH1* are well documented and explain approximately 40% of HDGC cases (Hansford et al., 2015). An explanation for the remaining 60% of HDGC cases remains largely elusive. A small number of additional HDGC predisposition genes have been proposed, but still require further validation. There is also emerging evidence that other cancer predisposition syndromes may commonly present with the same characteristics as HDGC.

5.1.2.1 Additional genes implicated in HDGC

As previously outlined in Section 1.1.4.3, further to germline *CDH1* variants, three studies have proposed a total of five additional HDGC genes, specifically *CTNNA1* (Majewski et al., 2013), *MAP3K6* (Gaston et al., 2014), *INSR* (Donner, Kiviluoto, Ristimäki, Aaltonen, & Vahteristo, 2015), *FBXO24*, (Donner et al.,

2015), and *DOT1L* (Donner et al., 2015). These genes and the studies which proposed them are discussed in detail below.

Pathogenic germline variants in α-catenin (*CTNNA1*) were first described by Majewski et al. (2013) in a Dutch family with several cases of diffuse gastric cancer over four generations. In this family, a truncating germline *CTNNA1* variant was identified in two family members diagnosed with diffuse gastric cancer, and four first degree relatives that had intramucosal signet ring cells identified during endoscopic screening. Cases with the *CTNNA1* variant developed diffuse gastric cancer late in life (generally > 50 years of age), suggesting this variant has a low to moderate penetrance. α-catenin was detectable in normal epithelial structures, but was completely absent in the tumour cells from two affected family members. Additionally, α-catenin expression was lost in foci of signet ring cells detected in biopsy samples obtained from a first-degree relative that carried the *CTNNA1* variant.

α-catenin is involved in the cell-adhesion complex and facilitates adhesion and communication between neighbouring epithelial cells. As previously described in Section 1.1.4, α-catenin binds and secures interaction between β-catenin, E-cadherin, and the actin cytoskeleton. Through its interactions with the adherens junction, α-catenin directly regulates the assembly and organisation of actin-filaments within the cell-adhesion complex (Gall & Frampton, 2013; Sun, Zhang, & Ma, 2014). *CTNNA1* is mutated in a variety of cancer cell lines and primary tumours, and loss of α-catenin expression has been reported in sporadic gastric cancers (Bignell et al., 2010; Shiozaki et al., 1994). In animal models, loss of α-catenin induces hyperproliferation and impaired apoptosis in skin cells (Vasioukhin, Bauer, Degenstein, Wise, & Fuchs, 2001). In addition to the family described by Majewski et al. (2013), germline variants in *CTNNA1* have been discovered in a small number of

additional HDGC families, supporting it's inclusion as an HDGC gene, and highlighting the importance of the adherens junction complex in HDGC (Hansford et al., 2015; Schuetz et al., 2012).

Gaston et al. (2014) first described germline variants in *MAP3K6* in a large family from Maritime Canada with a history of gastric cancer. This family presented with many features associated with HDGC, but with late onset of disease (generally greater than 50 years of age). In this family, four affected cases and five of 27 currently unaffected relatives harboured a truncating *MAP3K6* variant. A somatic mutation in *MAP3K6* and hypermethylation of *MAP3K6* were detected in tumour DNA from two variant carriers. In screening of an additional 115 individuals from unrelated *CDH1* variant-negative gastric cancer families, four additional *MAP3K6* variants (one truncating and three missense variants) were identified. It is noteworthy that in this family, the *MAP3K6* variant did not completely segregate with disease in this family and appears at a moderate frequency in the ExAc database (MAF 0.005), suggesting that this variant may be low penetrance.

MAP3K6 encodes a serine/threonine protein kinase that forms a component of protein kinase-mediated signal transduction cascades (Iriyama et al., 2009). The encoded kinase participates in the regulation of vascular endothelial growth factor expression, apoptosis, and inflammation (Iriyama et al., 2009). *MAP3K6* is known to act as a tumour suppressor gene in epithelial cells and is mutated in some gastric tumours and gastric cancer cell lines (Zang et al., 2011). Further evidence of cancer risk will be required for *MAP3K6* to be considered a putative HDGC gene.

Donner et al. (2015) studied a family in which there there were six cases of diffuse gastric cancer over two generations. Exome sequencing was used to identify shared rare deleterious variants in two cases that were then Sanger

sequenced in a third affected family member. Three candidate variants were shared by all three cases: *INSR* p.Glu1313Lys, *FBXO24* p.Arg81Pro, and *DOT1L* p.Pro1146Leu. Analysis of tumour samples from these three family members did not show loss of heterozygosity for any of the candidate genes, and screening of an additional 26 gastric cancer patients with a confirmed or suspected family history of diffuse gastric cancer did not identify any additional variants in *INSR, FBXO24*, and *DOT1L*.

Of particular is the *INSR* variant. *INSR* encodes a transmembrane insulin receptor that is activated by insulin, IGF-I, and IGF-II (Ebina et al., 1985). *INSR* belongs to the tyrosine kinase growth factor receptor family and functions as an enzyme that transfers phosphate groups from ATP to tyrosine on intracellular proteins (Arcidiacono et al., 2012). Metabolically, *INSR* plays a key role in insulin signalling and is primarily involved in glucose metabolism. *INSR* has been implicated in the regulation of cellular growth and differentiation, and under degenerate conditions, is thought to contribute to cancer progression (Ebina et al., 1985; Malaguarnera et al., 2012; Taniguchi, Emanuelli, & Kahn, 2006). Similar to *MAP3K6*, more evidence will be required before *INSR*, *DOTL1*, or *FBXO24* can be confirmed as HDGC genes.

5.1.2.2 Gastric cancer as a part of other cancer predisposition syndromes

Gastric cancer is a part of a number of other familial cancer syndromes (previously described in detail in Section 1.1.3.4). Accordingly, some studies have reported families who meet the clinical criteria for HDGC and carry pathogenic variants in genes that are more strongly associated with other cancer types.

In a study of *CDH1* variant-negative HDGC probands, Hansford et al. (2015) used a NGS targeted panel approach to identify pathogenic variants in genes known to be associated with hereditary cancer syndromes, for which gastric cancer is a part of the wider tumour spectrum. Overall 11% of *CDH1* pathogenic variant-negative HDGC probands carried candidate germline variants in other cancer predisposition genes. Specifically, the authors identified truncating variants in *ATM*, *BRCA2*, *CTNNA1*, *MSR1*, *PALB2*, *PRSS1*, *SDHB*, and *STK11* (Hansford et al., 2015).

Similarly, Sahasrabudhe et al. (2016) used a combination of whole-exome sequencing, targeted amplicon sequencing of candidate genes, and genotyping of specific variants, to identify causal variants in known DNA damage repair genes. Overall, 6.5% (2/31) of cases in this study that fitted the clinical criteria for HDGC carried a pathogenic variant in either *PALB2*, *RAD51C*, or *BRCA1* (Sahasrabudhe et al., 2016), genes that are more associated with breast cancer risk. Strikingly, a further 9/331 (2.8%) patients with sporadic gastric cancer were found to have pathogenic variants in *PALB2*, *RAD51C*, or *BRCA1*.

Some studies have also noted pathogenic germline variants in cancer predisposition genes in unselected gastric cancer cases. In a study of the germline variants in data generated by The Cancer Genome Atlas (TCGA), Lu et al. (2015) reported 11% of gastric cancer cases carried rare truncating variants in the tumour suppressor genes *ATM*, *BRCA2*, *BRIP1*, *EME2*, *PALB2*, and *XRCC2*. Interestingly, the authors also reported a bimodal distribution for age of onset in their gastric cancer cohort and an association between pathogenic *ATM* variants and early-onset gastric cancer ($p < 0.05$), an observation that has also been reported by others (Helgason et al., 2015; Huang et al., 2015). Cancer predisposition syndromes that are characterised by variants in these genes are described in detail in Section 1.1.3.4.

5.1.3 Overlap of somatic and germline cancer genes

Lauren's classification of gastric cancer is largely based on histological assessment of glandular formation in the intestinal-type versus early loss in cell-to-cell adhesion in the diffuse-type (Lauren, 1965). As previously described in Section 1.1.1.1, in addition to histological classification, recent molecular profiling of gastric tumours found four specific molecular profiles with unique mutational signatures: Epstein-Barr virus (EBV)-positive, microsatellite instable (MSI), chromosomal instable (CSI), and genomically stable (GS) (Bass et al., 2014). Of particular interest is the GS subtype, which was enriched for diffuse-type tumours with poor cell differentiation and defective cell adhesion (Bass et al., 2014). As HDGC is characterised by diffuse-type tumours, it is logical to consider that the genes frequently mutated in the GS subtype may also be genes that are important in the development of hereditary diffuse gastric cancer risk

Analysis of the GS tumours found that *CDH1* was somatically mutated in up to 37% of cases (Bass et al., 2014; Chen et al., 2015; Wang et al., 2014; Wong et al., 2014). Furthermore, pathway enrichment analysis using the genes that were most frequently altered by mutations, hyper-methylation, and copy number changes, found that the adherens junctions, focal adhesions, and WNT signaling pathways were over-represented. Of note, the genes *CTNNA1*, *CTNNA2*, and *CTNNB1* were all frequently mutated in GS tumours. Together, these findings highlight the importance of the adherens junction complex and cell adhesion in the diffuse gastric cancer (Wang et al., 2014). Unexpectedly, aberrations in Rho signalling were also enriched in GS tumours. Ras homolog gene family, member A (RhoA), encoded by the gene *RHOA*, was reported to be mutated in up to 25.3% of GS tumours (Bass et al., 2014; Wang et al., 2014).

Additionally, fusions involving Rho-family GTPase-activating proteins (RhoGAPs) were reported in approximately 15% of GS tumours (Bass et al., 2014).

RhoA is a member of the Rho GTPase family, which belongs to the Ras superfamily of small GTP-binding proteins (Wennerberg, Rossman, & Der, 2005). RhoA is a multifunctional protein that, through the action of various effector proteins, regulates a range of biological functions including actin organisation, intracellular transport, cell migration, cell cycle, cell proliferation, cell adhesion, oncogenic transformation, and tissue repair (Ridley et al., 2003; Thumkeo, Watanabe, & Narumiya, 2013). RhoA exists in two states: active GTP-bound and inactive GDP-bound. The activity state of RhoA is regulated by a group of activating guanine exchange factors (GEFs), that control the change of GTP to GDP, and two groups of deactivators, GTPase activating proteins (GAPs) and guanine dissociation inhibitors (GDIs), that increase the rate of GTP hydrolysis and inhibits the release of GDP, respectively (Ridley et al., 2003).

RHOA mutations are highly enriched in GS tumours, but have also been identified in EBV-positive and MSI tumours at a lower frequency, suggesting there may be some overlap between molecular subtypes (Bass et al., 2014; Chen et al., 2015; Wang et al., 2014; Wong et al., 2014). Somatic *RHOA* mutations in gastric tumours show clustering in the GTP domains and effector binding regions of the protein. Functional characterisation of two hotspot mutations (p.Tyr42Cys and p.Leu57Val) has shown that these mutations result in abnormal *RHOA* proteins that promote evasion from anoikis (Wang et al., 2014).

Regulation of Rho signalling was further implicated as a key regulator of carcinogenesis of diffuse-type tumours by the discovery of recurrent genomic

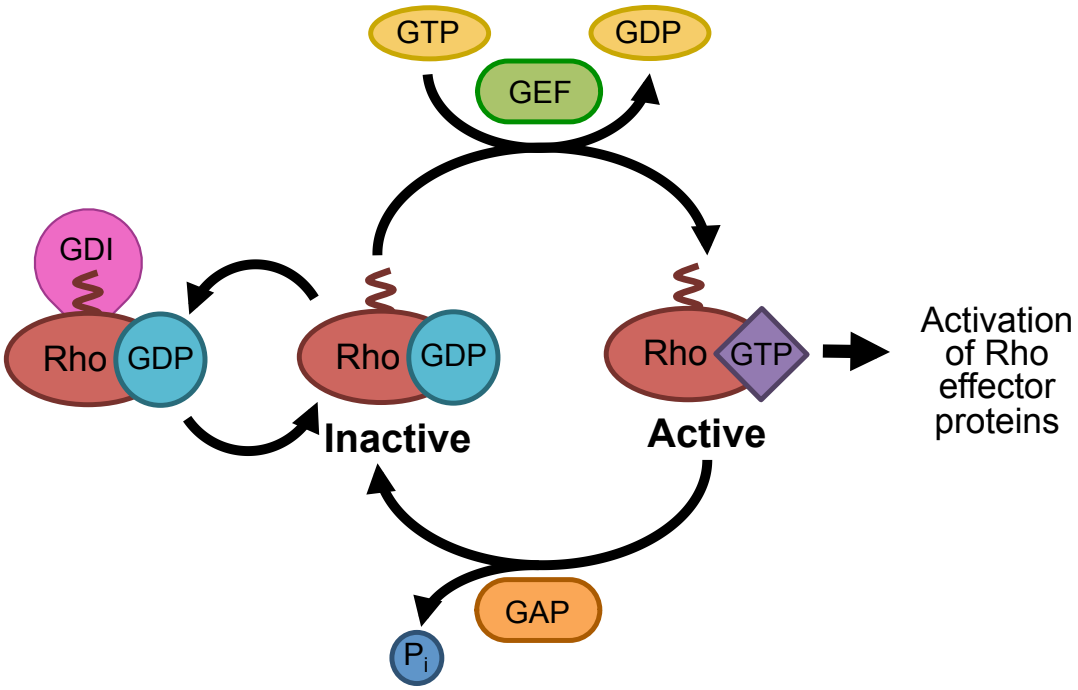


Figure 5.1: The Rho-GTPase activation cycle. Regulation of small RhoGTPases (RhoA; red) by guanine nucleotide exchange factors (GEFs; green) and GTPase-activating proteins (GAPs; orange). GEFs replaces GDP with GTP to activate the signalling function of the GTPase. Conversely, GAPs stimulate hydrolysis of GTP into GDP to inactivate the GTPase. Guanine dissociation inhibitors (GDIs) inhibit the release of GDP. Abbreviations: GAP, GTPase-activating protein; GEF, guanine exchange factor; GDI, gaunine dissociation inhibitors; GDP, Guanosine diphosphate; GTP, Guanosine triphosphate; P, phosphate.

fusions between *CLDN18* and two different RhoGAPs (*ARHGAP26* and *ARHGAP6*) in the GS subtype (Bass et al., 2014). *CLDN18* is a component of tight junction adhesion structures and has a role in maintaining cell polarity and cell signalling (Morita, Furuse, Fujimoto, & Tsukita, 1999). Functional analysis of *CLDN18-ARHGAP26* showed that this fusion inhibits RhoA and impairs epithelial integrity by reducing cell–cell and cell–extracellular matrix adhesion (Yao et al., 2015). As RhoGAPs function to convert the active GTP form of RhoA to its inactive GDP form, it is plausible that the fusion with *CLDN18*, a membrane protein highly expressed in stomach, may result in increased presence of these ARHGAPs in proximity to the cell membrane, increasing GTPase activity and inactivation of RhoA at cell junctions.

Furthermore, *RHOA* mutations do not overlap with *CLDN18-ARHGAP* fusions, suggesting they may have similar roles in the carcinogenic process (Bass et al., 2014).

Hypothetically, mutations in guanine exchange factors, that normally activate RhoA, may prevent the normal activation of RhoA effector proteins. Mutations in these guanine exchange factors may thereby cause a similar effect to *CLDN18-ARHGAP* fusions and mutated *RHOA*, and inhibit or impair effective RhoA signalling. Somatic nonsynonymous mutations in guanine exchange factors (e.g., *VAV1*, *VAV3*, and *ARGEF13*) have been identified in both diffuse- and intestinal-type tumours, suggesting they may be important in tumourgenesis (Kakiuchi et al., 2014).

5.2 Results

Approximately 60% of gastric cancer patients who meet clinical criteria for HDGC testing do not carry a pathogenic variant in germline *CDH1*. It is probable that a proportion of these patients carry a predisposing variant in another gene. In this chapter, whole-exome sequencing was used to search for germline variants that may be predisposing to diffuse gastric cancer in a series of *CDH1* variant-negative HDGC patients.

5.2.1 Study cohort

Our whole-exome sequencing cohort comprised fourteen diffuse gastric cancer patients from twelve different families (Table 5.1). Ten of the 14 cases have one or more family members that have been diagnosed with gastric cancer or breast cancer. Unfortunately, the subtypes of the majority of breast

and gastric cancers in the extended family members of these cases could not be verified. Cases that were confirmed were all lobular breast cancers or diffuse gastric cancers. All patients in this study had previously tested negative for germline *CDH1* variants and meet one of the clinical criteria for HDGC *CDH1* screening (van der Post et al., 2015). Eleven patients were single probands from HDGC families, and three patients were from the same family (Family 9). Family 9 is an Ecuadorian family in which three siblings were diagnosed with diffuse gastric cancer aged 32, 53, and 56 years. The average age of the cases in this study at the time of diagnosis was 37 years (range 22-56 years).

5.2.2 Characterisation of exome sequencing data

Sequence data was generated on an Illumina HiSeq in two sequencing runs. The first run sequenced 11 patients in October 2013. The remaining three patients were not available at that time, therefore they were sequenced separately in July 2014. Basic read and mapping information is shown in Table 5.2.

For the first run, eleven libraries were prepared using TruSeq Exome Enrichment Library Preparation Kit and were sequenced on one lane of an Illumina HiSeq. There was considerable difference in the representation of these twelve libraries. The total number of reads ranged between 10.6 million and 39.6 million per sample. The number of on-target reads ranged between 6.0 million and 39.4 million reads per sample, and the average depth of sequencing at target regions ranged between 7.8 and 52.0 reads.

For the second run, three libraries were prepared using a Nextera Exome Enrichment Library Preparation kit and also sequenced on one lane of an Illumina HiSeq. Compared to the libraries on run one, the representation for

Patient	Description	Family
X3279	Diagnosed with diffuse gastric cancer aged 33 years. Sister diagnosed with gastric cancer (unspecified type) age 35 years. Mother diagnosed with abdominal carcinoma age 78 years.	1
X5799	Diagnosed with diffuse gastric cancer aged 50 years. Father diagnosed with gastric cancer (unspecified type) aged 30 years, uncle diagnosed with gastric cancer (unspecified type) aged 62 years, sister diagnosed with cervical cancer aged 55 years, and grandfather diagnosed with prostate cancer aged 81 years.	2
X7045	Diagnosed with diffuse gastric cancer aged 36 years. Aunt diagnosed with breast cancer (unspecified type).	3
X7223	Diagnosed with diffuse gastric cancer aged 32 years. Two sisters diagnosed with diffuse gastric cancer aged 28 and 36 years, one brother diagnosed with diffuse gastric cancer aged 40 years, mother diagnosed with gastric cancer (unspecified type) aged 41 years, and uncle diagnosed with gastric cancer (unspecified type) aged 55 years.	4
X8288	Diagnosed with diffuse gastric cancer aged 22 years. No reported history.	5
X8289	Diagnosed with diffuse gastric cancer aged 31 years. No reported history.	6
X8706	Diagnosed with diffuse gastric cancer aged 36 years. No reported history.	7
X8744	Diagnosed with diffuse gastric cancer aged 31 years. Sister diagnosed with diffuse gastric cancer aged 33 years.	8
Y128	Diagnosed with diffuse gastric cancer aged 32 years. Two siblings diagnosed with diffuse gastric cancer aged 53 and 56 years.	9
Y129	Diagnosed with diffuse gastric cancer aged 53 years. Two siblings diagnosed with diffuse gastric cancer aged 32 and 56 years.	9
Y130	Diagnosed with diffuse gastric cancer aged 56 years. Two siblings diagnosed with diffuse gastric cancer aged 32 and 53 years.	9
Y712	Diagnosed with diffuse gastric cancer aged 45 years. Previously diagnosed with bilateral breast cancer (unspecified type). Father diagnosed with gastric cancer (age unspecified).	10
Y713	Diagnosed with diffuse gastric cancer aged 29 years. Father diagnosed with pancreatic cancer aged 55 years.	11
Y714	Diagnosed with diffuse gastric cancer aged 36 years. Aunt diagnosed with breast cancer aged 60 years.	12

Table 5.1: Exome sequencing study cohort. Patient identifier, description, and family identifier for each study participant. The vertical line highlights the three siblings from Family 9.

these three libraries were much more even. The total number of reads ranged between 89.1 million and 90.9 million per sample. The number of on-target reads ranged between 54.4 million and 56.2 million per sample, and the average depth of sequencing at target regions ranged between 63.6 and 66.3 reads.

5.2.3 Variant filtering

Exome data was filtered to identify rare variants that may be of significance for diffuse gastric cancer (Figure 5.2). Before filtering there were 1,001,969 variant sites (894,699 SNPs and 107,270 indels). First, variant sites were filtered for quality by removing variants that did not pass variant quality score recalibration (VQSR). An additional hard filter was used to remove low coverage variant (< 5 reads), that are difficult to interpret and more prone to bias. Variants were then filtered for effect, where synonymous, intragenic, and intronic variants were removed, retaining only nonsense, missense, frameshift, and splice site variants. Finally, variant sites were filtered for rarity by removing variants which were common (MAF > 0.001) in any of the three major population databases (ESP6500, 1000 Genomes, and ExAc). Post filtering there were 2,227 rare variants (2,022 SNPs and 205 indels).

5.2.4 Prioritising variants

Following variant filtering, the list of 2,227 variant sites was examined to identify genes with a potential role in the carcinogenesis of diffuse gastric cancer. Although 224 genes contained multiple rare variants, as with many small scale exome-sequencing based studies, there was no single gene or variant which presented as a clear candidate for further analysis. To reduce the

Sample ID	HiSeq run	Capture Kit	Total Reads	Mappable reads	On-target reads	On-target reads (%)	Coverage 4X (%)	Coverage 10X (%)	Coverage 20X (%)	Average depth[a]
X3279	1	TruSeq	38,638,490	38,350,627	26,103,528	67.6	91.6	84.8	68.8	34.4
X5799	1	TruSeq	35,677,016	35,354,489	23,437,715	65.7	92.1	84.6	65.1	30.9
X7045	1	TruSeq	10,579,073	10,443,878	6,087,397	57.5	67.4	24.9	4.0	7.80
X7223	1	TruSeq	39,494,271	39,202,369	25,612,221	64.9	92.7	86.3	70.4	34.0
X8288	1	TruSeq	7,147,094	7,141,253	5,087,405	71.2	53.4	23.7	2.80	12.4
X8289	1	TruSeq	36,048,225	35,684,186	22,544,964	62.5	92.4	83.6	60.7	29.8
X8706	1	TruSeq	23,160,051	22,988,533	14,364,078	62.0	88.7	70.6	33.8	18.9
X8744	1	TruSeq	57,225,126	57,185,344	37,563,015	65.6	94.0	90.2	82.2	49.6
Y128	1	TruSeq	36,890,873	36,858,828	25,019,070	67.8	92.3	85.6	68.6	33.0
Y129	1	TruSeq	24,977,830	24,960,112	17,551,781	70.3	86.4	73.2	48.7	23.1
Y130	1	TruSeq	58,883,315	58,840,655	39,418,919	66.9	94.0	90.3	82.6	52.0
Y712	2	Nextera	90,861,920	90,804,300	56,206,822	61.9	92.4	86.6	76.1	66.3
Y713	2	Nextera	89,143,544	89,085,784	55,605,232	62.4	92.1	85.8	74.8	65.3
Y714	2	Nextera	89,317,408	89,262,789	54,368,183	60.9	92.1	85.8	74.6	63.6

Table 5.2: Summary of exome sequencing data. Basic details of sequence data generated for whole exome libraries. [a]Average depth of reads at target sites.

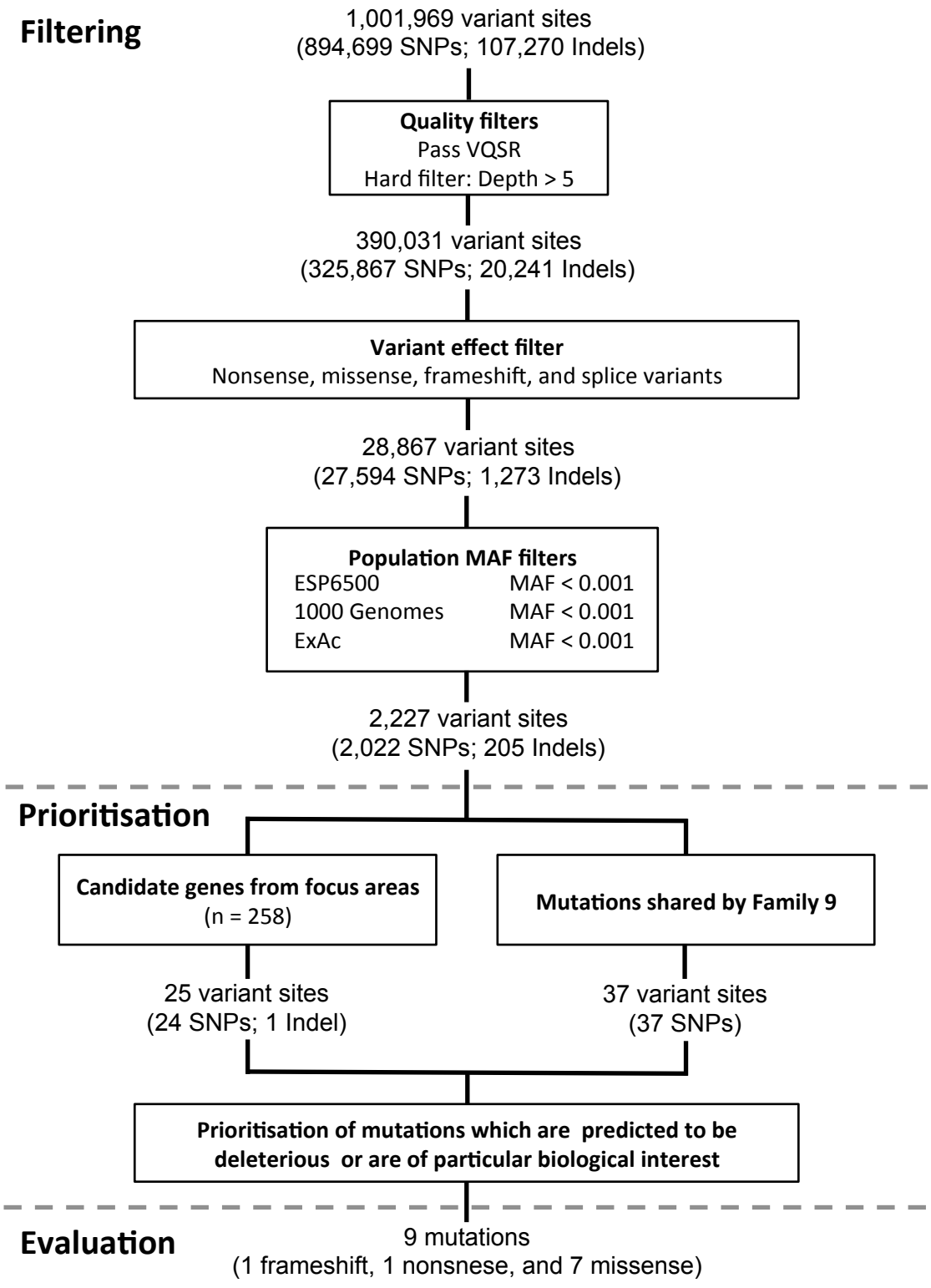


Figure 5.2: Schematic diagram of variant filtering, prioritisation, and evaluation of variants. Variants were filtered for quality, location, and frequency in population databases. Variants which were identified in focus areas, or were shared by all three members of Family 9 were prioritised for further evaluation. Abbreviations: ESP6500, Exome Sequencing Project; ExAc, Exome Aggregation Consortium; MAF, Minor allele frequency; VQSR, Variant Quality Score Recalibration.

Adherens junction[1]

ACP1, ACTB, ACTG1, ACTN1, ACTN2, ACTN3, ACTN4, AFDN, BAIAP2, CDC42, CDH1, CREBBP, CSNK2A1, CSNK2A2, CSNK2B, CTNNA1*, CTNNA2*, CTNNA3, CTNNB1*, CTNND1, EGFR*, EP300, ERBB2*, FARP2, FER, FGFR1, FYN, IGF1R, INSR*, IQGAP1, LEF1, LMO7, MAP3K7, MAPK1, MAPK3, MET*, NECTIN1, NECTIN2, NECTIN3, NECTIN4, NLK, PARD3, PTPN1, PTPN6, PTPRB, PTPRF, PTPRJ, PTPRM, RAC1, RAC2, RAC3, RHOA*, SMAD2, SMAD3, SMAD4*, SNAI1, SNAI2, SORBS1, SRC, SSX2IP, TCF7, TCF7L1, TCF7L2, TGFBR1*, TGFBR2*, TJP1, VCL, WAS*, WASF1, WASF2, WASF3, WASL, YES1*

Frequently mutated in sporadic gastric cancer[2]

APC, ARID1A*,ASTN1, BNC2, CDH1*, CTNNA2, CTNNB1, DCLK1, DLGAP2, EPB41L3, ERBB2, ERBB4, GLI3, KRAS, LRFN5, MACF1, MUC6, NRG2, OPRK1, PIK3CA, PTPRC, RASA1, RHOA, RIMS2, RNF43, SMAD4*, SPTA1, SYNE1, TGFBR2*, THBS1, THSD7B, TLR4, TP53*, TSHZ3, WDFY4, ZIC4*

Homologous recombination and non-homologous end-joining[3]

ATM, BABAM1, BARD1, BLM*, BRCA1*, BRCA2*, BRCC3, BRE, BRIP1A*, DCLRE1C, DNTT, EME1, FAM175A, FEN1, LIG4, MRE11, MUS81, NBN*, NHEJ1, PLAB2*, POLD1*, POLD2, POLD3, POLD4, POLL, POLM, PRKDC, RAD50, RAD51, RAD51B, RAD51C*, RAD51D*, RAD52, RAD54B, RAD54L, RBBP8, RPA1, RPA2, RPA3, RPA4, SEM1, SSBP1, SYCP3, TOP3A, TOPBP1, UIMC1, XRCC2, XRCC3, XRCC4, XRCC5, XRCC6*

Previously proposed to cause HDGC or gastrointestinal tumours[4]

AKAP12, AKR7A3, APC, ARID1A*, ATM*, BCL2L10, BMPR1A*, BRCA1*, BRCA2*, CASP10, CDH1*, CDKN2A*, CFTR, CHEK2*, CTHRC1, CTNNA1*, DOT1L, EPCAM, FAT4, FBXO24, FHIT, FOXF1, GAB2, GREM1, HSPA5, IDH1, IDH2, ITIH2, MAP3K6, MCCC1, MLH1*, MSH2*, MSH3, MSH6*, MSR1, MUTYH*, PALB2*, PMS1, PMS2*, PRR5, PRSS1*, PSCA, PTEN*, PXN, SCARF2, SCG5, SDHB, SDHC, SDHD, SLC22A4, SMAD4*, SPINK1, STK11*, TGFBR2*, TP53**

Guanine Exchange Factors[5]

ABR, AKAP13, ARHGEF1, ARHGEF10, ARHGEF10L, ARHGEF11, ARHGEF12, ARHGEF15, ARHGEF16, ARHGEF17, ARHGEF18, ARHGEF19, ARHGEF2, ARHGEF26, ARHGEF3, ARHGEF33, ARHGEF35, ARHGEF37, ARHGEF38, ARHGEF39, ARHGEF4, ARHGEF40, ARHGEF5, ARHGEF6, ARHGEF7, ARHGEF9, CDC42, ECT2, FGD1, FGD2, FGD3, FGD4, GNA13, ITSN1, KALRN, MCF2, MCF2L, NET1, NGEF, OBSCN, PLEKHG2, PLEKHG5, PREX1, RAC1*, RAC2*, RAC3*, RASGRF2, RHOA*, RHOB, RHOBTB1, RHOBTB2, RHOC, RHOD, RHOF, RHOG, RHOH, RHOJ, RHOQ, RHOT1, RHOT2, RHOU, RHOV, SOS1, SOS2, TIAM1, TIAM2, TRIO, VAV1, VAV2, VAV3*

Table 5.3: Candidate genes for the five focus areas. [1]Gene list downloaded from KEGG, accession number hsa04520. [2]Genes list curated from Bass et al. (2014), Chen et al. (2015), Wang et al. (2014), and Wong et al. (2014). [3]Gene list downloaded from KEGG, accession numbers hsa03450 and hsa03450. [4]Gene list curated from Donner et al. (2015), Gaston et al. (2014), Hansford et al. (2015) and Majewski et al. (2013). [5] Gene list downloaded from Reactome, accession number R-HSA-194849.3. *Gene a part of multiple lists.

list of variant sites to a manageable size for a comprehensive evaluation, two strategies were used. The first was the prioritisation of variants in genes involved in specific cellular functions (focus areas), and the second was the prioritisation of variants which were shared by all three members of Family 9.

5.2.4.1 Focus areas

After a review of published literature, I identified five focus areas that I believe to be important for the carcinogenesis of diffuse gastric cancer. The focus areas were: (1) genes involved in the adherens junctions, (2) genes which are frequently mutated in sporadic gastric cancers, (3) genes involved in DNA homologous recombination and non-homologous end-joining, (4) genes which have previously been implicated in hereditary gastrointestinal cancers and HDGC, and (5) guanine exchange factors involved in RhoA regulation. A candidate gene list encompassing the genes important to these five focus areas was curated from recent publications and online databases (Table 5.3). In total, I identified 258 candidate genes, 21 of which were common to two or more focus areas.

Using the candidate gene list, I prioritised 25 variants (one frameshift variant, one nonsense variant, and 23 missense variants) in 22 different candidate genes for further consideration (Table 5.4). Three different variants were identified in one candidate gene (*SYNE1*: c.14273T>C, c.6254T>C, and c.4427C>T), and two different variants in another (*AKAP12*: c.3925A>G and c.4383T>G). Variants in genes previously proposed to cause gastrointestinal cancers (focus area 4) were most common and made up ten of the 25 short listed variants. Variants in guanine exchange factors (focus area 5) were least common and only contributed three variants to this prioritised list. Of note,

three of these focus area variants were shared by multiple members of Family 9 (*SYNE1* c.6254T>C, 2/3 family members; *LMO7* c.3823C>T, 3/3 family members; and *IGF1R* c.2683C>T, 3/3 family members).

In addition to the nonsense and frameshift variants in *PMS1* and *ATM*, respectively, 7/32 missense variants were predicted to be deleterious by Condel (Table 5.5). Four of the deleterious variants were in genes that have previously been proposed to increase gastric cancer risk (*APC*, *ATM*, *PMS1*, and *TP53*), while three were in genes that had not previously been associated with gastric cancer risk (*FARP2*, *FGD4*, and *LMO7*).

5.2.4.2 Variants shared by Family 9

Variants that were shared by all three affected members of Family 9 were also short listed for further evaluation. There were no rare nonsense or frameshift variants shared by these individuals, however there were 37 rare missense variants. The effect of these variants were predicted using Condel (Table 5.6). Variants in the genes *MRPS22*, *SRP72*, *ANK2*, *NADKD1*, *ZNF474*, *MLL5*, *PLEC*, *OR51A7*, *TRIM66*, *LMO7*, and *TMPRSS6* were predicted to be deleterious. The function of these genes were investigated for a potential link to the carcinogenesis of diffuse gastric cancer.

The most compelling gene and variants in the aforementioned gene list was *LMO7* c.3823C>T. *LMO7*, a candidate gene which belonged to focus area 1 (adherens junction). The cellular function of LMO7 discussed in detail in Section 5.2.5.2. *IGF1R* is also a candidate gene (focus area 1), and the c.2683C>T variant was shared by all three members of Family 9. However, the *IGF1R* c.2683C>T variant was predicted to be neutral by Condel,

Patient	Chr	Position[a]	SNP ID	Gene	Focus area	Nucleotide change	Protein change	Class	Ref/Alt reads	Depth
X3279	1	45800164	.	*MUTYH*	4	c.56G>A	p.Arg19Gln	Missense	24, 22	46
X8706	1	46714199	.	*RAD54L*	3	c.19C>T	p.Pro7Ser	Missense	30, 26	56
X7045	2	190708804	rs121434628	*PMS1*	4	c.697C>T	p.Gln233*	Nonsense	8, 15	23
X8289	2	242312646	.	*FARP2*	1	c.124A>G	p.Lys42Glu	Missense	4, 5	9
X8706	5	112175379	.	*APC*	2, 4	c.4088A>T	p.Lys1363Ile	Missense	13, 21	34
X8706	5	131911536	.	*RAD50*	3	c.281T>C	p.Ile94Thr	Missense	13, 7	20
X8744	6	151673451	rs370121685	*AKAP12*	4	c.3925A>G	p.Thr1309Ala	Missense	23, 31	54
Y714	6	151673909	rs368535187	*AKAP12*	4	c.4383T>G	p.Asn1461Lys	Missense	25, 25	50
Y713	6	152651547	.	*SYNE1*	2	c.14273T>C	p.Leu4758Ser	Missense	50, 50	100
Y128	6	152730821	.	*SYNE1*	2	c..6254T>C	p.Ile2085Thr	Missense	17, 21	38
Y129	6	152730821	.	*SYNE1*	2	c..6254T>C	p.Ile2085Thr	Missense	13, 19	32
Y130	6	152754964	.	*SYNE1*	2	c.4427C>T	p.Ser1476Leu	Missense	13, 10	23
X8289	7	100187684	.	*FBXO24*	4	c.238T>C	p.Phe80Leu	Missense	24, 13	37
X8706	8	95399302	.	*RAD54B*	3	c.1895A>G	p.Asn632Ser	Missense	10, 9	19
Y8706	10	5495515	rs151234884	*NET1*	5	c.760G>A	p.Val254Met	Missense	6, 4	10
X8289	11	108119751	.	*ATM*	3, 4	c.1157delG	p.Arg386fs	Frameshift	5, 5	10
Y713	12	32777929	rs141237776	*FGD4*	5	c.1562A>G	p.Asp521Gly	Missense	24, 22	46
X8706	12	120652959	.	*PXN*	4	c1093G>A	p.Ala365	Missense	8, 9	17
Y128	13	76427385	rs141386159	*LMO7*	1	c.3823C>T	pArg1275Trp	Missense	16, 25	41
Y129	13	76427385	rs141386159	*LMO7*	1	c.3823C>T	pArg1275Trp	Missense	5, 11	16
Y130	13	76427385	rs141386159	*LMO7*	1	c.3823C>T	pArg1275Trp	Missense	48, 42	90
Y128	14	21555571	.	*ARHGEF40*	5	c.4336G>T	p.Ala1446Ser	Missense	13, 7	20
Y128	15	99467814	.	*IGF1R*	1	c.2683C>T	p.Arg895Trp	Missense	19, 25	44
Y129	15	99467814	.	*IGF1R*	1	c.2683C>T	p.Arg895Trp	Missense	13, 10	23
Y130	15	99467814	.	*IGF1R*	1	c.2683C>T	p.Arg895Trp	Missense	43, 33	76

Table continued on next page...

Patient	Chr	Position[a]	SNP ID	Gene	Focus area	Nucleotide change	Protein change	Class	Ref/Alt reads	Depth
X5799	16	23647293	.	*PALB2*	3, 4	c.574G>A	p.Val192Ile	Missense	13, 17	30
X8744	17	1775808	rs370337203	*RPA1*	3	c.442A>G	p.Ser148Gly	Missense	13, 7	20
Y712	17	7577568	.	*TP53*	2, 4	c.713G>A	p.Cys238Tyr	Missense	49, 57	106
X8289	17	37881095	rs146603731	*ERBB2*	1, 2	c.2424C>A	p.Asp808Glu	Missense	9, 16	25
Y128	17	56435497	.	*RNF43*	2	c.1640A>C	p.His547Pro	Missense	7, 12	19

Table 5.4: Germline variants in candidate genes. Focus areas are: (1) genes involved in the adherens junctions, (2) genes which are frequently mutated in sporadic gastric cancers, (3) genes involved in DNA homologous recombination and non-homologous end-joining, (4) genes which have previously been proposed to increase the risk of gastrointestinal cancers, and (5) guanine exchange factors involved in RhoA regulation. The vertical line highlights variants shared by multiple siblings from Family 9. All variants had the maximum quality score of 99. [a]Reference sequence for variant position is Hg19. Abbreviations: Chr, Chromosome; Ref/Alt, Reference allele, Alternate allele.

Patient	Chr	Position[a]	Gene	Focus area	Nucleotide change	Protein change	SIFT	PPH2	MA	FATHMM	Condel score	Condel prediction
X3279	1	45800164	*MUTYH*	4	c.56G>A	p.Arg19Gln	0.20	0.27	1.15	3.17	0.47	Neutral
X8706	1	46714199	*RAD54L*	3	c.19C>T	p.Pro7Ser	0.01	1.00	0.00	-2.66	0.47	Neutral
X8289	2	242312646	*FARP2*	1	c.124A>G	p.Lys42Glu	0.00	0.19	2.00	-1.63	0.56	Deleterious
X8706	5	112175379	*APC*	2,4	c.4088A>T	p.Lys1363Ile	0.14	0.99	0.98	-2.64	0.52	Deleterious
X8706	5	131911536	*RAD50*	3	c.281T>C	p.Ile94Thr	0.05	0.09	-0.38	3.45	0.33	Neutral
X8744	6	151673451	*AKAP12*	4	c.3925A>G	p.Thr1309Ala	0.70	0.01	0.70	3.13	0.36	Neutral
Y714	6	151673909	*AKAP12*	4	c.4383T>G	p.Asn1461Lys	0.91	0.01	0.35	3.27	0.36	Neutral
Y713	6	152651547	*SYNE1*	2	c.14273T>C	p.Leu4758Ser	0.72	0.9	1.65	0.65	0.47	Neutral
Y128	6	152730821	*SYNE1*	2	c..6254T>C	p.Ile2085Thr	0.23	0.09	1.85	1.32	0.42	Neutral
Y129	6	152730821	*SYNE1*	2	c..6254T>C	p.Ile2085Thr	0.23	0.09	1.85	1.32	0.42	Neutral
Y130	6	152754964	*SYNE1*	2	c.4427C>T	p.Ser1476Leu	0.23	0.03	2.05	0.59	0.49	Neutral
X8706	7	100187684	*FBXO24*	4	c.238T>C	p.Phe80Leu	0.76	0.05	-2.6	2.76	0.24	Neutral
X8706	8	95399302	*RAD54B*	3	c.1895A>G	p.Asn632Ser	0.95	0.00	0.29	-0.9	0.48	Neutral
Y8288	10	5495515	*NET1*	5	c.760G>A	p.Val254Met	0.05	0.04	1.01	-0.09	0.49	Neutral
Y713	12	32777929	*FGD4*	5	c.1562A>G	p.Asp521Gly	0.04	0.47	1.70	-2.31	0.54	Deleterious
X8706	12	120652959	*PXN*	4	c1093G>A	p.Ala365	0.42	0.04	0.76	0.56	0.46	Neutral
Y128	13	76427385	*LMO7*	1	c.3823C>T	pArg1275Trp	0.01	0.95	2.28	0.21	0.54	Deleterious
Y129	13	76427385	*LMO7*	1	c.3823C>T	pArg1275Trp	0.01	0.95	2.28	0.21	0.54	Deleterious
Y130	13	76427385	*LMO7*	1	c.3823C>T	pArg1275Trp	0.01	0.95	2.28	0.21	0.54	Deleterious
Y128	14	21555571	*ARHGEF40*	5	c.4336G>T	p.Ala1446Ser	0.41	0.37	0.55	4.28	0.33	Neutral
Y128	15	99467814	*IGF1R*	1	c.2683C>T	p.Arg895Trp	0.02	0.85	1.53	0.40	0.48	Neutral
Y129	15	99467814	*IGF1R*	1	c.2683C>T	p.Arg895Trp	0.02	0.85	1.53	0.40	0.48	Neutral
Y130	15	99467814	*IGF1R*	1	c.2683C>T	p.Arg895Trp	0.02	0.85	1.53	0.40	0.48	Neutral
X5799	16	23647293	*PALB2*	3,4	c.574G>A	p.Val192Ile	0.49	0.04	1.25	2.40	0.39	Neutral
X8744	17	1775808	*RPA1*	3	c.442A>G	p.Ser148Gly	0.45	0.08	2.13	0.87	0.49	Neutral

Table continued on next page...

Patient	Chr	Position[a]	Gene	Focus area	Nucleotide change	Protein change	SIFT	PPH2	MA	FATHMM	Condel score	Condel prediction
Y712	17	7577568	*TP53*	2,4	c.713G>A	p.Cys238Tyr	0.00	1.00	3.37	-9.96	0.78	Deleterious
X8289	17	37881095	*ERBB2*	1,2	c.2424C>A	p.Asp808Glu	0.13	0.65	0.94	-0.08	0.49	Neutral
Y128	17	56435497	*RNF43*	2	c.1640A>C	p.His547Pro	0.00	0.79	1.10	1.98	0.40	Neutral

Table 5.5: Functional prediction of rare missense variants in candidate genes. Functional predictions were generated using Condel. Pathogenicity scores for each of the *in silico* prediction tools used by Condel are shown. A Condel score of > 0.522 was used to define deleterious variants. The vertical line highlights variants shared by multiple siblings from Family 9. [a]Reference sequence for variant position is Hg19. Abbreviations: Chr, Chromosome; MA, Mutation assessor; PPH2, Polyphen2.

Chr	Position[a]	SNP ID	Nucleotide change	Protein Change	Gene	SIFT	PPH2	MA	FATHMM	Condel score	Condel prediction
2	20146313	.	*WDR35*	c.1576A>G	p.Ile526Val	0.83	0.01	0.73	-0.16	0.48	Neutral
3	127642421	.	*KBTBD12*	c.517G>A	p.Glu173Lys	0.08	0.51	1.81	-0.32	0.51	Neutral
3	139074541	.	*MRPS22*	c.896A>C	p.Asn299Thr	0.23	0.09	1.24	-1.62	0.53	Deleterious
3	194373822	.	*LSG1*	c.809A>C	p.His270Pro	0.25	0.00	0.00	1.02	0.41	Neutral
4	15818139	.	*CD38*	c.239T>C	p.Val80Ala	0.08	0.10	1.32	2.55	0.38	Neutral
4	48169969	rs199797945	*TEC*	c.497G>A	p.Arg166Gln	0.30	0.05	0.70	2.31	0.39	Neutral
4	57337893	.	*SRP72*	c.238C>T	p.Leu80Phe	0.36	0.48	1.46	-1.04	0.54	Deleterious
4	75248428	.	*EREG*	c.345C>A	p.Ser115Arg	0.35	0.75	1.78	2.57	0.36	Neutral
4	114286269	.	*ANK2*	c.4681C>T	p.Pro3622Ser	0.35	0.10	1.10	-3.13	0.53	Deleterious
5	36219773	.	*NADKD1*	c.569G>T	p.Gly190Val	0.00	1.00	2.68	-2.04	0.63	Deleterious
5	121488178	.	*ZNF474*	c.493C>G	p.Leu165Val	0.05	1.00	3.12	0.07	0.65	Deleterious
5	149914468	.	*NDST1*	c.1136C>T	p.Ser379Leu	0.08	0.01	0.81	0.88	0.45	Neutral
5	176919615	.	*PDLIM7*	c.160G>A	p.Asp54Asn	0.03	0.01	-0.47	2.98	0.34	Neutral
6	116951628	.	*RSPH4A*	c.1829G>C	p.Arg610Pro	0.01	0.86	.	1.98	0.31	Neutral
7	100017371	.	*ZCWPW1*	c.164T>C	p.Leu55Pro	0.01	0.90	1.53	0.83	0.46	Neutral
7	104746043	.	*MLL5*	c.2354C>T	p.Ser785Phe	0.01	0.55	1.04	-2.99	0.53	Deleterious
7	105664913	.	*CDHR3*	c.2163A>G	p.Ile721Met	0.20	0.00	0.60	0.44	0.46	Neutral
8	2832039	.	*CSMD1*	c.8674A>G	p.Ile2754Val	0.72	1.00	-0.08	0.07	0.43	Neutral
8	144996983	.	*PLEC*	c.7525C>T	p.Arg2509W	0.00	0.99	1.74	-1.16	0.55	Deleterious
9	102780561	.	*ERP44*	c.622A>C	p.Asn208His	0.25	0.01	2.28	1.50	0.46	Neutral
11	428546	.	*ANO9*	c.1114G>A	p.Glu372Lys	0.45	0.16	0.59	-0.18	0.48	Neutral
11	4929194	.	*OR51A7*	c.595G>A	p.Gly199Ser	0.02	0.76	2.30	-0.54	0.58	Deleterious
11	8646619	.	*TRIM66*	c.2032C>G	p.Pro678Ala	0.48	0.09	2.02	-0.14	0.53	Deleterious
11	59271250	.	*OR4D11*	c.202A>G	p.Ile68Val	0.26	0.00	-0.60	6.66	0.26	Neutral
13	52439707	.	*CCDC70*	c.193T>C	p.Phe65Lys	1.00	0.20	2.14	2.12	0.41	Neutral

Table continued on next page…

Chr	Position[a]	SNP ID	Nucleotide change	Protein Change	Gene	SIFT	PPH2	MA	FATHMM	Condel score	Condel prediction
13	76427385	rs141386159	*LMO7*	c.3823C>T	p.Arg1275Trp	0.01	0.95	2.28	0.21	0.54	Deleterious
13	96086274	rs199666320	*CLDN10*	c.187C>T	p.Pro63Ser	0.10	0.22	.	-2.28	0.47	Neutral
15	99467814	.	*IGF1R*	c.2683C>T	p.Arg895Trp	0.02	0.85	1.53	0.40	0.48	Neutral
16	22358755	.	*CDR2*	c.896C>G	p.Pro299Arg	0.53	0.68	1.39	0.65	0.47	Neutral
17	11795208	.	*DNAH9*	c.11227A>G	p.Ile3743Val	0.49	0.00	0.22	0.39	0.44	Neutral
17	48460536	.	*LRRC59*	c.737C>T	p.Ala246Val	0.18	0.00	1.50	1.04	0.45	Neutral
18	67695973	.	*RTTN*	c.5810A>G	p.Asn1937Ser	0.24	0.50	2.00	0.77	0.48	Neutral
20	45315465	.	*TP53RK*	c.689A>C	p.Lys230Thr	0.10	0.05	1.94	2.00	0.38	Neutral
22	24724835	.	*SPECC1L*	c.2168A>G	p.Asp723Gly	0.04	0.98	1.70	0.15	0.49	Neutral
22	31485905	.	*SMTN*	c.692C>T	p.Pro231Leu	0.68	0.00	0.15	-0.31	0.45	Neutral
22	37470698	.	*TMPRSS6*	c.1420T>C	p.Cys474Arg	0.00	0.75	3.47	-0.57	0.72	Deleterious
22	37892522	.	*CARD10*	c.1993G>C	p.Glu665Gln	0.19	0.01	0.00	0.96	0.42	Neutral

Table 5.6: Functional prediction of missense variants shared by Family 9. Functional predictions of missense muations were generated by Condel. Pathogenicity scores for each of the *in silico* prediction tools used by Condel are shown. A Condel score of > 0.522 was used to define deleterious variants. [a]Reference sequence for variant position is Hg19. Abbreviations: Chr, Chromosome; MA, Mutation assessor; PPH2, Polyphen 2.

suggesting it was not likely to contribute to the diffuse gastric cancers in this family.

5.2.5 Evaluation of prioritised variants

We identified nine variants which were of functional and/or biological interest with consideration of the predicted effect and known biological function. The nine different variants included six variants in genes previously associated with an increased risk cancer risk (*APC* c.4088A>T, *ATM* c.1157delG, *MUTYH* c.56G>A, *PALB2* c.574G>A, *PMS1* c.697C>T, *TP53* c.713G>A), and three variants in novel candidate diffuse gastric cancer predisposition genes (*FARP2* c.124A>G, *FGD4* c.1562A>G, and *LMO7* c.3823C>T). Patient X8706 carried a deleterious variants in both *FARP2* and *ATM*. The variant in *LMO7* was shared by the three siblings from Family 9 (Y128, Y129, and Y130). Although there were multiple different variants in *SYNE1* and *AKAP12*, they were not considered for further analysis as none of these variants were predicted to be deleterious. The nine genes and the specific variants identified in this study are evaluated in detail below.

5.2.5.1 Variants in known cancer risk genes

APC c.4088A>T

APC encodes a tumour suppressor involved in cell adhesion, cell migration, organisation of the actin and microtubule networks, as well as spindle formation and chromosome segregation (Barth, Näthke, & Nelson, 1997). Variants in *APC* are causative of a series of autosomal dominant APC-associated polyposis conditions including FAP, attenuated FAP, and GAPPS (previously described in Sections 1.1.3 and 1.1.3.4, respectively). The

penetrance of colonic adenomatous polyposis and colon cancer is nearly 100% in FAP families, and approximately 70% in attenuated FAP families (Neklason et al., 2008). The penetrance of gastrointestinal polyps and cancer in GAPPS families is currently unknown. *APC* is known to be frequently mutated in sporadic gastric cancers, however less than 3% of *APC* pathogenic variant carriers develop gastric cancer (Iwama et al., 2004).

The *APC* c.4088A>T variant substitutes a lysine for an isoleucine at codon 1363 of the *APC* protein (p.Lys1363Ile). The lysine residue is highly conserved across species and there is a moderate physicochemical difference between lysine and isoleucine. *APC* c.4088A>T is present in population ExAC databases (MAF 0.0001). To our knowledge, *APC* c.4088A>T variant has not been reported in anyone with an *APC*-related syndrome. Condel predicts this missense change to be deleterious. ClinVar currently describes *APC* c.4088A>T as a variant of uncertain significance.

***ATM* c.1157delG**

ATM encodes a 350kD kinase involved in DNA double-stranded break-repair, activation of cell-cycle checkpoints, and induction of apoptosis (Savitsky et al., 1995). Biallelic pathogenic variants in *ATM* cause ataxia-telangiectasia (A-T), a syndrome characterised by immune deficiency, progressive dysfunction of the cerebellum, and an increased risk of cancer (McKinnon, 2004; Savitsky et al., 1995). Lymphomas and leukemias are most common for A-T cases less than 20 years of age, however adults are susceptible to both lymphoid tumours and a variety of solid tumours including breast, liver, gastric, and esophageal carcinomas (Rothblum-Oviatt et al., 2016).

It has also been shown that pathogenic monoallelic *ATM* variants increase breast cancer risk (Bernstein et al., 2006). Furthermore, it is suggested monoallelic *ATM* variants increase risk of colorectal, gastric, and pancreatic

cancers (Bernstein et al., 2006; Renwick et al., 2006; D. Thompson et al., 2005). The estimated penetrance of breast cancer in three hereditary breast cancer families with a monoallelic truncating *ATM* variants was approximately 60% by 70 years of age (Chenevix-Trench et al., 2002). Recently, a series of population based studies have shown a significant enrichment of deleterious germline *ATM* variants in gastric cancer patients, affirming pathogenic *ATM* variants increase gastric cancer risk (Helgason et al., 2015; Huang et al., 2015).

ATM c.1157delG causes a frameshift variant which is predicted to cause an early truncation of the *ATM* protein. This variant has not been reported in any literature, population databases, or ClinVar. According to the ACMG/AMP guidelines for the interpretation of sequence variants (Richards et al., 2015), as *ATM* c.1157delG causes a frameshift and early truncation of the *ATM* protein it is classified as pathogenic.

***MUTYH* c.56G>A**

MUTYH encodes a DNA glycosylase involved in repair of oxidative DNA damage. Pathogenic variants in the *MUTYH* gene cause an autosomal recessive form of FAP known as MUTYH-associated polyposis (MAP) (Cheadle & Sampson, 2007). MAP is caused by biallelic pathogenic variants in germline *MUTYH* and is characterised by an extreme lifetime risk of colorectal cancer (almost 100% penetrance) (Cheadle & Sampson, 2007). Although monoallelic variants are not thought to cause high penetrance MAP, it has been proposed monoallelic *MUTYH* variants may act as low-penetrance susceptibility modifiers for colorectal cancer risk (Morak, Laner, Bacher, Keiling, & Holinski-Feder, 2010). Patients with MAP do not appear to have a rate of gastric cancer any higher than the general population (Cheadle & Sampson, 2007; Vogt et al., 2009).

The *MUTYH* c.56G>A variant replaces arginine with glutamine at codon 19 of

the *MUTYH* protein (p.Arg19Gln). The arginine residue is moderately conserved and there is a small physicochemical difference between arginine and glutamine. The glutamine amino acid residue is found in multiple mammalian species, suggesting that this missense change does not adversely affect protein function. *MUTYH* c.56G>A is present in the ExAC database (MAF 0.00006) and has been reported in an individual with breast cancer and individuals who had previously been referred for Lynch syndrome testing (Out et al., 2012; Tricarico et al., 2011; Yurgelun et al., 2015). Condel predicts the effect of this missense change to be neutral. ClinVar classifies *MUTYH* c.56G>A as a variant of uncertain significance.

***PALB2* c.574G>A**

PALB2 is important for genome maintenance and repair of DNA double stranded breaks (Rahman et al., 2007). Biallelic variants in *PALB2* are causative of hereditary Fanconi anemia and are associated with a high-risk of breast cancer (Fernandes et al., 2014). Monoallelic or heterozygous germline variants in *PALB2* have been implicated in a small number of hereditary breast and pancreatic cancers, and are associated with a moderate cancer risk (Fernandes et al., 2014). Recently pathogenic germline variants in *PALB2* were also identified in a series of studies that examined germline DNA from gastric cancer patients (Hansford et al., 2015; Lu et al., 2015; Sahasrabudhe et al., 2016). The frequent discovery of deleterious variants in *PALB2* in gastric cancer patients suggests that they may increase gastric cancer risk.

PALB2 c.574G>A causes a valine to isoleucine substitution at codon 192 of the *PALB2* protein (p.Val192Ile). The lysine residue is weakly conserved across species and there is only a small physicochemical difference between valine and isoleucine. *PALB2* c.574G>A variant is predicted to be neutral by Condel and has not been reported in the literature, population databases, or ClinVar.

According to the ACMG/AMP guidelines for the interpretation of sequence variants (Richards et al., 2015), *PALB2* c.574G>A is a variant of uncertain significance.

***PMS1* c.697C>T**

PMS1 has been commonly associated with Lynch Syndrome (syndrome described in Section 1.1.3.4), although a robust link between deleterious *PMS1* variants and cancer risk is yet to be established. On re-examination of the only Lynch syndrome family which has been described with a deleterious *PMS1* variant, a pathogenic *MSH2* variant was also discovered (Nicolaides et al., 1994). *MSH2* is also involved in DNA mismatch repair, and segregation analysis showed that only the *MSH2* variant co-segregated with colon cancer in the Lynch syndrome family (Liu et al., 2001). Additional analysis of a large series of Lynch syndrome families failed to find any *PMS1* variants and $PMS1^{-/-}$ mice did not show any cancer phenotype (Prolla et al., 1998).

PMS1 c.697C>T causes a truncating p.Gln233* variant. *PMS1* c.697C>T is rare in the ExAc population database (MAF 0.00003), but is absent from ClinVar and, to our knowledge, has not been reported in any cancer families. Given there is no irrefutable evidence *PMS1* is a cancer predisposition gene, this nonsense variant is unlikely to be of consequence.

***TP53* c.713G>A**

Sometimes described as the guardian of the genome, *TP53* is tumour suppressor gene that plays a role in apoptosis, cell-cycle regulation, genomic stability, and inhibition of angiogenesis (Bieging, Mello, & Attardi, 2014; Levine, 1997). *TP53* is frequently mutated in human cancers, and pathogenic germline variants in *TP53* are causative of Li-Fraumeni syndrome (Malkin et al., 1990; Varley, 2003). Although sarcomas, brain tumours, leukemias, breast cancers, and adrenal cortical carcinomas are typically recognised as

Li-Fraumeni syndrome-associated tumours, gastric cancer is also a part of the wider tumour spectrum (Bouaoun et al., 2016). The frequency of gastric cancer in families carrying pathogenic *TP53* variants ranges from 1·8% to 4·9%, and up to 40% of families with a pathogenic *TP53* variant will present with at least one gastric cancer diagnosis (Bouaoun et al., 2016; Masciari et al., 2011). It remains unclear if Li-Fraumeni is associated with a particular histological subtype of gastric cancer.

The c.713G>A variant replaces cysteine with tyrosine at codon 238 of the *TP53* protein (p.Cys238Tyr). The cysteine codon is highly conserved across species and there is a large physicochemical difference between cysteine and tyrosine. The cysteine residue at codon 238 is involved in coordinating a zinc ion in the DNA-binding domain of the *TP53* protein. *TP53* c.713G>A has been reported in individuals with Hodgkin's disease and breast cancer (Alsner, Yilmaz, Guldberg, Hansen, & Overgaard, 2000; Nichols et al., 2003). *In vitro* studies have shown that this missense change results in significantly decreased transactivation activity of *TP53*, and that it may act in a dominant negative fashion to reduce the transactivation activity of the wild-type allele (Monti et al., 2011). Additionally, other amino acid substitutions at this codon have been reported in families with Li-Fraumeni syndrome, all of which have also been described as severe deficiency alleles (Monti et al., 2011). ClinVar classifies *TP53* c.713G>A as likely deleterious.

5.2.5.2 Variants in novel candidate diffuse gastric cancer genes

The novel candidate diffuse gastric cancer genes *FARP2*, *FGD4*, and *LMO7* were examined in CBioPortal (Gao et al., 2013) for additional evidence *FARP2*, *FGD4*, and *LMO7* are involved in diffuse gastric cancer carcinogenesis.

Individual somatic mutations have been found throughout the coding regions of all three genes for various cancer types, including gastric cancer. However, none of these genes are frequently mutated in gastric cancer, nor were these specific mutations enriched in diffuse-type tumours.

***FARP2* c.124A>G**

FARP2 was identified as a candidate gene in the adherens junction pathway (focus area 1). *FARP2* encodes a guanine nucleotide exchange factor that is thought to be important for the remodelling of the actin-cytoskeleton (Fukuhara et al., 2004; Toyofuku et al., 2005). FARP2 activates Rac1 and Cdc42 in response to upstream signals, thereby regulating processes such as neuronal axon guidance and bone homeostasis (Toyofuku et al., 2005). Motivated by the abundant expression of *FARP2* in neurons in the adult brain, functional studies of FARP2 have primarily focused on its role in the regulation of neuronal development and morphology (Kawakita et al., 2003; Kubo et al., 2002). How the GEF activity of FARP2 is regulated remains poorly understood (He, Kuo, Rosche, & Zhang, 2013). *FARP2* is moderately expressed in the stomach (GTEx Consortium, 2015). There are no studies implicating *FARP2* as a cancer predisposition gene.

***FGD4* c.1562A>G**

FGD4 was identified as a candidate gene with Rho guanine exchange factor activity (focus area 5). *FGD4* encodes a protein known to be expressed in the nervous system where it regulates cell myelin production and facilitates the attachment of the actin cytoskeleton to the cell membrane (Y. Kim et al., 2002). Pathogenic variants in *FGD4* are known to cause autosomal recessive Charcot-Marie-Tooth disease, a disorder of the peripheral nervous system, characterised by progressive weakness and atrophy (Delague et al., 2007). *FGD4* is highly expressed in the stomach (GTEx Consortium, 2015). There is

no prior evidence that *FGD4* is associated with gastric cancer risk.

***LMO7* c.3823C>T**

LMO7 encodes a widely expressed protein-protein recognition module that is characterised by its PDZ and LIM domains. LMO7 is an important protein in the adherens junction (focus area 1), that in conjunction with α-actin, is known to connect the E-cadherin-catenin and nectin-afadin complexes (B. Harris & Lim, 2001; Kadrmas & Beckerle, 2004; Ooshio et al., 2004). The nectin-afadin complex is involved in the activation of Cdc42 and Rac small G proteins, that are known to enhance the formation of adherens junctions through reorganisation of the actin cytoskeleton (Ooshio et al., 2004). LMO7 is exclusively located at the apical surface of epithelial cells and is thought to assist in the formation and maintenance of epithelial architecture (Ooshio et al., 2004). *LMO7* is highly expressed in the stomach (GTEx Consortium, 2015). *LMO7* is not known to be important to the carcinogenesis of gastric cancer, but has previously been implicated as a cause of hereditary breast cancer in humans and lung cancer in mice (Kainu et al., 2000; Tanaka-Okamoto et al., 2009).

5.2.6 Summary of prioritised variants

After the in-depth evaluation of the nine prioritised variants (*APC* c.4088A>T, *ATM* c.1157delG, *FARP2* c.124A>G, *FGD4* c.1562A>G, *LMO7* c.3823C>T, *MUTYH* c.56G>A, *PALB2* c.574G>A, *PMS1* c.697C>T, and *TP53* c.713G>A), I concluded that both the *ATM* c.1157delG and *TP53* c.713G>A variants were deleterious and most probably predisposed to the diffuse gastric cancers diagnosed in patients X8289 and Y712, respectively. The *APC* c.4088A>T, *PALB2* c.574G>A, and *MUTYH* c.56G>A variants are variants of unknown

significance and will need to be reviewed again in light of future data. There was no evidence to suggest the nonsense *PMS1* variant increases cancer risk.

All three variants in the novel candidate diffuse gastric cancer genes *FARP2* (c.124A>G), *FGD4* (c.1562A>G), and *LMO7* (c.3823C>T) are compelling candidates for further studies. As *LMO7* is a gene important for cell-cell adhesion and the c.3823C>T variant in this gene is shared by all three members of Family 9, it is a particularly interesting candidate for functional studies. Further evidence will be required before any of these novel candidate diffuse gastric cancer genes are considered as HDGC predisposition genes. Further evidence supporting the inclusion of these genes as HDGC genes includes the re-occurrence of deleterious germline variant in these genes in further diffuse gastric cancer families, loss of heterozygosity in tumour samples, and functional *in vitro* analyses showing variants in these genes can induce cancer phenotypes.

As pathogenic variants in genes which are associated with increased cancer risk have clinical implications for the patients in the current study, I successfully validated each of the six variants I identified in genes associated with cancer risk using Sanger sequencing (Figure 5.3). Furthermore, as the variants in novel candidate diffuse gastric cancer genes may be of interest to future studies, I also attempted to validate these variants. The *FARP2* c.124A>G variant carried by patient Y8289 and the *LMO7* c.3823C>T variant carried by patients Y128 and Y130 were validated successfully (Figure 5.4). Due to insufficient sample however, I was unable to validate the *FGD4* variant carried by patient Y713 and the *LMO7* variant carried by Y129 with Sanger sequencing. Given the depth and quality of NGS reads for these variants, it is highly likely that these variants are real.

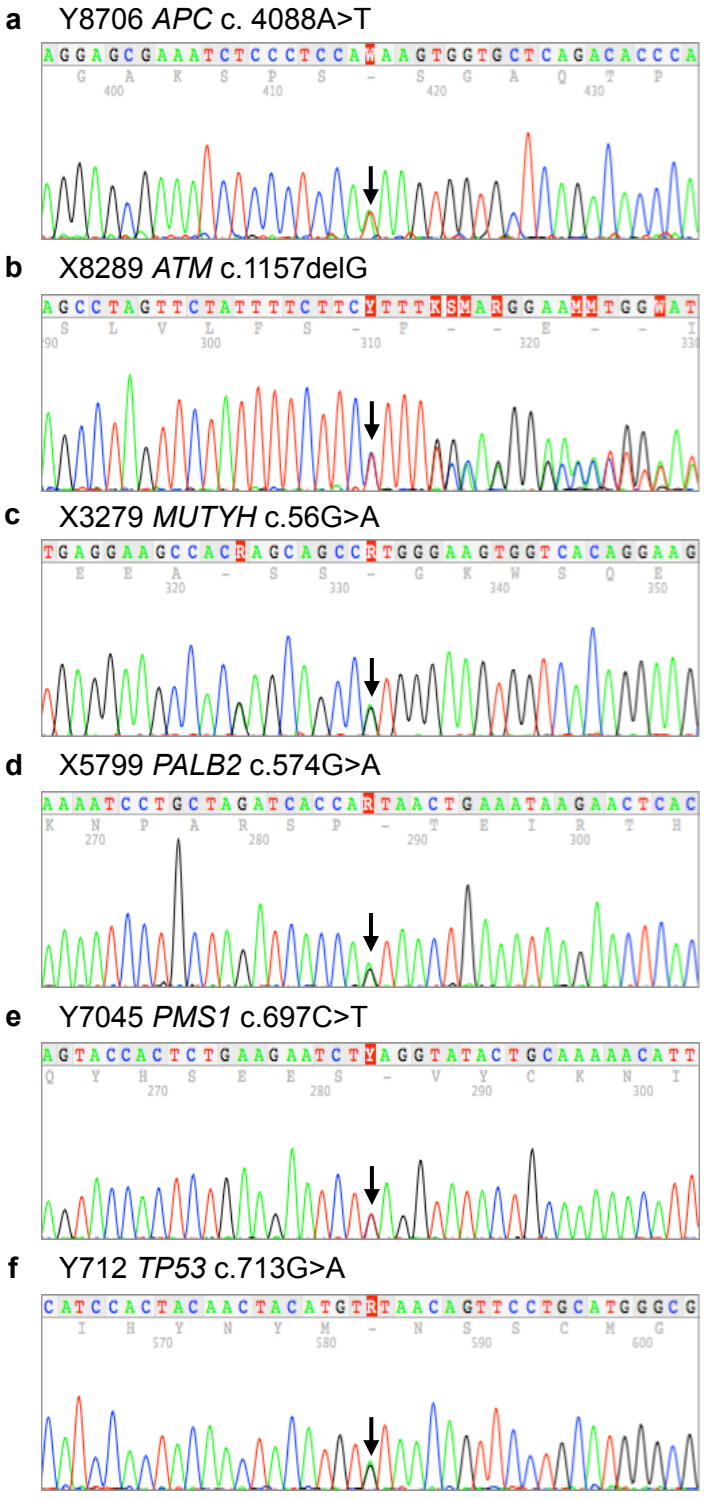


Figure 5.3: Validation of variants in genes associated with increased cancer risk. Sanger sequence validation of (a) Y8706 *APC* c.4088A>T, (b) X8289 *ATM* c.1157delG, (c) X3279 *MUTYH* c.56G>A, (d) X5799 *PALB2* c.574G>A, (e) Y7045 *PMS1* c.697C>T, and (f) Y712 *TP53* c.713G>A. Sanger sequences for a,c,d,e, and f are shown in the forward direction. Due to a repetitive sequence upstream of the *ATM* variant, Sanger sequence for (b) (*ATM* c.1157delG) is shown in reverse direction.

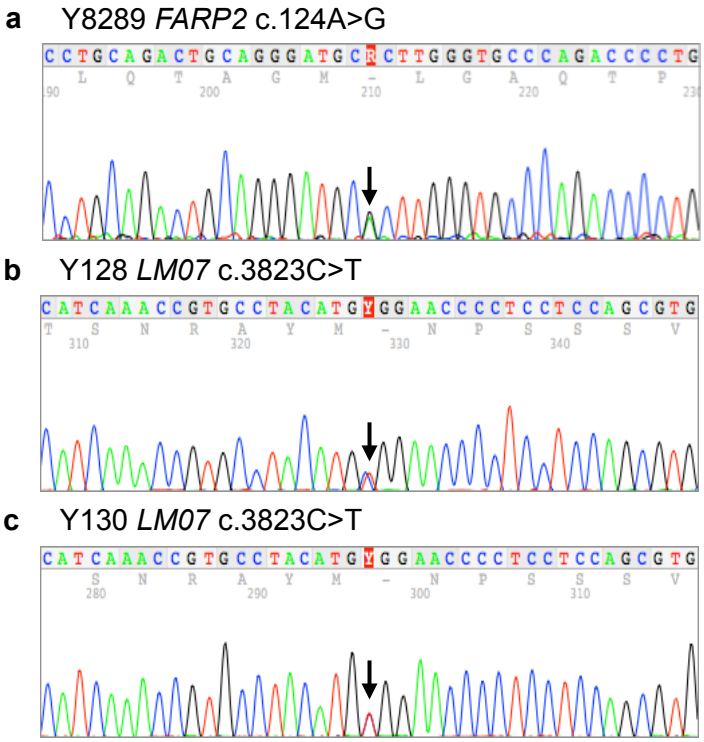


Figure 5.4: Validation of variants in candidate HDGC genes. Sanger sequence validation of (a) Y8289 *FARP2* c.124A>G, (b) Y128 *LMO7* c.3823C>T, and (c) Y130 *LMO7* c.3823C>T. All Sanger sequences are shown in the forward direction.

5.3 Discussion

In this study, whole-exome sequencing was used to search for deleterious variants in 14 diffuse gastric cancer patients that meet the clinical criteria for HDGC, but did not carry a pathogenic germline variant in *CDH1*. Variants in these cases were filtered and then prioritised for further evaluation. I identified pathogenic variants in *ATM* and *TP53*, two genes that are not associated with clinical HDGC, but are known to increase gastric cancer risk. Additional variants of interest were identified in *FARP2*, *FGD4*, and *LMO7*, genes that are important in the coordination of the actin cytoskeleton and/or cell adhesion, two cellular functions which are known to be dysregulated in diffuse-type gastric tumours (Bass et al., 2014). Until now, *FARP2*, *FGD4*, and *LMO7* were not previously associated with diffuse gastric cancer risk.

5.3.1 Pathogenic variants in *ATM* and *TP53*

Two patients in this study were found to carry deleterious variants in *ATM* and *TP53*, two genes associated with a moderate-risk of gastric cancer. Patient X8289 was diagnosed with early-onset diffuse gastric cancer (diagnosis at age 36 years) and carries the *ATM* variant, whereas patient Y712 was diagnosed with bilateral breast cancer and diffuse gastric cancer (diagnosis at age 45 years) and carries the *TP53* variant.

As previously described, monoallelic loss-off-function variants in germline *ATM* are most commonly associated with a moderately increased risk of breast cancer, but are also associated with an increased risk of various other cancer types, including gastric cancer. Unfortunately, our understanding of the gastric cancer risk associated with deleterious *ATM* variants is still

developing. Thus far, studies have shown that deleterious *ATM* variants increase gastric cancer risk more than a four-fold (OR 4.74, 95% CI 3.0–7.4) (Helgason et al., 2015) and that *ATM* variant carriers are diagnosed with gastric cancer at a significantly earlier age than non-carriers (Helgason et al., 2015; Lu et al., 2015). Additionally, *ATM* is known to be frequently mutated in gastric cell lines and some gastric tumours (Zhang et al., 2004). To date, no studies have found an association between deleterious *ATM* variants and a particular subtype of gastric cancer, or estimated the cumulative incidence of gastric cancer in *ATM* variant carriers (Helgason et al., 2015).

Similar to pathogenic *ATM* variant carriers, gastric cancer is diagnosed significantly earlier in *TP53* variant carriers than non-carriers in the general population (Masciari et al., 2011). Furthermore, although gastric cancer is not typically associated with *TP53* variants, gastric cancers make up 1.8-4.9% of all cancers diagnosed in Li-Fraumeni syndrome families (Malkin et al., 1990; Masciari et al., 2011). Little is known about the pathological features of these cancers, however, both intestinal- and diffuse-type gastric cancer have been diagnosed in Li-Fraumeni syndrome families, suggesting that these variants are not associated with a specific histological subtype (Masciari et al., 2011).

Notably, before patient Y712 was diagnosed with gastric cancer, she had previously been diagnosed with bilateral breast cancer. The identification of a pathogenic *TP53* variant in a patient with a history of multiple malignancies and breast cancer is not unusual. Breast cancer is the most common tumour type in women with Li-Fraumeni syndrome and the frequency of pathogenic *TP53* variants in early-onset breast cancers patients (aged less that 30 years at diagnosis) ranges from 1-7% (Gonzalez et al., 2009; Lalloo et al., 2006; Mouchawar et al., 2010). Strikingly, half of all Li-Fraumeni syndrome patients will develop at least one Li-Fraumeni syndrome-associated cancer by the age

of 30 years. Furthermore, approximately one third of Li-Fraumeni syndrome cancer patients will develop multiple primary cancers over their lifetime (Sorrell, Espenschied, Culver, & Weitzel, 2013).

As a result of our study and the identification of the deleterious variants in *ATM* and *TP53*, both patient X8289 and Y712, and their families, will receive genetic counselling. The management guidelines for both *TP53* and *ATM* variant carriers recommend comprehensive annual physical examination, screening for breast cancer from an early age, and to consider risk-reducing prophylactic mastectomy (Bevers et al., 2009; Daly et al., 2010). Additionally, pathogenic *ATM* variant carriers are advised to consider options for pancreatic cancer screening (Bevers et al., 2009; Canto et al., 2013). As gastric cancer is not a common malignancy associated with either *ATM* or *TP53*, neither set of management guidelines for variant carriers recommends endoscopic screening for variant carriers. Despite this, emerging evidence suggests that pathogenic variants in these genes increase the risk of early-onset gastric cancers, therefore I suggest that pathogenic variant carriers may benefit from regular endoscopic screening from an early age.

There is some evidence that gastric cancer is more prevalent in Li-Fraumeni syndrome families from Asian populations compared to Caucasian populations (Ariffin et al., 2015; I. Kim et al., 2005; Yamada et al., 2007). Although there are both genetic and environmental differences between Asian and Caucasian populations, one possible explanation for this difference is that there is a synergistic effect between pathogenic *TP53* variants and *H. pylori* infections. *H. pylori* is an environmental risk factor that is known to be more common in Asian populations (Miwa, Go, & Sato, 2002). Both *in vitro* and *in vivo* studies have shown that *H. pylori* infections can cause DNA damage and promote the expression of genes in DNA repair pathways, including *ATM* and

TP53 (Kalisperati et al., 2017; Koeppel, Garcia-Alcalde, Glowinski, Schlaermann, & Meyer, 2015; Toller et al., 2011). It is possible that by having a mutated *TP53* or *ATM*, the DNA damage response is not as effective in responding to *H. pylori* induced DNA damage. Consequently, pathogenic variant carriers may accumulate mutations caused by *H. pylori* and have a higher risk of tumour formation. Currently no studies have directly assessed *H. pylori* infection related carcinogenesis in families with *ATM* or *TP53* variants. If the risk of gastric cancer associated with *H. pylori* is higher for those with a pathogenic *ATM* or *TP53* variant, targeted screening and eradication of *H. pylori* may reduce the burden of gastric cancer.

5.3.2 Evidence LMO7 is a tumour suppressor

In addition to variants in genes previously associated with cancer risk, I identified three rare variants in candidate genes which were predicted to be deleterious. Of particular interest, the *LMO7* c.3823C>T variant was predicted to be deleterious and was shared by all three members of Family 9. *LMO7* has previously been implicated as a tumour suppressor in both lung cancer in mice and breast cancer in humans (Kainu et al., 2000; Tanaka-Okamoto et al., 2009).

In a study by Tanaka-Okamoto et al. (2009), *LMO7*-deficient mice developed irregular epithelial lesions in their terminal respiratory bronchioles, that tended to progress to lung adenocarcinoma (Tanaka-Okamoto et al., 2009). *LMO7*-deficient epithelial cells developed a protruding phenotype which characteristically associated with in-folding of the basement membrane. The cumulative incidence of lung cancer in $LMO7^{-/-}$ and $LMO7^{+/-}$ mice was 22% and 13%, respectively. Notably, it took more than 90 weeks for $LMO7^{-/-}$ mice to develop lung cancer, suggesting that *LMO7* acts as a tumour-suppressor,

and that secondary events are required for carcinogenesis (Tanaka-Okamoto et al., 2009). *LMO7* expression has been shown to be inversely correlated with the development and prognosis of human lung adenocarcinoma (Nakamura et al., 2011), supporting its role as a potential tumour suppressor in humans.

Tanaka-Okamoto et al. (2009) propose a model for how *LMO7* deficiency could be inducing lung adenocarcinoma. In their model, LMO7 is localised to the apical membrane of normal epithelial cells, and controls the spread of epithelial sheets by balancing tension between apical and basal cell surfaces. In epithelial cells lacking functional LMO7, it is proposed that the forces in the apical and basal cell surfaces are unbalanced, causing the basement membrane of epithelial cells to begin in-folding and form apical protrusions. Due to limited space in the epithelial plane, some epithelial cells may disrupt epithelial surface integrity and be displaced laterally. Tanaka-Okamoto et al. (2009) conclude that, in addition to cancer-promoting events, displaced *LMO7* deficient cells may progress to lung adenocarcinoma.

The model proposed by Tanaka-Okamoto et al. (2009) is strikingly similar to a proposed model for the initiation of diffuse gastric cancer in pathogenic *CDH1* variant carriers (Humar & Guilford, 2008). In the Humar and Guilford model, multifocal SRCC are initiated in *CDH1* variant carriers when loss of E-cadherin leads to the loss of cell-cell adhesion and mitotic spindle orientation in gastric epithelial cells. It is hypothesised that when these cells divide, a proportion of daughter cells are displaced outside of the epithelial plane into the lamina propria, a proportion of which undergo an EMT and acquire the ability to invade surrounding tissues (Humar & Guilford, 2008). Although neither model has been proven correct, the similarities between the models are clear.

LMO7 may have a role in maintaining epithelial integrity and protecting cells from invasion of microbes and extracellular toxins. In a study by Lim, Kim,

and Kim (2003), E-cadherin expression was decreased in gastric cells cultured with *H. pylori*. In contrast, the expression of *LMO7*, as well as other genes that regulate cell–cell adhesion and the cell–extracellular matrix, were up-regulated (Lim et al., 2003). The importance of this upregulation is yet to be established, however, the upregulation of *LMO7* in response to a known gastric cancer risk factor is of interest. It may be that increased *LMO7* expression is a compensatory mechanism for decreased E-cadherin, in which LMO7 stabilises cells in response to decreased cell-cell adhesion. Alternatively, over expression of *LMO7* may lead to disruption of intracellular signaling, cell differentiation, or cell adhesion, and contribute to the carcinogenic process.

5.3.3 Whole-exome sequencing studies for the identification of variants which predispose to diffuse gastric cancer

The exome represents less than 2% of the genome and contains approximately 85% of known disease-related variants (Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014). Whole-exome sequencing is a cost-effective way to identify variants in the exome and has been extremely successful in the discovery of variants and genes important in hereditary disorders (Van Dijk et al., 2014). In this study whole-exome sequencing was utilised to search for variants that may increase diffuse gastric cancer risk.

While sequencing exomes from patients with a salient family history or early-onset of cancer facilities the identification of novel variants, their discovery among the tens of thousands of variants in an exome can be overwhelming and difficult to interpret. Population based variant databases such as ESP6500, ExAC, and 1000 Genomes are useful in providing population-specific allele frequencies that can assist filtering for rare variants.

However, even after the exclusion of common variants, there are often still too many variants to conclusively identify causative variants. Therefore, additional filtering and prioritising strategies are required.

The small scale of the current whole-exome sequencing study meant it was not well powered for the statistical based discovery of re-occurring rare variants that predispose to diffuse gastric cancer (Lee, Abecasis, Boehnke, & Lin, 2014). To combat this, two separate strategies were used. The first was the prioritisation of variants in genes from focus areas. The second was the prioritisation of variants that were shared by all three members of Family 9. Naturally, both of these strategies have limitations that have been minimised in the current study.

Candidate gene studies are most notably limited by the extent of knowledge available on the disease of interest. Because of limited knowledge, candidate gene studies have a significant risk of overlooking potentially causative variants in genes that are not currently implicated in disease pathways. In the current study, I minimised the risk by using broad focus areas to identify candidate genes. Focus areas were identified from current literature regarding diffuse gastric cancer carcinogenesis, and included a breadth of genes that are not currently known to increase diffuse gastric cancer risk.

The prioritisation of variants shared by all three members of Family 9 also has a risk of overlooking potentially causative variants. By electing to only evaluate variants shared by all three family members, I did not address the possibility that one or more of these family members were a sporadic case, or that the exome sequencing may not have covered a specific locus in all three family members. Therefore, variants of interest may have been missed. However, as variants in disease pathways would have been identified in our focus areas, even a single member of Family 9 with a variant in any gene

related to one of our selected disease pathways would still have been identified by our analyses.

One of the main limitations of whole-exome sequencing compared to whole-genome sequencing is the limited coverage of variants outside the coding regions. Intronic and intergenic regions are known to encompass a large number of disease associated variants (Manolio et al., 2009). Notably, most GWAS loci lie in non-coding regions and results from the Encyclopedia of DNA Elements (ENCODE) Project suggest that many non-coding regions have important biological functions (Lee et al., 2014). Despite this limitation, the relative cost effectiveness and focus on a high-value portion of the genome suggest that exome sequencing will remain an important experimental approach for the identification of rare variants. As more intronic and intergenic variants are associated with disease, future studies will need to consider whole-genome sequencing as a way to capture variation in these areas.

Additional issues also remain persistent in whole-exome sequencing studies including strand bias, the level and variability of coverage across target regions, and false positive calls (Bertier, Hétu, & Joly, 2016). Many of these issues stem from low sequencing depth and the subsequent processing steps undertaken to generate and process variant calls. To minimise the impact of these factors in the current study, the GATK best practise guidelines were followed using the most up-to-date bioinformatic tools.

The current study was unable to definitively identify any rare variants that increase gastric cancer risk outside of known cancer predisposition genes. Ideally, if tumour samples were available, loss of heterozygosity would be looked for in patients with rare variants in novel candidate genes. Furthermore, if a large series of *CDH1* variant-negative HDGC cases were

available, novel candidate genes could be sequenced to identify further variant carriers. If variants in these genes were found in any large families, segregation analysis could be carried out to determine penetrance of candidate variants. Functional *in vitro* assays will also be useful for establishing whether any of these rare variants disrupt protein structure, trafficking, or signalling.

5.3.4 Conclusions

In conclusion, whole-exome sequencing is a powerful tool for the identification of novel variants in HDGC *CDH1* variant-negative patients. In our study I identified pathogenic variants in *ATM* and *TP53*, two genes that are not associated with HDGC, but are known to increase gastric cancer risk. Additional variants of interest were identified in *FARP2*, *FGD4*, and *LMO7*, genes which are linked to cellular processes that may be important in diffuse gastric cancers carcinogenesis. In particular, I believe *LMO7* is an interesting candidate for further studies. It is clear from this study and other HDGC studies that there is no other common gene for HDGC, but many families may carry private variants in genes rarely associated with disease. Identifying such genes, and quantifying their risk, remains one of the great challenges of human genetics.

Chapter 6

General Discussion and Conclusions

6.1 General discussion

Although the identification of a predisposing variant does not directly remove the associated risks, it enables risk stratification within a family and creates opportunities for targeted intervention. This concept was illustrated by all three of the cohorts sequenced in this thesis. By identifying one pathogenic *CDH1* variant carrier in the Chilean gastric cancer cohort, a further five asymptomatic family members were found to carry the same pathogenic variant. Several lives will almost certainly be saved by the increased surveillance and prophylactic surgery available to the family as variant carriers. Similarly, by exome sequencing a cohort of *CDH1* variant-negative gastric cancer patients, two individuals were found to carry clear pathogenic variants in *ATM* and *TP53*, and are now being counselled for the risks associated with variants in these genes. Undoubtedly, they too will benefit from the enhanced surveillance and interventions for the cancers associated with their specific variants.

The benefits of the subsequent interventions available to pathogenic *CDH1* variant carriers were also evident in the current thesis. Each of the 15 known *CDH1* variant carriers in the Māori gastric cancer cohort were all still alive five years post diagnosis. In contrast, the two variant carriers who did not appear to be aware of their carrier status both died shortly after diagnosis. Clearly, the translation of genetic screening and the identification of those who are at an extreme risk of diffuse gastric cancer into targeted interventions is saving lives.

Despite the benefits of genetic screening and subsequent interventions, the identification of a variant in a high-penetrance genes, such as *CDH1*, can occasionally lead to more questions than answers. When a clear pathogenic *CDH1* variant is found in the context of its classically associated phenotype,

management recommendations are generally well delineated. However, when a variant of uncertain significance is identified in an individual without a strong family history of disease, the quantification of risk and subsequent management can be unsettling.

Variants of uncertain significance in *CDH1* are typically rare missense variants and splice site variants that do not clearly disrupt E-cadherin function. Variants of uncertain significance are particularly hard to manage. As a prophylactic gastrectomy may be unwarranted and regular endoscopic screening may miss early-stage foci, those that are found to carry variants of uncertain significance are left with major uncertainty about their risk. Genetic parameters such as the variant frequency in healthy control populations, recurrence of the variant in independent HDGC families, and co-segregation of the pathogenic variant and cancer within a pedigree are all helpful in assessing risk. However when the variant is *de novo*, a low proportion of family members with the variant are affected by disease, the variant is not common in the healthy population, and the variant is absent from disease databases, *in silico* and *in vitro* assays are required to provide further information on pathogenicity.

In silico assays use various parameters to assess if a variant is deleterious. These parameters include the degree of interspecies conservation of the mutated site, the effect of the variant on splicing, and the impact of the variant on the protein structure to assess pathogenicity (Leong, Stuckey, Lai, Skinner, & Love, 2015). *In silico* prediction tools are quick and easy to use, and can be useful as additional evidence of pathogenicity. However these tools are prone to over estimating pathogenicity and are limited to assessment of protein structure (B. A. Thompson et al., 2013). In contrast, *in vitro* assays can be time consuming and technically demanding. In addition to being able to determine

the effect of a variant on protein structure, *in vitro* assays are able to interrogate the functional effect of variants on cellular trafficking and signalling (Figueiredo et al., 2013). Therefore, a combination of *in vitro* and *in silico* techniques seems to be the best approach to guide the genetic counselling, surveillance, and interventions offered to the carriers of variants of uncertain significance in the *CDH1* gene. Currently, several variants of uncertain significance in *CDH1* are being evaluated by *in vitro* assays at the Institute of Molecular Pathology and Immunology of the University of Porto (Portugal) (van der Post et al., 2015). Results from *in vitro* assays will undoubtedly improve our understanding of variants of uncertain significance and improve the management of patients with these variants.

The rates of variants of uncertain significance are higher in ethnic minorities and genes that are rarely sequenced (Ricker et al., 2016; Susswein et al., 2015; Yorczyk, Robinson, & Ross, 2015). However as more cancer patients and healthy population controls are sequenced, many variants of uncertain significance are being reclassified. Notably, with widespread *BRCA1* and *BRCA2* testing, the prevalence of variants of uncertain significance in these two genes has declined from approximately 20% to 2% of individuals tested (Eggington et al., 2014; Frank et al., 2002). As seen in the Māori population studied in this thesis, the frequency of specific variants can be quite different to those in large population databases. This type of information is critical in resolving the effect of variants that may be reasonably common in some minorities.

The inclusion of *CDH1* on multi-gene cancer panels has complicated the interpretation of hereditary risk. A number of recent studies using gene panels have identified unexpected and clearly deleterious *CDH1* variants in individuals whose families do not have the expected pattern of HDGC

associated cancers (Huynh & Laukaitis, 2016; Lajus & Sales, 2015; Xie et al., 2011). None of the families in these studies have a history of diffuse gastric cancer, however some, but not all, had a limited history of lobular breast cancer (Huynh & Laukaitis, 2016; Lajus & Sales, 2015; Xie et al., 2011). As more families without a classical history of HDGC are being identified, it is becoming apparent that the penetrance and effect of *CDH1* variants may be much more variable than previously thought.

One possible reason for the the lack of diffuse gastric cancer in families with pathogenic *CDH1* variants is that some *CDH1* variants may have a reduced penetrance of gastric tumours. Specific variants in cancer predisposition genes are known to be associated with a reduced penetrance, as well as a bias towards specific cancer phenotypes. For example, *BRCA1* p.Arg1699Gln is characterised by a cumulative risk of breast and ovarian cancer by the age of 70 years of only 24% (Spurdle et al., 2012), much lower than the average for pathogenic *BRCA1* variant carriers (71%) (van der Kolk et al., 2010). Furthermore, compared to the average risk estimates for breast and ovarian cancer for *BRCA1* variant carriers, the *BRCA1* p.Arg1699Gln variant is associated with a comparatively lower breast cancer risk, but increased ovarian cancer risk (Spurdle et al., 2012). Currently, the reasons for the reduced penetrance and cancer phenotypes associated with the *BRCA1* p.Arg1699Gln are unclear (Spurdle et al., 2012), and at present there are no *CDH1* variants which have been definitively shown to be associated with a reduced breast and/or gastric cancer penetrance.

An alternative explanation for a lack of diffuse gastric cancer in families with pathogenic *CDH1* variants is that *CDH1* variants are likely subject to genetic modifiers and/or environmental interactions which affect their penetrance. Both genetic and environmental factors are known to affect the penetrance of

variants in other cancer predisposition genes. For instance, breast feeding is known to reduce the risk of breast and ovarian cancer for *BRCA1* variant carriers (Cullinane et al., 2005; McLaughlin et al., 2007), and carrier status of certain SNPs is predictive of the risk of developing breast cancer for *BRCA1* variant carriers (Kuchenbaecker et al., 2017). Notably, factors that modify the penetrance of pathogenic variants have been identified for various cancer predisposition syndromes including hereditary breast and ovarian cancers (Muranen et al., 2016; Peterlongo et al., 2015), Lynch syndrome (Bellido et al., 2013), and Li-Fraumeni syndrome (Said & Malkin, 2015). At present there are no genetic or environmental factors which are known to affect the penetrance of *CDH1* variants.

NGS of multi-gene panels offers the rapid and reliable identification of variants in hundreds of genes across many samples in parallel, saving time and reducing costs associated with running multiple separate assays (Kamps et al., 2017; Kapoor et al., 2015; Sun et al., 2015). Gene panels are equally as effective for identifying variants as single gene tests, but also have the capacity to identify variants in multiple genes at the same time (Kamps et al., 2017; Kapoor et al., 2015). As seen in the current thesis, variants in genes associated with a moderate gastric cancer risk (such as *ATM* and *TP53*) can be a cause of familial gastric cancer in some *CDH1* variant-negative HDGC families. Because of this, multi-gene panel testing should be considered for all gastric cancer cases that fit the clinical criteria for HDGC testing.

Similar to the custom gene panel used by Hansford et al. (2015), a clinical gene panel for screening of those who meet the clinical criteria for HDGC testing should include genes previously shown to cause HDGC (*CDH1* and *CTNNA1*), as well as genes which have been associated with gastrointestinal cancer risk (genes described by Hansford et al. (2015)). In addition to the 40%

of HDGC cases that would be expected to carry a pathogenic *CDH1* variants, approximately 11% of *CDH1* variant-negative cases would be expected to carry a pathogenic variants in the genes tested on a clinical panel (Hansford et al., 2015). Genes that have been proposed to cause HDGC but are still awaiting extensive validation (such as *MAP3K6*, *INSR*, *FBXO24*, and *DOT1L*) could be integrated into gene panels in a research capacity, but should not be used to guide clinical management of gastric cancer patients and their families until they have been definitively shown to increase gastric cancer risk. By using a NGS gene panel, additional variants in genes associated with gastric cancer risk will be identified and will enable risk stratification and targeted interventions for those at risk.

Despite the potential benefits of gene panel testing, it is also important to consider that gene panel testing can complicate the management of at risk families. A major concern for large sequencing panels is the discovery of variants of uncertain significance, which as previously discussed, do not always contribute to risk assessment and may prompt anxiety and over-treatment. Currently, variants of uncertain significance make up approximately 40% of variants identified in broad cancer predisposition gene panels (Frey et al., 2015; Tung et al., 2015; Yurgelun et al., 2015). Furthermore, it is well know that as more genes are tested on gene panels, more variants of uncertain significance are identified (Sun et al., 2015). Because of the uncertainty associated with variants of unknown significance, despite the clear benefits of gene panels, the difficulty in managing patients who are found to carry variants of unknown significance need to be considered before panels are integrated into clinical care.

6.2 Future directions

The Māori gastric cancer and healthy control cohort described in this thesis is a significant asset that can be utilised for further studies examining the factors that influence gastric cancer risk. Additional research should examine the impact of variants in other genes that are associated with gastric cancer risk on the Māori gastric cancer population. Furthermore, in conjunction with the environmental exposure data collected for this study, there is potential to examine the DNA samples from this study for specific polymorphisms which are thought to modify environmental risk (Section 3.1.5). One possible avenue of research is to examine polymorphisms in the human interleukin-1 gene cluster which are associated with increased rates of pre-malignant lesions in response to *H. pylori* infection (El-Omar et al., 2000; Tu et al., 2008). As *H. pylori* is known to be an major contributor to the high incidence of gastric cancer for Māori, research regarding this risk factor is a priority.

Considering the positive impact genetic screening and targeted interventions has had on the New Zealand Māori gastric cancer population, it would be valuable to identify other populations that are affected by pathogenic *CDH1* variants. Interestingly, both the Pacific Island population and the indigenous Mapuche population in Chile are thought to have a high incidence of diffuse gastric cancer (Biggar et al., 2011; Heise et al., 2009; Tapia et al., 2010). Despite the high incidence, to our knowledge, pathogenic *CDH1* variants have never been documented in either group. Given the major impact *CDH1* variants have on the Māori gastric cancer population, I hypothesise that *CDH1* variants are a prevalent cause of diffuse gastric cancer in both of the Pacific Island and Mapuche populations. If the same prevalence of germline *CDH1* variants in the Pacific Island and Mapuche populations are the same as the Māori

population, regular endoscopic screening and prophylactic surgeries could be used to reduce the overall mortality rate from diffuse gastric cancers by 10%.

Future research should also focus on identifying and understanding the genetic and environmental factors that influence the penetrance of *CDH1* variants. Due to the relative rarity of *CDH1* variants in most populations, future research will require large collaborative efforts to catalogue genetic variation and exposure to environmental risk in HDGC families worldwide. These studies will be complicated by the endoscopic screening and preventative prophylactic surgery offered to known pathogenic *CDH1* variant carriers. Cases diagnosed via these methods are typically found to carry early stage T1a tumours that can remain indolent for an unknown length of time (Guilford et al., 2010; van der Post et al., 2015). Because of this, it is unclear if and when, these tumours would have progressed, limiting the utility of the genetic and environmental exposure information from these variant carriers.

A transgenic mouse model with a knockdown of *CDH1* could also be used to test the impact of different environmental factors on the progression of gastric tumours in pathogenic *CDH1* variant carriers. Previous $Cdh1^{+/-}$ heterozygous mice mouse models required treatment of with the carcinogen N-methyl-N-nitrosourea to promote carcinogenesis (Humar et al., 2009). More recently, an engineered mouse model with an inducible Cre-loxP *Cdh1* knockout in the gastric gland's parietal cell lineage was developed. However, Cre-loxP *Cdh1* knockout mice only developed early stage diffuse gastric cancer that did not progress to advanced disease (Mimata, Fukamachi, Eishi, & Yuasa, 2011). I have hypothesised that this model would be improved if the knockout of *Cdh1* could be induced in the proliferative stem cell compartment of the gastric gland (from which the cancer arises) rather than being restricted to the parietal cell lineage. For this reason, a Cre/loxP inducible mouse that

will use the *CD44* promoter to knockout *Cdh1* and red reporter construct (tdTomato) in gastric gland stem cells (CD44-Cre/$Cdh1^{loxP/loxP}$/tdTomato) have been ordered and will arrive in this laboratory soon.

To generate a heterozygous mouse model, I will breed a version of the CD44-Cre/$Cdh1^{loxP/loxP}$/tdTomato mouse that contains one wildtype *CDH1* allele (i.e. CD44-Cre/$Cdh1^{loxP/+}$/tdTomato). I hypothesise that certain environmental exposures will sensitise the *Cdh1* heterozygous mice to developing diffuse gastric cancer via the downregulation of *CDH1* expression or disruption of the functional E-cadherin. To begin to explore our hypothesis, a small number of environmental factors that associated with an increased risk of sporadic diffuse gastric cancer in the Māori have been identified. By comparing the number of diffuse tumours that develop in CD44-Cre/$Cdh1^{loxP/+}$/tdTomato mice, with and without the environmental exposures, I hope to be able to characterise the risk associated with such factors.

As demonstrated in this thesis, there are significant difficulties in identifying genes that are predisposing to diffuse gastric cancer in single probands and small families using whole-exome sequencing. If there are further genes or specific variants that are increasing diffuse gastric cancer risk, it is likely that they are rare. To identify such variants, large cohorts of *CDH1* variant-negative families will need to be sequenced in depth. When candidate genes and variants are identified, further functional analysis will be required to assess their associated risk. By identifying additional genes that are increasing diffuse gastric cancer risk, as with germline *CDH1* variants, screening and interventions can be introduced to improve patient management and save lives.

Beyond studies that identify and characterise cancer predisposing variants, the

development of synthetic lethal drugs that target E-cadherin-deficient cells should be considered a priority. *In vitro* E-cadherin loss is known to create a series of vulnerabilities in E-cadherin deficient cells (Telford et al., 2015). By targeting these vulnerabilities using drugs, tumours caused by E-cadherin loss can be weakened or destroyed. Being able to target E-cadherin deficient tumours would be particularly useful in treating those that have been identified with variants of uncertain significance, in which endoscopic screening may miss tumours but prophylactic surgery is unwarranted. If the development of drugs targeting these vulnerabilities are successful, they could be used to circumvent the need for prophylactic surgery for all pathogenic *CDH1* variant carriers and treat all sporadic E-cadherin-deficient tumours.

6.3 Concluding remarks

In summary, the current thesis describes a series of studies investigating genetic predisposition to gastric cancer. First, pathogenic *CDH1* variants were shown to be a major contributor to the high incidence of early-onset diffuse gastric cancer in the Māori population. Second, pathogenic *CDH1* variants were shown to be a rare cause of gastric cancer in a cohort of Chilean gastric cancer patients. Third, *CDH1* variant-negative HDGC patients were shown to carry pathogenic variants in genes that are not associated with HDGC, but are associated with increased gastric cancer risk, as well rare variants in novel genes important in diffuse gastric cancer carcinogenesis pathways. Importantly, as a direct result of this thesis, a series of gastric cancer patients and their families are now receiving additional targeted care and have a better prognosis than if they were to present with late stage disease.

Taken together, the studies described here demonstrate the variable frequency

of germline *CDH1* variants in different populations, the absence of other commonly mutated genes in familial diffuse gastric cancers, and the importance of genetic screening and targeted interventions for those found to be at a heightened risk.

References

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, *7*(4), 248–249.

Aird, I., Bentall, H. H., & Roberts, J. F. (1953). Relationship between cancer of stomach and the ABO blood groups. *British Medical Journal*, *1*(4814), 799.

Al-Moundhri, M. S., Al-Khanbashi, M., Al-Kindi, M., Al-Nabhani, M., Burney, I. A., Al-Farsi, A., … Al-Bahrani, B. (2010). Association of E-cadherin (CDH1) gene polymorphisms and gastric cancer risk. *World J Gastroenterol*, *16*(27), 3432–3436.

Alemán, J. O., Eusebi, L. H., Ricciardiello, L., Patidar, K., Sanyal, A. J., & Holt, P. R. (2014). Mechanisms of obesity-induced gastrointestinal neoplasia. *Gastroenterology*, *146*(2), 357–373.

Allen, B. A., & Terdiman, J. P. (2003). Hereditary polyposis syndromes and hereditary non-polyposis colorectal cancer. *Best practice & research Clinical gastroenterology*, *17*(2), 237–258.

Alsner, J., Yilmaz, M., Guldberg, P., Hansen, L. L., & Overgaard, J. (2000). Heterogeneity in the clinical phenotype of TP53 mutations in breast cancer patients. *Clinical Cancer Research*, *6*(10), 3923–3931.

Arcidiacono, B., Iiritano, S., Nocera, A., Possidente, K., Nevolo, M. T., Ventura, V., … Brunetti, A. (2012). Insulin resistance and cancer risk: an overview of the pathogenetic mechanisms. *Experimental diabetes research*, *2012*.

Ariffin, H., Chan, A., Oh, L., Abd-Ghafar, S., Ong, G., Mohamed, M., … Shamsani, J. (2015). Frequent occurrence of gastric cancer in Asian kindreds with Li–Fraumeni syndrome. *Clinical genetics*, *88*(5), 450–455.

Ascaño, J. J., Frierson, H., Moskaluk, C. A., Harper, J. C., Roviello, F., Jackson, C. E., … Powell, S. M. (2001). Inactivation of the E-cadherin gene in sporadic diffuse-type gastric cancer. *Modern Pathology*, *14*(10), 942–949.

Avizienyte, E., Launonen, V., Salovaara, R., Kiviluoto, T., & Aaltonen, L. A. (2001). E-cadherinis not frequently mutated in hereditary gastric cancer. *Journal of medical genetics*, *38*(1), 49–52.

Baker, M. G., Goodyear, R., Telfar Barnard, L., & Howden-Chapman, P. (2006). *The distribution of household crowding in New Zealand: an analysis based on 1991 to 2006 Census data*. Wellington.

Barth, A. I., Näthke, I. S., & Nelson, W. J. (1997). Cadherins, catenins and APC protein: interplay between cytoskeletal complexes and signaling pathways. *Current opinion in cell biology*, *9*(5), 683–690.

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., ... Weisenberger, D. J. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202.

Bellido, F., Guinó, E., Jagmohan-Changur, S., Seguí, N., Pineda, M., Navarro, M., ... Capellá, G. (2013). Genetic variant in the telomerase gene modifies cancer risk in Lynch syndrome. *European Journal of Human Genetics*, *21*(5), 511.

Berlth, F., Bollschweiler, E., Drebber, U., Hoelscher, A. H., & Moenig, S. (2014). Pathohistological classification systems in gastric cancer: diagnostic relevance and prognostic value. *World J Gastroenterol*, *20*(19), 5679–5684.

Bernstein, J., Teraoka, S., Southey, M., Jenkins, M., Andrulis, I., Knight, J., ... West, D. (2006). Population-based estimates of breast cancer risks associated with ATM gene variants c. 7271T> G and c. 1066-6T> G (IVS10-6T> G) from the Breast Cancer Family Registry. *Human mutation*, *27*(11), 1122.

Bertier, G., Hétu, M., & Joly, Y. (2016). Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. *BMC medical genomics*, *9*(1), 52.

Berx, G., Staes, K., van Hengel, J., Molemans, F., Bussemakers, M. J., van Bokhoven, A., & Van Roy, F. (1995). Cloning and characterization of the human invasion suppressor gene E-cadherin (CDH1). *Genomics*, *26*(2), 281–289.

Bevers, T. B., Anderson, B. O., Bonaccio, E., Buys, S., Daly, M. B., Dempsey, P. J., ... Heerdt, A. S. (2009). Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network*, *7*(10), 1060–1096.

Bieging, K. T., Mello, S. S., & Attardi, L. D. (2014). Unravelling mechanisms of p53-mediated tumour suppression. *Nature Reviews Cancer*, *14*(5), 359–370.

Biggar, M., Srinivasa, S., Wickramarachchi, B., Babor, R., Poole, G. H., & Hill, A. G. (2011). Gastric cancer location and histological subtype in Pacific people and Maori defies international trends. *The New Zealand Medical Journal (Online)*, *124*(1331).

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., ... Widaa, S. (2010). Signatures of mutation and selection in the cancer genome. *Nature*, *463*(7283), 893–898.

Blair, V. R. (2012). Familial gastric cancer: genetics, diagnosis, and management. *Surgical oncology clinics of North America*, *21*(1), 35–56.

Bodian, D. L., McCutcheon, J. N., Kothiyal, P., Huddleston, K. C., Iyer, R. K., Vockley, J. G., & Niederhuber, J. E. (2014). Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: implications for individual genome sequencing. *PloS one*, *9*(4), e94554.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*.

Boquet, P., & Ricci, V. (2012). Intoxication strategy of Helicobacter pylori VacA toxin. *Trends in microbiology*, *20*(4), 165–174.

Bosman, F. T., Carneiro, F., Hruban, R. H., & Theise, N. D. (2010). *WHO classification of tumours of the digestive system.* (No. Ed. 4). World Health Organization.

Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J., & Olivier, M. (2016). TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Human mutation*, *37*(9), 865–876.

Breast Cancer Linkage Consortium. (1999). Cancer risks in BRCA2 mutation carriers. *Journal of the National Cancer Institute*, *91*(15), 1310–1316.

Brooks-Wilson, A., Kaurah, P., Suriano, G., Leach, S., Senz, J., Grehan, N., ... Kelsey, M. (2004). Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria. *Journal of medical genetics*, *41*(7), 508–517.

Brown, L. M. (2000). Helicobacter pylori: epidemiology and routes of transmission. *Epidemiologic reviews*, *22*(2), 283–297.

Caglevic, C., Silva, S., Mahave, M., Rolfo, C., & Gallardo, J. (2016). The current situation for gastric cancer in Chile. *ecancermedicalscience*, *10*.

Caldas, C., Carneiro, F., Lynch, H. T., Yokota, J., Wiesner, G. L., Powell, S. M., ... Luben, R. (1999). Familial gastric cancer: overview and guidelines for management. *Journal of medical genetics*, *36*(12), 873–880.

Canto, M. I., Harinck, F., Hruban, R. H., Offerhaus, G. J., Poley, J., Kamel, I., ... Levy, M. J. (2013). International Cancer of the Pancreas Screening (CAPS) Consortium summit on the management of patients with increased risk for familial pancreatic cancer. *Gut*, *62*(3), 339–347.

Capelle, L. G., Van Grieken, N. C., Lingsma, H. F., Steyerberg, E. W., Klokman, W. J., Bruno, M. J., ... Kuipers, E. J. (2010). Risk and epidemiological time trends of gastric cancer in Lynch syndrome carriers in the Netherlands. *Gastroenterology*, *138*(2), 487–492.

Castillo Taucher, S. (2015). Medical genetics and genomic medicine in Chile: opportunities for improvement. *Molecular genetics & genomic medicine*, *3*(4), 243–247.

Charlton, A., Blair, V., Shaw, D., Parry, S., Guilford, P., & Martin, I. (2004). Hereditary diffuse gastric cancer: predominance of multiple foci of signet ring cell carcinoma in distal stomach and transitional zone. *Gut*, *53*(6), 814–820.

Cheadle, J. P., & Sampson, J. R. (2007). MUTYH-associated polyposis—from defect in base excision repair to clinical genetic testing. *DNA repair*, *6*(3), 274–279.

Chen, K., Yang, D., Li, X., Sun, B., Song, F., Cao, W., ... Zhao, Y. (2015). Mutational landscape of gastric adenocarcinoma in Chinese: implications for prognosis and therapy. *Proceedings of the National Academy of Sciences*, *112*(4), 1107–1112.

Chenevix-Trench, G., Spurdle, A. B., Gatei, M., Kelly, H., Marsh, A., Chen, X., ... Scott, C. (2002). Dominant negative ATM mutations in breast cancer

families. *Journal of the National Cancer Institute*, *94*(3), 205–215.
Chun, N., & Ford, J. M. (2012). Genetic testing by cancer site: stomach. *The Cancer Journal*, *18*(4), 355–363.
Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S., … Ruden, D. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80-92.
Clifford, R. J., Edmonson, M. N., Nguyen, C., & Buetow, K. H. (2004). Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, *20*(7), 1006–1014.
Collier, S., & Sater, W. F. (2004). *A history of Chile, 1808-2002* (Vol. 82). Cambridge University Press.
Cooper, D. N., & Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human genetics*, *78*(2), 151–155.
Correa, P. (1992). Human gastric carcinogenesis: a multistep and multifactorial process—first American Cancer Society award lecture on cancer epidemiology and prevention. *Cancer research*, *52*(24), 6735–6740.
Corso, G., Marrelli, D., Pascale, V., Vindigni, C., & Roviello, F. (2012). Frequency of CDH1 germline mutations in gastric carcinoma coming from high-and low-risk areas: metanalysis and systematic review of the literature. *BMC cancer*, *12*(1), 1.
Cover, T. L., & Peek, R. M., Jr. (2013). Diet, microbial virulence, and Helicobacter pylori-induced gastric cancer. *Gut microbes*, *4*(6), 482–493.
Crew, K. D., & Neugut, A. I. (2006). Epidemiology of gastric cancer. *World Journal of Gastroenterology*, *12*(3), 354.
Crovetto, M., & Uauy, R. (2013). Recommendations for cancer prevention of World Cancer Research Fund (WCRF): situational analysis for Chile [Journal Article]. *Rev Med Chil*, *141*.
Crow, J. (2013). *The Mapuche in Modern Chile*. University Press of Florida.
Cullinane, C. A., Lubinski, J., Neuhausen, S. L., Ghadirian, P., Lynch, H. T., Isaacs, C., … Friedman, E. (2005). Effect of pregnancy as a risk factor for breast cancer in BRCA1/BRCA2 mutation carriers. *International journal of cancer*, *117*(6), 988–991.
da Silva Tatley, F., Aldwell, F. E., Dunbier, A. K., & Guilford, P. J. (2003). N-terminal E-cadherin peptides act as decoy receptors for Listeria monocytogenes. *Infection and immunity*, *71*(3), 1580–1583.
Daly, M. B., Axilbund, J. E., Buys, S., Crawford, B., Farrell, C. D., Friedman, S., … Kaklamani, V. (2010). Genetic/familial high-risk assessment: breast and ovarian. *Journal of the National Comprehensive Cancer Network*, *8*(5), 562–594.
Delague, V., Jacquier, A., Hamadouche, T., Poitelon, Y., Baudot, C., Boccaccio, I., … Mègarbanè, A. (2007). Mutations in FGD4 encoding the Rho GDP/GTP exchange factor FRABIN cause autosomal recessive Charcot-Marie-Tooth type 4H. *The American Journal of Human Genetics*, *81*(1), 1–16.
de la Jara, J. J., Bastias, G., Ferreccio, C., Moscoso, C., Sagues, S., Cid, C., …

Velasquez, E. V. (2015). A snapshot of cancer in Chile: analytical frameworks for developing a cancer policy. *Biological research*, *48*(1), 10.
Desmet, F., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., & Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*, *37*(9), e67–e67.
Donner, I., Kiviluoto, T., Ristimäki, A., Aaltonen, L. A., & Vahteristo, P. (2015). Exome sequencing reveals three novel candidate predisposition genes for diffuse gastric cancer. *Familial cancer*, *14*(2), 241–246.
Ebina, Y., Ellis, L., Jarnagin, K., Edery, M., Graf, L., Clauser, E., ... Roth, R. A. (1985). The human insulin receptor cDNA: the structural basis for hormone-activated transmembrane signalling. *Cell*, *40*(4), 747–758.
Eggington, J., Bowles, K., Moyes, K., Manley, S., Esterling, L., Sizemore, S., ... Pruss, D. (2014). A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. *Clinical genetics*, *86*(3), 229–237.
El-Omar, E. M., Carrington, M., Chow, W., & McColl, K. E. (2000). Interleukin-1 polymorphisms associated with increased risk of gastric cancer. *Nature*, *404*(6776), 398–402.
Ellison-Loschmann, L., Sporle, A., Corbin, M., Cheng, S., Harawira, P., Gray, M., ... Pearce, N. (2017). Risk of stomach cancer in aotearoa/new zealand: A māori population based case-control study. *PloS one*, *12*(7), e0181581.
Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, *6*(5), e19379.
Eusebi, L. H., Zagari, R. M., & Bazzoli, F. (2014). Epidemiology of Helicobacter pylori infection. *Helicobacter*, *19*(s1), 1–5.
Everhart, J. E. (2000). Recent developments in the epidemiology of Helicobacter pylori. *Gastroenterology Clinics of North America*, *29*(3), 559–578.
Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, *136*(5), E359–E386.
Fernandes, P. H., Saam, J., Peterson, J., Hughes, E., Kaldate, R., Cummings, S., ... Roa, B. B. (2014). Comprehensive sequencing of PALB2 in patients with breast cancer suggests PALB2 mutations explain a subset of hereditary breast cancer. *Cancer*, *120*(7), 963–967.
Fernández, F. L. (2005). The technical composition of the three cultural areas of the American continent at the beginning of the twenty-first century. *Convergence. Journal of Social Sciences*.
Figueiredo, J., Söderberg, O., Simoes-Correia, J., Grannas, K., Suriano, G., & Seruca, R. (2013). The importance of E-cadherin binding partners to evaluate the pathogenicity of E-cadherin missense mutations associated to HDGC. *European Journal of Human Genetics*, *21*(3), 301–309.
Fitzgerald, R. C., Hardwick, R., Huntsman, D., Carneiro, F., Guilford, P., Blair,

V., ... Dwerryhouse, S. (2010). Hereditary diffuse gastric cancer: updated consensus guidelines for clinical management and directions for future research. *Journal of medical genetics*, *47*(7), 436–444.

Fock, K. (2014). Review article: the epidemiology and prevention of gastric cancer. *Alimentary pharmacology & therapeutics*, *40*(3), 250–260.

Forman, D., & Burley, V. (2006). Gastric cancer: global pattern of the disease and an overview of environmental risk factors. *Best practice & research Clinical gastroenterology*, *20*(4), 633–649.

Franceschi, S., & La Vecchia, C. (1994). Alcohol and the risk of cancers of the stomach and colon-rectum. *Digestive Diseases*, *12*(5), 276–289.

Frank, T. S., Deffenbaugh, A. M., Reid, J. E., Hulick, M., Ward, B. E., Lingenfelter, B., ... Critchfield, G. C. (2002). Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. *Journal of Clinical Oncology*, *20*(6), 1480–1490.

Freedman, N. D., Subar, A. F., Hollenbeck, A. R., Leitzmann, M. F., Schatzkin, A., & Abnet, C. C. (2008). Fruit and vegetable intake and gastric cancer risk in a large United States prospective cohort study. *Cancer causes & control*, *19*(5), 459–467.

Frey, M. K., Kim, S. H., Bassett, R. Y., Martineau, J., Dalton, E., Chern, J.-Y., & Blank, S. V. (2015). Rescreening for genetic mutations using multi-gene panel testing in patients who previously underwent non-informative genetic screening. *Gynecologic oncology*, *139*(2), 211–215.

Friedenson, B. (2004). BRCA1 and BRCA2 pathways and the risk of cancers other than breast or ovarian. *MedGenMed: Medscape general medicine*, *7*(2), 60–60.

Fukuhara, T., Shimizu, K., Kawakatsu, T., Fukuyama, T., Minami, Y., Honda, T., ... Takai, Y. (2004). Activation of Cdc42 by trans interactions of the cell adhesion molecules nectins through c-Src and Cdc42-GEF FRG. *Journal of Cell Biology*, *166*(3), 393–405.

Gall, T. M., & Frampton, A. E. (2013). Gene of the month: E-cadherin (CDH1). *Journal of clinical pathology*, jclinpath–2013.

Gamboa-Dominguez, A., Ubbelohde, T., Saqui-Salces, M., Romano-Mazzoti, L., Cervantes, M., Domínguez-Fonseca, C., ... Ruíz-Palacios, G. M. (2007). Salt and stress synergize H. pylori-induced gastric lesions, cell proliferation, and p21 expression in Mongolian gerbils. *Digestive diseases and sciences*, *52*(6), 1517–1526.

Gancz, H., Jones, K. R., & Merrell, D. S. (2008). Sodium chloride affects Helicobacter pylori growth and gene expression. *Journal of bacteriology*, *190*(11), 4100–4105.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., ... Cerami, E. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, *6*(269), pl1.

Garziera, M., Canzonieri, V., Cannizzaro, R., Geremia, S., Caggiari, L., De Zorzi, M., ... De Paoli, P. (2013). Identification and characterization of CDH1 germline variants in sporadic gastric cancer patients and in individuals at risk of gastric cancer. *PLoS One*, *8*(10), e77035.

Gaston, D., Hansford, S., Oliveira, C., Nightingale, M., Pinheiro, H., Macgillivray, C., ... Huang, W. Y. (2014). Germline mutations in MAP3K6 are associated with familial gastric cancer. *PLoS Genet*, *10*(10), e1004669.

Gayther, S. A., Gorringe, K. L., Ramus, S. J., Huntsman, D., Roviello, F., Grehan, N., ... MacLeod, P. (1998). Identification of germ-line E-cadherin mutations in gastric cancer families of European origin. *Cancer research*, *58*(18), 4086–4089.

Ge, S., Feng, X., Shen, L., Wei, Z., Zhu, Q., & Sun, J. (2012). Association between habitual dietary salt intake and risk of gastric cancer: a systematic review of observational studies. *Gastroenterology research and practice*, *2012*.

Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., & Chess, A. (2007). Widespread monoallelic expression on human autosomes. *Science*, *318*(5853), 1136–1140.

Gonzalez, C. A., Lujan-Barroso, L., Jenab, M., Duell, E. J., Agudo, A., Tjønneland, A., ... Kaaks, R. (2012). Fruit and vegetable intake and the risk of gastric adenocarcinoma: A reanalysis of the european prospective investigation into cancer and nutrition (EPIC-EURGAST) study after a longer follow-up. *International Journal of Cancer*, *131*(12), 2910–2919.

Gonzalez, C. A., Sala, N., & Capellá, G. (2002). Genetic susceptibility and gastric cancer risk. *International journal of cancer*, *100*(3), 249–260.

González, J. R., Armengol, L., Solé, X., Guinó, E., Mercader, J. M., Estivill, X., & Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, *23*(5), 654–655.

Gonzalez, K. D., Noltner, K. A., Buzin, C. H., Gu, D., Wen-Fong, C. Y., Nguyen, V. Q., ... Weitzel, J. N. (2009). Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *Journal of Clinical Oncology*, *27*(8), 1250–1256.

González-Pérez, A., & López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *The American Journal of Human Genetics*, *88*(4), 440–449.

Goss, P. E., Lee, B. L., Badovinac-Crnjevic, T., Strasser-Weippl, K., Chavarri-Guerra, Y., & St Louis, J. (2013). Planning cancer control in Latin America and the Caribbean. *Lancet Oncology*, *14*.

Grady, W., Willis, J., Guilford, P., Dunbier, A., Toro, T., Lynch, H., ... eong-Jin, K. (2000). Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nature genetics*, *26*(1), 16–17.

Grodecká, L., Kramárek, M., Lockerová, P., Kováčová, T., Ravčuková, B., Richterová, R., ... Freiberger, T. (2014). No major effect of the CDH1 c. 2440-6C> G mutation on splicing detected in last exon-specific splicing minigene assay. *Genes, Chromosomes and Cancer*, *53*(9), 798–801.

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660.

Guggenheim, D. E., & Shah, M. A. (2013). Gastric cancer epidemiology and

risk factors. *Journal of surgical oncology*, *107*(3), 230–236.
Guilford, P., Hopkins, J., Grady, W., Markowitz, S., Willis, J., Lynch, H., … Toro, T. (1999). E-cadherin germline mutations define an inherited cancer syndrome dominated by diffuse gastric cancer. *Human mutation*, *14*(3), 249.
Guilford, P., Hopkins, J., Harraway, J., McLeod, M., McLeod, N., Harawira, P., … Reeve, A. E. (1998). E-cadherin germline mutations in familial gastric cancer. *Nature*, *392*(6674), 402–405.
Guilford, P., Humar, B., & Blair, V. (2010). Hereditary diffuse gastric cancer: translation of CDH1 germline mutations into clinical practice. *Gastric cancer*, *13*(1), 1–10.
Hamon, M., Bierne, H., & Cossart, P. (2006). Listeria monocytogenes: a multifaceted model. *Nature Reviews Microbiology*, *4*(6), 423–434.
Hansford, S., Kaurah, P., Li-Chang, H., Woo, M., Senz, J., Pinheiro, H., … Santos, T. A. (2015). Hereditary diffuse gastric cancer syndrome: CDH1 mutations and beyond. *JAMA oncology*, *1*(1), 23–32.
Harris, B., & Lim, W. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *Journal of cell science*, *114*(18), 3219–3231.
Harris, R., Tobias, M., Jeffreys, M., Waldegrave, K., Karlsen, S., & Nazroo, J. (2006a). Effects of self-reported racial discrimination and deprivation on Māori health and inequalities in New Zealand: cross-sectional study. *The Lancet*, *367*(9527), 2005–2009.
Harris, R., Tobias, M., Jeffreys, M., Waldegrave, K., Karlsen, S., & Nazroo, J. (2006b). Racism and health: The relationship between experience of racial discrimination and health in New Zealand. *Social science & medicine*, *63*(6), 1428–1441.
He, X., Kuo, Y., Rosche, T. J., & Zhang, X. (2013). Structural basis for autoinhibition of the guanine nucleotide exchange factor FARP2. *Structure*, *21*(3), 355–364.
Heise, K., Bertran, E., Andia, M. E., & Ferreccio, C. (2009). Incidence and survival of stomach cancer in a high-risk population of Chile. *World J Gastroenterol*, *15*(15), 1854–62.
Helgason, H., Rafnar, T., Olafsdottir, H. S., Jonasson, J. G., Sigurdsson, A., Stacey, S. N., … le Roux, L. (2015). Loss-of-function variants in ATM confer risk of gastric cancer. *Nature genetics*, *47*(8), 906–910.
Helicobacter and Cancer Collaborative Group. (2001). Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. *Gut*, *49*(3), 347–353.
Howe, J., Sayed, M., Ahmed, A., Ringold, J., Larsen-Haidle, J., Merg, A., … Tinley, S. T. (2004). The prevalence of MADH4 and BMPR1A mutations in juvenile polyposis and absence of BMPR2, BMPR1B, and ACVR1 mutations. *Journal of medical genetics*, *41*(7), 484–491.
Hoyo, C., Cook, M. B., Kamangar, F., Freedman, N. D., Whiteman, D. C., Bernstein, L., … Wu, A. H. (2012). Body mass index in relation to oesophageal and oesophagogastric junction adenocarcinomas: a pooled analysis from the International BEACON Consortium. *International*

journal of epidemiology, *41*(6), 1706–1718.
Hu, B., El Hajj, N., Sittler, S., Lammert, N., Barnes, R., & Meloni-Ehrig, A. (2012). Gastric cancer: Classification, histology and application of molecular pathology. *J Gastrointest Oncol*, *3*(3), 251–61.
Huang, D., Tao, H., He, X., Long, M., Yu, S., Xia, Y., … Yan, H. (2015). Prevalence of deleterious ATM germline mutations in gastric cancer patients. *Oncotarget*, *6*(38), 40953.
Humar, B., Blair, V., Charlton, A., More, H., Martin, I., & Guilford, P. (2009). E-cadherin deficiency initiates gastric signet-ring cell carcinoma in mice and man. *Cancer research*, *69*(5), 2050–2056.
Humar, B., Fukuzawa, R., Blair, V., Dunbier, A., More, H., Charlton, A., … Guilford, P. (2007). Destabilized adhesion in the gastric proliferative zone and c-Src kinase activation mark the development of early diffuse gastric cancer. *Cancer research*, *67*(6), 2480–2489.
Humar, B., Graziano, F., Cascinu, S., Catalano, V., Ruzzo, A. M., Magnani, M., … Guilford, P. (2002). Association of CDH1 haplotypes with susceptibility to sporadic diffuse gastric cancer. *Oncogene*, *21*(53), 8192.
Humar, B., & Guilford, P. (2008). Hereditary diffuse gastric cancer and lost cell polarity: a short path to cancer. *Future Medicine*.
Humar, B., & Guilford, P. (2009). Hereditary diffuse gastric cancer: a manifestation of lost cell polarity. *Cancer science*, *100*(7), 1151–1157.
Humar, B., Toro, T., Graziano, F., Müller, H., Dobbie, Z., Kwang-Yang, H., … Parry, S. (2002). Novel germline CDH1 mutations in hereditary diffuse gastric cancer families. *Human mutation*, *19*(5), 518–525.
Huynh, J. M., & Laukaitis, C. M. (2016). Panel testing reveals nonsense and missense CDH1 mutations in families without hereditary diffuse gastric cancer. *Molecular genetics & genomic medicine*.
IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. (2010a). IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. *IARC monographs on the evaluation of carcinogenic risks to humans/World Health Organization, International Agency for Research on Cancer*, *94*, v.
IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. (2010b). IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. *IARC monographs on the evaluation of carcinogenic risks to humans/World Health Organization, International Agency for Research on Cancer*, *94*, v.
International Agency for Research on Cancer and World Health Organization. (1988). Alcohol drinking. *IARC Monographs on the evaluation of carcinogenic risks to humans and their*(supplements I).
Iriyama, T., Takeda, K., Nakamura, H., Morimoto, Y., Kuroiwa, T., Mizukami, J., … Saegusa, K. (2009). ASK1 and ASK2 differentially regulate the counteracting roles of apoptosis and inflammation in tumorigenesis. *The EMBO journal*, *28*(7), 843–853.
Iwama, T., Tamura, K., Morita, T., Hirai, T., Hasegawa, H., Koizumi, K., … Utsunomiya, J. (2004). A clinical overview of familial adenomatous

polyposis derived from the database of the Polyposis Registry of Japan. *International journal of clinical oncology*, *9*(4), 308–316.

Jaime, F., Villagrán, A., Serrano, C., Cerda, J., & Harris, P. R. (2013). Frequency of Helicobacter pylori infection in 144 school age Chilean children. *Revista medica de Chile*, *141*(10), 1249–1254.

Jakubowska, A., Scott, R., Menkiszak, J., Gronwald, J., Byrski, T., Huzarski, T., … Starzynska, T. (2003). A high frequency of BRCA2 gene mutations in Polish families with ovarian and stomach cancer. *European journal of human genetics*, *11*(12), 955–958.

Jeanes, A., Gottardi, C., & Yap, A. (2008). Cadherins and cancer: how does cadherin dysfunction promote tumor progression&quest. *Oncogene*, *27*(55), 6920–6929.

Jiang, B., Zhu, K., Shao, H., Bao, C., Ou, J., & Sun, W. (2015). Lack of association between the CDH1 polymorphism and gastric cancer susceptibility: a meta-analysis. *Scientific reports*, *5*.

Jones, E. (1964). Familial gastric cancer. *The New Zealand medical journal*, *63*, 287–296.

Jonsson, B., Bergh, A., Stattin, P., Emmanuelsson, M., & Grönberg, H. (2002). Germline mutations in E-cadherin do not explain association of hereditary prostate cancer, gastric cancer and breast cancer. *International journal of cancer*, *98*(6), 838–843.

Kadrmas, J. L., & Beckerle, M. C. (2004). The LIM domain: from the cytoskeleton to the nucleus. *Nature reviews Molecular cell biology*, *5*(11), 920–931.

Kainu, T., Juo, S. H., Desper, R., Schäffer, A. A., Gillanders, E., Rozenblum, E., … Bailey-Wilson, J. (2000). Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus. *Proceedings of the National Academy of Sciences*, *97*(17), 9603–9608.

Kakiuchi, M., Nishizawa, T., Ueda, H., Gotoh, K., Tanaka, A., Hayashi, A., … Ichimura, T. (2014). Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. *Nature genetics*, *46*(6), 583–587.

Kalisperati, P., Spanou, E., Pateras, I. S., Korkolopoulou, P., Varvarigou, A., Karavokyros, I., … Sougioultzis, S. (2017). Inflammation, DNA Damage, Helicobacter pylori and Gastric Tumorigenesis. *Frontiers in Genetics*, *8*.

Kamps, R., Brandão, R. D., Bosch, B. J., Paulussen, A. D., Xanthoulea, S., Blok, M. J., & Romano, A. (2017). Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *International Journal of Molecular Sciences*, *18*(2), 308.

Kaneko, S., & Yoshimura, T. (2001). Time trend analysis of gastric cancer incidence in Japan by histological types, 1975-1989. *British journal of cancer*, *84*(3), 400.

Kapoor, N. S., Curcio, L. D., Blakemore, C. A., Bremner, A. K., McFarland, R. E., West, J. G., & Banks, K. C. (2015). Multigene panel testing detects equal rates of pathogenic BRCA1/2 mutations and has a higher diagnostic yield compared to limited BRCA1/2 analysis alone in patients at risk for hereditary breast cancer. *Annals of surgical oncology*, *22*(10),

3282–3288.
Karimi, P., Islami, F., Anandasabapathy, S., Freedman, N. D., & Kamangar, F. (2014). Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. *Cancer Epidemiology Biomarkers & Prevention*, *23*(5), 700–713.
Kaurah, P., MacMillan, A., Boyd, N., Senz, J., De Luca, A., Chun, N., … Nikkel, S. (2007). Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. *Jama*, *297*(21), 2360–2372.
Keller, G., Vogelsang, H., Becker, I., Plaschke, S., Ott, K., Suriano, G., … Doring, C. (2004). Germline mutations of the E-cadherin (CDH1) and TP53 genes, rather than of RUNX3 and HPP1, contribute to genetic predisposition in German gastric cancer patients. *Journal of medical genetics*, *41*(6), e89–e89.
Kelley, J. R., & Duggan, J. M. (2003). Gastric cancer epidemiology and risk factors. *Journal of clinical epidemiology*, *56*(1), 1–9.
Kim, I., Kang, H., Shin, Y., Yoo, B., Yang, H., & Park, J. (2005). Familial gastric cancers with Li-Fraumeni Syndrome: a case repast. *World Journal of Gastroenterology: WJG*, *11*(26), 4124.
Kim, S., Chung, J., Jeong, T., Park, Y., Lee, J. H., Ahn, J. Y., … Lee, G. H. (2013). Searching for E-cadherin gene mutations in early onset diffuse gastric cancer and hereditary diffuse gastric cancer in Korean patients. *Familial cancer*, *12*(3), 503–507.
Kim, Y., Ikeda, W., Nakanishi, H., Tanaka, Y., Takekuni, K., Itoh, S., … Takai, Y. (2002). Association of frabin with specific actin and membrane structures. *Genes to Cells*, *7*(4), 413–420.
King, M., Marks, J. H., & Mandell, J. B. (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, *302*(5645), 643–646.
Kluijt, I., Sijmons, R. H., Hoogerbrugge, N., Plukker, J. T., de Jong, D., van Krieken, J. H., … Cats, A. (2012). Familial gastric cancer: guidelines for diagnosis, treatment and periodic surveillance. *Familial cancer*, *11*(3), 363–369.
Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38.
Koeppel, M., Garcia-Alcalde, F., Glowinski, F., Schlaermann, P., & Meyer, T. F. (2015). Helicobacter pylori infection causes characteristic DNA damage patterns in human cells. *Cell reports*, *11*(11), 1703–1713.
Kraus, C., Rau, T. T., Lux, P., Erlenbach-Wünsch, K., Löhr, S., Krumbiegel, M., … Stürzl, M. (2015). Comprehensive screening for mutations associated with colorectal cancer in unselected cases reveals penetrant and nonpenetrant mutations. *International journal of cancer*, *136*(6).
Kuchenbaecker, K. B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Healey, S., … Ramus, S. J. (2017). Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *JNCI: Journal of the National Cancer Institute*, *109*(7), djw302.

Ladeiras-Lopes, R., Pereira, A. K., Nogueira, A., Pinheiro-Torres, T., Pinto, I., Santos-Pereira, R., & Lunet, N. (2008). Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. *Cancer causes & control*, *19*(7), 689–701.

Lajus, T. B. P., & Sales, R. M. D. (2015). CDH1 germ-line missense mutation identified by multigene sequencing in a family with no history of diffuse gastric cancer. *Gene*, *568*(2), 215–219.

Lalloo, F., Varley, J., Moran, A., Ellis, D., O'Dair, L., Pharoah, P., ... Bulman, B. (2006). BRCA1, BRCA2 and TP53 mutations in very early-onset breast cancer with associated risks to relatives. *European journal of cancer*, *42*(8), 1143–1150.

Larue, L., Ohsugi, M., Hirchenhain, J., & Kemler, R. (1994). E-cadherin null mutant embryos fail to form a trophectoderm epithelium. *Proceedings of the National Academy of Sciences*, *91*(17), 8263–8267.

Lauren, P. (1965). The two histological main types of gastric carcinoma, an attempt at a histoclinical classification. *Acta Pathol. Microbiol. Scand.*, *64*, 31–49.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, *95*(1), 5–23.

Lee, S., Kang, D., Shim, K., Choe, J., Hong, W., & Choi, H. (2003). Effect of diet and Helicobacter pylori infection to the risk of early gastric cancer. *Journal of epidemiology*, *13*(3), 162–168.

Lei, H., Sjöberg-Margolin, S., Salahshor, S., Werelius, B., Jandakova, E., Hemminki, K., ... Vořechovskỳ, I. (2002). CDH1 mutations are present in both ductal and lobular breast cancer, but promoter allelic variants show no detectable breast cancer risk. *International journal of cancer*, *98*(2), 199–204.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Leong, I. U., Stuckey, A., Lai, D., Skinner, J. R., & Love, D. R. (2015). Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. *BMC medical genetics*, *16*(1), 34.

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *cell*, *88*(3), 323–331.

Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., ... Halle, D. (1997). Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. *American journal of human genetics*, *60*(5), 1059.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, J., Woods, S. L., Healey, S., Beesley, J., Chen, X., Lee, J. S., ... Pearson, J. (2016). Point mutations in exon 1B of APC reveal gastric adenocarcinoma

and proximal polyposis of the stomach as a familial adenomatous polyposis variant. *The American Journal of Human Genetics*, *98*(5), 830–842.
Li, L., Chui, R. M., Sasaki, M., Nakajima, K., Perinchery, G., Au, H. C., … Dahiya, R. (2000). A single nucleotide polymorphism in the E-cadherin gene promoter alters transcriptional activities. *Cancer research*, *60*(4), 873–876.
Lim, J. W., Kim, H., & Kim, K. H. (2003). Cell adhesion-related gene expression by Helicobacter pylori in gastric epithelial AGS cells. *The international journal of biochemistry & cell biology*, *35*(8), 1284–1296.
Lipton, L., & Tomlinson, I. (2006). The genetics of FAP and FAP-like syndromes. *Familial cancer*, *5*(3), 221–226.
Liu, T., Yan, H., Kuismanen, S., Percesepe, A., Bisgaard, M., Pedroni, M., … Peltomaki, P. (2001). The role of hPMS1 and hPMS2 in predisposing to colorectal cancer. *Cancer research*, *61*(21), 7798–7802.
Loh, J. T., Torres, V. J., & Cover, T. L. (2007). Regulation of Helicobacter pylori cagA expression in response to salt. *Cancer research*, *67*(10), 4709–4715.
Lu, C., Xie, M., Wendl, M. C., Wang, J., McLellan, M. D., Leiserson, M. D., … Ning, J. (2015). Patterns and functional implications of rare germline variants across 12 cancer types. *Nature communications*, *6*.
Lynch, H. T., Grady, W., Suriano, G., & Huntsman, D. (2005). Gastric cancer: new genetic developments. *Journal of surgical oncology*, *90*(3), 114–133.
López-Rodríguez, G., Galván-García, M., & Muzzo, B. S. (2009). Excreción urinaria de sodio en niños y adultos de una comuna de la Región Metropolitana de Santiago de Chile [Journal Article]. *Rev Chil Nutr*, *36*.
Majewski, I. J., Kluijt, I., Cats, A., Scerri, T. S., Jong, D., Kluin, R. J., … Winter, M. (2013). An α-E-catenin (CTNNA1) mutation in hereditary diffuse gastric cancer. *The Journal of pathology*, *229*(4), 621–629.
Malaguarnera, R., Sacco, A., Voci, C., Pandini, G., Vigneri, R., & Belfiore, A. (2012). Proinsulin binds with high affinity the insulin receptor isoform A and predominantly activates the mitogenic pathway. *Endocrinology*, *153*(5), 2152–2163.
Malkin, D., Li, F. P., Strong, L. C., Fraumeni Jr, J. F., Nelson, C. E., Kim, D. H., … Friend, S. H. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, *250*(4985), 1233–1238.
Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Cho, J. H. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.
Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, *24*(3), 133–141.
Marmot, M., Atinmo, T., Byers, T., Chen, J., Hirohata, T., Jackson, A., … Mann, J. (2007). Food, nutrition, physical activity, and the prevention of cancer: a global perspective.
Masciari, S., Dewanwala, A., Stoffel, E. M., Lauwers, G. Y., Zheng, H., Achatz, M. I., … Verselis, S. J. (2011). Gastric cancer in individuals with Li-Fraumeni syndrome. *Genetics in Medicine*, *13*(7), 651–657.

McDonald, A. M., Sarfati, D., Baker, M. G., & Blakely, T. (2015). Trends in Helicobacter pylori Infection Among Māori, Pacific, and European Birth Cohorts in New Zealand. *Helicobacter*, *20*(2), 139–145.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297–1303.

McKinnon, P. J. (2004). ATM and ataxia telangiectasia. *EMBO reports*, *5*(8), 772–776.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069–2070.

McLaughlin, J. R., Dryer, D., Mao, Y., Marrett, L., Morrison, H., Schacter, B., & Villeneuve, G. (2006). Canadian cancer statistics 2006. *Toronto (Canada): National Cancer Institute of Canada*.

McLaughlin, J. R., Risch, H. A., Lubinski, J., Moller, P., Ghadirian, P., Lynch, H., … Offit, K. (2007). Reproductive risk factors for ovarian cancer in carriers of BRCA1 or BRCA2 mutations: a case-control study. *The Lancet Oncology*, *8*(1), 26–34.

Medina-Franco, H., Ramos-De la Medina, A., Vizcaino, G., & Medina-Franco, J. L. (2007). Single nucleotide polymorphisms in the promoter region of the E-cadherin gene in gastric cancer: case-control study in a young Mexican population. *Annals of surgical oncology*, *14*(8), 2246–2249.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, *11*(1), 31–46.

Mimata, A., Fukamachi, H., Eishi, Y., & Yuasa, Y. (2011). Loss of E-cadherin in mouse gastric epithelial cells induces signet ring-like cells, a possible precursor lesion of diffuse gastric cancer. *Cancer science*, *102*(5), 942–950.

Ministry of Health. (2015). Annual Update of Key Results 2014/15: New Zealand Health Survey. *Wellington, New Zealand: Ministry of Health*.

Ministry of Health. (2016a). Annual Update of Key Results 2015/16: New Zealand Health Survey. *Wellington, New Zealand: Ministry of Health*.

Ministry of Health. (2016b). Cancer: New registrations and deaths 2013. *Wellington, New Zealand: Ministry of Health*.

Miwa, H., Go, M. F., & Sato, N. (2002). H. pylori and gastric cancer: the Asian enigma. *The American journal of gastroenterology*, *97*(5), 1106–1112.

Modan, B., Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., … Tucker, M. A. (2001). Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine*, *345*(4), 235–240.

Molinaro, V., Pensotti, V., Marabelli, M., Feroce, I., Barile, M., Pozzi, S., … Ranzani, G. N. (2014). Complementary molecular approaches reveal heterogeneous CDH1 germline defects in Italian patients with hereditary diffuse gastric cancer (HDGC) syndrome. *Genes, Chromosomes and Cancer*, *53*(5), 432–445.

Monti, P., Perfumo, C., Bisio, A., Ciribilli, Y., Menichini, P., Russo, D., …

Fronza, G. (2011). Dominant-negative features of mutant TP53 in germline carriers have limited impact on cancer outcomes. *Molecular Cancer Research*, *9*(3), 271–279.

Morak, M., Laner, A., Bacher, U., Keiling, C., & Holinski-Feder, E. (2010). MUTYH-associated polyposis–variability of the clinical phenotype in patients with biallelic and monoallelic MUTYH mutations and report on novel mutations. *Clinical genetics*, *78*(4), 353–363.

More, H., Humar, B., Weber, W., Ward, R., Christian, A., Lintott, C., … Harlan, M. (2007). Identification of seven novel germline mutations in the human E-cadherin (CDH1) Gene. *Human mutation*, *28*(2), 203–203.

Morita, K., Furuse, M., Fujimoto, K., & Tsukita, S. (1999). Claudin multigene family encoding four-transmembrane domain protein components of tight junction strands. *Proceedings of the National Academy of Sciences*, *96*(2), 511–516.

Moslehi, R., Chu, W., Karlan, B., Fishman, D., Risch, H., Fields, A., … Schwartz, P. (2000). BRCA1 and BRCA2 mutation analysis of 208 Ashkenazi Jewish women with ovarian cancer. *The American Journal of Human Genetics*, *66*(4), 1259–1272.

Mouchawar, J., Korch, C., Byers, T., Pitts, T. M., Li, E., McCredie, M. R., … Southey, M. C. (2010). Population-based estimate of the contribution of TP53 mutations to subgroups of early-onset breast cancer: Australian Breast Cancer Family Study. *Cancer research*, *70*(12), 4795–4800.

Muranen, T. A., Greco, D., Blomqvist, C., Aittomäki, K., Khan, S., Hogervorst, F., … Luben, R. (2016). Genetic modifiers of CHEK2* 1100delC associated breast cancer risk. *Genetics in medicine: official journal of the American College of Medical Genetics*.

Nakamura, H., Hori, K., Tanaka-Okamoto, M., Higashiyama, M., Itoh, Y., Inoue, M., … Miyoshi, J. (2011). Decreased expression of LMO7 and its clinicopathological significance in human lung adenocarcinoma. *Experimental and therapeutic medicine*, *2*(6), 1053–1057.

Neklason, D. W., Stevens, J., Boucher, K. M., Kerber, R. A., Matsunami, N., Barlow, J., … Burt, R. W. (2008). American founder mutation for attenuated familial adenomatous polyposis. *Clinical Gastroenterology and Hepatology*, *6*(1), 46–52.

Nevalainen, T., Laurén, P., & Gavin, J. (1988). The intestinal and diffuse types of gastric carcinoma in Maori and non-Maori patients in Auckland. *Scandinavian journal of gastroenterology*, *23*(5), 591–594.

Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., … Abraham, J. P. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *384*(9945), 766–781.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, *31*(13), 3812–3814.

Nichols, K. E., Heath, J. A., Friedman, D., Biegel, J. A., Ganguly, A., Mauch, P., & Diller, L. (2003). TP53, BRCA1, and BRCA2 tumor suppressor genes

are not commonly mutated in survivors of Hodgkin's disease with second primary neoplasms. *Journal of clinical oncology*, *21*(24), 4505–4509.
Nicolaides, N. C., Papadopoulos, N., Liu, B., Wei, Y., Carter, K. C., Ruben, S. M., ... Fraser, C. M. (1994). Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature*, *371*(6492), 75.
Oliveira, C., de Bruin, J., Nabais, S., Ligtenberg, M., Moutinho, C., Nagengast, F. M., ... Carneiro, F. (2004). Intragenic deletion of CDH1 as the inactivating mechanism of the wild-type allele in an HDGC tumour. *Oncogene*, *23*(12), 2236–2240.
Oliveira, C., Ferreira, P., Nabais, S., Campos, L., Ferreira, A., Cirnes, L., ... Dias, L. M. (2004). E-Cadherin (CDH1) and p53 rather than SMAD4 and Caspase-10 germline mutations contribute to genetic predisposition in Portuguese gastric cancer patients. *European journal of cancer*, *40*(12), 1897–1903.
Oliveira, C., Senz, J., Kaurah, P., Pinheiro, H., Sanges, R., Haegert, A., ... Keller, G. (2009). Germline CDH1 deletions in hereditary diffuse gastric cancer families. *Human molecular genetics*, *18*(9), 1545–1555.
Oliveira, C., Seruca, R., & Carneiro, F. (2009). Hereditary gastric cancer. *Best Practice & Research Clinical Gastroenterology*, *23*(2), 147–157.
Oliver, W. H. (1981). *The Oxford History of New Zealand*. Oxford University Press, USA.
Onder, T. T., Gupta, P. B., Mani, S. A., Yang, J., Lander, E. S., & Weinberg, R. A. (2008). Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. *Cancer research*, *68*(10), 3645–3654.
Ooshio, T., Irie, K., Morimoto, K., Fukuhara, A., Imai, T., & Takai, Y. (2004). Involvement of LMO7 in the association of two cell-cell adhesion molecules, nectin and E-cadherin, through afadin and α-actinin in epithelial cells. *Journal of Biological Chemistry*, *279*(30), 31365–31373.
Out, A. A., Wasielewski, M., Huijts, P. E., van Minderhout, I. J., Houwing-Duistermaat, J. J., Tops, C. M., ... van Asperen, C. J. (2012). MUTYH gene variants and breast cancer in a Dutch case-control study. *Breast cancer research and treatment*, *134*(1), 219–227.
Papusheva, E., & Heisenberg, C. (2010). Spatial organization of adhesion: force-dependent regulation and function in tissue morphogenesis. *The EMBO journal*, *29*(16), 2753–2768.
Paredes, J., Figueiredo, J., Albergaria, A., Oliveira, P., Carvalho, J., Ribeiro, A. S., ... Pinheiro, H. (2012). Epithelial E-and P-cadherins: role and clinical significance in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, *1826*(2), 297–311.
Park, B., Shin, A., Park, S. K., Ko, K., Ma, S. H., Lee, E., ... Yoo, K. Y. (2011). Ecological study for refrigerator use, salt, vegetable, and fruit intakes, and gastric cancer. *Cancer Causes & Control*, *22*(11), 1497–1502.
Park, J., Park, S., Hong, K., Han, Y., Jang, S., Kim, E., & Hahm, K. (2014). Special Licorice Extracts Containing Lowered Glycyrrhizin and Enhanced Licochalcone A Prevented Helicobacter pylori-Initiated, Salt Diet-Promoted Gastric Tumorigenesis. *Helicobacter*, *19*(3), 221–236.

Parsonnet, J., Friedman, G., Orentreich, N., & Vogelman, H. (1997). Risk for gastric cancer in people with CagA positive or CagA negative Helicobacter pylori infection. *Gut*, *40*(3), 297–301.

Parsonnet, J., Vandersteen, D., Goates, J., Sibley, R. K., Pritikin, J., & Chang, Y. (1991). Helicobacter pylori infection in intestinal- and diffuse-type gastric adenocarcinomas. *Journal of the National Cancer Institute*, *83*(9), 640–643.

Peek, R. M., & Crabtree, J. E. (2006). Helicobacter infection and gastric neoplasia. *The Journal of pathology*, *208*(2), 233–248.

Peleteiro, B., Lopes, C., Figueiredo, C., & Lunet, N. (2011). Salt intake and gastric cancer risk according to Helicobacter pylori infection, smoking, tumour site and histological type. *British journal of cancer*, *104*(1), 198–207.

Pentecost, M., Otto, G., Theriot, J. A., & Amieva, M. R. (2006). Listeria monocytogenes invades the epithelial junctions at sites of cell extrusion. *PLoS Pathog*, *2*(1), e3.

Peterlongo, P., Chang-Claude, J., Moysich, K. B., Rudolph, A., Schmutzler, R. K., Simard, J., ... Wilkening, S. (2015). Candidate genetic modifiers for breast and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers. *Cancer Epidemiology and Prevention Biomarkers*, *24*(1), 308–316.

Pharoah, P., Guilford, P., Caldas, C., International Gastric Cancer Linkage Consortium, & International Gastric Cancer Linkage Consortium. (2001). Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology*, *121*(6), 1348–1353.

Pinheiro, H., Bordeira-Carriço, R., Seixas, S., Carvalho, J., Senz, J., Oliveira, P., ... Seruca, R. (2010). Allele-specific CDH1 downregulation and hereditary diffuse gastric cancer. *Human molecular genetics*, *19*(5), 943–952.

Polk, D. B., & Peek, R. M. (2010). Helicobacter pylori: gastric cancer and beyond. *Nature Reviews Cancer*, *10*(6), 403–414.

Polkowski, W., van Sandick, J. W., Offerhaus, G. J. A., ten Kate, F. J., Mulder, J., Obertop, H., & van Lanschot, J. J. B. (1999). Prognostic value of Lauren classification and c-erbB-2 oncogene overexpression in adenocarcinoma of the esophagus and gastroesophageal junction. *Annals of Surgical Oncology*, *6*(3), 290–297.

Prolla, T. A., Baker, S. M., Harris, A. C., Tsao, J., Yao, X., Bronner, C. E., ... Shibata, D. (1998). Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DMA mismatch repair. *Nature genetics*, *18*(3), 276–279.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature*, *505*(7483), 302–308.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., ... Jayatilake, H. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature genetics*, *39*(2), 165–167.

Rahner, N., Steinke, V., Schlegelberger, B., Olschwang, S., Eisinger, F., & Hutter,

P. (2010). Clinical utility gene card for: Lynch syndrome (MLH1, MSH2, MSH6, PMS2). *Eur J Hum Genet*, *18*(9).
Reese, M. G., Eeckman, F. H., Kulp, D., & Haussler, D. (1997). Improved splice site detection in Genie. *Journal of computational biology*, *4*(3), 311–323.
Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., … le Roux, L. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature genetics*, *38*(8), 873–875.
Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, gkr407.
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Voelkerding, K. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–423.
Ricker, C., Culver, J. O., Lowstuter, K., Sturgeon, D., Sturgeon, J. D., Chanock, C. R., … Gruber, S. B. (2016). Increased yield of actionable mutations using multi-gene panels to assess hereditary cancer susceptibility in an ethnically diverse clinical cohort. *Cancer genetics*, *209*(4), 130–137.
Ridley, A. J., Schwartz, M. A., Burridge, K., Firtel, R. A., Ginsberg, M. H., Borisy, G., … Horwitz, A. R. (2003). Cell migration: integrating signals from front to back. *Science*, *302*(5651), 1704–1709.
Risinger, J. I., Berchuck, A., Kohler, M. F., & Boyd, J. (1994). Mutations of the E-cadherin gene in human gynecologic cancers. *Nature genetics*, *7*(1), 98–102.
Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, *29*(1), 24–26.
Robson, B., Purdie, G., & Cormack, D. (2010). Unequal impact II: Māori and non-Māori cancer statistics by deprivation and rural-urban Status, 2002-2006. *Wellington: Ministry of Health*.
Rogers, W. M., Dobo, E., Norton, J. A., Van Dam, J., Jeffrey, R. B., Huntsman, D. G., … Longacre, T. A. (2008). Risk-reducing total gastrectomy for germline mutations in E-cadherin (CDH1): pathologic findings with clinical implications. *The American journal of surgical pathology*, *32*(6), 799–809.
Rothblum-Oviatt, C., Wright, J., Lefton-Greif, M. A., McGrath-Morrow, S. A., Crawford, T. O., & Lederman, H. M. (2016). Ataxia telangiectasia: a review. *Orphanet Journal of Rare Diseases*, *11*(1), 159.
Sadler, K., Nicholson, S., Steer, T., Gill, V., Bates, B., & Tipping, S. (2011). *Assessment of Dietary Sodium Levels Among Adults (aged 19–64) in England* [Book]. UK: Department of Health.
Sahasrabudhe, R., Lott, P., Bohorquez, M., Toal, T., Estrada, A. P., Suarez, J. J., … Mantilla, A. (2016). Germline Mutations in PALB2, BRCA1, and RAD51C, Which Regulate DNA Recombination Repair, in Patients with Gastric Cancer. *Gastroenterology*.

Said, B. I., & Malkin, D. (2015). A functional variant in miR-605 modifies the age of onset in Li-Fraumeni syndrome. *Cancer genetics*, *208*(1), 47–51.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467.

Savitsky, K., Sfez, S., Tagle, D. A., Ziv, Y., Sartiel, A., Collins, F. S., ... Rotman, G. (1995). The complete sequence of the coding region of the ATM gene reveals similarity to cell cycle regulators in different species. *Human Molecular Genetics*, *4*(11), 2025–2032.

Schuetz, J. M., Leach, S., Kaurah, P., Jeyes, J., Butterfield, Y., Huntsman, D., & Brooks-Wilson, A. R. (2012). Catenin family genes are not commonly mutated in hereditary diffuse gastric cancer. *Cancer Epidemiology Biomarkers & Prevention*, *21*(12), 2272–2274.

Service, N. Z. H. I. (2006). Cancer patient survival: covering the period 1994 to 2003. *Wellington: Ministry of Health*.

Shibata, D., & Weiss, L. (1992). Epstein-Barr virus-associated gastric adenocarcinoma. *The American journal of pathology*, *140*(4), 769.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., ... Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, *34*(1), 57–65.

Shiozaki, H., Iihara, K., Oka, H., Kadowaki, T., Matsui, S., Gofuku, J., ... Mori, T. (1994). Immunohistochemical detection of alpha-catenin expression in human cancers. *The American journal of pathology*, *144*(4), 667.

Signal, V., Sarfati, D., Cunningham, R., Gurney, J., Koea, J., & Ellison-Loschmann, L. (2015). Indigenous inequities in the presentation and management of stomach cancer in New Zealand: a country with universal health care coverage. *Gastric Cancer*, *18*(3), 571–579.

Skierucha, M., Milne, A. N., Offerhaus, G. J. A., Polkowski, W. P., Maciejewski, R., & Sitarz, R. (2016). Molecular alterations in gastric cancer with special reference to the early-onset subtype. *World journal of gastroenterology*, *22*(8), 2460.

Smith, A. M., Heisler, L. E., Onge, R. P. S., Farias-Hesson, E., Wallace, I. M., Bodeau, J., ... Nislow, C. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic acids research*, gkq368.

Sorrell, A. D., Espenschied, C. R., Culver, J. O., & Weitzel, J. N. (2013). Tumor protein p53 (TP53) testing and Li-Fraumeni syndrome. *Molecular diagnosis & therapy*, *17*(1), 31-47.

Spigelman, A., Talbot, I., Williams, C., Domizio, P., & Phillips, R. (1989). Upper gastrointestinal cancer in patients with familial adenomatous polyposis. *The Lancet*, *334*(8666), 783–785.

Spurdle, A. B., Whiley, P. J., Thompson, B., Feng, B., Healey, S., Brown, M. A., ... Van den Ouweland, A. M. (2012). BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk. *Journal of medical genetics*, *49*(8), 525–532.

Stalnikowicz, R., & Benbassat, J. (1990). Risk of gastric cancer after gastric surgery for benign disorders. *Archives of internal medicine*, *150*(10), 2022–2026.

Statistics New Zealand. (2013a). Census QuickStats about culture and identity. *Statistics New Zealand*.

Statistics New Zealand. (2013b). Census quickstats about Māori. *Wellington, New Zealand: Statistics New Zealand*.

Stone, E. A., & Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome research*, *15*(7), 978–986.

Strong, V. E., D'Amico, T. A., Kleinberg, L., & Ajani, J. (2013). Impact of the 7th Edition AJCC staging classification on the NCCN clinical practice guidelines in oncology for gastric and esophageal cancers. *Journal of the National Comprehensive Cancer Network*, *11*(1), 60–66.

Sun, Y., Ruivenkamp, C. A., Hoffer, M. J., Vrijenhoek, T., Kriek, M., Asperen, C. J., … Santen, G. W. (2015). Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Human mutation*, *36*(6), 648–655.

Sun, Y., Zhang, J., & Ma, L. (2014). α-catenin: A tumor suppressor beyond adherens junctions. *Cell cycle*, *13*(15), 2334–2339.

Suriano, G., Oliveira, C., Ferreira, P., Machado, J. C., Bordin, M. C., De Wever, O., … Richards, F. M. (2003). Identification of CDH1 germline missense mutations associated with functional inactivation of the E-cadherin protein in young gastric cancer probands. *Human molecular genetics*, *12*(5), 575–582.

Suriano, G., Seixas, S., Rocha, J., & Seruca, R. (2006). A model to infer the pathogenic significance of CDH1 germline missense variants. *Journal of Molecular Medicine*, *84*(12), 1023–1031.

Suriano, G., Yew, S., Ferreira, P., Senz, J., Kaurah, P., Ford, J. M., … Oliveira, M. J. (2005). Characterization of a recurrent germ line mutation of the E-cadherin gene: implications for genetic testing and clinical management. *Clinical cancer research*, *11*(15), 5401–5409.

Susswein, L. R., Marshall, M. L., Nusbaum, R., Postula, K. J. V., Weissman, S. M., Yackowski, L., … Gibellini, F. (2015). Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. *Genetics in Medicine*, *18*(8), 823–832.

Takahashi, M., Kokubo, T., Furukawa, F., Kurokawa, Y., & Hayashi, Y. (1984). Effects of sodium chloride, saccharin, phenobarbital and aspirin on gastric carcinogenesis in rats after initiation with N-methyl-N-nitro-N-nitrosoguanidine. *GANN Japanese Journal of Cancer Research*, *75*(6), 494–501.

Tanaka-Okamoto, M., Hori, K., Ishizaki, H., Hosoi, A., Itoh, Y., Wei, M., … Miyoshi, J. (2009). Increased susceptibility to spontaneous lung cancer in mice lacking LIM-domain only 7. *Cancer science*, *100*(4), 608–616.

Taniguchi, C. M., Emanuelli, B., & Kahn, C. R. (2006). Critical nodes in signalling pathways: insights into insulin action. *Nature reviews Molecular cell biology*, *7*(2), 85–96.

Tapia, E., Roa, S., Carlos, J., Manterola, D., Villaseca, H., Gutierrez, M., ... Guzman, G. (2010). Changes in morphological an clinical characteristics of gastric cancer from 1986 to 2005. *Revista Chilena De Cirugia*, *62*(2), 125–130.

Tatematsu, M., Takahashi, M., Fukushima, S., Hananouchi, M., & Shirai, T. (1975). Effects in rats of sodium chloride on experimental gastric cancers induced by N-methyl-N-nitro-N-nitrosoguanidine or 4-nitroquinoline-1-oxide. *Journal of the National Cancer Institute*, *55*(1), 101–106.

Telford, B. J., Chen, A., Beetham, H., Frick, J., Brew, T. P., Gould, C. M., ... Guilford, P. (2015). Synthetic Lethal Screens Identify Vulnerabilities in GPCR Signaling and Cytoskeletal Organization in E-Cadherin-Deficient Cells. *Molecular cancer therapeutics*, *14*(5), 1213–1223.

Thompson, B. A., Greenblatt, M. S., Vallee, M. P., Herkert, J. C., Tessereau, C., Young, E. L., ... Mooney, S. D. (2013). Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. *Human mutation*, *34*(1), 255–265.

Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., ... Easton, D. F. (2005). Cancer risks and mortality in heterozygous ATM mutation carriers. *Journal of the National Cancer Institute*, *97*(11), 813–822.

Thompson, D. E., Mabuchi, K., Ron, E., Soda, M., Tokunaga, M., Ochikubo, S., ... Preston, D. L. (1994). Cancer incidence in atomic bomb survivors. Part II: Solid tumors, 1958-1987. *Radiation research*, *137*(2s), S17–S67.

Thumkeo, D., Watanabe, S., & Narumiya, S. (2013). Physiological roles of Rho and Rho effectors in mammals. *European journal of cell biology*, *92*(10), 303–315.

Tiwari, N., Gheldof, A., Tatari, M., & Christofori, G. (2012). EMT as the ultimate survival mechanism of cancer cells. In *Seminars in cancer biology* (Vol. 22, pp. 194–207).

Toller, I. M., Neelsen, K. J., Steger, M., Hartung, M. L., Hottiger, M. O., Stucki, M., ... Muller, A. (2011). Carcinogenic bacterial pathogen Helicobacter pylori triggers DNA double-strand breaks and a DNA damage response in its host cells. *Proceedings of the National Academy of Sciences*, *108*(36), 14944–14949.

Toyofuku, T., Yoshida, J., Sugimoto, T., Zhang, H., Kumanogoh, A., Hori, M., & Kikutani, H. (2005). FARP2 triggers signals for Sema3A-mediated axonal repulsion. *Nature neuroscience*, *8*(12), 1712-1719.

Tramacere, I., Negri, E., Pelucchi, C., Bagnardi, V., Rota, M., Scotti, L., ... Boffetta, P. (2011). A meta-analysis on alcohol drinking and gastric cancer risk. *Annals of oncology*, mdr135.

Tricarico, R., Crucianelli, F., Alvau, A., Orlando, C., Sestini, R., Tonelli, F., ... Genuardi, M. (2011). High resolution melting analysis for a rapid identification of heterozygous and homozygous sequence changes in the MUTYH gene. *BMC cancer*, *11*(1), 305.

Tsugane, S., & Sasazuki, S. (2007). Diet and the risk of gastric cancer: review of epidemiological evidence. *Gastric cancer*, *10*(2), 75–83.

Tu, S., Bhagat, G., Cui, G., Takaishi, S., Kurt-Jones, E. A., Rickman, B., … Wang, T. C. (2008). Overexpression of interleukin-1β induces gastric inflammation and cancer and mobilizes myeloid-derived suppressor cells in mice. *Cancer cell*, *14*(5), 408–419.

Tung, N., Battelli, C., Allen, B., Kaldate, R., Bhatnagar, S., Bowles, K., … Krejdovsky, J. (2015). Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. *Cancer*, *121*(1), 25–33.

Tung, N., Lin, N. U., Kidd, J., Allen, B. A., Singh, N., Wenstrup, R. J., … Garber, J. E. (2016). Frequency of germline mutations in 25 cancer susceptibility genes in a sequential series of patients with breast cancer. *Journal of Clinical Oncology*, *34*(13), 1460–1468.

Utsunomiya, J., Gocho, H., Miyanaga, T., Hamaguchi, E., & Kashimure, A. (1975). Peutz-Jeghers syndrome: its natural course and management. *The Johns Hopkins Medical Journal*, *136*(2), 71–82.

van der Post, R. S., Vogelaar, I. P., Carneiro, F., Guilford, P., Huntsman, D., Hoogerbrugge, N., … Bardram, L. (2015). Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline CDH1 mutation carriers. *Journal of medical genetics*, *52*(6), 361–374.

van der Kolk, D. M., de Bock, G. H., Leegte, B. K., Schaapveld, M., Mourits, M. J., de Vries, J., … Oosterwijk, J. C. (2010). Penetrance of breast cancer, ovarian cancer and contralateral breast cancer in BRCA1 and BRCA2 families: high cancer incidence at older age. *Breast cancer research and treatment*, *124*(3), 643–651.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, *30*(9), 418–426.

Van Lier, M., Wagner, A., Mathus-Vliegen, E., Kuipers, E., Steyerberg, E., & Van Leerdam, M. (2010). High cancer risk in Peutz–Jeghers syndrome: a systematic review and surveillance recommendations. *The American journal of gastroenterology*, *105*(6), 1258–1264.

Van Roy, F., & Berx, G. (2008). The cell-cell adhesion molecule E-cadherin. *Cellular and molecular life sciences*, *65*(23), 3756–3788.

Varley, J. (2003). Germline TP53 mutations and Li-Fraumeni syndrome. *Human mutation*, *21*(3), 313–320.

Vasen, H. F., Moeslein, G., Alonso, A., Aretz, S., Bernstein, I., Bertario, L., … Colas, C. (2008). Guidelines for the clinical management of familial adenomatous polyposis (FAP). *Gut*, *57*(5), 704–713.

Vasioukhin, V., Bauer, C., Degenstein, L., Wise, B., & Fuchs, E. (2001). Hyperproliferation and defects in epithelial polarity upon conditional ablation of α-catenin in skin. *Cell*, *104*(4), 605–617.

Vleminckx, K., Vakaet, L., Mareel, M., Fiers, W., & Van Roy, F. (1991). Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell*, *66*(1), 107–119.

Vogt, S., Jones, N., Christian, D., Engel, C., Nielsen, M., Kaufmann, A., … Hes, F. J. (2009). Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. *Gastroenterology*, *137*(6), 1976–1985.

Wan, M., Lee, S. S. J., Zhang, X., Houwink-Manville, I., Song, H., Amir, R. E., ... Zoghbi, H. Y. (1999). Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. *The American Journal of Human Genetics*, *65*(6), 1520–1529.

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H., Shi, S. T., ... Chan, K. H. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics*, *46*(6), 573–582.

Ward, E., Jemal, A., Cokkinides, V., Singh, G. K., Cardinez, C., Ghafoor, A., & Thun, M. (2004). Cancer disparities by race/ethnicity and socioeconomic status. *CA: a cancer journal for clinicians*, *54*(2), 78–93.

Warner, E., Foulkes, W., Goodwin, P., Meschino, W., Blondal, J., Paterson, C., ... Di Prospero, L. (1999). Prevalence and penetrance of BRCA1 and BRCA2 gene mutations in unselected Ashkenazi Jewish women with breast cancer. *Journal of the National Cancer Institute*, *91*(14), 1241–1247.

Washington, K. (2010, Dec 01). 7th edition of the ajcc cancer staging manual: Stomach. *Annals of Surgical Oncology*, *17*(12), 3077–3079. Retrieved from `https://doi.org/10.1245/s10434-010-1362-z` doi: 10.1245/s10434-010-1362-z

Weber, G. F., Bjerke, M. A., & DeSimone, D. W. (2011). Integrins and cadherins join forces to form adhesive networks. *J Cell Sci*, *124*(8), 1183–1193.

Wennerberg, K., Rossman, K. L., & Der, C. J. (2005). The Ras superfamily at a glance. *J cell Sci*, *118*(5), 843–846.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wiseman, M. (2008). The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. *Proceedings of the Nutrition Society*, *67*(03), 253–256.

Wong, S. S., Kim, K., Ting, J. C., Yu, K., Fu, J., Liu, S., ... Wang, J. (2014). Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing. *Nature communications*, *5*, 5477.

Worthley, D., Phillips, K., Wayte, N., Schrader, K., Healey, S., Kaurah, P., ... Cullen, D. (2012). Gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS): a new autosomal dominant syndrome. *Gut*, *61*(5), 774–779.

Wu, M., Huang, S., Chang, Y., Lin, M., Shun, C., Chang, M., ... Lin, J. (2002). Association of the- 160CA promoter polymorphism of E-cadherin gene with gastric carcinoma risk. *Cancer*, *94*(5), 1443–1448.

Xie, Z. M., Li, L. S., Laquet, C., Penault-Llorca, F., Uhrhammer, N., Xie, X. M., & Bignon, Y. J. (2011). Germline mutations of the E-cadherin gene in families with inherited invasive lobular breast carcinoma but no diffuse gastric cancer. *Cancer*, *117*(14), 3112–3117.

Yaghoobi, M., Bijarchi, R., & Narod, S. (2010). Family history and the risk of gastric cancer. *British journal of cancer*, *102*(2), 237–242.

Yamada, H., Shinmura, K., Ito, H., Kasami, M., Sasaki, N., Shima, H., ... Tsuneyoshi, T. (2011). Germline alterations in the CDH1 gene in familial

gastric cancer in the Japanese population. *Cancer science*, *102*(10), 1782–1788.

Yamada, H., Shinmura, K., Okudela, K., Goto, M., Suzuki, M., Kuriki, K., … Sugimura, H. (2007). Identification and characterization of a novel germ line p53 mutation in familial gastric cancer in the Japanese population. *Carcinogenesis*, *28*(9), 2013–2018.

Yang, P., Zhou, Y., Chen, B., Wan, H., Jia, G., Bai, H., & Wu, X. (2009). Overweight, obesity and gastric cancer risk: results from a meta-analysis of cohort studies. *European journal of cancer*, *45*(16), 2867–2873.

Yao, F., Kausalya, J. P., Sia, Y. Y., Teo, A. S., Lee, W. H., Ong, A. G., … Liu, X. (2015). Recurrent fusion genes in gastric cancer: CLDN18-ARHGAP26 induces loss of epithelial integrity. *Cell reports*, *12*(2), 272–285.

Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology*, *11*(2-3), 377–394.

Yorczyk, A., Robinson, L., & Ross, T. (2015). Use of panel tests in place of single gene tests in the cancer genetics clinic. *Clinical genetics*, *88*(3), 278–282.

Yurgelun, M. B., Allen, B., Kaldate, R. R., Bowles, K. R., Judkins, T., Kaushik, P., … Syngal, S. (2015). Identification of a variety of mutations in cancer predisposition genes in patients with suspected Lynch syndrome. *Gastroenterology*, *149*(3), 604–613.

Zang, Z. J., Ong, C. K., Cutcutache, I., Yu, W., Zhang, S. L., Huang, D., … Lim, S. (2011). Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing. *Cancer research*, *71*(1), 29–39.

Zhang, L., Jia, G., Li, W., Guo, R.-F., Cui, J., Yang, L., & Lu, Y. (2004). Alteration of the ATM gene occurs in gastric cancer cell lines and primary tumors associated with cellular response to DNA damage. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *557*(1), 41–51.

Appendix A

Sequencing Primers

Target region	Primers		Product size (bp)	Annealing temp (°C)
Promoter	Forward	ACGACGCTCTTCCGATCTTCGAACCCAGTGGAATCAGAAC	300	64
	Reverse	CGTGTGCTCTTCCGATCTACAGGTGCTTTGCAGTTCCG		
Exon 1	Forward	ACGACGCTCTTCCGATCTGAACTGCAAAGCACCTGTGA	286	65
	Reverse	CGTGTGCTCTTCCGATCTGTGACGACGGGAGAGGAAG		
Exon 2	Forward	ACGACGCTCTTCCGATCTTTTCGGTGAGCAGGAGGGAA	291	65
	Reverse	CGTGTGCTCTTCCGATCTGGTGTGGGAGTGCAATTTCT		
Exon 3	Forward	ACGACGCTCTTCCGATCTCGCTCTTTGGAGAAGGAATG	410	65
	Reverse	CGTGTGCTCTTCCGATCTCGGTACCAAGGCTGAGAAAC		
Exon 4	Forward	ACGACGCTCTTCCGATCTTGATTGGTCATTTTGGTGGA	460	65
	Reverse	CGTGTGCTCTTCCGATCTGAATTAGTAAAGAAGGATCCCAAC		
Exon 5	Forward	ACGACGCTCTTCCGATCTAGTGTTGGGATCCTTCTT	393	65
	Reverse	CGTGTGCTCTTCCGATCTCCCATCACTTCTCCTTAGCA		
Exon 6	Forward	ACGACGCTCTTCCGATCTCAGCAGCACATGTGTGAGAAAAGTC	440	64
	Reverse	CGTGTGCTCTTCCGATCTGGAAGGATCAGCTTTAGTTACAC		
Exon 7	Forward	ACGACGCTCTTCCGATCTCCAGTCCCAAAGTGCAGCTTGTCT	292	66
	Reverse	CGTGTGCTCTTCCGATCTCACCCTCTGGATCCTCCTGA		
Exon 8	Forward	ACGACGCTCTTCCGATCTGTTCCGTGCCTAGAAGACA	452	66
	Reverse	CGTGTGCTCTTCCGATCTGCCATCTCAAGATGCTTGCT		
Exon 9	Forward	ACGACGCTCTTCCGATCTTGACACATCTCTTTGCTCTGC	318	66
	Reverse	CGTGTGCTCTTCCGATCTAGAAGATACCAGGGGACAAGG		
Exon 10	Forward	ACGACGCTCTTCCGATCTAACCACAGTTACTTTTGCACC	445	66
	Reverse	CGTGTGCTCTTCCGATCTAACCAGTTGCTGCAAGTCAG		
Exon 11	Forward	ACGACGCTCTTCCGATCTTTCTAAAAGCCAGAGCTTGTCC	277	65
	Reverse	CGTGTGCTCTTCCGATCTGAGGGGCAAGGAACTGAACT		

Table continued on next page...

Target region	Primers		Product size (bp)	Annealing temp (°C)
Exon 12	Forward	ACGACGCTCTTCCGATCTACCACTGAAGAGCCAGGAC	444	65
	Reverse	CGTGTGCTCTTCCGATCTGAAATTGAAAGGTGGGGATCT		
Exon 13	Forward	ACGACGCTCTTCCGATCTCGGGTGTCTTTAGTTCACTAGC	484	65
	Reverse	CGTGTGCTCTTCCGATCTTGGGAGTCTCTTTCCCACAT		
Exon 14	Forward	ACGACGCTCTTCCGATCTGTGATAGCTGCTGCTTCTGG	324	65
	Reverse	CGTGTGCTCTTCCGATCTTGTTTCAAATGCCTACCTC		
Exon 15	Forward	ACGACGCTCTTCCGATCTACATAGCCCTGTGTGTATGAC	289	65
	Reverse	CGTGTGCTCTTCCGATCTAGAGATGAGCCATGCTTTGG		
Exon 16	Forward	ACGACGCTCTTCCGATCTGATGACAGGTGTGCCCTTC	363	65
	Reverse	CGTGTGCTCTTCCGATCTCAGCAACGTGATTTCTGCAT		

Table A.1: *CDH1* amplicon specific primers for next generation sequencing. Amplicon specific primers designed with non-specific 18 bp sequence at 5′ end of primers. Non-specific sequence for forward and reverse primers are ACGACGCTCTTCCGATCT and CGTGTGCTCTTCCGATCT, respectively. All primers are shown in 5′ to 3′ orientation. Abbreviations: bp, base pairs; temp, temperature.

Primer	Sequence (5′>3′)	Index
FPI1	AATGATACGGCGACCACCGAGATCTACAC**CGTGAT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGTGAT
FPI2	AATGATACGGCGACCACCGAGATCTACAC**ACATCG**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACATCG
FPI3	AATGATACGGCGACCACCGAGATCTACAC**GCCTAA**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCCTAA
FPI4	AATGATACGGCGACCACCGAGATCTACAC**TGGTCA**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGGTCA
FPI5	AATGATACGGCGACCACCGAGATCTACAC**CACTGT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CACTGT
FPI6	AATGATACGGCGACCACCGAGATCTACAC**ATTGGC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATTGGC
FPI7	AATGATACGGCGACCACCGAGATCTACAC**GATCTG**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GATCTG
FPI8	AATGATACGGCGACCACCGAGATCTACAC**TCAAGT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCAAGT
FPI9	AATGATACGGCGACCACCGAGATCTACAC**CTGATC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CTGATC
FPI10	AATGATACGGCGACCACCGAGATCTACAC**AAGCTA**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAGCTA
FPI11	AATGATACGGCGACCACCGAGATCTACAC**GTAGCC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GTAGCC
FPI12	AATGATACGGCGACCACCGAGATCTACAC**TACAAG**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TACAAG
FPI13	AATGATACGGCGACCACCGAGATCTACAC**TTGACT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TTGACT
FPI14	AATGATACGGCGACCACCGAGATCTACAC**GGAACT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGAACT
FPI15	AATGATACGGCGACCACCGAGATCTACAC**TGACAT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGACAT
FPI16	AATGATACGGCGACCACCGAGATCTACAC**GGACGG**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGACGG
FPI17	AATGATACGGCGACCACCGAGATCTACAC**CTCTAC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CTCTAC
FPI18	AATGATACGGCGACCACCGAGATCTACAC**GCGGAC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCGGAC
FPI19	AATGATACGGCGACCACCGAGATCTACAC**TTTCAC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TTTCAC
FPI20	AATGATACGGCGACCACCGAGATCTACAC**GGCCAC**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGCCAC

Table A.2: Forward adapter primers for creating next generation sequencing libraries. Index sequence shown in bold.

Primer	Sequence (5′>3′)	Index
RPI1	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CGTGAT
RPI2	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	ACATCG
RPI3	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GCCTAA
RPI4	CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TGGTCA
RPI5	CAAGCAGAAGACGGCATACGAGAT**CACTGT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CACTGT
RPI6	CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	ATTGGC
RPI7	CAAGCAGAAGACGGCATACGAGAT**GATCTG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GATCTG
RPI8	CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TCAAGT
RPI9	CAAGCAGAAGACGGCATACGAGAT**CTGATC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CTGATC
RPI10	CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AAGCTA
RPI11	CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GTAGCC
RPI12	CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TACAAG
RPI13	CAAGCAGAAGACGGCATACGAGAT**TTGACT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TTGACT
RPI14	CAAGCAGAAGACGGCATACGAGAT**GGAACT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GGAACT
RPI15	CAAGCAGAAGACGGCATACGAGAT**TGACAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TGACAT
RPI16	CAAGCAGAAGACGGCATACGAGAT**GGACGG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GGACGG
RPI17	CAAGCAGAAGACGGCATACGAGAT**CTCTAC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CTCTAC
RPI18	CAAGCAGAAGACGGCATACGAGAT**GCGGAC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GCGGAC
RPI19	CAAGCAGAAGACGGCATACGAGAT**TTTCAC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TTTCAC
RPI20	CAAGCAGAAGACGGCATACGAGAT**GGCCAC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GGCCAC
RPI21	CAAGCAGAAGACGGCATACGAGAT**CGAAAC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CGAAAC
RPI22	CAAGCAGAAGACGGCATACGAGAT**CGTACG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CGTACG
RPI23	CAAGCAGAAGACGGCATACGAGAT**CCACTC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CCACTC
RPI24	CAAGCAGAAGACGGCATACGAGAT**GCTACC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GCTACC
RPI25	CAAGCAGAAGACGGCATACGAGAT**ATCAGT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	ATCAGT
RPI26	CAAGCAGAAGACGGCATACGAGAT**GCTCAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	GCTCAT
RPI27	CAAGCAGAAGACGGCATACGAGAT**AGGAAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AGGAAT
RPI28	CAAGCAGAAGACGGCATACGAGAT**CTTTTG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CTTTTG
RPI29	CAAGCAGAAGACGGCATACGAGAT**TAGTTG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TAGTTG
RPI30	CAAGCAGAAGACGGCATACGAGAT**CCGGTG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	CCGGTG

Table A.3: Reverse adapter primers for creating next generation sequencing libraries. Index sequence shown in bold.

Target	Forward primer (5′ > 3′)	Reverse primer (5′ > 3′)	Product size (bp)	Annealing temp (°C)
Promoter	TCGAACCCAGTGGAATCAGAAC	ACAGGTGCTTTGCAGTTCCG	264	60
Exon 1	GAACTGCAAAGCACCTGTGA	GTGACGACGGGAGAGGAAG	250	60
Exon 2	TTTCGGTGAGCAGGAGGGAA	GGTGTGGGAGTGCAATTTCT	255	58
Exon 3	AGCCCAAAATGTCAACGGTA	CGCTCTTTGGAGAAGGAATG	389	58
Exon 4	GAATTAGTAAAGGATCCCAAC	TGATTGGTCATTTGGTGGA	424	55
Exon 5	TGGGTTATTCCTCCCATCAG	TCCTGGTTGTATGAGGATGC	816	56
Exon 6	CAGCAGCACATGTGTGAGAAAAGTC	GGAAGGATCAGCTTTAGTTACAC	404	60
Exon 7	CCAGTCCCAAAGTGCAGCTTGTCT	CACCCTCTGGATCCTCCTGA	256	62
Exon 8	GTTCCGTGCCTAGAAGACA	GCCATCTCAAGATGCTTGCT	416	58
Exon 9	TGACACATCTCTTTGCTCTGC	AGAAGATACCAGGGGACAAGG	282	60
Exon 10	TGAAAGTCATGGCAGAAACC	GCAGCACATCAGACCAAGAA	590	58
Exon 11	TTCTAAAAGCCAGAGCTTGTCC	GAGGGGCAAGGAACTGAACT	241	60
Exon 12	GACCACTGAAGAGCCAGGAC	ATGGCAGTTGGAGCAAAGTT	374	60
Exon 13	CGGGTGTCTTTAGTTCACTAGC	TGGGAGTCTCTTTCCCACAT	448	58
Exon 14	CCTCCATCACACCAAACATGC	GAGAATGGCAATGAGGAAGGC	586	60
Exon 15	TCCAGTGCTATTTGGGAGAC	TGACACAACTCCTCCTGAGC	386	58
Exon 16	GATGACAGGTGTGCCCTTC	CAGCAACGTGATTTCTGCAT	327	58

Table A.4: Primers for Sanger sequencing of germline *CDH1*. Abbreviations: bp, base pairs; temp, temperature.

Target	Forward primer (5′>3′)	Reverse primer (5′>3′)	Product size (bp)	Annealing temp
APC	AGAGGCAGAATCAGCTCCATCC	CACTCAGGCTGGATGAACAAG	846	60
ATM	GAGTTCCCTGTATCTTCATTTC	CAGAGAGCCAGTCCTCCTTC	841	58
FARP2	TGATTAAGACTCACAGAAGTTGCAG	CTCTTAGCCAGCCTCTCACG	597	60
IGF1R	GCATAGGCCAGGTCAGTTCT	CTGGCTCTCCAGTGGTTACA	700	62
LMO7	CCCCATCATCATACATCAAGGAAC	TGAGAAGCATTCGTGAGCTG	814	62
MUTYH	CACAAATGCTCCCAAACAGGAGG	GTCCCAATGTCACACAGCAATAC	847	62
PALB2	CCCATCTTGATGAAGAAACTGGAG	GCCTTGTGCCTCCAAACTTACAG	623	62
PMS1	TGGTACACTGCAACATATACAGC	GCTGGGGAAAGAGGTTTAACA	489	60
TP53	CTCTGGGAGGAGGGGTTAAG	TCTACTCCCAACCACCCTTG	921	60

Table A.5: Primers for validation of variants identified by whole-exome sequencing. Abbreviations: bp, base pairs; temp, temperature.

Appendix B

Validation of rare variants in Māori case-control cohort

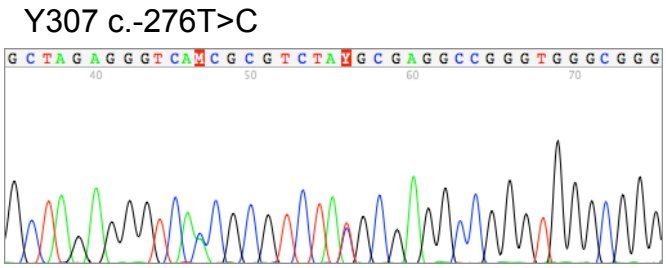


Figure B.1: Validation of upstream c.-276T>C variant. Sanger sequence in forward direction for control Y307.

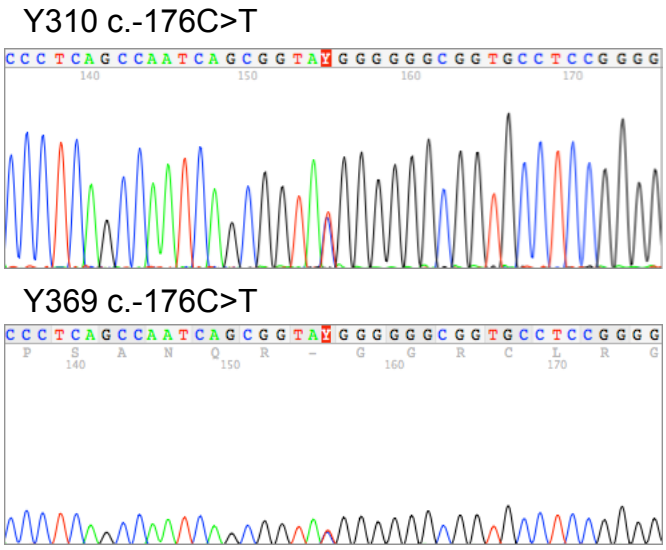


Figure B.2: Validation of upstream c.-176C>T variant. Sanger sequence in forward direction for controls Y310 and Y369.

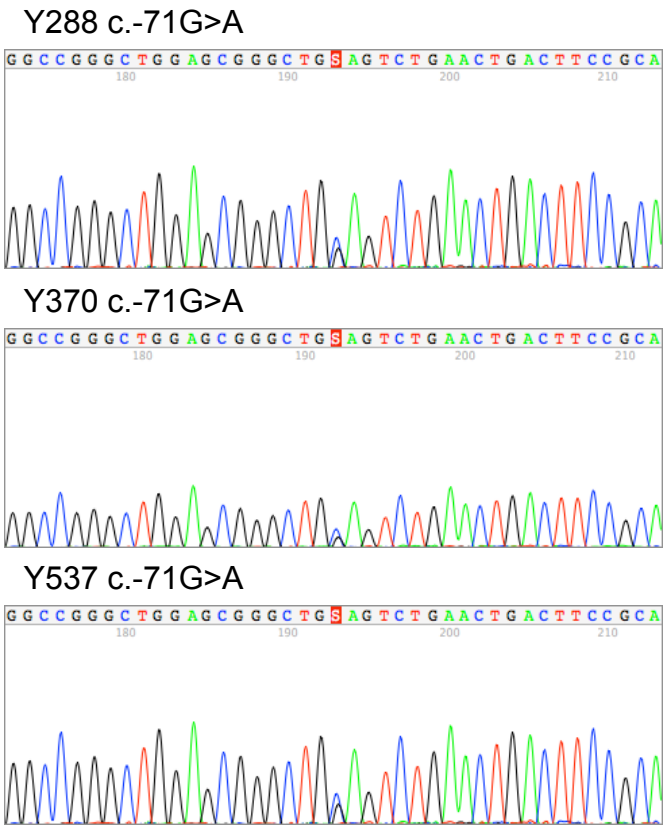


Figure B.3: Validation of *CDH1* 5′ UTR c.-71G>A variant. Sanger sequence in forward direction for controls Y288, Y370, and Y537.

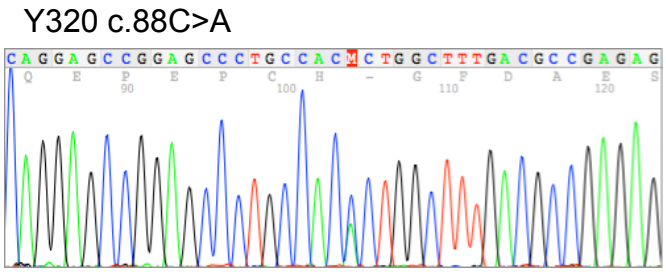


Figure B.4: Validation of *CDH1* c.88C>A variant (p.Pro30Thr). Sanger sequence in forward direction for control Y320.

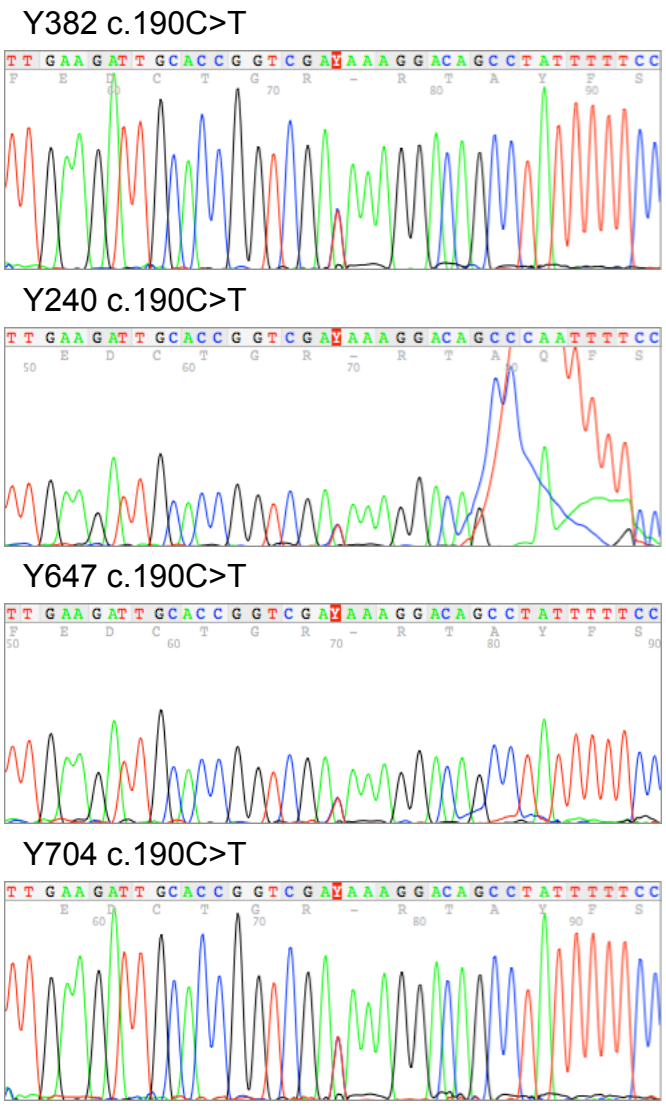


Figure B.5: Validation of *CDH1* c.190C>T variant (p.Gln64*). Sanger sequence in forward direction for cases Y240, Y382, Y647, and Y704.

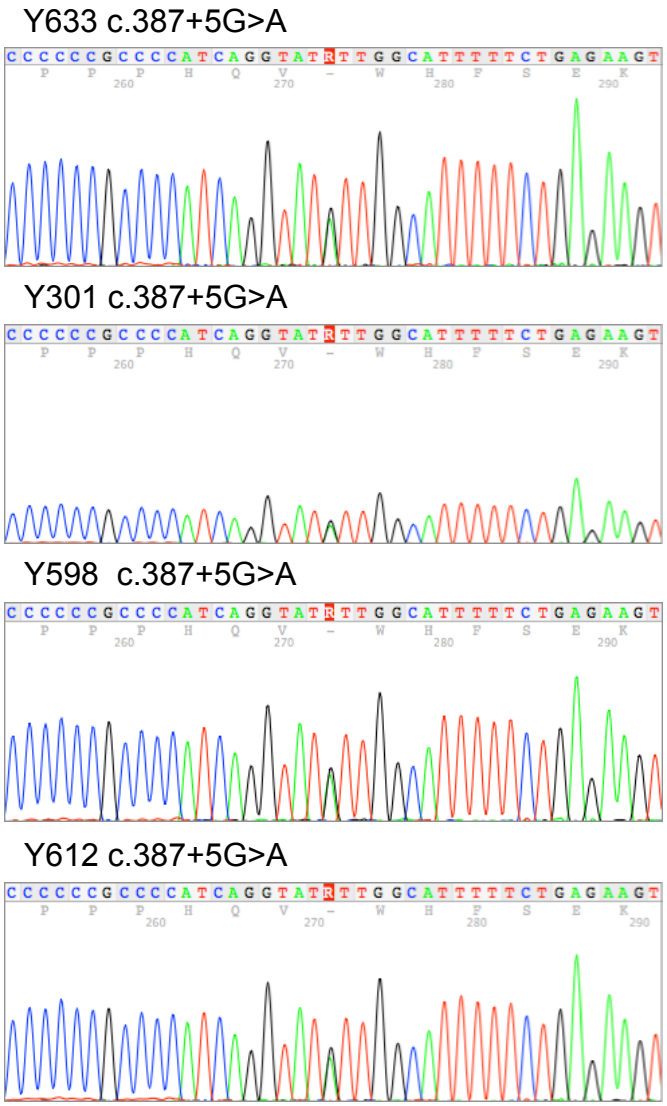


Figure B.6: Validation of *CDH1* c.387+5A>G variant. Sanger sequence in forward direction for case Y633 and controls Y301, Y598, and, Y612.

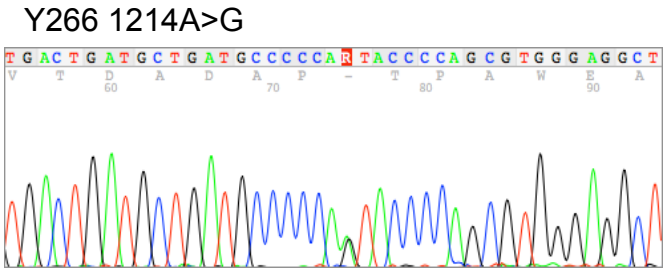


Figure B.7: Validation of *CDH1* c.1214A>G variant (p.Asn405Ser). Sanger sequence in forward direction for control Y266.

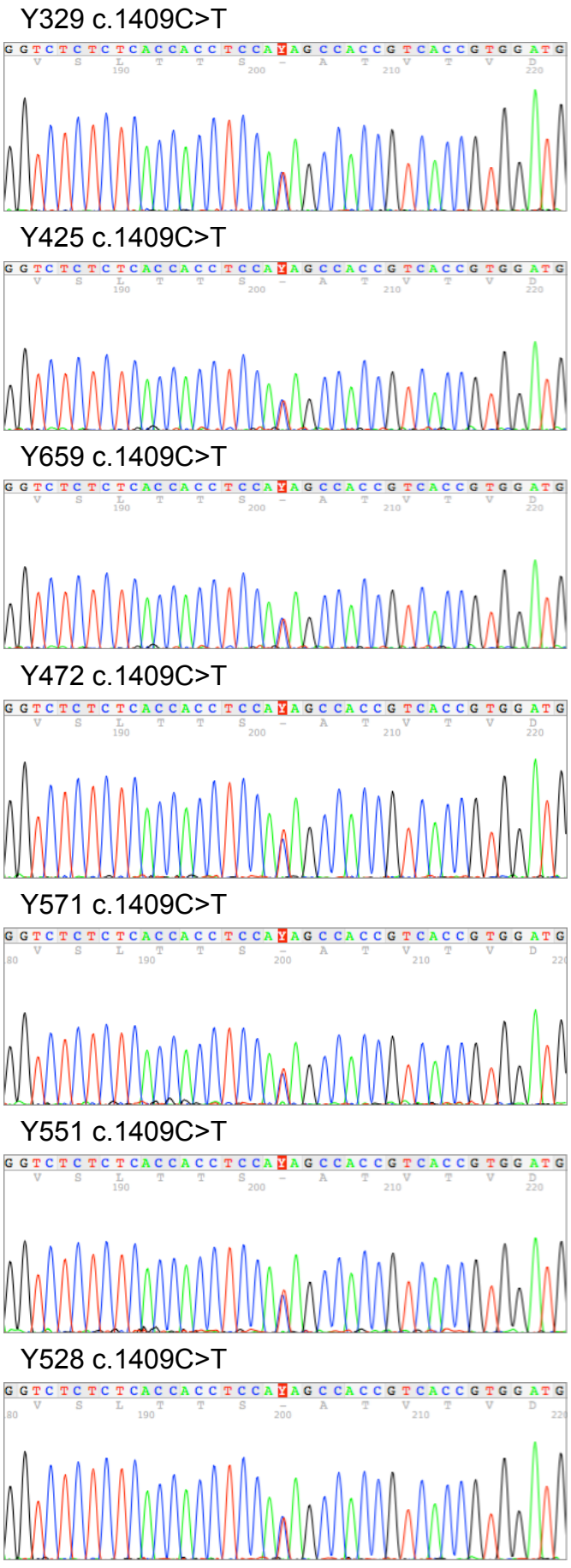


Figure B.8: Validation of *CDH1* c.1409C>T variant (p.Thr470Ile). Sanger sequence in forward direction for cases Y329, Y425 ,and Y659, and controls Y472, Y571, Y551, and Y528.

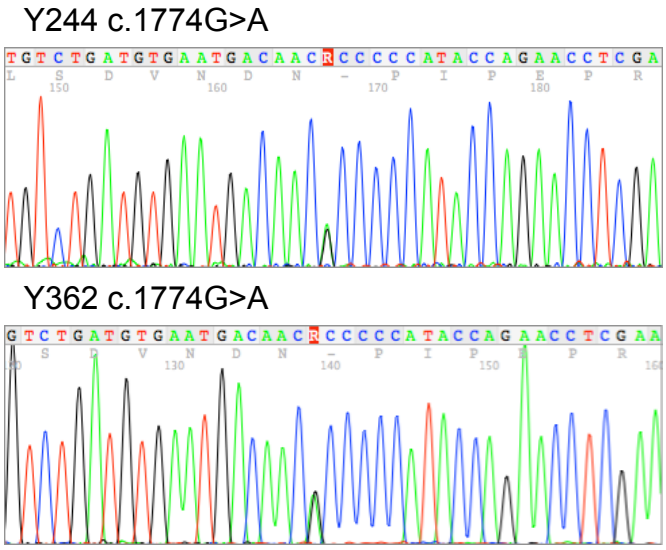


Figure B.9: Validation of *CDH1* c.1774G>A variant (p.Ala592Thr). Sanger sequence in forward direction for cases Y244 and control Y362.

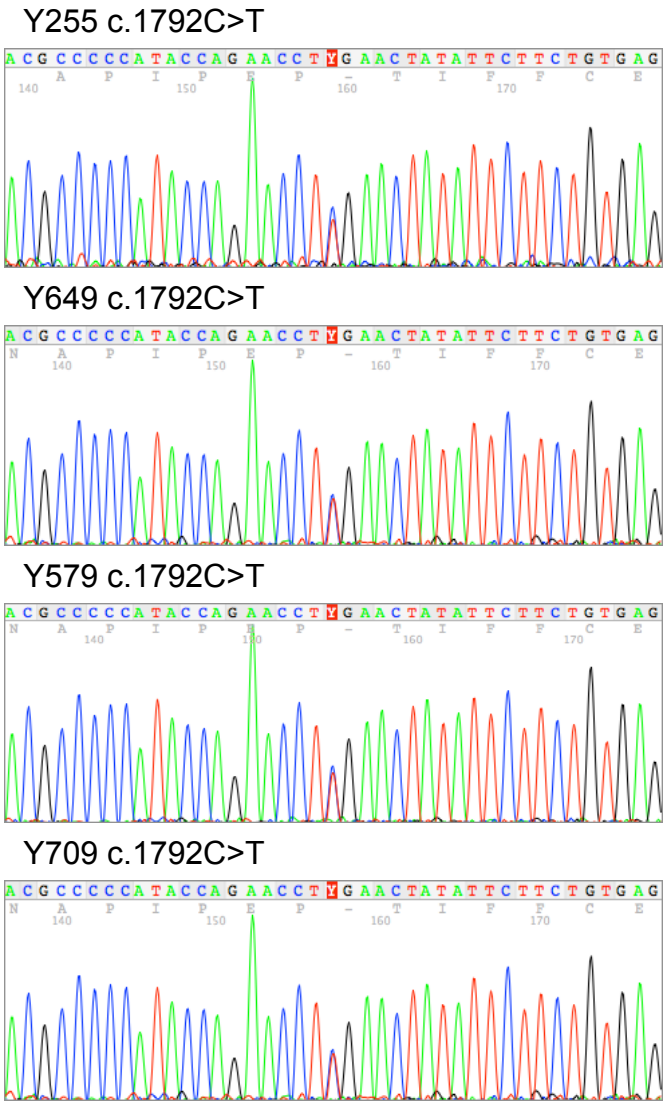


Figure B.10: Validation of *CDH1 CDH1* c.1792C>T variant (p.Arg598*). Sanger sequence in forward direction for cases Y255, Y579, Y649, and Y709

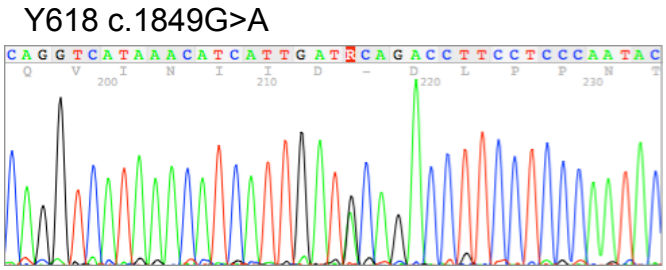


Figure B.11: Validation of *CDH1* c.1849G>A variant (p.Ala617Thr). Sanger sequence in forward direction for case Y618

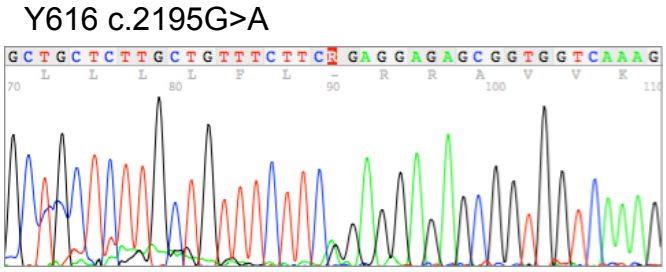


Figure B.12: Validation of *CDH1* c.2195G>A variant (p.Arg732Gln). Sanger sequence in forward direction for cases Y616.

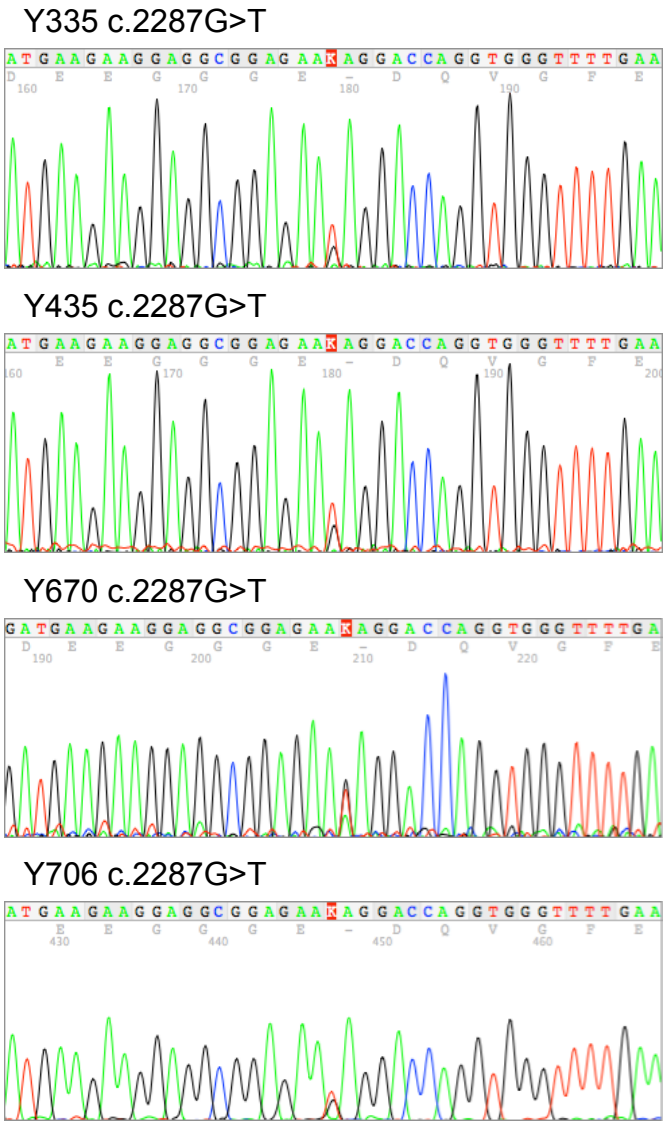


Figure B.13: Validation of *CDH1* c.2287G>T variant (p.Glu763*). Sanger sequence in forward direction for cases Y335, Y435, Y670, and Y706.

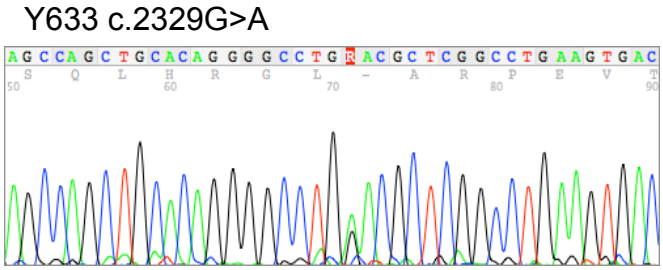


Figure B.14: Validation of *CDH1* c.2329G>A variant (p.Asp777Asn). Sanger sequence in forward direction for case Y633.

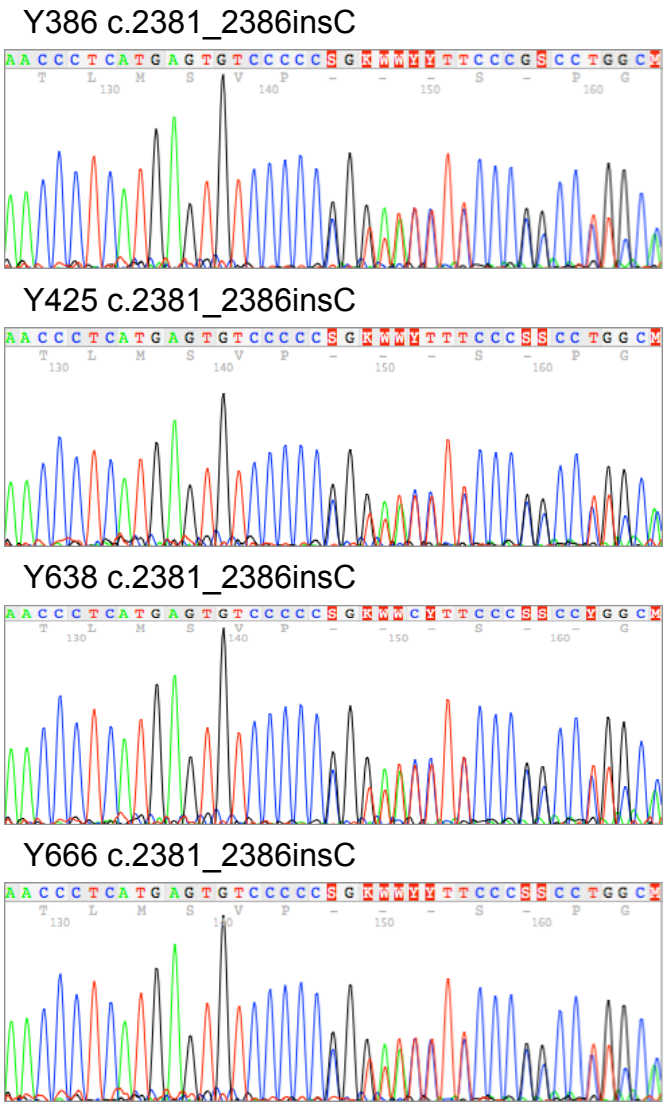


Figure B.15: Validation of *CDH1* c.2381_ 2386insC variant (p.Arg796fs). Sanger sequence in forward direction for cases Y386, Y425, Y638, and Y666.

Appendix C

MLPA ratios

Case	Exon 1	Exon2(1)	Exon2(2)	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	Exon 14	Exon 15	Exon 16
Y214	0.96	0.96	0.93	1.10	0.96	1.10	1.05	0.98	0.94	0.96	0.91	0.96	1.06	0.89	0.93	1.13	1.01
	(0.07)	(0.08)	(0.08)	(0.12)	(0.08)	(0.12)	(0.10)	(0.07)	(0.09)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.08)	(0.04)	(0.08)
Y215	1.01	0.95	1.00	1.03	1.08	1.05	0.98	1.01	1.05	1.03	1.00	1.05	1.04	1.07	1.00	1.01	0.97
	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.09)	(0.07)	(0.05)	(0.08)	(0.05)	(0.05)	(0.07)	(0.06)	(0.09)	(0.07)	(0.08)	(0.05)
Y219	0.96	0.89	0.89	1.23	0.95	1.19	1.07	0.96	0.87	0.96	0.89	0.91	1.04	0.88	0.93	1.09	1.02
	(0.13)	(0.12)	(0.12)	(0.15)	(0.11)	(0.14)	(0.11)	(0.10)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)	(0.12)	(0.11)	(0.12)	(0.10)
Y222	0.76	0.93	0.84	1.05	1.03	1.13	0.98	0.98	0.97	1.01	0.99	1.01	0.90	0.98	1.06	1.12	1.09
	(0.09)	(0.10)	(0.10)	(0.11)	(0.09)	(0.13)	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)	(0.07)	(0.10)	(0.09)	(0.11)	(0.09)
Y224	0.84	0.96	0.96	1.03	1.04	1.01	1.03	1.02	1.08	0.98	1.02	0.97	0.98	1.00	0.96	1.04	1.01
	(0.08)	(0.09)	(0.09)	(0.09)	(0.08)	(0.09)	(0.06)	(0.05)	(0.08)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.05)
Y225	1.02	0.96	1.06	1.09	1.07	1.03	0.94	1.00	1.05	1.03	1.03	1.02	1.00	1.05	1.04	1.00	1.00
	(0.10)	(0.09)	(0.09)	(0.10)	(0.08)	(0.09)	(0.06)	(0.06)	(0.08)	(0.06)	(0.06)	(0.07)	(0.06)	(0.09)	(0.08)	(0.07)	(0.06)
Y226	1.05	0.97	1.06	1.08	1.08	1.02	1.05	1.04	1.06	1.03	1.03	1.02	1.01	1.09	1.05	1.02	0.98
	(0.10)	(0.10)	(0.10)	(0.09)	(0.07)	(0.08)	(0.07)	(0.06)	(0.08)	(0.06)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y232	0.96	0.96	1.02	1.04	1.07	0.98	1.02	1.00	1.02	0.99	0.99	0.98	1.00	1.03	0.99	0.99	1.00
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y234	1.05	1.08	1.03	1.02	1.08	0.93	0.94	1.04	1.04	1.03	1.01	1.10	1.01	1.06	1.04	0.96	0.99
	(0.09)	(0.09)	(0.09)	(0.09)	(0.07)	(0.08)	(0.06)	(0.06)	(0.08)	(0.05)	(0.05)	(0.06)	(0.06)	(0.09)	(0.07)	(0.07)	(0.05)
Y235	0.71	0.88	0.86	1.08	0.90	1.16	1.11	0.91	1.01	0.98	1.05	0.92	1.04	0.84	0.81	1.08	0.97
	(0.08)	(0.10)	(0.08)	(0.12)	(0.10)	(0.13)	(0.10)	(0.07)	(0.10)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.08)	(0.10)	(0.08)
Y236	0.81	1.12	0.97	0.90	1.01	0.96	0.94	0.97	0.92	0.93	0.96	0.99	0.97	1.02	0.99	0.95	0.98
	(0.08)	(0.10)	(0.10)	(0.07)	(0.07)	(0.08)	(0.07)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.06)	(0.05)
Y244	1.24	0.97	1.12	1.00	0.98	0.87	1.03	1.09	1.10	0.96	0.93	1.06	0.88	1.21	1.03	0.71	1.03
	(0.15)	(0.13)	(0.13)	(0.11)	(0.12)	(0.10)	(0.11)	(0.11)	(0.12)	(0.10)	(0.09)	(0.11)	(0.09)	(0.15)	(0.12)	(0.07)	(0.11)
Y245	0.96	1.09	0.98	0.98	1.05	0.97	0.99	0.97	0.94	0.95	0.94	0.99	0.99	1.04	1.01	0.99	1.03
	(0.09)	(0.10)	(0.10)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y246	1.05	0.97	1.05	1.06	1.02	0.97	0.95	1.03	1.03	1.04	1.04	1.08	0.98	1.07	1.05	0.98	1.02
	(0.09)	(0.09)	(0.09)	(0.09)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.05)	(0.07)	(0.05)	(0.10)	(0.07)	(0.07)	(0.05)
Y254	1.04	1.00	1.05	1.06	1.03	0.96	0.98	1.00	1.00	1.02	0.98	1.07	1.00	1.09	1.04	0.96	1.01
	(0.10)	(0.09)	(0.09)	(0.09)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.04)	(0.07)	(0.05)	(0.10)	(0.07)	(0.06)	(0.05)
Y256	0.95	1.10	0.99	0.99	1.02	0.93	1.00	0.95	0.95	0.92	0.92	0.97	0.98	1.02	0.98	0.96	1.03
	(0.08)	(0.10)	(0.10)	(0.08)	(0.06)	(0.07)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.09)	(0.06)	(0.06)	(0.05)
Y260	0.92	1.15	1.02	0.91	1.08	1.03	0.98	0.96	0.95	0.99	0.95	1.01	1.09	1.02	0.98	0.95	1.08
	(0.10)	(0.13)	(0.13)	(0.09)	(0.09)	(0.09)	(0.08)	(0.06)	(0.08)	(0.07)	(0.06)	(0.08)	(0.08)	(0.10)	(0.08)	(0.07)	(0.07)
Y263	0.71	0.93	0.82	1.03	0.86	1.12	1.08	0.90	0.97	0.94	0.99	0.86	1.03	0.81	0.80	1.06	0.94
	(0.07)	(0.11)	(0.11)	(0.12)	(0.08)	(0.12)	(0.10)	(0.07)	(0.09)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.08)	(0.10)	(0.08)
Y268	1.03	1.05	1.07	0.98	1.01	0.96	0.95	1.06	1.02	1.04	1.04	1.11	0.97	1.10	1.10	1.00	1.02
	(0.09)	(0.10)	(0.10)	(0.08)	(0.08)	(0.07)	(0.05)	(0.05)	(0.08)	(0.05)	(0.05)	(0.07)	(0.05)	(0.10)	(0.08)	(0.07)	(0.05)
Y272	1.04	1.05	1.05	1.01	1.03	1.04	1.00	1.00	0.99	1.00	1.01	1.10	1.04	1.06	1.06	1.02	1.02
	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.09)	(0.06)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.06)	(0.09)	(0.07)	(0.08)	(0.05)

Table continued on next page…

Case	Exon 1	Exon 2(1)	Exon 2(2)	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	Exon 14	Exon 15	Exon 16
Y298	1.14	0.84	1.01	1.09	0.98	1.24	1.05	1.02	0.92	1.07	1.05	1.01	1.00	0.98	1.07	1.14	1.04
	(0.14)	(0.09)	(0.09)	(0.15)	(0.09)	(0.15)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.11)	(0.10)	(0.13)	(0.09)
Y323	1.01	1.04	1.07	1.03	1.08	0.97	0.97	1.00	1.01	1.02	1.01	1.05	1.01	1.01	1.02	0.96	1.03
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.07)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.05)	(0.04)	(0.08)	(0.06)	(0.07)	(0.05)
Y324	1.05	0.99	1.07	1.03	1.03	0.97	0.94	1.02	1.05	1.01	0.99	1.03	0.97	1.00	1.02	0.98	0.95
	(0.10)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.05)	(0.05)	(0.08)	(0.05)	(0.04)	(0.06)	(0.04)	(0.09)	(0.07)	(0.07)	(0.04)
Y329	0.80	1.08	0.96	0.87	0.99	0.94	0.92	1.03	0.96	0.99	0.99	1.03	0.94	1.03	1.02	0.97	1.02
	(0.07)	(0.10)	(0.10)	(0.07)	(0.07)	(0.07)	(0.05)	(0.06)	(0.07)	(0.04)	(0.04)	(0.06)	(0.04)	(0.08)	(0.07)	(0.06)	(0.05)
Y331	0.71	1.01	0.90	0.98	0.95	0.98	0.99	0.95	1.01	0.99	1.02	0.97	0.97	0.93	0.94	1.06	1.00
	(0.07)	(0.09)	(0.09)	(0.08)	(0.07)	(0.09)	(0.06)	(0.06)	(0.08)	(0.06)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.08)	(0.06)
Y334	1.07	1.04	1.05	0.99	1.02	0.94	0.96	1.00	1.00	1.02	1.00	1.07	1.02	1.07	1.05	0.96	1.00
	(0.10)	(0.09)	(0.09)	(0.08)	(0.06)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.04)	(0.10)	(0.07)	(0.07)	(0.05)
Y337	1.14	0.84	1.01	1.22	0.94	1.22	1.03	1.02	0.93	1.08	1.03	1.04	1.00	0.93	0.99	1.13	0.99
	(0.15)	(0.10)	(0.10)	(0.15)	(0.10)	(0.15)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.13)	(0.10)
Y339	1.03	1.08	1.02	1.00	1.08	0.97	0.95	1.01	0.99	0.99	1.00	1.03	1.01	1.05	1.02	0.97	1.02
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.08)	(0.06)	(0.06)	(0.05)
Y340	0.99	1.05	1.02	1.02	1.05	0.98	1.01	1.00	1.00	1.02	0.99	1.03	1.02	1.03	1.01	1.00	0.98
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.04)
Y343	0.92	1.14	0.99	0.90	1.01	0.95	0.98	0.98	0.97	0.98	0.95	1.00	0.99	1.02	1.03	1.00	0.99
	(0.09)	(0.11)	(0.11)	(0.08)	(0.07)	(0.08)	(0.07)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.08)	(0.06)	(0.07)	(0.05)
Y347	1.09	0.83	0.99	1.15	0.94	1.24	1.07	1.02	0.98	1.14	1.01	1.06	1.05	0.94	0.98	1.17	1.02
	(0.15)	(0.11)	(0.11)	(0.14)	(0.10)	(0.15)	(0.12)	(0.11)	(0.11)	(0.12)	(0.10)	(0.12)	(0.11)	(0.12)	(0.11)	(0.14)	(0.11)
Y350	1.16	0.84	1.02	1.22	0.96	1.24	1.02	1.05	0.94	1.14	1.00	1.07	0.97	0.92	1.00	1.12	1.01
	(0.15)	(0.10)	(0.10)	(0.15)	(0.10)	(0.15)	(0.11)	(0.11)	(0.10)	(0.10)	(0.09)	(0.11)	(0.09)	(0.10)	(0.10)	(0.13)	(0.10)
Y358	0.94	1.02	0.96	0.94	0.99	0.91	0.97	0.96	0.97	0.96	0.94	1.00	1.00	0.98	0.96	0.99	0.99
	(0.08)	(0.09)	(0.09)	(0.07)	(0.06)	(0.07)	(0.07)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.05)
Y365	1.06	1.03	1.02	1.06	1.05	1.00	0.97	1.02	1.00	1.01	1.00	1.06	1.07	1.08	1.05	0.98	1.05
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)	(0.04)	(0.06)	(0.05)	(0.09)	(0.06)	(0.06)	(0.05)
Y371	1.12	0.82	1.01	1.23	0.96	1.21	1.03	1.06	0.98	1.19	1.04	1.11	0.95	0.95	1.02	1.20	1.02
	(0.14)	(0.10)	(0.10)	(0.15)	(0.11)	(0.15)	(0.11)	(0.10)	(0.11)	(0.12)	(0.11)	(0.12)	(0.10)	(0.11)	(0.11)	(0.13)	(0.10)
Y393	0.99	0.89	0.96	1.23	0.96	1.21	1.01	1.04	0.97	1.13	0.98	1.07	0.97	0.96	1.03	1.08	0.97
	(0.13)	(0.10)	(0.10)	(0.13)	(0.09)	(0.15)	(0.10)	(0.09)	(0.10)	(0.09)	(0.08)	(0.10)	(0.09)	(0.10)	(0.10)	(0.12)	(0.08)
Y395	1.00	1.05	0.98	0.90	0.96	0.98	1.03	0.96	0.99	0.99	1.01	1.00	1.02	1.00	0.94	0.98	0.99
	(0.09)	(0.10)	(0.10)	(0.07)	(0.07)	(0.08)	(0.07)	(0.05)	(0.07)	(0.05)	(0.04)	(0.06)	(0.05)	(0.08)	(0.06)	(0.07)	(0.05)
Y396	1.03	0.94	0.97	1.23	0.98	1.19	1.05	1.01	0.91	1.03	0.97	0.98	1.05	0.92	0.97	1.06	1.02
	(0.13)	(0.12)	(0.12)	(0.13)	(0.10)	(0.13)	(0.10)	(0.09)	(0.10)	(0.09)	(0.08)	(0.10)	(0.09)	(0.11)	(0.10)	(0.10)	(0.09)
Y399	0.97	1.18	1.01	0.89	1.00	0.87	0.95	1.00	0.92	0.97	0.96	1.01	0.99	1.07	1.02	0.95	1.02
	(0.09)	(0.11)	(0.11)	(0.08)	(0.07)	(0.08)	(0.07)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y402	1.06	1.05	1.07	1.00	1.09	0.92	0.96	1.03	1.04	1.02	1.00	1.02	1.01	1.07	1.07	0.95	0.98
	(0.10)	(0.11)	(0.11)	(0.08)	(0.08)	(0.08)	(0.06)	(0.06)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.10)	(0.08)	(0.07)	(0.05)
Y416	1.12	0.85	0.98	1.21	0.98	1.24	1.04	1.01	0.91	1.02	1.02	0.92	0.99	0.90	0.98	1.10	0.94
	(0.14)	(0.11)	(0.11)	(0.15)	(0.11)	(0.14)	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.11)	(0.10)	(0.12)	(0.09)

Table continued on next page...

Case	Exon 1	Exon 2(1)	Exon 2(2)	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	Exon 14	Exon 15	Exon 16
Y436	1.00	1.11	1.04	1.01	1.06	1.05	1.03	0.98	1.04	1.01	1.00	0.97	1.04	0.95	1.00	0.98	1.00
	(0.09)	(0.10)	(0.10)	(0.09)	(0.07)	(0.09)	(0.08)	(0.06)	(0.08)	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	(0.06)
Y437	0.98	1.04	0.99	1.00	1.02	1.01	1.00	0.91	0.96	0.97	1.00	0.99	1.06	1.00	0.97	1.05	0.97
	(0.09)	(0.09)	(0.09)	(0.07)	(0.07)	(0.08)	(0.06)	(0.04)	(0.07)	(0.05)	(0.05)	(0.05)	(0.06)	(0.09)	(0.06)	(0.07)	(0.04)
Y445	0.95	0.98	1.00	1.01	0.97	0.91	0.94	0.99	0.96	1.00	1.01	1.04	0.98	1.00	1.00	0.96	0.98
	(0.09)	(0.09)	(0.09)	(0.08)	(0.06)	(0.07)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.10)	(0.06)	(0.06)	(0.04)
Y446	1.03	0.91	0.93	1.23	0.99	1.20	1.03	0.94	0.89	0.95	0.93	0.87	1.03	0.88	0.93	1.12	1.00
	(0.14)	(0.12)	(0.12)	(0.15)	(0.12)	(0.15)	(0.12)	(0.10)	(0.11)	(0.10)	(0.10)	(0.10)	(0.11)	(0.11)	(0.11)	(0.13)	(0.11)
Y455	0.89	1.04	0.99	0.93	1.00	0.96	0.95	1.02	0.99	1.00	0.98	1.02	0.97	0.99	1.00	0.96	1.02
	(0.08)	(0.10)	(0.10)	(0.07)	(0.06)	(0.07)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.09)	(0.06)	(0.06)	(0.04)
Y456	1.04	1.03	1.03	0.96	1.06	0.95	1.05	1.01	1.09	1.00	1.01	0.98	1.08	1.04	1.03	0.98	1.05
	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.07)	(0.06)	(0.05)	(0.08)	(0.05)	(0.05)	(0.06)	(0.06)	(0.09)	(0.07)	(0.07)	(0.06)
Y464	0.79	1.01	0.97	0.89	0.98	0.95	0.95	1.00	1.00	0.97	0.97	0.98	0.96	1.00	0.97	0.96	0.98
	(0.07)	(0.10)	(0.10)	(0.07)	(0.07)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)	(0.05)	(0.06)	(0.05)	(0.08)	(0.07)	(0.07)	(0.05)
Y507	0.97	1.02	0.99	0.99	1.07	0.99	0.95	1.01	1.04	0.99	1.02	0.98	1.04	1.04	1.03	1.01	0.99
	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)	(0.05)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y520	0.85	1.00	0.96	0.96	1.04	0.96	0.97	1.00	0.99	0.97	0.97	0.95	1.01	0.97	0.96	0.98	0.98
	(0.07)	(0.09)	(0.09)	(0.07)	(0.06)	(0.07)	(0.07)	(0.05)	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.07)	(0.06)	(0.07)	(0.05)
Y557	1.05	0.85	0.94	1.24	0.94	1.21	1.06	0.97	0.90	0.99	0.94	0.92	1.03	0.89	0.92	1.12	1.00
	(0.15)	(0.11)	(0.11)	(0.16)	(0.11)	(0.14)	(0.12)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.12)	(0.11)	(0.12)	(0.11)
Y566	0.76	0.93	0.89	0.92	0.94	1.00	0.97	0.96	1.02	0.99	1.03	0.94	0.95	0.91	0.91	1.03	0.91
	(0.07)	(0.11)	(0.11)	(0.07)	(0.07)	(0.09)	(0.07)	(0.06)	(0.09)	(0.06)	(0.06)	(0.07)	(0.06)	(0.08)	(0.07)	(0.07)	(0.05)
Y606	1.13	0.82	1.00	1.05	0.96	1.20	1.07	1.03	0.92	1.15	1.03	1.08	1.01	0.91	0.99	1.16	1.00
	(0.15)	(0.11)	(0.11)	(0.15)	(0.10)	(0.14)	(0.11)	(0.10)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.12)	(0.10)
Y617	0.76	0.77	0.76	0.96	0.84	1.21	1.11	0.85	1.01	0.94	0.96	0.82	0.98	0.76	0.76	1.10	0.95
	(0.07)	(0.11)	(0.11)	(0.12)	(0.10)	(0.15)	(0.13)	(0.09)	(0.13)	(0.10)	(0.10)	(0.09)	(0.10)	(0.10)	(0.08)	(0.13)	(0.10)
Y618	1.15	1.10	1.04	1.04	1.11	1.01	1.03	1.01	0.98	1.00	0.94	1.02	0.97	1.09	1.04	1.01	1.10
	(0.13)	(0.13)	(0.13)	(0.10)	(0.10)	(0.11)	(0.09)	(0.08)	(0.09)	(0.08)	(0.07)	(0.08)	(0.07)	(0.11)	(0.09)	(0.09)	(0.09)
Y619	0.93	0.93	0.97	0.95	0.92	1.01	0.94	0.97	1.03	1.00	1.01	1.00	0.98	0.95	0.94	0.99	0.97
	(0.08)	(0.10)	(0.10)	(0.07)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.04)	(0.06)	(0.05)	(0.08)	(0.06)	(0.07)	(0.05)
Y623	0.87	0.88	0.94	0.96	0.90	1.03	1.11	0.91	1.02	0.99	1.04	0.95	1.02	0.89	0.88	1.07	0.92
	(0.08)	(0.09)	(0.09)	(0.09)	(0.07)	(0.09)	(0.07)	(0.05)	(0.07)	(0.06)	(0.05)	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	(0.05)
Y626	1.02	1.04	1.02	1.00	1.03	1.01	0.96	0.97	1.03	0.98	1.02	1.00	1.01	1.03	1.03	1.02	1.03
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.04)	(0.06)	(0.05)	(0.08)	(0.07)	(0.07)	(0.05)
Y627	1.00	1.04	1.01	0.98	1.05	1.00	1.04	0.99	1.05	0.99	1.01	1.00	1.04	1.08	1.02	1.02	1.06
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.06)	(0.08)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y630	1.11	0.89	1.01	1.24	1.02	1.19	1.04	1.01	0.92	1.01	0.97	0.94	1.02	0.91	0.97	1.11	1.00
	(0.15)	(0.11)	(0.11)	(0.15)	(0.11)	(0.14)	(0.11)	(0.10)	(0.10)	(0.10)	(0.09)	(0.10)	(0.09)	(0.11)	(0.10)	(0.12)	(0.10)
Y633	1.07	1.03	1.07	0.97	0.97	0.98	0.98	1.04	1.07	1.02	1.02	1.00	0.94	1.07	1.08	0.95	0.96
	(0.10)	(0.10)	(0.10)	(0.08)	(0.07)	(0.10)	(0.07)	(0.06)	(0.09)	(0.05)	(0.06)	(0.06)	(0.05)	(0.09)	(0.08)	(0.08)	(0.05)
Y637	1.02	1.01	1.00	0.94	0.99	0.94	1.01	0.99	1.05	1.00	1.02	1.00	0.99	1.03	1.03	0.96	1.01
	(0.10)	(0.09)	(0.09)	(0.07)	(0.07)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)

Table continued on next page...

Case	Exon 1	Exon 2(1)	Exon 2(2)	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	Exon 14	Exon 15	Exon 16
Y639	0.90	0.94	0.96	0.96	0.96	1.06	1.04	0.95	1.03	0.99	1.05	0.99	1.00	0.98	0.95	1.03	0.93
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.09)	(0.07)	(0.06)	(0.08)	(0.05)	(0.05)	(0.06)	(0.06)	(0.08)	(0.06)	(0.07)	(0.05)
Y640	0.96	1.04	1.00	0.92	1.03	0.98	1.04	0.99	1.06	1.00	1.01	0.99	1.02	0.97	1.00	0.96	0.98
	(0.09)	(0.10)	(0.10)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)	(0.05)	(0.06)	(0.05)	(0.08)	(0.06)	(0.07)	(0.05)
Y643	1.06	1.05	1.02	0.96	0.98	0.92	0.98	0.99	1.05	1.00	0.99	1.03	0.98	1.07	1.07	0.96	1.01
	(0.10)	(0.09)	(0.09)	(0.07)	(0.07)	(0.07)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y646	1.03	1.03	1.00	0.97	1.03	0.97	1.01	1.00	1.05	0.99	0.98	0.97	0.98	1.00	1.00	1.00	1.02
	(0.10)	(0.09)	(0.09)	(0.07)	(0.06)	(0.08)	(0.07)	(0.06)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.08)	(0.06)	(0.08)	(0.05)
Y650	0.91	0.99	0.99	0.99	0.98	1.04	1.00	1.00	1.05	0.99	1.02	0.97	1.00	1.00	1.02	1.03	0.99
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.09)	(0.07)	(0.06)	(0.08)	(0.05)	(0.05)	(0.07)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y654	1.01	1.03	1.02	0.99	1.04	1.01	1.02	1.01	1.05	1.00	1.00	1.00	1.05	1.04	1.01	1.05	0.98
	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)	(0.04)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.05)
Y659	1.05	0.84	0.95	1.21	0.95	1.23	1.04	1.02	0.94	1.10	0.97	1.03	0.99	0.91	0.99	1.17	1.02
	(0.14)	(0.09)	(0.09)	(0.15)	(0.09)	(0.15)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.11)	(0.09)	(0.11)	(0.10)	(0.12)	(0.09)
Y660	1.04	1.04	1.01	0.97	0.99	0.94	1.01	1.00	1.05	1.01	1.02	1.03	1.01	1.08	1.03	0.98	1.03
	(0.09)	(0.09)	(0.09)	(0.07)	(0.06)	(0.09)	(0.06)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.07)	(0.05)
Y669	1.14	1.02	1.03	0.95	0.99	0.96	0.97	1.10	1.02	1.00	1.01	1.03	0.95	1.08	1.10	0.97	1.04
	(0.11)	(0.10)	(0.10)	(0.08)	(0.08)	(0.09)	(0.07)	(0.07)	(0.08)	(0.06)	(0.06)	(0.07)	(0.06)	(0.11)	(0.08)	(0.07)	(0.06)
Y702	0.77	0.77	0.86	1.00	0.89	1.04	0.99	0.98	1.05	1.01	1.00	0.96	0.95	0.90	0.91	1.02	0.99
	(0.09)	(0.09)	(0.09)	(0.10)	(0.08)	(0.10)	(0.09)	(0.09)	(0.11)	(0.08)	(0.08)	(0.09)	(0.08)	(0.09)	(0.08)	(0.10)	(0.08)
Y703	0.98	1.07	1.01	0.99	1.05	0.97	1.02	1.04	1.06	1.02	0.99	0.98	1.00	1.02	1.01	0.93	1.01
	(0.10)	(0.11)	(0.11)	(0.08)	(0.07)	(0.08)	(0.06)	(0.06)	(0.08)	(0.05)	(0.05)	(0.07)	(0.05)	(0.08)	(0.07)	(0.05)	(0.05)
Y705	1.12	0.79	0.97	1.23	0.99	1.16	1.00	1.05	0.94	1.01	1.01	0.93	0.95	0.92	0.97	1.10	1.03
	(0.14)	(0.09)	(0.09)	(0.15)	(0.09)	(0.14)	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.09)	(0.08)	(0.10)	(0.09)	(0.11)	(0.09)
Y707	0.83	1.01	1.01	1.02	1.05	1.07	1.04	0.98	1.05	0.98	1.00	0.93	1.03	0.92	0.96	0.98	1.03
	(0.08)	(0.11)	(0.11)	(0.09)	(0.09)	(0.10)	(0.07)	(0.06)	(0.09)	(0.06)	(0.06)	(0.07)	(0.07)	(0.10)	(0.08)	(0.08)	(0.07)
Y708	0.77	0.95	0.94	1.04	0.92	1.05	1.00	0.95	0.99	1.00	1.02	1.02	0.92	0.97	0.95	1.01	0.95
	(0.07)	(0.09)	(0.09)	(0.08)	(0.06)	(0.09)	(0.07)	(0.06)	(0.08)	(0.05)	(0.06)	(0.07)	(0.05)	(0.09)	(0.06)	(0.07)	(0.05)
Y710	1.02	1.12	1.06	1.05	1.06	1.00	1.00	1.03	1.01	1.02	0.98	1.03	1.00	1.07	1.05	0.96	1.06
	(0.11)	(0.11)	(0.11)	(0.09)	(0.08)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)	(0.05)	(0.07)	(0.05)	(0.09)	(0.07)	(0.06)	(0.06)

Table C.1: MLPA results for Māori gastric cancer cases. Cases did not carry clear pathogenic variants. The dosage quotient (DQ) and standard deviation (shown in brackets) of *CDH1* MLPA probes. DQ values ranged between 0.7-1.3 are considered normal.

	Exon 1	Exon 2 (1)	Exon 2 (2)	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	Exon 14	Exon 15	Exon 16
Y247	1.05	0.99	1.07	1.03	1.03	0.97	0.94	1.02	1.05	1.01	0.99	1.03	0.97	1.03	1.02	0.98	0.95
	-0.07	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.05)	(0.05)	(0.08)	(0.05)	(0.04)	(0.06)	(0.04)	(0.09)	(0.07)	(0.07)	(0.04)
Y278	0.92	1.08	0.96	0.92	0.99	0.94	0.92	1.03	0.96	0.99	0.99	1.03	0.94	1.03	1.02	0.97	1.02
	(0.07)	-0.07	(0.07)	(0.07)	(0.07)	(0.07)	(0.05)	(0.06)	(0.07)	(0.04)	(0.04)	(0.06)	(0.04)	(0.08)	(0.07)	(0.06)	(0.05)
Y297	0.90	1.01	0.90	0.98	0.95	0.98	0.99	0.95	1.01	0.99	1.02	0.97	0.97	0.93	0.94	1.06	1.00
	(0.07)	(0.09)	(0.09)	(0.08)	(0.07)	(0.09)	(0.06)	(0.06)	(0.08)	(0.06)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)	(0.08)	(0.06)
Y357	1.07	1.04	1.15	0.80	0.96	0.94	0.96	1.00	1.00	1.02	1.100	1.07	1.02	1.07	1.05	0.96	1100
	(0.08)	(0.13)	(0.09)	(0.10)	(0.10)	(0.08)	(0.09)	(0.15)	(0.07)	(0.13)	(0.12)	(0.12)	(0.14)	(0.05)	(0.07)	(0.13)	(0.12)
Y364	1.10	0.94	1.1	1.08	0.94	1.05	1.03	1.02	0.93	1.08	1.03	1.04	1.00	0.93	0.99	1.10	0.99
	(0.09)	(0.09)	(0.09)	(0.08)	(0.06)	(0.08)	(0.06)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.04)
Y486	1.03	1.08	1.02	1.00	1.08	0.97	0.95	1.01	0.99	0.99	1.00	1.03	1.01	1.05	1.02	0.97	1.02
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.08)	(0.06)	(0.06)	(0.05)
Y537	0.99	1.05	1.02	1.02	1.05	0.98	1.01	1.00	1.00	1.02	0.99	1.03	1.02	1.03	1.01	1.00	0.98
	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.07)	(0.04)	(0.04)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.04)
Y548	0.92	1.07	0.99	0.9	1.01	0.95	0.98	0.98	1.00	0.98	0.95	1.00	0.99	1.02	1.03	1.00	0.99
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)	(0.07)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.08)	(0.06)	(0.07)	(0.05)
Y585	0.94	1.02	0.96	0.94	0.99	0.91	0.97	0.96	0.97	0.96	0.94	1.00	1.00	0.98	0.96	0.99	0.99
	(0.08)	(0.09)	(0.09)	(0.07)	(0.06)	(0.07)	(0.07)	(0.05)	(0.07)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.06)	(0.07)	(0.05)
Y638	1.16	1.13	1.11	1.08	1.06	1.05	0.97	1.02	1.00	1.12	1.09	1.11	1.10	1.11	0.85	0.98	1.13
	(0.09)	(0.09)	(0.10)	(0.08)	(0.07)	(0.08)	(0.17)	(0.05)	(0.13)	(0.12)	(0.14)	(0.06)	(0.15)	(0.09)	(0.16)	(0.06)	(0.12)

Table C.2: MLPA results for Māori from control cohort used as reference samples. The dosage quotient (DQ) and standard deviation (shown in brackets) of *CDH1* MLPA probes. A standard deviation of less than 0.10 and DQ values ranging between 0.85-1.15 for both target and reference probes were considered acceptable for reference samples.

Appendix D

Validation of rare variants in Chilean gastric cancer cohort

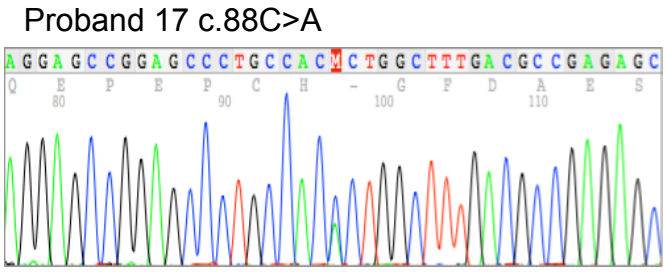


Figure D.1: Validation of *CDH1* c.88C>A variant (p.Pro30Thr). Sanger sequence in forward direction for Proband 17.

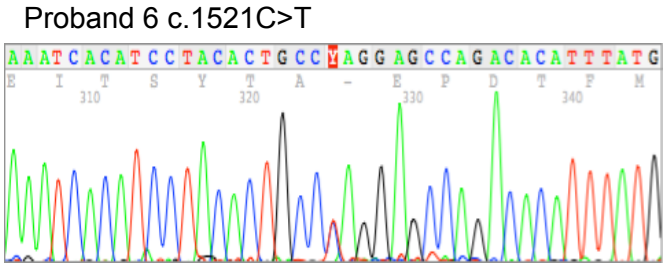


Figure D.2: Validation of *CDH1* c.1521C>T variant (p.Gln511*). Sanger sequence in forward direction for Proband 6.

Appendix E

Screening of the extended family of Proband 6

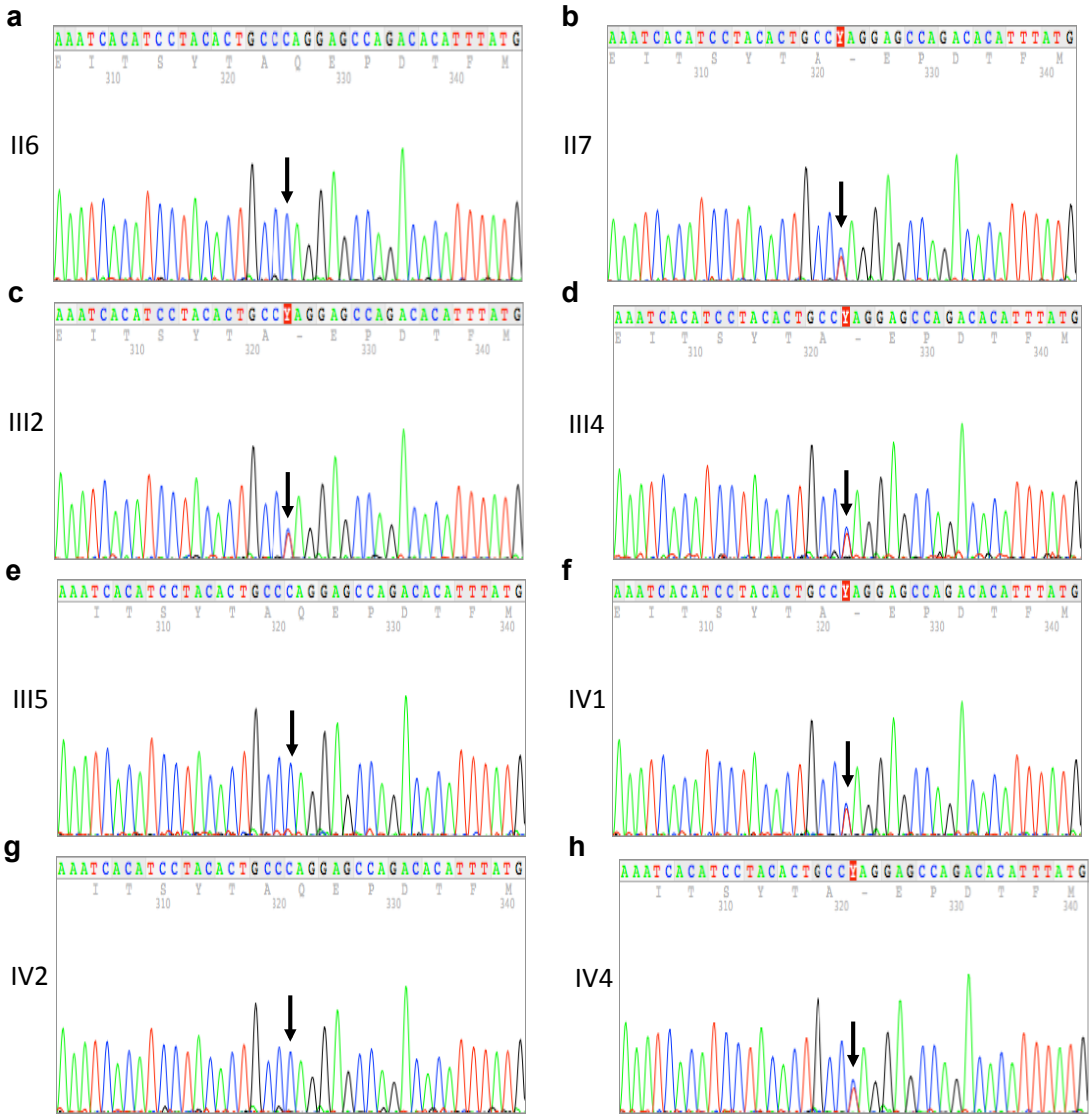


Figure E.1: Sanger sequence analysis of proband 6 family members. Sanger sequence of *CDH1* exon 10 showing carrier status of c.1521C>T. Chromatogram sequence for: a) II6, maternal grandfather ; b) II7 maternal grandmother; c) III2, maternal aunt; d) III3, mother; e) III4, father; f) IV1, maternal cousin; g) IV2, maternal cousin; and h) IV4, sibling, of proband 6 (IV5).