Spring 2017

Identification of Clinically Relevant Regions of the Human $CDH1$ Gene for Determining Predisposition to Hereditary Diffuse Gastric Cancer

James Stephenson  
*University of North Georgia, jdstep3558@ung.edu*

Follow this and additional works at: [https://digitalcommons.northgeorgia.edu/honors_theses](https://digitalcommons.northgeorgia.edu/honors_theses)  
Part of the Biology Commons

Recommended Citation
[https://digitalcommons.northgeorgia.edu/honors_theses/20](https://digitalcommons.northgeorgia.edu/honors_theses/20)

This Honors Thesis is brought to you for free and open access by the Honors Program at Nighthawks Open Institutional Repository. It has been accepted for inclusion in Honors Theses by an authorized administrator of Nighthawks Open Institutional Repository.
Identification of Clinically Relevant Regions of the Human CDH1 Gene for Determining Predisposition to Hereditary Diffuse Gastric Cancer

A Thesis Submitted to
the faculty of the University of North Georgia
In Partial Fulfillment
Of the Requirements for the Degree
Bachelor of Science in Biology
With Honors

James Stephenson
Spring 2017
Accepted by the Honors Faculty

of the University of North Georgia

in partial fulfillment of the requirements for the title of

Honors Program Graduate

Thesis Committee:

Holly Carpenter

Thesis Chair

Ryan A. Smith

Committee Member

Nancy Pelkon

Committee Member

Jana L. H.

Committee Member

Honors Program Director
Introduction

Genetic testing has opened new doors into understanding a patient’s future health. The human genome is extremely vast and complicated. Testing an entire genome for the innumerable genetic variants that have a bearing on the development of cancer is costly both in time\(^25\) and efficiency. Testing for just a single gene results in seeing a whole group of variations, which may or may not be pathogenic\(^11\). Pathogenic, in this case, is defined as a variant known to cause or increase risk of a disease, the degree of which is determined by how co-segregated it is with the disease\(^16,17\). One of the genes that is examined when looking at hereditary diffuse gastric cancer (HDGC) often is Cadherin-1, or CDH1.

The human CDH1 gene codes for epithelial cadherin, or E-cadherin. The E-cadherin family are transmembrane proteins involved in signaling pathways within the cell, cell maturation, cell movement, and cellular adhesion in the epithelial tissue in many organs. Due to the protein’s functions, the CHD1 gene is classified as a tumor suppressor gene, meaning it prevents cells from undergoing unregulated rapid division, which results in tumor formation. In the case of CDH1, the formation of tumors is usually the result of the loss of contact inhibition. A variation in either the gene or a regulatory factor, such as micro RNA\(^32\), can result in a change of the gene’s expression, resulting in the absence or truncation of a protein. When the gene that codes for E-cadherin contains a pathogenic variant, researchers have found associations in genotype with an increased risk of lobular breast cancers in females\(^15\) and hereditary diffuse gastric cancer in both sexes\(^34\). Cancer is a multigenic disease\(^14\), meaning it requires multiple genetic abnormalities to occur. This is
why a variant does not guarantee that an individual will develop cancer, only increasing
the likelihood of it developing.

The CDH1 gene is located on chromosome 16 and contains just over 98,000 base
pairs\textsuperscript{19}. Each active region in a gene codes for amino acids that are used to build proteins
that have specific functions in the body\textsuperscript{10}. Being able to distinguish variants in these regions
can show us how the protein is altered and the result of a modification. From that
information, a variant can be classified as pathogenic or not. With such a large number of
base pairs and variants in each, the question becomes, which regions of the CDH1 gene are
clinically relevant?

\textit{Architecture of CDH1}

The CDH1 gene is made up of five distinct regions\textsuperscript{5,33} (Figure 2). Each region plays
a different role in how E-cadherin is formed. The first region codes for the signaling peptide
that E-cadherin 1 uses as a part of pathways involving contact inhibition through adherens
junctions\textsuperscript{8}. This region is comprised of exon one and partially exon two, and it is the second
smallest region.

The next region of CDH1 is the precursor region. It is comprised of the remaining
portion of exon two, exon three, and part of four. The coding in this region is responsible
for building the precursor protein for E-cadherin. This precursor will remain inactive until
it has undergone cleavage and other posttranscriptional modifications\textsuperscript{26}. It remains inactive
due to the risk of detrimental morphogenesis\textsuperscript{9} of the tissue due to overabundant E-cadherin.
Changes in this region can result in the precursor being truncated, incorrectly cleaved, or
absent\textsuperscript{26}. These changes would result in E-cadherin being unable to adhere and form adherens junctions with other cells’ E-cadherins.

The largest region of \textit{CDH1} is the extracellular domain. It is made up of the latter portion of exon four and exons five through thirteen. The extracellular domain of the E-cadherin protein lies outside of the cell to interact with the extracellular domains of other cells’ E-cadherin to adhere to each other and form adheren junctions\textsuperscript{8}. The Ca\textsuperscript{2+} pockets that mediate e-cadherin’s adhesion are highly conserved\textsuperscript{28}, meaning that mutations to this region are very detrimental. Changes to this region would result in a change in E-cadherin adhesive abilities and impaired contact inhibition of the cells. If cellular adhesion is lost, the cancer cells are more likely to undergo metastasis due to their new ability to easily break off from one another.
Figure 1. Ribbon structure of two E-cadherin extracellular domains interacting with each other, mediated by calcium ions\textsuperscript{27}.

Following the extracellular domain is the transmembrane region. It is the shortest region, being only comprised of parts of exons thirteen and fourteen. This region is responsible for connecting the extracellular domain of the E-cadherin protein with the cytoplasmic region across a cell’s membrane. Any changes to this region will inhibit the structure’s ability to hold the extracellular domain and cytoplasmic regions together, thus not allowing e-cadherin to be attached to the cell or actin cytoskeleton. Given its size, role in adherens junctions and highly conserved nature, it is expected this region will contain the most pathogenic variants.

Figure 2. Illustration of the extracellular domain forming an adheren junction (AJ) with another E-cadherin 1, and the cytoplasmic anchor binding $\alpha$-catenin (green square) to attach to the actin cytoskeleton and the binding of $\beta$-catenin (blue circle) to the cytoplasmic anchor\textsuperscript{28}. 

The final region of \textit{CDH1} is the E-cadherin cytoplasmic domain that anchors the cadherin to the parent cell through association with the actin cytoskeleton via binding to \( \alpha \)-catenin. The cytoplasmic region is also responsible for binding \( \beta \)-catenin. It is comprised of the remaining exons, 14 through 16. Variations in this region can result in it being truncated and reduce its ability to anchor the protein to the cell’s cytoskeleton. These variants may also cause more \( \beta \)-catenin to be released and enter the nucleus causing a transactivation of genes associated with cellular division, such as \textit{Myc}.

Figure 3. Functional regions of the CDH1 gene and the exons that are contained in each\textsuperscript{5}. Figure is not made to scale.

\textit{Clinical Relevance of CDH1}

Variations in \textit{CDH1} are known to be associated with increased risks for gastric cancers and lobular breast cancers for years\textsuperscript{33}. In a 2015 study by Wei Zeng, \textit{et al.}, the researchers confirmed that \textit{CDH1} had clinicopathological significance. This conclusion was reached by studying methylation levels of \textit{CDH1} in cancerous and non-cancerous cells. Cancerous cells had significantly higher hypermethylation levels, meaning more of
the gene was inactivated, resulting in an interference in E-cadherin 1’s functions. Cells whose cadherins were not functioning properly exhibited malignancy.

Variants in \( CDH1 \) are considered autosomal dominant due to the fact only one copy of the variant is needed to be pathogenic. Heterozygosity is rarely lost. The downregulation of the second, normal allele is thought to be caused by hypermethylation of that allele. Downregulation of E-cadherin is also seen in malignant cells when the epithelial growth factor receptor, to which E-cadherin is coupled, is stimulated by epithelial growth factor or TGF-\( \alpha \).

E-cadherin 1 is also seen to inhibit many proteins involved DNA replication and cell proliferation, such as Geminin and Cdc6. The loss of E-cadherin 1 would result in a loss of inhibition of these proteins, causing the cell to have an increased rate of division.

**Importance of Truncation**

Protein truncation has been widely documented in proteogenomic studies in cancer in a wide variety of proteins. Truncating a protein means to shorten a protein by removing amino acids starting at either end. While studying TP53, another tumor suppressor, Marlon Lindenbergh-van der Plas et al., suggested that truncation can be used as a strong prognosticator for determining if a variant is pathogenic or not.

**Demographics**

Gastric cancers are the third leading cause of cancer-related deaths worldwide. Two-thirds of these come from eastern Asia, eastern Europe, and South America. The case fatality rate in these regions is very high at 78% and only drops slightly to 65% in the
industrialized world. This distribution may be attributed to *H. pylori*, a causative agent of cancer.

_Difference Between Hereditary Diffuse Gastric Cancer and Familial Diffuse Gastric Cancer_

_Familial diffuse gastric cancer (FDGC) is extremely similar to HDGC. The main difference between the two is that HDGC is purely genetic. The development of FDGC is dependent on both environmental and genetic factors_. It is included in this study due to its relation to HDGC and its genetic component.

**Methods**

Genetic variants of the CHD1 gene were gathered from the ClinVar database. They were filtered by looking for variants in _CDH1_ associated with HDGC or FDGC and labeled as pathogenic, likely pathogenic, or conflicting. In addition to looking at the literature, the number of submissions, most recent submission, and summaries given were examined. Many submissions from Ambry Genetics lacked literature, but they gave a detailed chart of their criteria for determining if a variant was pathogenic.

When I read the literature, I focused on looking for two main criteria. First, the variant had to be present in multiple generations. In the generations where the variant was present, individuals diagnosed with HDGC must have been about age 50 or below. Once an individual has reached the age of 50, the chances of developing gastric cancer due to a genetic cause decreases. This means the cause of cancer would be more likely arise from a somatic mutation or DNA damage once an individual has reached age 50. One factor that
strengthened my determination of pathogenicity was if multiple individuals within a
generation were diagnosed with gastric cancer\textsuperscript{3}. Once these factors were evaluated, how
the variant affected \textit{CDH1} and its product was examined.

Once the variants were determined to be pathogenic, their location in terms of exon
or intron was noted. If the exon or intron location was not present in the literature, the
variants’ base location on chromosome 16 was input into the UCSC Genome Browser and
their locations were determined from there.

\textit{Non-pathogenic Variants}

Not all of the variants listed by ClinVar are expected to be pathogenic. Variants
that have only been seen in a single individual were designated as non-pathogenic in terms
of HDGC because it is unknown if the variant is hereditary. If a variant on the list was not
directly listed as being associated with HDGC or FDGC, it was also marked as non-
pathogenic due to it being more strongly associated with lobular breast cancer or not
directly related to HDGC at all. A variant that was also not strongly segregated with HDGC
was also marked as non-pathogenic. It has been seen in multiple unrelated individuals and
affects protein function and mRNA splicing, but its correlation to cancer is weak\textsuperscript{23}.

\textit{Example of Pathogenic Variant}

The variant, NM_004360.4(CDH1):c.1003C>T (p.Arg335Ter), has been identified
as pathogenic by multiple submitters on ClinVar. This variant results in a change from
cytosine to thymine at coding locus 1003, which results in a change from Arginine to a
premature stop codon (Ter) at codon 335. This will result in loss of E-cadherin function
due to either protein truncation or nonsense-mediated mRNA decay\textsuperscript{21}.
Results

Of the 68 variants listed by the ClinVar database, 51 variants were deemed as pathogenic (Figure 3). Exons contained more pathogenic variations with 41, and introns only had 10 variants. Exon size did not play a role in the number of pathogenic variants they contained. This is shown by the largest exon, 16, containing only one pathogenic variant, while the shortest exon, 2, had seven pathogenic variants. Exon 5 contained no reported pathogenic variants. Most pathogenic variants were reported or suggested to truncate E-cadherin 1.

By region, the extracellular domain contained the most pathogenic variants, with a total of approximately 30 variants from both exons and introns. The signaling and precursor regions both contain approximately eight pathogenic variants each. The transmembrane region contains approximately seven pathogenic regions. The cytoplasmic anchoring region only contains about four pathogenic variants.

Figure 4. Scaled diagram of the CDH1 gene by exon. The introns are not to scale due to their immense size in comparison to the exons. The number of pathogenic variants for exons are listed below the gene and the pathogenic variants for the intron are listed above the gene.
Importance of intron variants

Introns do not code for a protein; however, they play an important role in splicing. Variants in these regions alter splice donor sites. This can cause either an abnormal protein produced, or nonsense-mediated mRNA decay to occur\(^2^2\).

Mass Deletions

Mass deletions were omitted from my results. Cancers resulting from mass deletions may be due to the deletions affecting other genes and cannot be solely attributed to the deletion in \textit{CDH1} resulting in the cancer. Therefore, the deletion itself may be clinically relevant, but just not solely in terms of \textit{CDH1}. An example of this is seen in variant nsv513771 where exons 1 and 2 were deleted, but so was the entire \textit{CDH3} gene\(^2^4,2^5\) which is involved in loss of heterozygosity events associated with several cancers that \textit{CDH1} is associated with\(^7\).

Discussion

As expected, the extracellular domain contained the most pathogenic variants. Containing over half of the pathogenic variants suggest that the extracellular region of the \textit{CDH1} is sensitive to changes in amino acid sequence. The same can be inferred about exon 2 of the gene given that it contains the most variants out of the individual exons. The coding sequences for the end of the signaling region to the beginning of the precursor region would be sensitive to changes in their amino acid sequences as well.
Clinical Use

As direct-to-consumer (DTC) genetic testing has become more accessible, the legality of the “diagnosis” an individual can perceive from their results has been called into question. Currently, genetic testing companies are not legally allowed to provide medical advice with the raw data collected from the results of the individual. It is up to a medical professional to interpret these results.

The field of genetic testing is relatively new in the clinical world and not all medical professionals are well versed in interpreting genetic results. Studies like mine can aid doctors in diagnosing patients with genetic conditions and diseases by providing them with a map of which areas are sensitive to variation.

Limitations

The variants in this study were drawn from an expansive, but singular database. It relies on submissions to compile the variants. Variants included in literature outside the database were not able to be included as a result.

Further Research

Research into literature outside the database could greatly expand the content of this list as well as improve its accuracy. Encouraging submissions to databases like ClinVar would immensely streamline research in the future. One step further would be to consolidate all data into a singular database.

Ambry Genetics only provided results of clinical data, so research into any submissions without literature would strengthen the position on if the variants are pathogenic
or not. Any variant with a single submitter would benefit from having more submissions, re-evaluating their conclusions.
Bibliography


7. CDH3 cadherin 3 [Homo sapiens (human)] - Gene - NCBI. National Center for Biotechnology Information. [accessed 2017 May 3].


