UNIVERSITY OF CALIFORNIA,
IRVINE

CDH1 Germline Mutations in the Prevalence of Gastric Cancer in Historically Underrepresented Racial and Ethnic Groups in Healthcare

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE
in Biomedical and Translational Science

by

Aaqil M. Khan

Thesis Committee:
Clinical Professor Maheswari Senthil, Chair
Professor Sherrie Kaplan
Assistant Professor Robert Wilson

2024

DEDICATION

This work is dedicated to my parents, whose sacrifices allow me to continually push the boundaries of what is achievable,

My mother,
who first showed me the world within a microscope. You are my biggest supporter, my inspiration, my champion, and my source of guidance. I could ask for nothing more.

My father,
who stressed the importance of a curious mind and reminded me that there is more to life than my textbooks, thank you for your never-ending wisdom.

My brother,
who stands by my side in every endeavor,

The Patients,
who have given us the ultimate gift–
a piece of themselves
to advance scientific knowledge and discovery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Upon the completion of this thesis, I complete yet another chapter of life. I would like to take a moment to dwell upon those who guided me and inspired me on this wonderful journey – I am truly thankful.

I would like to thank my committee chair, Dr. Maheswari Senthil, for being an excellent mentor and inspiring me to pursue a career in surgical oncology. Dr. Senthil, your visionary approach to patient care and relentless nature will always motivate me to give science and medicine nothing short of my very best. Words cannot express my appreciation for the knowledge you have bestowed upon me.

I would like to thank my committee members, Dr. Sherrie Kaplan, and Dr. Robert Wilson for being cornerstones of the MS-BATS program and providing me with a refined thought process in practicing translational science. Additionally, I would like to extend my thanks to the faculty and staff of the MS-BATS program– Dr. Greenfield, Dr. Kelly, Kaelyn and Zee, your help has been immeasurable in my success. A special thanks to Leora Fellus, of the graduate division for her help as well.

I would like to thank Jonathan LoTempio, Chris Hatch, Eduard Mas, and Girish Senthil for being so supportive in helping me process and analyze the AllofUs genomic data. I truly would not have been able to do this without your guidance. Chris, I will always cherish our keyboard smashing sessions trying to elucidate these variants.

I would like to thank Dr. Vinodh Kumar Radhakrishnan, for always pushing me to be a better scientist and reminding me that my contributions – as big or little as they are, make a difference.

I would like to thank Dr. Elena Grintsevich for instilling the love of research in me. Our investigations into actin made me realize what I wanted to do for the rest of my life.

I would like to thank my younger sibling unit, comprised of Ishaq and Cassandra, for tirelessly listening to my scientific jargon rants.

I would like to thank my fellow lab mates– both at the Senthil Lab and Hughes Lab, for being so supportive of all of my projects and endeavors. You guys are amazing.

I would like to thank Andy Chan for all of his support through this process.

Finally, I would like to thank my peers from the MS-BATS program who truly became family. Katie, Lauren, Rafael, Muhammed, Sultan, Jeff, Hridhay, Otilio, Jiali, we became more than classmates– we became family. I look forward to seeing all of you grow and make this world a better place. Thank you for your presence in my life.

ABSTRACT OF THE THESIS

CDH1 Germline Mutations in the Prevalence of Gastric Cancer in Historically Underrepresented Racial and Ethnic Groups in Healthcare

by

Aaqil M. Khan

Master of Biomedical and Translational Science

University of California, Irvine, 2024

Dr. Maheswari Senthil, Chair

Gastric cancer (GC) is a disease that has high incidence and mortality for Hispanic Americans at disproportionate levels. Previous studies have indicated that Hispanic gastric cancer patients may have high frequencies of CDH1 germline mutations affecting cell adherence and contact inhibition protein expression. The All of Us (AoU) Research Program is a National Institutes of Health initiative to develop a million-patient cohort of Americans from all racial and ethnic backgrounds to further precision medicine. In this thesis, the AoU Whole Genome Sequencing (WGS) dataset was utilized to assess the frequency and pathogenicity of CDH1 mutations in both Hispanic and non-Hispanic gastric cancer patients and determine the presence of a different mutational landscape within their CDH1 genes.

Chapter 1: Introduction

Gastric cancer, a malignant neoplasm of the stomach, is the 5th most common diagnosed cancer and 5th most common cause of cancer death globally[1]. Patients with metastatic gastric cancer experience high mortality rates with a relative 5-year survival rate of 36.4%[2]. Within the United States, numerous prior studies have shown that the Hispanic population experiences disproportionately high levels of gastric cancer, often characterized by more aggressive molecular subtypes and diagnosis at a younger age than non-Hispanic counterparts[3,4]. Gastric cancers have demonstrated familial clustering, with hereditary diffuse gastric cancer (HDGC) accounting for 1-3% of annual gastric cancer cases[5]. Furthermore, previous studies have implicated mutations in cell adhesion proteins as contributing to pathogenicity and as such, increased levels of disease, with some of these mutations being present within germline cells[4]. This study aims to use the large dataset of Hispanic individuals within the All of Us Research Program to assess the distribution and incidence of CDH1 germline variants within Hispanic gastric cancer patients.

Chapter 2: Background

Epidemiology

The overall incidence of gastric cancer has decreased significantly over the past 30 years, with GLOBOCAN 2020 data reporting the areas of highest incidence occurring in Latin America, Asia, and the Middle East (Figure 1)[6]. Developing nations with high rates of *Heliobacter pylori* have the highest rates of gastric adenocarcinoma. Studies have indicated that socio-economic and epigenetic factors play a role in the development of disease[7].

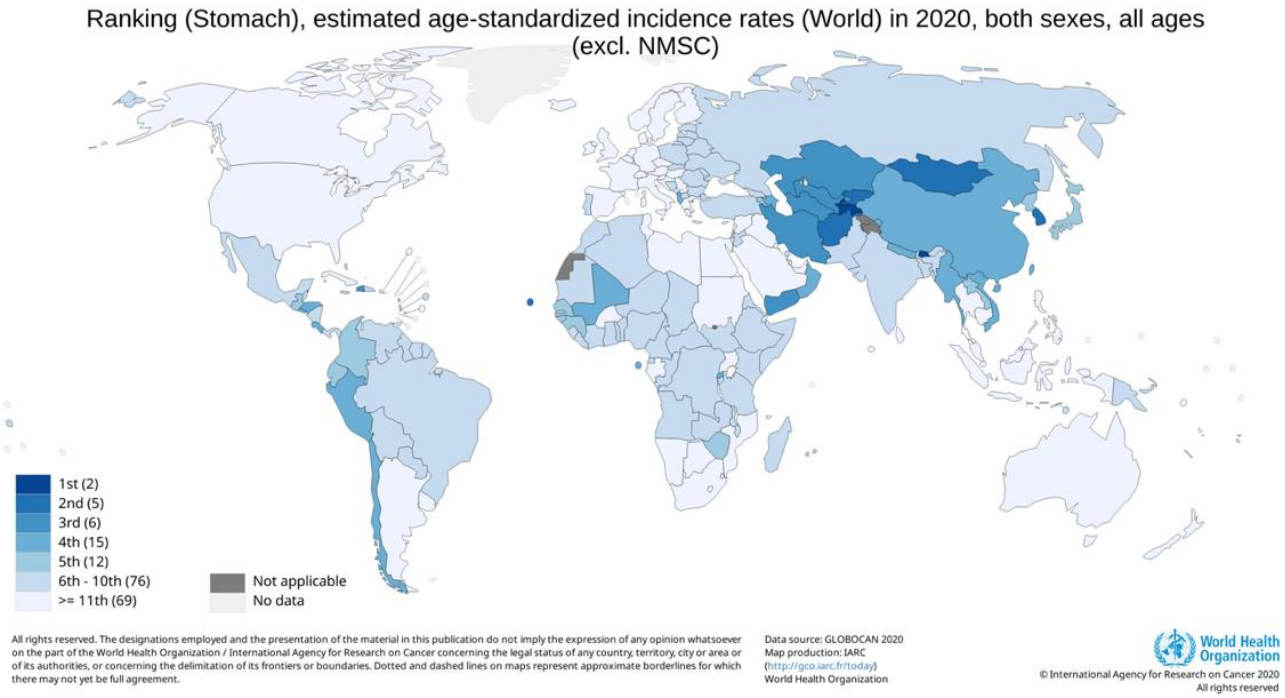

Figure 2.1 – GLOBOCAN 2020[6] data indicating the areas of highest incidence for gastric cancers.

In the United States, approximately 26,500 individuals will be diagnosed with gastric cancer this year[2]. Miller et al. reports increased incidence in Hispanic males and females (1.62 and 2.22, respectively) compared to their non-Hispanic White counterparts[8]. Mortality rates are also much higher, with Hispanic males and females experiencing 2.04 and 2.58 higher ratios of mortality than non-Hispanic Whites[8]. In Southern California, Hispanic patients experience higher rates of gastric cancer at an age-adjusted incidence rate of 10.7 cases, compared to the baseline rate of 8.6 cases for all races[9].

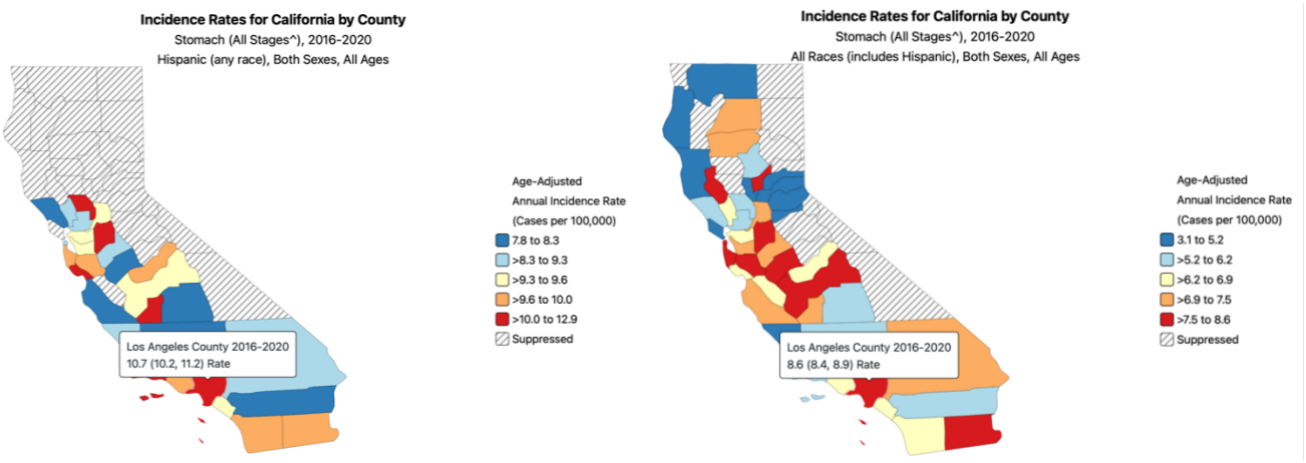

Figure 2.2 – Gastric cancer incidence rates for Hispanic and all races in Southern California. Source: SEER Data 2020[2].

Healthcare Disparities in Cancer

Vast healthcare disparities still exist within the United States for patients of different socioeconomic, demographic, and ethnic groups. Hispanic patients are more likely to present with gastric adenocarcinoma, cervical cancer, and hepatobiliary cancers than non-Hispanic White groups, with poorer 5-year mortality rates[3,10]. Other contributors to health disparity include lack of access to tertiary care centers and health insurance accessibility and thus, prolonged care[11]. Furthermore, many screening programs, including the Cancer Genome Atlas, lack a robust Hispanic patient population within their studies[12]. Thus, this raises the need for further investigation into causality of gastric cancer within Hispanic patients and germline variants may play a role in this.

Pathophysiology of Hereditary Diffused Gastric Adenocarcinoma

Patients often present with unexplained weight loss, abdominal pain, nausea, and dysphagia, prompting further investigation. The standard clinical approach involves esophagogastroduodenoscopy (EGD) with biopsies to look for abnormal tissue. Biopsies

are pathologically analyzed for the presence of abnormal cellular structure, such as *in situ* signet ring cells or pagetoid spread of signet ring cells[13]. Figure 2.3 depicts signet cell morphology associated with diffuse type gastric adenocarcinoma[14].

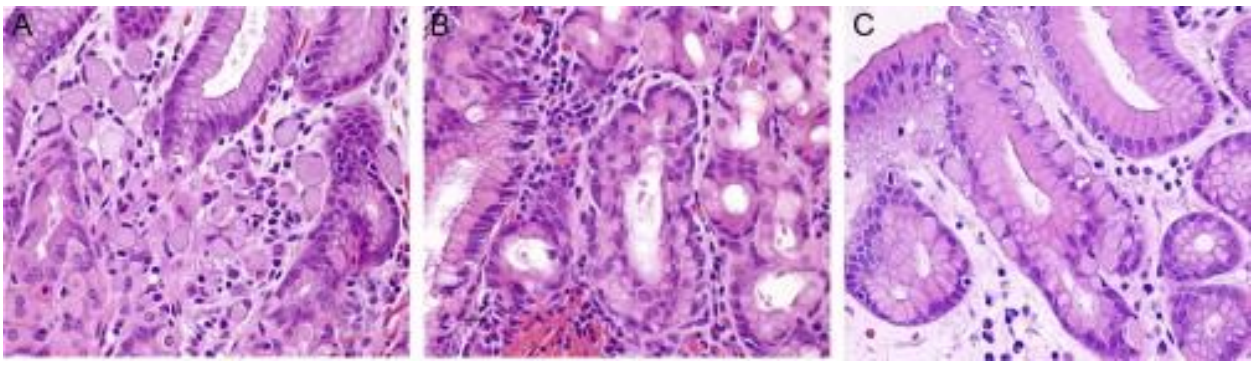

Figure 2.3 – (A) Intramucosal signet-ring cells in diffuse cancer. (B) Signet-ring cell carcinoma *in situ*. (C) Pagetoid spread of cells with signet-ring morphology. Source: Ontilio et al. 2013[16].

Computed tomography (CT) scans can provide enhanced imaging of the extent of neoplastic tissue and possible sites of metastasis.

Hereditary diffused gastric adenocarcinoma (HDGC) is a subset of gastric cancers occurring in approximately 1-3% of cases. HDGC is characterized by mutations in the gene CDH1 leading to reduced cellular adhesion and contact inhibition[15]. CDH1 mutations have also been shown to drive signet ring cell formation and have been implicated in lobular breast cancer as well. Patients with mutations in the CDH1 gene have a 75% higher likelihood of passing it to subsequent generations[14]. Currently, genetic counseling is recommended to individuals who have two or more immediate family members diagnosed with diffuse gastric cancer[15]. A study conducted by Ontilio et al. identified a germline mutation in a 56-year-old patient with terminal gastric adenocarcinoma[16]. Genetic screening of his siblings and children found cancerous sites in two siblings and both of his children, prompting prophylactic gastrectomies[16]. Pathological examination of the gastric tissue revealed malignant cells in all four family members. This study serves as an example of the potential in utilizing hereditary genetic testing and germline screening in preventing aggressive gastric adenocarcinoma prior to metastatic spread. However, it must be noted that only genomic mutations within the CDH1 gene were reported.

Molecular Diagnosis

E-cadherin coded by the CDH1 gene on chromosome 16 is a calcium dependent protein that mediates cell-cell interactions, cell adhesion and contact inhibition, and is classified as a tumor-suppressor gene[17]. CDH1 contains 16 exons that code for an 882 amino acid long protein with five E-cadherin repeats. Abnormalities in E-cadherin expression can be linked to mutations within CDH1, epigenetic factors (such as promoter methylation) transcriptional silencing, and regulatory microRNA dysfunction[18]. Mutations occurring at a single point within the coding sequence are referred to as single nucleotide polymorphisms (SNPs) and may be substitutions, insertions, or deletions. SNPs can have downstream effects to the protein's tertiary structure, leading to loss of function and loss of tumor suppression[19]. A catalog of SNPs and other genomic variants is made available by the National Center for Biotechnology Information (NCBI) through the ClinVar database. To date, ClinVar has identified 4476 variants within the CDH1 gene.

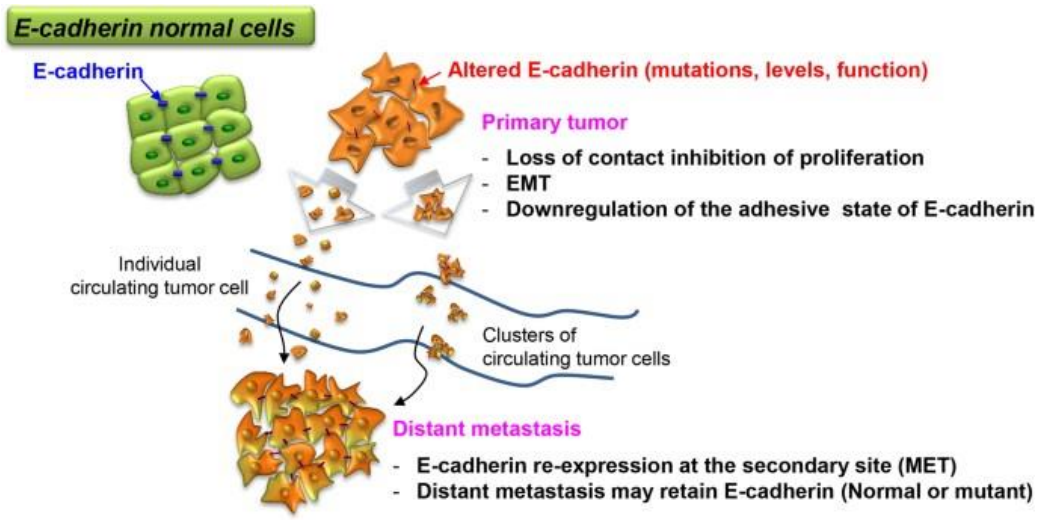

Figure 2.4 – A diagram depicting the downstream effects of E-Cadherin loss-of-function. Source: Shastry et al. 2007[19].

Improved Screening and Treatment Options

Genetic disease counseling is a rapidly emerging field that leverages the reduced cost and increased availability of gene sequencing for patient precision medicine. Discovery of the linkage between BRCA1 and BRCA2 mutations to ovarian and early onset breast cancer has influenced clinical decisions and allowed for prophylactic mastectomies and hysterectomies prior to cancer development[20]. Recently, the National Institutes of Health (NIH) has initiated the All of Us Research Program (AoU), a clinical research program aimed at preparing a cohort of 1 million individuals from diverse ethnic backgrounds representing the United States population[21]. The AoU program provides individualized patient medical histories as well as whole-genome sequencing to help further the personalized medicine initiative. Upon completion of data access training, any US researcher may utilize the AoU research data for medical research. Within this study, we utilized the AoU research program data to discern the incidence of CDH1 mutations within a gastric cancer cohort. Linked demographic and socioeconomic data allowed for preliminary analysis into sociodemographic factors that may play a role in disease. Potential genetic markers within the CDH1 gene may influence prophylactic gastrectomies to reduce the risk of diffuse gastric cancer in identified patients.

Chapter 3: Methods

Establishing the All of Us Cloud Platform

The All of Us Research Program provides summary statistics publicly to any user, available through the Data Browser. To access participant level data, researchers must undergo identity verification and research ethics trainings. Further training allows for access to the controlled tier of data, which contains de-identified participant electronic medical records, survey information, and medical histories, and whole-genome sequencing data. Both registered and controlled tiers must be accessed through the researcher workbench, a controlled virtual bioinformatics environment that allows for cohort selection, data aggregation, and bioinformatic analysis. For this study, the controlled tier was utilized to analyze genome wide data. Figure 3.1 depicts the different tiers of data access as well as the requirements for access.

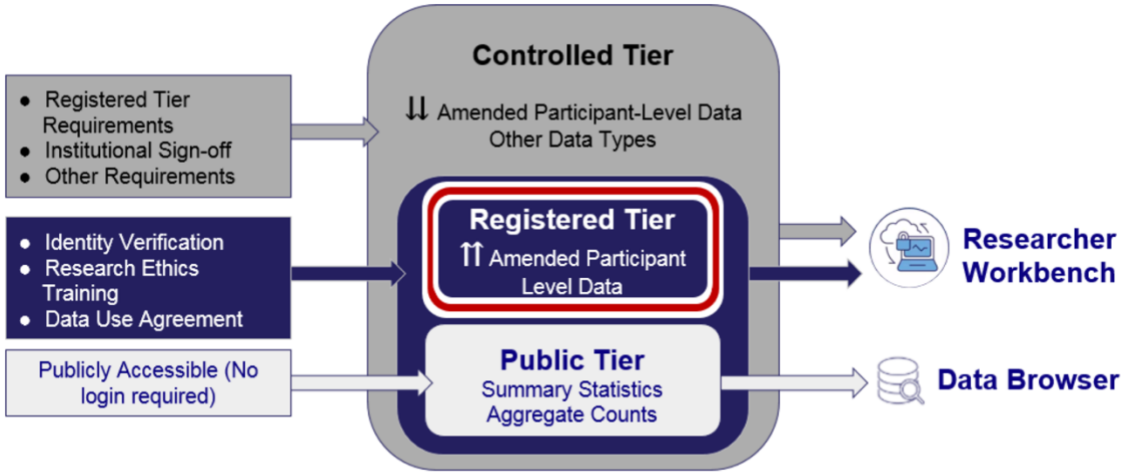

Figure 3.1 – Organization of data within the All of Us Research Program. Registered and Controlled tiers of data access require specific data handling training prior to access. Source: NIH 2024[21].

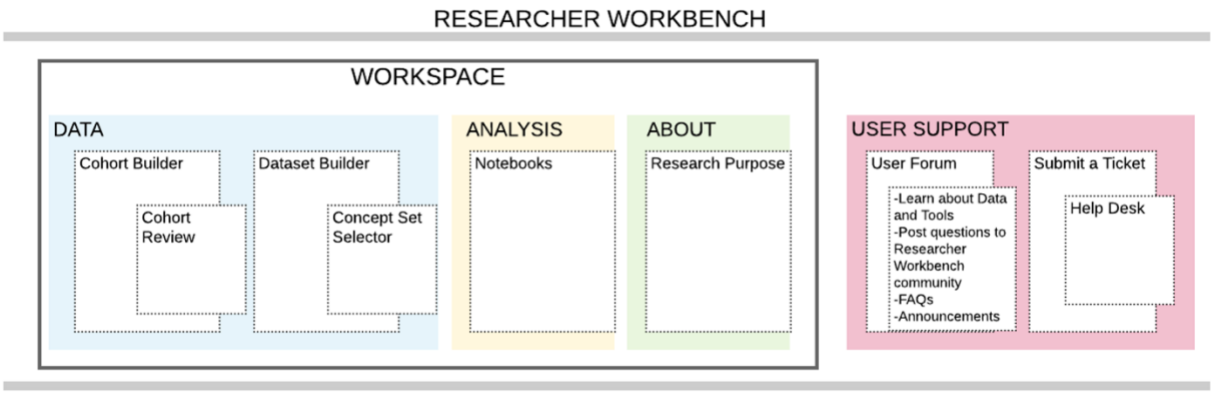

Figure 3.2 – Organization of the Researcher Workbench within the All of Us Research Program. Bioinformatic analysis is conducted using Python in Jupyter notebooks, located within the Analysis Workspace. Source: NIH 2024[21].

All of Us Research Program Sequence Alignment

To provide a robust whole genome sequencing data to Researcher Workbench users, the All of Us Research Program has implemented a number of quality control and quality assurance measures within their genomic processing pipeline[21]. Patient samples were prepared and sequenced using the Illumina Kapa HyperPrep kit and Illumina NovaSeq 6000 instrument. Initial quality control was conducted using the Illumina DRAGEN pipeline and used to assemble the Genomic Variant Store (GVS) following additional QC steps with reference to the Human Genome Reference Assembly GRCh38/hg19. GVS data was then parsed into Genomic Region Callsets that could be accessed through bioinformatics tools, such as Hail and Plink within the Researcher Workbench. The program reports consistent coverage and uniformity across the genome for all patient Whole Genome Sequencing (WGS) data on par with clinical-level data[21].

Jupyter Notebooks and Hail Matrix Tables

Bioinformatic analysis of WGS data by researchers using the Controlled Tier must be conducted within a Jupyter environment. Project Jupyter is an open-source python code

execution environment that has been incorporated within the All of Us Researcher Workbench. The Researcher Workbench contains tools to select cohorts and stratify patients by condition, social and economic background and conduct analysis within python notebooks. Analysis tools available to researchers within the workbench include Plink, Hail, R-Studio and recently SAS. For this study, Hail, an open-source python library for bioinformatic data exploration and analysis, was utilized to query the patient demographic and genomic data for the different approaches. Hail situated the data within a nested matrix, where variants were represented in row fields and patients/samples were represented in column fields. Data from Hail were exported to another open-source python program called Pandas that allowed for the manipulation of data tables, called dataframes within the python environment. Figure 3.3 depicts the layout of a matrix table. The general pipeline used to formulate results involved filtering and limiting the matrix table to the individual project aims, formulating Pandas dataframes with participant counts per variant, and performing analysis with reference to the number of individuals carrying that variant within the entire All of Us WGS cohort, gnomAD, and ClinVar.

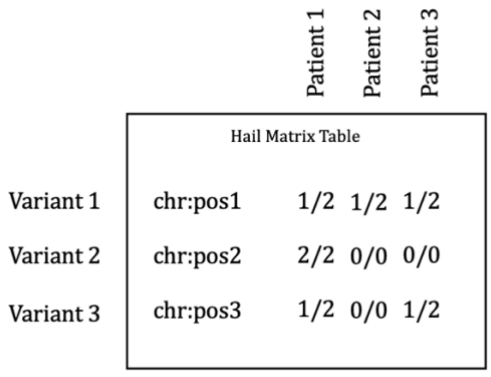

Figure 3.3 – Layout of a Hail Matrix table. Patients are assigned to column fields and variants within the table are listed in the row fields. Presence of each allele may be homozygous or heterozygous and represented by 1/2 or 2/2 within the intersection of patient and variant.

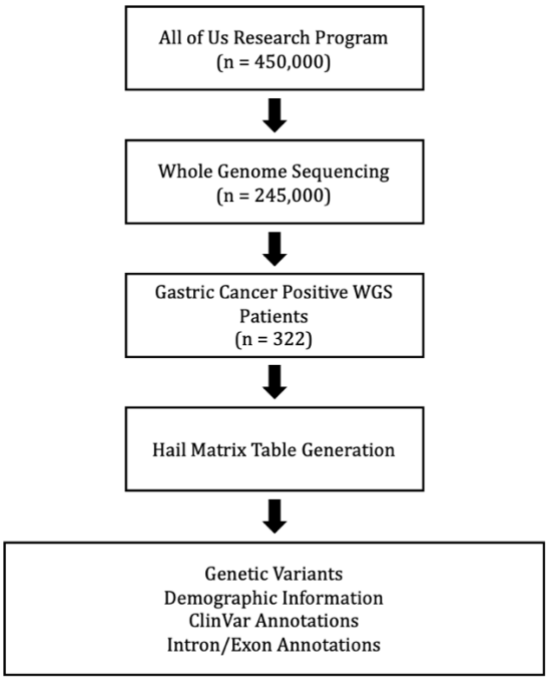

Figure 3.4 – Analysis workflow for determining CDH1 variants within the All of Us Research Program database.

Aim 1 – Identifying Variants within the Gastric Cancer+ Hispanic and Non-Hispanic Groups

The first aim was to determine the number of variants per person found within the matrix table for Hispanic and Non-Hispanic individuals with gastric cancer. Python code was utilized to determine the average, median, and range of mutations for individuals identifying as Hispanic or Non-Hispanic through the All of Us concept identifiers "race_concept_id" and "ethnicity_concept_id". Counts were formulated using concept identifiers and exported to Pandas for dataframe creation. Python addition algorithms were applied for unique concept identifiers corresponding to race and ethnicity and then the resultant dataframe was exported from the researcher workbench as a comma separated value (CSV) spreadsheet for final data preparation within Microsoft Excel. Concept identifiers used within this study and snippets of the code utilized to filter the matrix table and identify variants are present within Appendix I.

Aim 2 – Identifying Exon Variants in the ClinVar Database

The second aim of this project was to assess the exonic variants that were present within the CDH1 gene and reference them to ClinVar and gnomAD. This was accomplished by filtering the matrix table to exonic regions within CDH1 in reference to the positions listed for GRCh38/hg19 within UniProt[22]. Exonic locations are presented in Table 3.1. The matrix table was then cross-referenced using python to ensure that allele frequencies reported were homozygous or heterozygous and missing information was not included in patient counts prior to data frame creation. Exonic locations were manually curated referencing gnomAD and ClinVar for rs numbers and previous reports of pathogenicity. After referencing ClinVar and gnomAD, the All of Us Cohort Builder was used to determine descriptive statistics and racial/ethnic distribution of variant carriers for the AoU WGS cohort of patients.

Exon Number	**Genomic Locus GRCh38**
Exon 1	16:68,737,416 - 68,737,463
Exon 2	16:68,738,297 - 68,738,411
Exon 3	16:68,801,670 - 68,801,893
Exon 4	16:68,808,424 - 68,808,567
Exon 5	16:68,808,693 - 68,808,848
Exon 6	16:68,810,197 - 68,810,341
Exon 7	16:68,811,684 - 68,811,859
Exon 8	16:68,812,135 - 68,812,263
Exon 9	16:68,813,313 - 68,813,495
Exon 10	16:68,815,515 - 68,815,759
Exon 11	16:68,819,280 - 68,819,425
Exon 12	16:68,822,001 - 68,822,225
Exon 13	16:68,823,399 - 68,823,626
Exon 14	16:68,828,174 - 68,828,304

Exon 15	16:68,829,654 - 68,829,797
Exon 16	16:68,833,290 - 68,833,496

Table 3.1 – CDH1 exonic locations according to the Uniprot 2023 guidelines for reference genome GRCh38. Exon identification within the matrix table was manually limited to these locations.

Aim 3 – Correlation of previously reported CDH1 Intron Variants

The ClinVar reference for CDH1 including variants classified as "Pathogenic", "Likely Pathogenic", "Uncertain Significance" and "Conflicting Classifications" were downloaded from the NCBI website and imported as a Pandas dataframe to the All of Us environment. The genomic locus, reference, and alternate alleles were parsed between the gastric cancer positive (GC+) dataframe and the ClinVar dataframe to ensure data compatibility and manually referenced after merging. Python code was used to ensure that matching entries between the ClinVar dataframe and All of Us variants were carried over to the resultant dataframe. This dataframe was then exported as a CSV to Microsoft Excel. Prevalence rates within the All of Us WGS cohort were determined using the Cohort Builder.

Aim 4 – Correlation of previously reported CDH1 Variants by Wang et al.

Previously reported variants by Wang et al[4]. were then queried within the GC+ cohort and entire All of Us WGS cohort to determine if they were present within Hispanic GC+ patients at higher amounts than the non-Hispanic patients as hypothesized. This was accomplished by locating the genomic locus within GRCh38 using the Uniprot database, assembling an input dataframe and performing matching using a Python algorithm.

Aim 5 – Determining Novel Intron Variants found within the GC+ Cohort

Upon completion of the exon variant analysis, variants within the remaining list that had incidence rates of 1.5 times greater in Hispanic GC+ patients than non-Hispanic GC+

patients within the cohort were identified. These variants were then analyzed for previous entries within the gnomAD, ClinVar, and dbSNP databases, and manually curated to reflect population prevalence within the All of Us WGS cohort and gnomAD reference allele frequencies using the All of Us Cohort Builder.

Chapter 4: Results

Patient Demographics

A total of 322 patients within the All of Us Research program met the criteria for "Malignant tumor of the stomach" and whole genome sequencing. This cohort was comprised of 174 females (54%) and 145 males (45%). Two individuals declined to state their sex at birth. The average age was 69 years old ± 12 years (median = 71), and 52 (16%) individuals identified ethnically as Hispanic, while 263 (82%) identified as non-Hispanic. 7 individuals declined to state ethnicity or skipped the question (2%). The most common conditions apart from stomach cancer were essential hypertension (73% of cohort), abdominal pain (63% of cohort), hyperlipidemia (61% of cohort), anemia (59% of cohort), and gastroesophageal reflux disease without esophagitis (58% of cohort). A total of 245,388 patients were included in the All of Us WGS reference cohort, with 145,580 females (59%) and 94,760 males (39%), with an estimated average age of 56 years. There were 47,371 (20.1%) individuals who identified as Hispanic, with 188,650 (79.9%) individuals identifying as non-Hispanic.

Cohort Descriptive Statistics	**GC+ Cohort n = 322**	**AoU WGS Cohort n = 245,388**
Age		
18-39	5 (1.6)	50500 (20.6)
40-69	143 (44.4)	130040 (53)
70-89	166 (51.6)	62220 (25.4)
89+	14 (4.4)	2700 (1.1)
Sex at Birth		
Female	175 (54.3)	145580 (59.3)
Male	145 (45)	94760 (38.6)
Other	2 (0.6)	5080 (2.1)
Race		
White	196 (60.8)	129525 (52.8)
Asian	9 (2.7)	7647 (3.1)
Middle Eastern	5 (1.5)	1389 (0.5)
Black or African American	51 (15.8)	50969 (20.1)

Native Hawaiian or Pacific Islander	1 (0.3)	280 (11)
Skip/Not Included	57 (17.7)	51305 (20.9)
Multiple	3 (0.9)	4273 (1.74)
Ethnicity		
Hispanic	52 (16.4)	47371 (20.1)
Non-Hispanic	263 (81.6)	188650 (79.9)

Table 4.1 – Descriptive statistics of the All of Us WGS cohort (n = 245,388) as well as the GC+ cohort (n = 322).

Genomic Variants Present in the Gastric Cancer Positive Cohort

Within the GC+ cohort, 100% of the patients had variants within their CDH1 gene, with an average of 173 variants for individuals identifying as Hispanic with a median of 184 and a range of 40-260 variants per person. Non-Hispanic individuals had an average of 176 variants per person, with a median of 182 and a range of 52-305. Within the GC+ cohort, there were a total of 1332 variants found within the bounds of the CDH1 gene and ±1kb intergenic regions, with incidence in at least one individual. Of these, 662 of these variants were found in the group identifying as Hispanic and 1240 were found in the group identifying as non-Hispanic. Hispanic GC+ patients had 8 exon mutations, classified as benign by ClinVar and non-Hispanic patients had 22 mutations, 12 benign, 2 of conflicting classification, 2 of uncertain significance, and 2 pathogenic (Table 4.2). Furthermore, 48 patients (96%) of the Hispanic GC+ and 230 (87.4%) of the non-Hispanic GC+ patients exhibited exon mutations within their CDH1 gene (Table 4.3).

ClinVar Variant Frequencies	**Variants**	**Benign/Likely Benign**	**Conflicting Classifications**	**Uncertain Significance**	**Likely Pathogenic**	**Pathogenic**
All of Us GC+						
Total	1332	76	10	7	0	2
Intron	1309	64	8	5	0	0
Exon	23	12	2	2	0	2
Hispanic						
Total	662	46	3	0	0	0
Intron	654	38	3	0	0	0
Exon	8	8	0	0	0	0

Non-Hispanic						
Total	1240	71	7	7	0	2
Intron	1218	55	5	5	0	0
Exon	22	12	2	2	0	2

Table 4.2 – Frequencies of intron and exon variants within the GC+ patients, stratified by ClinVar pathogenicity.

Variant Frequency Statistics	**Hispanic GC+ n = 52 (%)**	**Non-Hispanic GC+ n = 322 (%)**
Total	52	263
Exon	48 (92)	230 (87.4)
Intron Only	4 (7.6)	33 (12.5)
Patients with 1 Exon Variant	37 (71.1)	189 (71.8)
Patients with 2 Exon Variants	7 (13.4)	32 (12.1)
Patients with 3-4 Exon Variants	4 (7.6)	9 (3.4)

Table 4.3 – Frequencies of intron and exon variants within the GC+ patients, stratified by ClinVar pathogenicity.

Exon variants were analyzed based on the type of mutation. Out of the 23 exon variants identified, 11 were synonymous mutations within the AoU GC+ cohort (Table 4.4). One mutation, chr16:68823538-T-C, was found in greater than 87% of all patients, regardless of cancer status or race/ethnicity, indicating that the majority of a given population may be carriers of benign exon CDH1 mutations. Although synonymous mutations retain the original amino acid structure, two were previously unreported and one had conflicting classifications. Three mutations were seen at higher incidences within Hispanic patients compared to non-Hispanic patients but were all classified as benign.

Synonymous Mutation (11)	**Alleles**	**rsID**	**ClinVar Pathogenicity**	**Hispanic n = 47371 (%)**	**Non-Hispanic n = 188650 (%)**	**GC+ Hispanic n = 52 (%)**	**GC+ non-Hispanic n = 263 (%)**
chr16:68737448	[‘G’,‘C’]	rs730881654	Benign	4 (0)	90 (0)	0 (0)	1 (0)
chr16:68801851	['G', 'A']	rs1801023	Benign	237 (1)	1250 (1)	0 (0)	1 (0)

chr16:68823538	['T', 'C']	rs1801552	Benign	42313 (89)	168595 (89)	48 (92)	229 (87)
chr16:68833484	['C', 'T']	rs2229044	Benign	1627 (3)	5759 (3)	4 (8)	7 (3)
chr16:68813447	['C', 'T']	rs61756284	Benign, Likely Benign	250 (1)	337 (0)	1 (2)	2 (1)
chr16:68833370	['C', 'T']	rs140328601	Benign, Likely Benign	128 (0)	224 (0)	0 (0)	1 (0)
chr16:68833487	['C', 'T']	rs141001592	Benign, Likely Benign	2 (0)	37 (0)	1 (2)	0 (0)
chr16:68815574	['A','G']	rs1597897867	Likely Benign	0 (0)	1 (0)	0 (0)	1 (0)
chr16:68811808*	['T', 'A']	rs549252135	Conflicting Classifications	3 (0)	6 (0)	0 (0)	1 (0)
chr16:68811784	['C','G']	rs35539711	Not Reported	543 (1)	3173 (2)	0 (0)	4 (2)
chr16:68819394	['G','C']	rs35741240	Not Reported	138 (0)	1083 (1)	0 (0)	1 (0)

Table 4.4 – Synonymous exon mutations found within the GC+ cohort of patients after ClinVar annotation. *Indicates variants present in Table VII, Table VIII.

There were an additional 10 missense mutations identified within the GC+ cohort (Table 4.5). Missense mutations occur when the amino acid sequences are altered due to the presence of a single nucleotide polymorphism and may have deleterious or pathogenic effects on the expression and function of E-cadherin. Within the missense mutations, 6 were classified as benign or likely benign, three had uncertain significance, one had conflicting classifications in ClinVar. Interestingly, of the missense mutations identified as benign, one mutation, chr16:68828262-C-T was present in 8% of the Hispanic GC+ cohort and 13.54% of the AoU WGS reference cohort, compared to 3% and 6.9% respectively. Three additional missense mutations had higher incidence within the GC+ Hispanic patients but were seen at low levels of incidence within the AoU WGS reference cohort. Two pathogenic variants were also identified within the exonic regions (Table 4.6), however, both occurred with very low frequency within the GC+ cohort and the AoU WGS cohort (<0.05%).

Missense Mutation (10)	Alleles	rsID	ClinVar Pathogenicity	Hispanic n = 47371 (%)	Non-Hispanic n = 188650 (%)	GC+ Hispanic n = 52 (%)	GC+ non-Hispanic n = 263 (%)
chr16:68811743	['G', 'A']	rs142822590	Benign	19 (0)	164 (0)	0 (0)	1 (0)
chr16:68822185	['C', 'T']	rs33969373	Benign	1841 (4)	7166 (4)	4 (8)	9 (3)
chr16:68828262	['C', 'T']	rs33964119	Benign	6414 (13.54)	13024 (6.9)	3 (5.77)	12 (4.56)
chr16:68822063	['G', 'A']	rs35187787	Benign	186 (0)	1345 (1)	0 (0)	2 (1)
chr16:68808832	['G', 'A']	rs201511530	Benign	4 (0)	17 (0)	1 (2)	0 (0)
chr16:68813472	['G', 'A']	rs199886166	Likely Benign	1 (0)	45 (0)	0 (0)	0 (0)
chr16:68822217*	['A', 'G']	rs1567512606	Uncertain Significance	0 (0)	3 (0)	0 (0)	1 (0)
chr16:68828273*	['A', 'G']	rs187289510	Uncertain Significance	0 (0)	2 (0)	0 (0)	0 (0)
chr16:68822138	['G', 'A']	rs33935154	Uncertain Significance	701 (1)	4506 (2)	2 (4)	3 (1)
chr16:68833365*	['G', 'A']	rs587780121	Not Reported, Conflicting Classifications	0 (0)	7 (0)	0 (0)	1 (0)

Table 4.5 – Frequencies of missense mutations found within the GC+ cohort after ClinVar annotation.

*Indicates variants present in Table VII, Table VIII.

Pathogenic (2)	Alleles	rsID	ClinVar Pathogenicity	Hispanic n = 47371 (%)	Non-Hispanic n = 188650 (%)	GC+ Hispanic n = 52 (%)	GC+ non-Hispanic n = 263 (%)
chr16:68801884	['GC', 'G']	rs1555514492	Pathogenic	0 (0)	1 (0)	0 (0)	1 (0)
chr16:68822081	['C', 'T']	rs121964877	Pathogenic	0 (0)	2 (0)	0 (0)	1 (0)

Table 4.6 – Frequencies of pathogenic found within the GC+ cohort after ClinVar annotation. Both variants are deleterious variants that have downstream loss-of-function effects.

ClinVar annotation was also applied to intronic regions within the CDH1 gene, yielding 7 variants of uncertain significance (Table 4.7) and 7 variants of conflicting classification (Table 4.8). There were no GC+ Hispanic patients with intronic variants of uncertain significance, and only one variant of conflicting classification was seen in a single Hispanic GC+ patient. Interestingly, three of the intron variants occurred at 5' or 3' untranslated regions (UTRs) within CDH1.

Intron Variants

Uncertain Significance (7)	**Alleles**	**Variant Type**	**Hispanic n = 47371 (%)**	**Non-Hispanic n = 188650 (%)**	**GC+ Hispanic n = 52 (%)**	**GC+ non-Hispanic n = 263 (%)**
chr16:68801969	[C, A]	Intron Variant	555 (1.2)	3146 (1.7)	0 (0)	5 (1.9)
chr16:68821840	[G, A]	Intron Variant	473 (1)	2037 (1.1)	0 (0)	3 (1.1)
chr16:68821874	[G, A]	Intron Variant	69 (0.1)	168 (0.1)	0 (0)	2 (0.8)
chr16:68822217*	[A, G]	Missense Variant, 5' UTR Variant	0 (0)	3 (0)	0 (0)	1 (0.4)
chr16:68823288	[C, A]	Intron Variant	238 (0.5)	1768 (0.9)	0 (0)	1 (0.4)
chr16:68828273*	[A, G]	Missense Variant	0 (0)	2 (0)	0 (0)	1 (0.4)
chr16:68833923	[T, A]	3' UTR Variant	13 (0)	264 (0.1)	0 (0)	1 (0.4)

Table 4.7 – Intron variants of uncertain significance found within the CDH1 gene with ClinVar annotation. 5' and 3' UTR variants may play a role in gene regulation and transcriptional processes. Table VII *Indicates exon variant.

Conflicting Classifications (7)	**Alleles**	**Variant Type**	**Hispanic n = 47371 (%)**	**Non-Hispanic n = 188650 (%)**	**GC+ Hispanic n = 52 (%)**	**GC+ non-Hispanic n = 263 (%)**
chr16:68737127	['GT', 'G']	deletion	2 (0)	32 (0)	0 (0)	1 (0.4)
chr16:68808403	['C', 'A']	Intron Variant	10 (0)	51 (0)	0 (0)	1 (0.4)
chr16:68811808*	['T', 'A']	Synonymous Variant, 5' UTR Variant	3 (0)	6 (0)	0 (0)	1 (0.4)
chr16:68819273	['C', 'T']	Intron Variant	0 (0)	9 (0)	0 (0)	1 (0.4)

chr16:68823374	['C', 'A']	Intron Variant	48 (0.1)	127 (0.1)	0 (0)	1 (0.4)
chr16:68833147	['G', 'A']	Intron Variant	730 (1.5)	5226 (2.8)	1 (1.9)	7 (2.7)
chr16:68833365*	['G', 'A']	Missense Variant	0 (0)	7 (0)	0 (0)	1 (0.4)

Table 4.8 – Intron variants of conflicting classification found within the CDH1 gene with ClinVar annotation. 5' and 3' UTR variants may play a role in gene regulation and transcriptional processes. *Indicates exon variant.

Of the 7 exon variants proposed by Wang et al[4], only 4 were seen in the All of Us WGS cohort (Table 4.9), and only one was seen in the GC+ cohort in 3 non-Hispanic individuals (1.1%). One variant, Exon3:c.286A>G, classified by ClinVar as benign was seen in 56 Hispanic WGS patients (0.12%), however none of these patients had gastric cancer. Functional assays by the Wang group suggested that although this variant had previously been classified as benign, it posed increased risk of cellular motility[4].

Patient ID	**Position Identifier**	**ClinVar Classification**	**Hispanic n = 47371 (%)**	**Non-Hispanic n = 188650 (%)**	**GC+ Hispanic n = 52 (%)**	**GC+ non-Hispanic n = 263 (%)**
P15	Exon3:c.286A>G	Benign	56 (0.12)	1 (0)	0	0
P20	Exon12:c.1849G>A	Benign	701 (1.48)	4506 (2.39)	2 (0.4)	3 (1.1)
P71	Exon16:c.2558C>T	Uncertain Significance	8 (0)	0 (0)	0 (0)	0 (0)
P50	Exon6:c.715G>A		0 (0)	1 (0)	0 (0)	0 (0)

Table 4.9 – Exon variants proposed by Wang et al[4]. that may be present in Hispanic gastric cancer patients.

Of the remaining intron variants, 13 presented at a frequency of 1.5 times or greater in Hispanic GC+ patients compared to non-Hispanic GC+ patients (Table 4.10). Several of these variants also demonstrated high frequency in Hispanic patients within the All of Us WGS population as well as high allele frequencies in the gnomAD Admixed American group, compared to the gnomAD aggregated allele frequencies. It is important to note, however, that these variants have no report of pathogenicity within ClinVar.

Intron Variants (13)	Variant Type	Hispanic n = 47371 (%)	Non-Hispanic n = 188650 (%)	GC+ Hispanic n = 52 (%)	GC+ non-Hispanic n = 263 (%)	gnomAD Admixed American Allele Frequency	gnomAD Allele Frequency	ClinVar Pathogenicity
16-68745340-A-T	Intron Variant	4975 (10.5)	561 (0.3)	7 (13.46)	0 (0)	0.04282	0.005675	Not Reported rs185033464
16-68746889-C-A	Intron Variant	2663 (5.62)	117 (0.06)	3 (5.77)	1 (0.38)	0.03088	0.003363	Not Reported rs191163372
16-68746948-C-CAACA	Intron Variant	2824 (5.96)	1019 (0.54)	4 (7.69)	0 (0)	0.03062	0.006271	Not Reported rs201828383
16-68755672-C-A	Intron Variant	1829 (3.86)	126 (0.07)	4 (7.69)	1 (0.38)	0.01481	0.001663	Not Reported rs189254840
16-68768520-A-G	Intron Variant	5779 (12.2)	1459 (0.77)	8 (15.38)	1 (0.38)	0.05756	0.009614	Not Reported rs12444784
16-68771844-G-A	Intron Variant	2828 (5.97)	990 (0.52)	4 (7.69)	0 (0)	0.03071	0.006266	Not Reported rs139474274
16-68782576-C-T	Intron Variant	457 (0.96)	14 (0.01)	2 (3.85)	0 (0)	0.002419	0.0002692	Not Reported rs188420567
16-68786814-C-A	Intron Variant	4906 (10.36)	308 (0.16)	6 (11.54)	0 (0)	0.04295	0.004885	Not Reported rs146972234
16-68789518-C-CT	Intron Variant	6068 (12.81)	687 (0.36)	8 (15.38)	0 (0)	0.06104	0.008348	Not Reported rs549644747
16-68819472-G-A	Intron Variant	3338 (7.05)	833 (0.44)	6 (11.54)	0 (0)	0.05328	0.003594	Not Reported rs35667437
16-68820350-TA-T	Intron Variant	4629 (9.77)	543 (0.29)	3 (5.77)	1 (0.38)	0.04913	0.006705	Not Reported rs202183535
16-68822855-G-A	Intron Variant	2486 (5.25)	115 (0.06)	2 (3.85)	1 (0.38)	0.03201	0.002847	Not Reported rs181878715
16-68832339-G-A	Intron Variant	201 (0.42)	7 (0)	2 (3.85)	0 (0)	0.001576	0.0001648	Not Reported rs148120621

Table 4.10 – Intron variants present at a frequency of 1.5x or higher in Hispanic GC+ patients, compared to non-Hispanic GC+ patients.

Chapter 5: Discussion and Future Work

Germline Mutations in Hispanic Individuals

Within this study, the CDH1 gene was analyzed in a cohort of gastric cancer positive patients to determine if Hispanic patients had higher frequencies of germline variants. Within the 23 exons identified, only 8 exon variants were present in 52 Hispanic patients, and all of these variants were classified as benign. This is in comparison to 22 exon variants in the non-Hispanic group. Interestingly, 92% of Hispanic and 87% of non-Hispanic individuals had germline exonic mutations, compared to previously reported 16% of patients in prior studies[4]. Furthermore, the frequency of Hispanic patients with 3-4 exon mutations was twice the frequency of non-Hispanic counterparts. Within the 8 exonic mutations, 4 were missense mutations that may play a loss-of-function role on E-cadherin due to changes in the amino acid sequence. Intron variants in ClinVar follow a similar trend with a single Hispanic patient harboring a previously reported intron variant. Although several intron variants were identified to have higher frequencies in Hispanic GC+ patients, the lack of pathogenicity classifications in ClinVar and other genomic databases limits the correlation with disease without further investigation on the evolutionary role of intronic sites. Within this study, the mutational landscape of CDH1 in Hispanic GC+ patients does not drastically differ from non-Hispanic GC+ patients as hypothesized. It does, however, does provide an insight on the depth of analysis that can be performed with a large dataset consisting of racial and ethnic groups previously underrepresented in medicine.

Limitations of Germline Screening

As germline screening continues to grow as a major part of patient precision medicine, it is important to understand the limitations of current approaches in disease correlation. Many studies asses the presence of germline mutations through whole exome sequencing pipelines, which assess only exonic locations for variants within the population being studied. Although exome sequencing allows for the elucidation of a majority of the variants

that may directly correlate with disease through amino acid changes, it does not take into account non-coding regions that play critical roles in gene regulation, such as splice donor/acceptor sites and 5'/3' intergenic regions. Furthermore, the protein expression of genes, is governed by several additional factors, such as epigenetic silencing and mRNA regulation. Although the All of Us Research Program has employed a robust preparative pipeline of raw genomic data to organized callsets that can be studied, it does not currently account for any of the epigenetic methylation or acetylation that critical genes like CDH1 may undergo in response to an individual's environment. Somatic mutations may also exist in many gastric cancer patients and have similar deleterious effects on protein function.

Reclassification of ClinVar Pathogenicity

Currently the criteria for pathogenicity level within ClinVar are selected using a set of algorithms that assess the downstream effect of each variant and the potential for loss of function on the resultant gene, then curated by an expert panel. Additionally, reclassification of these variants may occur in response to clinical data that shows tight correlation between disease status and genetic variant. For exon missense mutations identified in this study, *in silico* prediction of protein effects will allow for further context in pathogenicity. Programmatic tools, like *Sift* and *Polyphen2* are freely available and allow for deep computational modeling of protein structure and function. Identifying key splice donor and acceptor sites near the CDH1 coding regions can allow for further analysis on intronic variants that were identified. These can then be applied to functional cell-based assays that determine effects of CDH1 mutations through mutagenesis of highly specific variants. Additional exploration on the family history of GC+ patients by leveraging the diverse amount of data available within the All of Us Research Program can further aid in HDGC correlative analysis.

Clinical Interventions in Gastric Cancer Patients

Patient precision medicine involving germline mutations serves as an additional tool in the assessment of patients with gastric cancer. Presence of pathogenic mutations in patients may allow for modified treatment approaches including earlier screening EGDs and familial testing. This, however, addresses only one aspect of why Hispanic patients present with higher incidence and mortality rates of gastric cancer within the United States.

Chapter 6: Conclusion

This thesis examined the presence of germline CDH1 mutations in a cohort of gastric cancer patients with whole genome sequencing to determine the mutational landscape of CDH1 in Hispanic gastric cancer patients. Both Hispanic and non-Hispanic patients were found to have exon mutations within their CDH1 gene, at rates of 92% and 87% respectively. Of the 23 identified exonic variants, 8 were present in Hispanic GC+ patients, 4 of which were missense mutations that may be candidates for reclassification following further analysis. Variants of pathogenic, uncertain significance, or conflicting classifications of pathogenicity were not seen extensively amongst Hispanic patients, and a single intron variant of conflicting classification was seen in one Hispanic patient. Unclassified intron variants seen at high incidence warrant further investigation into protein regulatory functions. Somatic mutations and epigenetic alterations influenced by environmental factors may be an additional causal factor of the high rates of incidence and mortality observed in Hispanic gastric cancer patients.

REFERENCES

1. Smyth EC, Nilsson M, Grabsch HI, Van Grieken NC, Lordick F. Gastric cancer. *The Lancet*. 2020;396(10251):635-648. doi:10.1016/S0140-6736(20)31288-5

2. National Cancer Institute. *Stomach Cancer – Cancer Stat Facts*. https://seer.cancer.gov/statfacts/html/stomach.html

3. Yu J, Sullivan BG, Senthil GN, et al. Prevalence of Primary Liver Cancer is Affected by Place of Birth in Hispanic People Residing in the United States: All of Us Research Program Report. *The American Surgeon*. 2022;88(10):2565-2571. doi:10.1177/00031348221109465

4. Wang SC, Yeu Y, Hammer STG, et al. *Hispanic/Latino Gastric Adenocarcinoma Patients Have Distinct Molecular Profiles Including a High Rate of Germline* CDH1 *Mutations*. Cancer Biology; 2019. doi:10.1101/764779

5. Taja-Chayeb L, Vidal-Millán S, Trejo-Becerril C, et al. Hereditary diffuse gastric cancer (HDGC). An overview. *Clinics and Research in Hepatology and Gastroenterology*. 2022;46(4):101820. doi:10.1016/j.clinre.2021.101820

6. International Agency for Research on Cancer. Global Cancer Observatory - Stomach Cancer Fact Sheet. Accessed May 31, 2024. https://gco.iarc.who.int/media/globocan/factsheets/cancers/7-stomach-fact-sheet.pdf

7. Usui G, Matsusaka K, Huang KK, et al. Integrated environmental, lifestyle, and epigenetic risk prediction of primary gastric neoplasia using the longitudinally monitored cohorts. *eBioMedicine*. 2023;98:104844. doi:10.1016/j.ebiom.2023.104844

8. Miller KD, Ortiz AP, Pinheiro PS, et al. Cancer statistics for the US Hispanic/Latino population, 2021. *CA A Cancer J Clinicians*. 2021;71(6):466-487. doi:10.3322/caac.21695

9. State Cancer Profiles > Incidence Rates Table. Accessed May 31, 2024. https://statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=06&areatype=county&cancer=018&race=05&sex=0&age=001&stage=999&type=incd&sortVariableName=rate&sortOrder=asc#results

10. Choi H, Kim HJ, Yang J, et al. Acetylation changes tau interactome to degrade tau in Alzheimer's disease animal and organoid models. *Aging Cell*. 2020;19(1). doi:10.1111/acel.13081

11. Velasco-Mondragon E, Jimenez A, Palladino-Davis AG, Davis D, Escamilla-Cejudo JA. Hispanic health in the USA: a scoping review of the literature. *Public Health Rev*. 2016;37(1):31. doi:10.1186/s40985-016-0043-2

12. Sanjeevaiah A, Cheedella N, Hester C, Porembka MR. Gastric Cancer: Recent Molecular Classification Advances, Racial Disparity, and Management Implications. *JOP*. 2018;14(4):217-224. doi:10.1200/JOP.17.00025

13. Blair VR, McLeod M, Carneiro F, et al. Hereditary diffuse gastric cancer: updated clinical practice guidelines. *The Lancet Oncology*. 2020;21(8):e386-e397. doi:10.1016/S1470-2045(20)30219-9

14. Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F. E-Cadherin Alterations in Hereditary Disorders with Emphasis on Hereditary Diffuse Gastric Cancer. In: *Progress in Molecular Biology and Translational Science*. Vol 116. Elsevier; 2013:337-359. doi:10.1016/B978-0-12-394311-8.00015-7

15. Kluijt I, Siemerink EJM, Ausems MGEM, et al. *CDH1* -related hereditary diffuse gastric cancer syndrome: Clinical variations and implications for counseling. *Intl Journal of Cancer*. 2012;131(2):367-376. doi:10.1002/ijc.26398

16. Onitilo AA, Aryal G, Engel JM. Hereditary Diffuse Gastric Cancer: A Family Diagnosis and Treatment. *Clinical Medicine & Research*. 2013;11(1):36-41. doi:10.3121/cmr.2012.1071

17. Shenoy S. CDH1 (E-Cadherin) Mutation and Gastric Cancer: Genetics, Molecular Mechanisms and Guidelines for Management. *CMAR*. 2019;Volume 11:10477-10486. doi:10.2147/CMAR.S208818

18. Mendonsa AM, Na TY, Gumbiner BM. E-cadherin in contact inhibition and cancer. *Oncogene*. 2018;37(35):4769-4780. doi:10.1038/s41388-018-0304-2

19. Shastry BS. SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet*. 2007;52(11):871-880. doi:10.1007/s10038-007-0200-z

20. BRCA Gene Mutations: Cancer Risk and Genetic Testing Fact Sheet - NCI. Published November 25, 2020. Accessed May 31, 2024. https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet

21. The All of Us Research Program Genomics Investigators, Manuscript Writing Group, Bick AG, et al. Genomic data in the All of Us Research Program. *Nature*. Published online February 19, 2024. doi:10.1038/s41586-023-06957-x

22. The UniProt Consortium, Bateman A, Martin MJ, et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. 2023;51(D1):D523-D531. doi:10.1093/nar/gkac1052

APPENDIX

Python Code Snippets for Matrix Table Generation

```
#Cell 1
# Assuming 'survey_datetime' column exists and indicates the survey
completion time
latest_survey_df = dataset_93563537_survey_df.sort_values(by=['person_id',
'survey_datetime'], ascending=[True, False]) \

.drop_duplicates(subset=['person_id'])
```

```
#Cell 2
#This will merge the three dataframes with relevant information (survey,
zip/socioeconomic, and person)
# Merge the processed DataFrames
merged_df = dataset_93563537_person_df.merge(latest_survey_df,
on='person_id', how='outer') \

.merge(dataset_93563537_zip_code_socioeconomic_df, on='person_id',
how='outer')
```

```
#Cell 3
#Datatypes conversion so it matches the data in the matrixtable
# Convert datetime columns to strings
for col in merged_df.select_dtypes(include=['datetime64[ns, UTC]']).columns:
    merged_df[col] = merged_df[col].dt.strftime('%Y-%m-%d %H:%M:%S')

# Ensure all columns are in a compatible format
for col in merged_df.columns:
    if merged_df[col].apply(type).nunique() > 1:  # Mixed types in the column
        merged_df[col] = merged_df[col].astype(str)
```

```
#Cell 4
#Convert this to a matrixtable
ht_combined = hl.Table.from_pandas(merged_df, key='person_id')
```

```
#Cell 5
#Make sure the matrixtable has all of the columns we're looking for
ht_combined.describe()
'''
----------------------------------------
Global fields:
    None
----------------------------------------
Row fields:
    'person_id': int32
    'gender_concept_id': int32
    'gender': str
    'date_of_birth': str
    'race_concept_id': int32
    'race': str
    'ethnicity_concept_id': int32
    'ethnicity': str
```

```
    'sex_at_birth_concept_id': int32
    'sex_at_birth': str
    'survey_datetime': str
    'survey': str
    'question_concept_id': int32
    'question': str
    'answer_concept_id': str
    'answer': str
    'survey_version_concept_id': str
    'survey_version_name': str
    'observation_datetime': str
    'zip_code': str
    'assisted_income': float64
    'high_school_education': float64
    'median_income': float64
    'no_health_insurance': float64
    'poverty': float64
    'vacant_housing': float64
    'deprivation_index': float64
    'american_community_survey_year': int32
----------------------------------------
Key: ['person_id']
----------------------------------------

'''
```

```
#Cell 6
#enrich the genotypes matrixtable with the demographics/dataframe matrixtable
mt_enriched =
mt_93563537.annotate_cols(demographics=ht_combined[hl.int32(mt_93563537.s)])
mt_enriched.describe()

'''
----------------------------------------
Global fields:
    None
----------------------------------------
Column fields:
    's': str
    'demographics': struct {
        gender_concept_id: int32,
        gender: str,
        date_of_birth: str,
        race_concept_id: int32,
        race: str,
        ethnicity_concept_id: int32,
        ethnicity: str,
        sex_at_birth_concept_id: int32,
        sex_at_birth: str,
        survey_datetime: str,
        survey: str,
        question_concept_id: int32,
        question: str,
        answer_concept_id: str,
        answer: str,
        survey_version_concept_id: str,
```

```
        survey_version_name: str,
        observation_datetime: str,
        zip_code: str,
        assisted_income: float64,
        high_school_education: float64,
        median_income: float64,
        no_health_insurance: float64,
        poverty: float64,
        vacant_housing: float64,
        deprivation_index: float64,
        american_community_survey_year: int32
    }
----------------------------------------
Row fields:
    'locus': locus<GRCh38>
    'alleles': array<str>
    'rsid': str
    'qual': float64
    'filters': set<str>
    'info': struct {
        AC: array<int32>,
        AF: array<float64>,
        AN: int32,
        AS_QUALapprox: str,
        AS_VQSLOD: array<str>,
        AS_YNG: array<str>,
        QUALapprox: int32
    }
----------------------------------------
Entry fields:
    'AD': array<int32>
    'FT': str
    'GQ': int32
    'GT': call
    'RGQ': int32
----------------------------------------
Column key: ['s']
Row key: ['locus', 'alleles']
----------------------------------------
This is what we want!!

'''
```

```
#Cell 7
#Filter and create a new matrix table (chr_enriched_mt) with only CDH1 from
67,000,000 to 68,500,000
chromosome = 'chr16'
start_position = 68700000
end_position = 68850000

# Filter variants within the specified region
chr_enriched_mt = mt_enriched.filter_rows(
    (mt_enriched.locus.contig == chromosome) &
    (mt_enriched.locus.position >= start_position) &
    (mt_enriched.locus.position <= end_position)
)
```

```
#Cell 8
#Count the number of genetic variants that are present in this table:
num_rows = mt_enriched.count_rows()
num_rows2 = chr_enriched_mt.count_rows()
print(f"Number of rows (variants) in the MatrixTable before filtering:
{num_rows}")
print(f"Number of rows (variants) in the MatrixTable after filtering:
{num_rows2}")
```