Estimating the Prevalence of HHT Using Variant Effect Predictions

by

Carter White

Bachelor of Science, Michigan State University, 2022

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Public Health

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This essay is submitted

by

Carter White

on

December 15, 2023

and approved by

Essay Advisor: Beth Roman, PhD, Primary Faculty, Human Genetics, School of Public Health, University of Pittsburgh

Essay Reader: Jeremy Martinson, PhD, Primary Faculty, Infectious Disease and Microbiology, School of Public Health, University of Pittsburgh

Essay Reader: Anthony Anzell, PhD, Postdoctoral Fellow, Human Genetics, School of Public Health, University of Pittsburgh

Estimating the Prevalence of HHT Using Variant Effect Predictions

Carter White, MPH

University of Pittsburgh, 2023

Abstract

Hereditary Hemorrhagic Telangiectasia (HHT) is currently classified as a rare genetic condition that is characterized by potentially fatal blood vessel malformations. Upwards of 90% of HHT cases are caused by a pathogenic genetic variant in the *ENG* or *ACVRL1* gene. Current estimates, based on case counting methods, calculate the prevalence of HHT to be between 1 in 5,000 and 1 in 10,000 people worldwide (0.01% - 0.02%). Due to variable HHT symptom presentation and severity, along with physician unfamiliarity, HHT is likely underdiagnosed by providers and therefore past prevalence measures may be inaccurate. In this study I (1) gathered variant data from gnomAD, (2) predicted variant pathogenicity using a variant effect predictor (VEP) from Ensembl, and (3) calculated the prevalence of HHT pathogenic variants at different levels of restrictiveness using allele frequencies. Without excluding any variants from the VEP software, a prevalence of 2.75% was calculated. After adjusting for homozygous variants, variants with large individual allele frequencies, and variants reliably labeled benign via clinical observations, a prevalence of 0.3203% was calculated. Finally, I adjusted the calculation for all clinically benign variant annotations. The calculated prevalence of HHT is then 0.2945%, 14.7-fold higher than currently accepted values. While there are notable limitations to the variant database, VEP, and datapoints used, an increased prevalence value holds great public health significance for the HHT and medical communities. This increased prevalence estimate suggests that HHT may indeed be more common in the population than currently reported.

Table of Contents

List of Tables

List of Figures

Preface

I would like to acknowledge the kind guidance and contributions from Cure HHT, my essay readers, and my family and friends.

1.0 Introduction

Prevalence of a given disease is defined as the number of people that are living with the condition at any given time. Often, the prevalence of common diseases is measured by asking a sample of people whether they are affected by the specified condition. Other methods of prevalence calculation include searching medical records for disease diagnoses and using insurance codes to count the number of disease cases in a population. However, calculating the prevalence of rare diseases is significantly more challenging. In rare diseases, very large sample populations are necessary to identify even few cases of disease. The goal of this work is to use publicly available genomic sequence data to estimate the true prevalence of the inherited blood vessel disorder, hereditary hemorrhagic telangiectasia (HHT), which is thought to be underdiagnosed.

1.1 Background

1.1.1 Hereditary Hemorrhagic Telangiectasia

Hereditary hemorrhagic telangiectasia (HHT), or Osler-Weber-Rendu syndrome, is a blood vessel disorder that causes vascular malformations. Symptoms of HHT include frequent and/or severe nosebleeds (epistaxis); enlarged capillaries on the surface of the skin, in the nasal mucosa, and/or within the GI tract (telangiectasias); and direct connections between arteries and veins (arteriovenous malformations-AVMs) that can occur most problematically in the brain, lungs, and

liver[1,2]. In unscreened patients, HHT is associated with a decreased lifespan. However, HHT does not significantly reduce lifespan when patients are diagnosed, screened, monitored, and/or treated[3]. HHT is caused by genetic variants that are passed down in an autosomal dominant fashion. There is increasing evidence that HHT pathogenesis requires a "somatic second hit" mechanism to "activate" the condition, but autosomal dominant inheritance is still widely accepted and appropriate for prevalence estimates[4,5]. Upwards of 90% of HHT cases arise from variants in one of two genes, *ENG* or *ACVRL1 (ACVRL1* is sometimes referred to as *ALK1*)[4-7]. Fully functioning *ENG* and *ACVRL1* genes are very important in the formation of new blood vessels, and loss of function mutations lead to HHT. The *ENG* gene encodes the endoglin protein and the *ACVRL1* gene encodes the activin receptor-like kinase 1 protein. The protein products of *ENG* and *ACVRL1* can form a complex with each other that aids in the development and specialization of blood vessels[8]. HHT caused by variants in *ENG* is referred to as HHT1 (OMIM 187300) and HHT caused by *ACVRL1* variants is referred to as HHT2 (OMIM 600376). In the clinical setting, HHT1 and HHT2 have been calculated to have a similar prevalence[9-11]. It is also of note that hundreds of variants have been found in *ENG* and *ACVRL1* genes and pathogenic variants have been found across all gene exons[12]. Furthermore, there is no singular variant in either gene that currently stands out as the prevailing cause of HHT. Rather, many different variants within the genes have relatively low prevenances individually, which combine to reveal the total HHT prevalence. There can be other etiologies of HHT such as variants in the *SMAD4* gene, but those are excluded from this study due to their low contribution to overall HHT prevalence and their involvement in non-HHT conditions[13].

1.1.2 Prevalence of HHT

The currently accepted prevalence of HHT in most populations is between 1 in 10,000 and 1 in 5,000 individuals (0.01%-0.02%)[14-19]. However, most HHT physicians and researchers believe that this prevalence estimate is low due to factors such as variable age of onset, variable expressivity, and a lack of awareness of the disease within the medical community and population as a whole[20]. One of the largest HHT nonprofit organizations, Cure HHT, is leading the charge for improved HHT diagnosis and treatment, and they claim that "9 out of 10 HHT patients are undiagnosed." This claim is not backed by cited scientific evidence, but instead attempts to claim merit based on assumptions made about patterns in clinical diagnosis. The most common and visible symptoms for HHT, epistaxis and telangiectasias, can be overlooked by both patients and providers in a typical clinical visit. The potential dismissal of these symptoms and a lack of provider education on HHT lead to a long lag time between symptom onset and diagnosis, further supporting the idea that HHT is underdiagnosed[21]. If Cure HHT's underdiagnosis claim is to be true, this implies a 10-fold increase in disease prevalence to 0.2%.

Current methods of HHT prevalence estimation are based on clinical diagnosis, insurance billing codes, and/or other methods of case-counting. Use of case-counting methods in prevalence estimations of rare diseases are prone to significant bias. The case-counting methodology is based on a few important assumptions, two of which are critically relevant for calculating HHT's prevalence. The first assumption is that the condition in question is <u>easily</u> identified/diagnosed. HHT is in fact very easily diagnosed in adults. Diagnosis is based on consensus criteria, known as the "Curaçao criteria," developed by HHT experts and community members[22-24]. The Curacao criteria defines HHT through (1) epistaxis, (2) telangiectasias, (3) visceral lesions, and (4) a family history of HHT (through a first degree relative). An individual with at least three of the four

previously listed HHT diagnosis criteria is diagnosed as having "definite" HHT. An individual fulfilling two criteria is diagnosed with as having "possible or suspected" HHT, while those with less than two criteria are "unlikely" to have HHT[24]. The second critical assumption used in case-counting estimates is that the studied condition is reliably diagnosed. Even if the Curaçao criteria make HHT diagnosis relatively simple for providers, this does not mean that providers are reliably/consistently applying the criteria to diagnose patients. HHT is not well-covered in medical school curricula and providers are often unaware of HHT, leading to missed diagnoses.

To accurately estimate the prevalence of a rare disease that is suspected to be underdiagnosed, it may be best to bypass a reliance on the clinical diagnosis process. With recent innovations and advancements in genetics, the prevalence of diseases associated with known genes can be estimated or predicted using openly accessible databases and software programs. As stated previously, 90% of HHT cases are caused by genetic variants in just two genes that act in an autosomal dominant fashion. Additionally, HHT has a penetrance around 100% in adults and disease-causing variants can be found in any gene exon, or protein-coding region, within *ENG* or *ACVRL1*[4,20]. Inherited single-gene disorders in which a few genes cause 90% of disease cases that have 100% penetrance, such as HHT, provide an excellent opportunity to estimate disease prevalence using genetic data. If genetic information from an unbiased sample is queried, causal gene variants can be putatively identified, and variant frequencies within a population can be used to calculate disease prevalence.

1.1.3 Allele Frequencies and ClinVar

Because genetic variant analysis does not consider counted cases of diagnosed disease, a method to control in-silico miscalculations and outliers is required. Contrary to the case-counting

methods of past calculations, prevalence estimations that use genetic variants rely on population data such as allele counts and allele frequencies within a sample. Allele frequencies are therefore susceptible to bias based on the makeup and size of the population from which the data is derived. When studying a rare disease, exclusion of a certain variant based on allele frequency alone can be used to control an illogical overestimation of prevalence[25]. For example, a single variant that has a very large or disproportionate allele frequency can be presumed to be benign and eliminated from the disease prevalence calculation. Another way in which to control an in-silico analysis of disease prevalence is to attempt to control errors by using knowledge of clinical variant behavior. Variants that have been studied in a lab or seen in the clinical setting are the most reliable source of pathogenicity prediction and can help adjust generated prevalence estimations. For this study, verification of clinical patterns and behavior of genetic variants came through the use of ClinVar. ClinVar is an open access, NIH funded, NCBI-run online resource that allows for the assignment of pathogenicity labels to gene variants based on information submitted by clinicians[26]. ClinVar is a convenient resource because it is already integrated into the variant retrieval database used in this study, GnomAD. Other databases for clinically observed HHT-related variants including those in the *ACVRL1* and *ENG* genes are available through ARUP laboratories {HYPERLINK "https://arup.utah.edu/database/index.php"}. Unfortunately, it is difficult to match variants within the ARUP laboratory databases to the variants used in this study due to the differences in reference sequence and variant identifiers used. Therefore, clinical patterns were identified in this study based on data from *ACVRL1* and *ENG* searches within the ClinVar database.

1.1.4 GnomAD

In this study, genetic variant data were gathered from the large public database called "gnomAD." GnomAD is a genetic database that consolidates data from a wide variety of sources. The data in gnomAD come in large part from the 1,000 Genomes Project and the Exome Aggregation Consortium, along with direct contributions from more than 140 principal investigators and genome data from over 60 studies[27,28]. This database can claim relatively low levels of sampling bias for HHT because most data are gathered from case-control studies of common complex diseases. Additionally, quality control measures like removal of close relative data limits allele frequency inflation. GnomAD is one of the largest publicly available genetic datasets and has been continually updated and refined since its publication. There are currently two versions of gnomAD that are available for use, version 2.1.1 (v2.1.1) and version 3.1.1 (v3.1.1). V2.1.1 is aligned with the human genome assembly GRCh37/hg19 and contains data for a total of 125,748 exomes and 15,708 whole-genomes. V3.1.1 is aligned with the GRCh38 human genome assembly and includes 76,156 whole-genomes at the expense of v2.1.1's exome sequence data.

1.1.5 VEP: SIFT, PolyPhen, CADD

Verifying and predicting the pathogenicity of unknown genetic variants is a time consuming, complex, and imperfect process. The gold standard of such variant effect prediction is clinical observation and laboratory study, where genetic variants can be specifically studied in controlled environments. Unfortunately, laboratory pathogenicity confirmation of thousands of variants is not possible, realistic, or reasonable in modern science. For this reason, scientists must

rely on in-silico prediction models and algorithms to predict the pathogenicity of unknown or understudied genetic variants. Ensembl is a publicly available webpage that contains a user-friendly software interface, providing easy access to in-silico analyses tools for genetic variants. Formatted genetic variants entered into a variant effect predictor (VEP), provided by Ensembl, and can then be subject to pathogenicity predictions and scores[29]. The VEP software uses preloaded programs such as Sorting Intolerant From Tolerant (SIFT), PolyPhen, and Combined Annotation-Dependent Depletion (CADD) to provide pathogenicity predictions and scores for variants[29]. These three programs each predict the pathogenicity of variants in different ways. The SIFT and PolyPhen predictions methods can be similar/overlapping in their analysis type and are limited to producing pathogenicity predictions for missense variants only. Alternatively, the CADD prediction method is unique from SIFT/PolyPhen as it can provide scores for all single nucleotide polymorphisms (SNPs). When SIFT and PolyPhen are combined with CADD, the three methods can provide support for or against the pathogenic properties of almost any SNP. SIFT is an algorithm that was developed in the very early 2000s that uses the amino acid sequence of proteins to impute the functionality of a gene's protein product[30,31]. PolyPhen is a separate algorithm from SIFT, however, it works in a similar way as amino acid sequence changes caused by a variant are analyzed, and the chemical properties of the changed amino acid are considered. SIFT and PolyPhen also account for the conservation of a given amino acid across species to infer the "importance" of the amino acid within the final protein product. PolyPhen differs from SIFT in that PolyPhen considers both sequence and structural differences between the wild-type and variant proteins rather than just the sequence of amino acids[32]. On the other hand, CADD uses a completely different type of variant analysis that overlaps with neither SIFT nor PolyPhen. CADD uses machine learning programming that is trained to eliminate its susceptibility to ascertainment

bias[33,34]. Similarly to SIFT and PolyPhen, the CADD program is able to analyze the deleteriousness of missense protein coding variants. Unlike SIFT and PolyPhen, CADD can also calculate pathogenicity scores for many other types of genetic variants including synonymous variants, variants that affect splicing, transcription termination, frameshifts, etc. The Ensembl VEP software not only includes CADD score calculating capabilities, but it also provides a more easily interpretable version of the CADD score called, "CADD_PHRED". CADD_PHRED scores represent a logarithmic interpretation of the variant's pathogenicity, where a score of 15.0 translates to a 95% chance of pathogenicity, a score of 20.0 translates to 99%, and so on[34].

1.2 Aims

The first aim of this study was to gather *ENG* and *ACVRL1* genetic variants from the gnomAD database, run them through the Ensembl-VEP, and then use the predictions to categorize the variants as pathogenic or benign. The first aim involved a level of fact checking where the ClinVar database was used to verify whether certain variants have been correctly assigned "benign" or "pathogenic" designations through clinical submissions. The second aim of this study was to use the allele frequency data for the pathogenic-labeled variants to estimate the prevalence of HHT. Allele frequencies for each studied variant are presented in the gnomAD database as the number of variant alleles divided by the total number of alleles recorded in the individual study from which the variant came[28]. As mentioned previously, gnomAD is an aggregation database that is composed of many different studies. Therefore, variant allele frequencies were calculated in the original study the variant was recorded in. Finally, the third aim of the study was to critically

analyze the effectiveness and feasibility of using a genetic variant analysis methodology for the calculation of HHT prevalence and its potential for use in other rare inherited diseases.

1.3 Importance

The importance of this study is two-fold: To investigate the methodology of using genetic variant data to estimate the prevalence of a single-gene disease, and to provide an unbiased prevalence estimation for HHT. As previously described, HHT is likely to be underrecognized and underdiagnosed, meaning case-counting methods of prevalence estimates may not be the most effective or appropriate way to calculate HHT prevalence. The unbiased methods used in this study, which do not rely so strictly on clinical diagnosis, are relatively new and have not been extensively studied. There are indeed some examples of prevalence estimates for other diseases using a similar type of genetic variant analyses[35-39]. These studies focus on different diseases that do not share many features with HHT. For example, the studies follow recessively inherited disease and use Hardy-Weinberg assumptions to calculate prevalence. Notably, all studies cited calculated a higher prevalence for their respective diseases than what is found using other prevalence estimating methods, such as case-counting. Prevalence estimates of HHT using a genetic variant analysis will be useful to the greater scientific and epidemiological communities. If the use of genetic variant data provides a dramatically different prevalence calculation to the one that is currently used, this may call for a more critical or skeptical inspection of the genetic variant analysis methodology. On the other hand, if the original HHT prevalence calculations of ~1 in 5,000 are confirmed or supported by this method of calculation, then there can be more confidence given to those values and there may be a decreased concern about HHT underdiagnosis.

Moreover, there is a critical need for an accurate calculation of HHT prevalence within the medical and HHT communities. HHT is currently categorized as a "rare" disorder by the US's Orphan Drug Act due to current prevalence estimates of between 1 in 5,000 and 1 in 10,000 people, or under 200,000 individuals in the US[40]. With what the scientific community knows about HHT and its potential for underdiagnosis, along with the claims by the Cure HHT organization, there is suspicion that HHT has a higher prevalence than the reported 1 in 5,000 individuals. If the prevalence can be reliably and accurately calculated as considerably higher than 0.02%, there is massive potential for a domino effect of changing attitudes. If HHT prevalence is greater than around 3 in 5,000 (0.06%), HHT will no longer be classified as "rare" disease in the US[40]. A reclassification of HHT impacts patients and the public greatly because of an increased awareness, attention, and funding from the scientific community. Perhaps most importantly, with increased awareness about HHT and subsequent improvement of HHT diagnosis, there comes an increase in correctly identified HHT-affected individuals whose lives may be extended and/or improved.

2.0 Methods

2.1 Data Gathering

2.1.1 GnomAD Variant Retrieval

Figure 1 displays a graphical representation of the process used in this study to gather, analyze, and calculate the prevalence of HHT using genetic variant analysis. All genetic variant information and data on *ENG* and *ACVRL1* were gathered through the genome aggregation database (gnomAD). GnomAD v2.1.1 was used and terms "*ENG"* and "*ACVRL1"* were searched separately and independently. For both genes, the gnomAD search was composed of data from exomes, genomes, SNVs, and indels. Unfiltered variants were not included in this analysis, only variants that passed the gnomAD quality control process were included. Quality control was done by gnomAD in large part by use of a random forests model as well as a set threshold value for inbreeding coefficient, DNA read depth, genotype quality, and a minor allele balance for heterozygous genotypes[27]. Variants that did not pass the gnomAD quality control process are excluded by simply leaving the "Filtered Variants" box unchecked on the gnomAD v2.1.1 website. CSV files with all given variant information were downloaded from gnomAD and uploaded into an excel spreadsheet (Supplementary Table 1). The variants were labeled by gnomAD as "missense, frameshift, stop gained, stop lost, stop retained, start lost, splice acceptor, splice donor, splice region, in frame deletion, in frame insertion, synonymous, 3' untranslated region (UTR), or 5'UTR" in the "VEP annotation" column. For the variant analysis tables, information from the "VEP annotation" column was renamed, "gnomAD annotation," to make the source of variant type

labels clear. Intronic variants were also downloaded from gnomAD for both *ENG* and *ACVRL1* genes but were excluded from VEP analysis due to the unreliability of pathogenicity prediction models in non-coding regions of genes.

The variants were separated into five groups for analysis; "missense," "stop gain, frameshift, and start loss" (StopGain/FrameShift/StartLoss), "splice site and in frame deletion" (SpliceSite/InFrameDeletion), "synonymous," and "3'UTR and 5'UTR" (3'UTR/5'UTR). Though many datapoints are given through gnomAD, the only columns relevant to the HHT prevalence calculation were the "VEP [gnomAD] Annotation" for variant type categorization; "ClinVar Clinical Significance" for calculation adjustments; "Allele Frequency" for prevalence calculations; "rsIDs" and "HGVS Consequence" for variant identification; and "homozygote count" for calculation adjustments. No variants were intentionally excluded from the analysis in the initial variant gathering procedure. The only variants from gnomAD that were excluded from the dataset were left out based upon the VEP's inability to provide a pathogenicity prediction for that variant (Supplementary Table 2).

2.1.2 Variant Effect Prediction

For the prediction of variant pathogenicity, all variants were input into the VEP provided by Ensembl. To align variants accurately with gnomAD v2.1.1, the Ensembl website that is aligned with human genome assembly GRCh37 was used {HYPERLINK "https://grch37.ensembl.org/Homo_sapiens/Tools/VEP"}. Variants were analyzed in separate groups, "batches," consistent with gnomAD downloaded variant groups previously described. The variants were input into the VEP using the "rsID" number, provided by gnomAD, to differentiate variants. For the few variants without a listed rsID number, human genome variation society

(HGVS) notation was used (e.g. ACVRL1:c.335A>G). Default query parameters were used to ensure that consistent predictions and scores were given for all variants, with the addition of marking the "CADD" score box (Figure 2). To limit the output size of variant batches, "show one selected consequence" was selected in the "restrict results" section before the program was run. The selected "consequence" was decided by the VEP based on the Ensembl criteria {HYPERLINK "https://grch37.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequence_type_table"}. Variants that produced no prediction output or were analyzed by the Ensembl VEP under the incorrect "consequence," as compared to the gnomAD label, were re-analyzed separately. In the subsequent re-run "show all results" was selected in the "restrict results" section of the VEP query. If still no scores were produced with the correct variant consequence label, HGVS notation was used to input variants. The 16 variants that were not able to produce any VEP result after the previous measures were executed were excluded from this study (Supplementary Table 2). After all results were given, the output was downloaded as a .txt file, opened in excel, and then pasted alongside gnomAD data after verifying that variants were matched to their correct rows. The only columns relevant to this study from the Ensembl VEP were, "uploaded variation," "SIFT," "PolyPhen," "CADD_PHRED," "location," "allele," and "consequence."

2.1.3 Pathogenicity Calculation

To assign pathogenicity to the genetic variants analyzed through the VEP, threshold values for each prediction method were used for simple, binary classification. SIFT and PolyPhen score thresholds were used in accordance with their established and provided guidelines through the Ensembl VEP. Variants for which SIFT produced a score of less than 0.05 and/or were labeled "deleterious" were called pathogenic and those scored greater than 0.05 and/or were labeled

"tolerated" were called benign. Variants for which PolyPhen produced a "benign" prediction and a score of less than 0.445 and were labeled as benign. Variants where PolyPhen produced a "possibly/probably damaging" prediction and had scores greater than 0.445 were called pathogenic. For CADD scores, a CADD_PHRED threshold of 15.0 was used based on accepted guidelines that place given CADD scores in the top 5% of all possible reference genome single nucleotide variants[34]. Variants with a CADD_PHRED score below 15.0 were called benign, and scores above 15.0 were called pathogenic. When identifying variants to use for prevalence calculations, a "pathogenicity" column was added to the variant analysis sheet such that for each variant, an annotation could be added to display how many of the three prediction algorithms classified the variant as pathogenic (Supplementary Table 1). For example, if only a CADD_PHRED score of 16.0 was produced, the variant was called pathogenic and labeled "1/1"; If a variant produced a SIFT score of 0, a PolyPhen score of 0.993, and a CADD PHRED score of 13.0, then it would be labeled "2/3" and only used for prevalence calculations where 2/3 pathogenic predictions was deemed acceptable. Due to the use of three prediction methods, three separate calculations were made for evaluating the prevalence: "1 or more," "Majority," and "All." The "1 or more" category includes all variants where at least 1 of the algorithms predicted pathogenic effect (i.e. a pathogenicity score of 1/3, 2/3, 3/3, 1/2, 2/2, or 1/1). The "Majority" category includes all variants where two or three out of three predictions met the conditions for pathogenicity, or if only two predictions were given, at least one predictor called the variant pathogenic (i.e. a pathogenicity score of 2/3, 3/3, 1/2, 2/2, or 1/1). Finally, the "All" category only includes variants where all the given predictors indicated pathogenicity (i.e. a pathogenicity score of 3/3, 2/2, or 1/1).

2.2 Calculating Prevalence

2.2.1 Raw Calculation

The first raw prevalence of HHT was calculated by adding the allele frequencies of all the variants that were labeled "pathogenic" within their analysis categories. Allele frequencies were given by gnomAD, and pathogenicity labels were derived from the predictions/scores given to variants by the Ensembl VEP, based on the set thresholds previously described. The first allele frequency sum, referred to as "raw calculation" was calculated using all exonic variants obtained from gnomAD with at least one VEP score (Figure 3a).

2.2.2 Adjusted Calculation

The second calculation referred to as, "adjusted calculation," excludes variants for which there was a listed ClinVar clinical significance of "benign" or "likely benign" accompanied by a ClinVar-assigned 2-star reliability rating. Also excluded from the adjusted calculation were variants for which a homozygous variant was identified, or had a listed allele frequency greater than 0.0001 (Figure 3b). A ClinVar two-star reliability score is given according to the ClinVar website, {HYPERLINK "https://www.ncbi.nlm.nih.gov/clinvar/"}, to variants for which more than one clinical submission is received and the submitted variants are reported as "benign" or "likely benign" with no conflicting interpretations. The ClinVar star rating was searched for each variant that was called "benign" or "likely benign" via the gnomAD-produced "ClinVar clinical significance" column. In the one instance where an inconsistent label was observed, the clinical

significance label displayed through gnomAD was used and trusted in all calculations that included this variant for maximized consistency across all variants.

2.2.3 ClinVar Adjusted Calculation

The final prevalence calculation, referred to as, “ClinVar adjusted calculation,” excluded all variants that were excluded from the adjusted calculation plus variants with a “ClinVar clinical significance,” given by gnomAD of, “benign” or “likely benign,” regardless of star-rating (Figure 3c). This is the most conservative prevalence calculation and was used for most interpretations and conclusions.

3.0 Results

Supplementary Table 1 is a full display of all analyzed variants separated by gene and variant type. The table includes all relevant gnomAD and VEP data along with the created "pathogenicity" column.

3.1 Descriptive Data

Table 1 shows a descriptive display of variant data contained within the gnomAD database. The counts represent the total number of variants within a given variant type category as it appeared in gnomAD v2.1.1 at the time of data analysis (08/09/2023). Excluding intronic variants, 528 *ACVRL1* variants (42.4%) and 716 *ENG* variants (57.6%) were represented in the database for a total of 1244 genetic variants. Of the variants in the gnomAD database, VEPs were gathered for 1228 variants. 521 and 707, *ACVRL1* and *ENG* variants respectively. The variants that were not able to produce a VEP score were the only variants excluded from all calculations (Supplementary Table 2). For the purposes of this study, all statistics are measured based on the 1228 variants for which an effect prediction could be produced. Missense variants were the most common variant in both genes representing 49.5% of all variant types, followed by synonymous (29.7%), 3'UTR/5'UTR (10.6%), SpliceSite/InFrameDeletions (8.6%), and StopGain/Frameshift/StartLoss variants (1.6%).

3.2 Raw Calculation Prevalence

The first raw calculation of HHT prevalence is in table 2. The table shows how the data break down by gene (columns) and the restrictiveness of the pathogenicity prediction criteria (rows). The breakdown of variants by variant type and number of variants that contributed to the allele frequency value is also listed within the table. The least restrictive calculation, "1 or more," produces an allele frequency sum of 38.5% whereas the most restrictive estimate, "All," is summed to 2.75%, or around 137 in 5,000 individuals, exceeding the currently accepted HHT prevalence value by a factor of 137.

3.3 Adjusted Calculation Prevalence

Table 3 presents the adjusted HHT prevalence estimates after the pathogenicity predictions were controlled for clinically reliable designation of non-pathogenicity, the presence of homozygous individuals, and/or an allele frequency greater than 0.0001, as described in methods: adjusted calculation. These exclusionary criteria led to the exclusion of 25 additional variants (Supplementary Table 3).

The adjustment excluded six *ACVRL1* missense variants from the raw prevalence calculation. These six *ACVRL1* variants combined to make up 99.2% of the raw, "1 or more" missense calculation. Four of these variants had initially made up 52.4% of the raw "Majority" missense calculation and three of those made up 67.5% of the raw "All" missense calculation. Additionally, one *ACVRL1* 5'UTR variant was excluded from all of the raw calculation levels.

This 5'UTR variant had initially accounted for 95.6% of the raw *ACVRL1* 3'UTR/5'UTR pathogenic allele frequency calculation.

The adjustment step also excluded 18 total *ENG* variants from the calculation. 13 missense variants, three 5'UTR variants, and two 3'UTR variants. The 13 missense variants accounted for 88.5% of the raw "1 or more" missense calculation. 12 of these variants had initially made up 89.0% of the raw "Majority" missense calculations, and four of those made up 92.4% of the "All" missense calculation. The five 3'UTR/5'UTR variants contributed 96.7% of the raw UTR allele frequency for all three levels of prediction.

This adjusted calculation produces an HHT prevalence estimation of 0.663%, 0.520%, and 0.320% for the least to most restrictive pathogenic allele frequency sums. The most conservative, "All" estimate, suggests an HHT prevalence of 16.0 in 5,000 individuals, 16 times higher than the currently reported prevalence.

3.4 ClinVar Adjusted Calculation Prevalence

Table 4 is the final adjustment made to the HHT prevalence calculation. This table contains only the variants for which all prediction algorithms returned a designation of pathogenic, along with a ClinVar clinical significance notation of, "pathogenic," "likely pathogenic," "conflicting interpretations," "variant of unknown significance," or no ClinVar notation was given. When all criteria are met for this adjustment, nine additional variants are excluded (Supplementary Table 4) and the HHT prevalence estimate is calculated to be 0.2945%, or around 14.7 in 5,000. In this most-conservative calculation, the prevalence of HHT is estimated to be increased from currently reported calculations by a factor if around 15.

4.0 Discussion

HHT is a rare genetic disease that impacts the formation of blood vessels. Symptoms vary in presentation and severity with the most impactful lesions being nasal and gastrointestinal telangiectasias and the most life-threatening lesions being AVMs in the brain, lungs, or liver. Sudden bursting of AVMs can lead to serious and fatal outcomes. Current prevalence estimates for HHT range from around 1 in 10,000 to 1 in 5,000 cases worldwide[14-19]. Even though HHT can be easily diagnosed using the Curacao criteria, HHT is thought to be significantly underdiagnosed due to the lack of physician familiarity with HHT and overt symptoms that are not restricted to HHT[2]. Because current prevalence estimates are calculated using case-counting methods, there is a great potential for an underestimation of HHT cases and prevalence. In this study, the known genetic etiology of HHT was used to calculate the prevalence of HHT without reliance on clinical diagnosis.

4.1 Variant Processing

For the most consistency in analysis, variants for *ACVRL1* and *ENG* were gathered at the same timepoint and gnomAD v2.1.1 was used. GnomAD v2.1.1 was chosen for this study simply because it contains the largest variant sample size of the available gnomAD versions. Based on ClinVar data, HHT is largely caused by exonic variants in either *ACVRL1* or *ENG*, so exomes are equally as useful as genomes in terms of sample type used. GnomAD v2.1.1 had significantly more combined exomes and genomes available than v3.1.1. Therefore, the use of v2.1.1 allowed for the

analysis of more data. Larger sample sizes are more likely to be representative of the population and allows for a more reliable ascertainment of rare disease variants such as those involved in HHT.

After all variants were downloaded from gnomAD v2.1.1, they were separated into their respective analysis groups to both ease analysis and better investigate groupings of likely pathogenic variant types. The missense group was the largest group in terms of variant number and pathogenic allele frequency for both *ACVRL1* and *ENG* genes. The next group analyzed was StopGain/FrameShift/StartLoss because of the high likelihood of those variants causing protein loss of function, leading to disease. The three variant types were combined into one group due to the small number of variants within each of the individual variant type groups. Similarly, the SpliceSite/InFrameDeletion group was formed due to the relatively small number of the two variant type groups within the gnomAD database. The synonymous variant group was the second largest in terms of variant number but is perhaps the least likely variant type to affect protein function, so they were separated into their own category. The final category was 3'UTR/5'UTR which contained 3' and 5' UTR variants that, according to clinical patterns, are also unlikely to cause protein dysfunction. To support the creation of the outlined groupings, filters within the ClinVar website can be used to observe the number of pathogenic *ACVRL1* and *ENG* variants present in clinical data. On the ClinVar website, selecting filters "likely pathogenic," "pathogenic," and "multiple submitters," and only counting variants with one listed consequence type allows user to identify which variant types are most common in the searched gene. Most submissions in the "*ACVRL1*[gene]" search are missense variants followed by frameshift, nonsense (i.e. stop gained), and finally splice site variants. When the same filters are applied to an "*ENG*[gene]" search on ClinVar, pathogenic variants in order of most reported are frameshift, nonsense, splice

site, UTR, missense, and finally noncoding RNA (i.e. intron). ClinVar data therefore does not dispute the grouping of gnomAD variants and even offers some insight into how reliable the variant effect predictions are. If the proportion of pathogenic allele frequency each variant type contributes is compared to the proportion of clinically submitted variant types in ClinVar interesting patterns can be noted.

After obtaining the variants from gnomAD and separating them into groups the variants were input into a VEP openly accessible through Ensemble. An attempt to gather SIFT, PolyPhen, and CADD scores for all missense variants was made to provide some reliability and certainty to variants that were labeled benign or pathogenic. Unfortunately, very few missense variants overall produced a prediction from all three algorithms. This is in part due to results filtering by the VEP along with the lack of data provided by inputting only the rsID number into the VEP to specify variants.

4.2 The Raw Calculation Prevalence

Initial results produced what is likely a very large overestimation of HHT prevalence. The most restrictive criteria used for the first raw calculation produced a prevalence of 2.75%. With a significant number of HHT prevalence estimations revealing a worldwide prevalence less than or equal to 0.02% (1:5000) with complete penetrance, an estimation that is around 137-times higher seems unrealistic given HHT has persisted for so long in many populations. Several attempts were made to account for variables that may have led to the initial overestimation. The most probable cause for the significant overestimation in this calculation is that the VEP predictions were fully and solely trusted/relied upon. The VEP made predictions based on the rsID number provided,

meaning that scientifically logical exclusions such as homozygous variants were not explicitly excluded from the calculation, as the VEP did not have access to that information. It is very unlikely that a rare autosomal dominant disease has many, if any, affected homozygous individuals, especially given the etiology of HHT and its potential clinical consequences[41,42]. Therefore, it is important to exclude any variants that are observed as homozygous in individuals as the variant must not lead to gene loss of function. Also, prevalence was calculated using allele frequency based on allele counting, so individuals with two or more gene variants (such as homozygous individuals) were overrepresented in the final raw calculation.

4.3 The Adjusted Calculation Prevalence

First, adjustments were made to exclude homozygous variants, variants with a relatively "high" allele frequency (>0.0001), and variants with a ClinVar clinical significance of "benign" or "likely benign" with a 2-star reliability rating. Variants that were seen in homozygous individuals were excluded through this adjustment procedure to account for both the unlikelihood of an individual having two copies of a deleterious HHT variant as well as the overrepresentation that homozygotes receive in a calculation based upon allele frequencies. The removal of variants identified in homozygotes both provided the ability to rely on allele counting/allele frequency for accurate prevalence estimation as well as reduced the possibility that a truly benign variant was treated as pathogenic.

Additionally, in a similar way to homozygous individuals, alleles that are major contributors to the overall allele frequency are unlikely to be pathogenic. Because of the potential severity of HHT symptoms, a variant/allele's persistence in the population is an indication of its

clinical significance. An allele that reduces fitness, like one that causes HHT, is not likely to persist in the human population at a high prevalence. Therefore, an allele frequency cutoff was used in this study to control for alleles that are potentially individually prevalent in the population, as they are likely to not be harmful, and thus not cause HHT[25].

Furthermore, ClinVar clinical significance was also used for the adjusted calculation. ClinVar is a source of real-world evidence for the elucidation of genetic variant function. ClinVar was trusted as a reliable source of clinical variant annotation because annotations are submitted to the organization by clinicians. A variant with a 2-star rating in ClinVar indicates a variant that was submitted to the NCBI by two or more submitters (clinicians) and the submissions do not have conflicting interpretations of variant pathogenicity. Adjusting the HHT prevalence calculation after considering ClinVar annotations, the reliability of those ClinVar clinical significance notations, major allele frequency contributors, and homozygous variants significantly reduced the prevalence estimation. The adjustments excluded just 25 total variants from the prevalence estimate (7 *ACVRL1* and 18 *ENG* variants). The excluded variants reduced the most restrictive calculation by 8.6-fold and produced a prevalence estimation that is only about 16 times higher than currently accepted prevalence values. The still somewhat large difference in this prevalence estimate, as compared to currently reported values, warranted the adoption of further exclusionary measures.

4.4 The Final ClinVar Adjusted Prevalence

As a final calculation adjustment, only variants for which all algorithmic predictions indicated pathogenicity and "ClinVar clinical significance" indicated a variant as "pathogenic,"

"likely pathogenic," "variant of unknown significance," "conflicting interpretations" or an unlisted notation were included in the prevalence calculation. In other words, variants that were labeled by ClinVar as "benign" or "likely benign," no matter the star-rating, reliability, or consistency of the submission, were excluded from the calculation. Of the 1244 variants in gnomAD, ClinVar clinical significance is listed for 534 *ACVRL1* and *ENG* variants (42.9%), 322 of which are specified as either benign or pathogenic (Supplementary Table 5). The final adjustment, based on the ClinVar clinical significance label, led to the exclusion of nine additional *ENG* variants (Supplementary Table 4). This produced an estimated prevalence of 0.2945%, which is 14.7 times higher than currently reported estimates of HHT prevalence. The exclusion of all variants with mismatched VEP score and ClinVar clinical significance label, allowed for the clinical significance label to act as another (fourth) method of prediction independent of the SIFT, PolyPhen, or CADD scores. Using clinical significance as a method of pathogenicity prediction allowed for a more restrictive criteria to be applied to all variants, especially for those variants that were labeled pathogenic based on a minimal number VEP scores (e.g. only a CADD score was provided).

4.5 Genetic Analysis to Estimate Prevalence

The HHT prevalence estimates derived from genetic variant analysis indicate that the true prevalence of HHT may indeed be higher than the currently reported 1 in 5,000. However, the estimates reported in this study face some criticisms. One reason why the most conservative prevalence estimate of 0.2945% may not be trustworthy is due to the difference in prevalence between HHT caused by *ACVRL1* variants as opposed to *ENG* variants. In the clinical sphere HHT1, caused by *ENG* mutations, and HHT2, caused by the *ACVRL1* mutations, are observed at

a similar rate[9-11]. In this study, the separate calculation of pathogenic *ACVRL1* and *ENG* variants allows for the analysis of an estimated prevalence of HHT subtypes HHT1 and HHT2. In the most strictly adjusted prevalence estimate, *ENG* pathogenic variant prevalence is 1.4 times *ACVRL1* pathogenic variants prevalence. This difference in prevalence is not very large but may be an indication that the methods or sample used in this study are somewhat unreliable.

Another indication that the HHT prevalence estimation from this study may not be fully accurate is due to the difference between this study's calculated prevalence of 0.2945% and the currently accepted estimations around 0.02%. There have been many different studies in many different populations across the world that have reported an HHT prevalence of 0.02% or lower[14-19]. As previously mentioned, the leading nonprofit HHT organization, Cure HHT, claims that "9 out of 10 people with HHT are undiagnosed." Using the commonly reported 0.02% prevalence estimate and increasing it 10-fold, as Cure HHT suggests, finds that the true prevalence of HHT may be around 0.2%. The most restrictive pathogenicity labeling criteria used in this study produced a genetic variant-based prevalence estimate for HHT of 0.2945%. 0.2945% is not dramatically higher than the 0.2% anecdotal estimate suggested by Cure HHT. Although further work is necessary to refine the methods used in this study's prevalence estimate, reported findings suggest that Cure HHT's suggestion may not be too far from the true prevalence of HHT. Even if it is accepted that HHT prevalence is underestimated, it is still not clear whether a 15-fold increase, as this analysis suggests, is an accurate or reliable estimation of HHT prevalence.

4.6 Limitations

4.6.1 Limitations Leading to an Overestimation

There are many limitations to the genetic variant analysis used in this study. The limitations that may lead to an overestimation of disease prevalence include those inherent to the gnomAD database; those inherent to the VEP including SIFT, PolyPhen, and CADD; those inherent to the assumptions about HHT as a genetic disease; and those that come with the use of allele frequencies to calculate disease prevalence.

4.6.1.1 Limitations of GnomAD

GnomAD is a reliable database for the investigation of genetic variants, but like all aggregations databases, it can only tell users a limited amount of information and data from different sources are prone to inconsistency. Helpful information that was not provided by gnomAD included data on the number of individuals who had each variant or combination of variants (instead allele count is given) and phenotypic data on individuals with the given variant(s). One example of gnomAD data limitations is its apparent inconsistency with ClinVar clinical significance annotations. In the process of carrying out the adjusted and ClinVar adjusted prevalence estimates, variants with a VEP score indicating pathogenicity and a "benign" ClinVar clinical significance were checked using the ClinVar website. During this process, a variant (rs760001916) was found to have a different clinical significance label in ClinVar as compared to its "ClinVar clinical significance" label given by gnomAD. This discrepancy is very problematic as it calls into question all "ClinVar clinical significance" labels in gnomAD. For the scope of this study the ClinVar clinical significance label given by gnomAD was trusted and rs760001916 was

treated as if its clinical significance was “likely benign” with a less than two-star rating. The difference in this one variant’s clinical significance label may be a single anomaly due to updates in the ClinVar website after the time of gnomAD variant ascertainment, but it is unlikely that this variant is the only case of a differing clinical significance label.

The reliance on ClinVar and clinical patterns for calculation adjustments in this study revealed another limitation: the use of synonymous and UTR variants in the prevalence estimation. To date, there are no synonymous *ACVRL1* or *ENG* variants that are currently reported in ClinVar to be pathogenic or likely pathogenic with multiple submitters. Additionally, there are no pathogenic *ACVRL1* UTR variants within ClinVar. These facts may indicate the need to exclude variants that are predicted to be pathogenic within those variant type categories. For this study, all variants within the exonic and UTR regions of the selected genes were included in an attempt to reduce bias. If *ACVRL1* and *ENG* synonymous variants and *ACVRL1* UTR variants are excluded from the ClinVar adjusted prevalence estimation, this leads to a prevalence of 0.2814% (14.1:5000). The lack of clinical support for the pathogenicity of synonymous and UTR variant types highlights a larger issue that is the lack of pathogenic justification for variants beyond the VEP scores. For example, synonymous variants rs756285377 and rs1280679232 were both labeled pathogenic by the VEP. Without given algorithmic explanations/details and no laboratory or clinical evidence, the true pathogenicity of these variants cannot be confidently defended.

4.6.1.2 Limitations of the VEP

In-silico variant analysis relies on algorithms for pathogenicity prediction. In this study SIFT, PolyPhen, and CADD scores were generated to label variants as pathogenic, and each method has its own limitations. Due to the reliance of SIFT and PolyPhen on the conservation and properties of amino acids, only missense variants could be reliably analyzed and scored. This

means that for this study, CADD scores are the only predictor for the non-missense variant types. Furthermore, CADD's machine learning nature lends itself to bias and inaccuracy based on the data with which the program was trained. Perhaps the most important limitation associated with each VEP score is that they are calculated based on an algorithm and a set of biological/computational rules and patterns which are dynamic in the modern scientific era. On top of these individual limitations, combined analysis was difficult due to the rarity of variants that produced a predicted effect by all three algorithms. The majority of variants in the "All" category were based on one given pathogenicity prediction rather than all three or even two out of three predictors labeling the variant as pathogenic. The reliance on a single VEP score to label a variant as pathogenic or not can lead to an incorrect and unbalanced prevalence calculation. Additionally, the Ensembl-VEP used to produce SIFT, PolyPhen, and CADD scores for variants has its own parameters for analyzing and filtering data[29]. This became problematic when variant type labels given by the VEP were inconsistent with their gnomAD annotation. Inconsistencies between variant type labels between the VEP and gnomAD are possible due to the way in which variants were entered into the VEP. The use of rsID numbers as variant inputs for the VEP only allows the variant to be analyzed by the location and single nucleotide polymorphism at that site. This means that information related to isoforms of proteins or alternative splice products of the gene are unknown to the VEP. Variants labeled with a VEP consequence that was inconsistent with the gnomAD annotation were re-run and scored individually. However, this inconsistency may be an indication of other, less clearly visible, mistakes that are present when using a separate resource to gather (gnomAD) and analyze (VEP) genetic variants.

4.6.1.3 Limitations of the Known Genetics of HHT

Beyond the limitations of the variant predictions themselves, there are fundamental assumptions about HHT that must be made to calculate the prevalence of a genetic condition. For example, the use of HHT genetic variants from many locations within all exons assumes that mutations in any and all exonic locations of the affected gene have the ability to lead to HHT. Even given that clinical data shows the potential for HHT-causing variants in all *ACVRL1* and *ENG* exons, the assumption that any nucleotide change in any location within a gene can cause HHT cannot be verified. It was also assumed that all pathogenic variants share the ~100% penetrance that has been seen in HHT cases[20].

4.6.1.4 Limitations of Using Allele Frequencies

Furthermore, another limitation to this HHT prevalence estimation is the use of allele frequencies to calculate disease prevalence. When using allele frequencies alone, individuals with multiple variants cannot be accounted for. GnomAD attempts to assist with this issue by providing a "flag" datapoint that notes variants that may be seen in individuals alongside other variants, called "multi-nucleotide variant" (mnv). For this study the flag designation is not of much use due to the lack of information provided beyond a flag label itself. Specific flag label descriptions can be found using the gnomAD website and the variants with a reported flag are provided in Supplementary Table 6. The treatment of each variant allele count as a separate individual can lead to part of the overestimation seen in this study. The only way to avoid this limitation is to have the specific data on which variants are seen with which other variants to then not count the allele frequency twice for a variant combination that is only seen in one individual.

4.6.2 Limitations Leading to an Underestimation

Contrary to previous limitations, there are also some limitations of the methodology that would presumably lead to an underestimation of HHT prevalence that should be noted. One of these limitations is the exclusion of other genes and causes of HHT (i.e. the ~10% of HHT not caused by ACVRL1 and ENG variants). In opposition to a limitation that may have led to an overestimation, it is also possible that the inclusion of three different VEP algorithms provided an opportunity for variants to be labeled benign when they are truly pathogenic.

4.7 Future Directions

Future studies into HHT prevalence would benefit from many additional considerations. The most manageable consideration is to alter the pathogenicity prediction process. Alteration of the variant effect predictions can be done simply by altering the selectivity in the pathogenicity thresholds used, incorporating different or more prediction algorithms/methods, and/or changing the filtering settings on the VEP used. For example, a CADD PHRED threshold of 20.0 can be used and if variants are run in smaller batches with different VEP filtering settings, all three predictors (SFIT, PolyPhen, and CADD) may produce a pathogenicity score for all missense variants. Future studies that aim to calculate the prevalence of a rare disease using genetic variant data would also likely benefit from more specific data on the frequency of the genetic variants in individuals. There are also different databases that can be used to gather genetic variant data, such as the “All of Us” cohort. Different genetic variant databases may be better equipped to address occurrences of analyzed variants in individuals and note some phenotypic and clinical patterns

seen in the individuals with a specific variant. Additionally, samples more representative of true population demographics will provide a more in depth understanding of HHT prevalence across the globe (Supplementary Table 7).

5.0 Conclusion

In conclusion, using a genetic variant analysis to estimate the prevalence of HHT produced a prevalence of 0.2945%, a much higher prevalence than has been previously reported (0.02%). The use of gnomAD for the ascertainment of genetic variants and their allele frequencies as well as Ensembl for a VEP to produce variant SIFT, PolyPhen, and CADD scores led to important opportunities and limitations. Because of the magnitude of difference in this study's prevalence estimation and the methodological limitations as compared to previous work, even the lowest 0.2945% prevalence value is likely to be an overestimation of true HHT prevalence. The methodology used in this study was intended to control for factors that would lead to an overestimation of prevalence, but even so, a prevalence estimate 14.7 times higher than what is currently accepted was calculated. Given this effort, the results of this study indicate that further analysis of HHT prevalence is warranted, and perhaps genetic variant analysis methodologies overall require a more controlled approach. Genetic variant analyses are best supplemented with more complete clinical data and ethnically diverse samples. As for the prevalence of HHT, there does seem to be sufficient evidence to suggest that the prevalence of HHT is greater than the current 1 in 5,000 estimates, but more investigation is required to find the true prevalence of HHT.

6.0 Figures and Tables

6.1 Figures

6.1.1 Figure 1

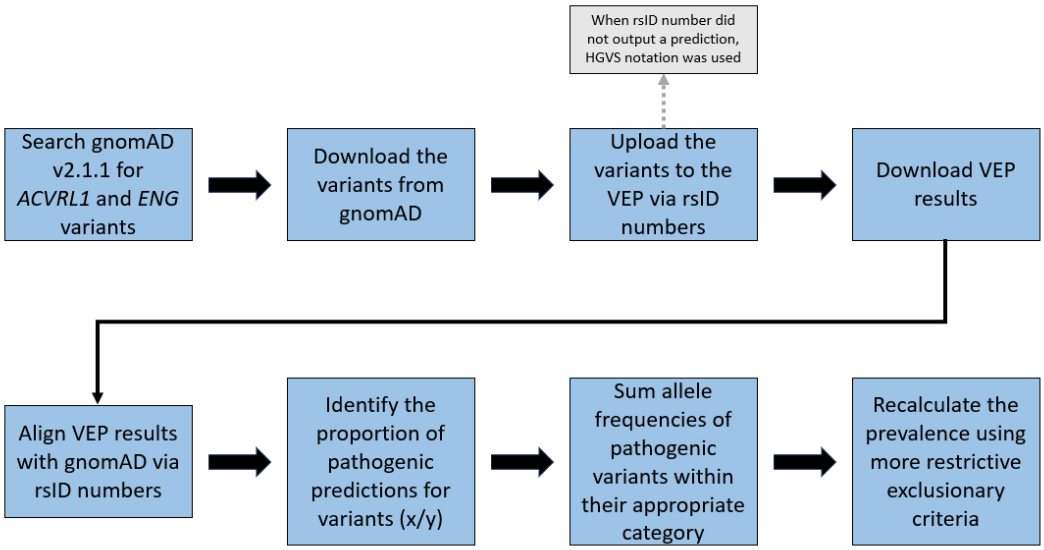
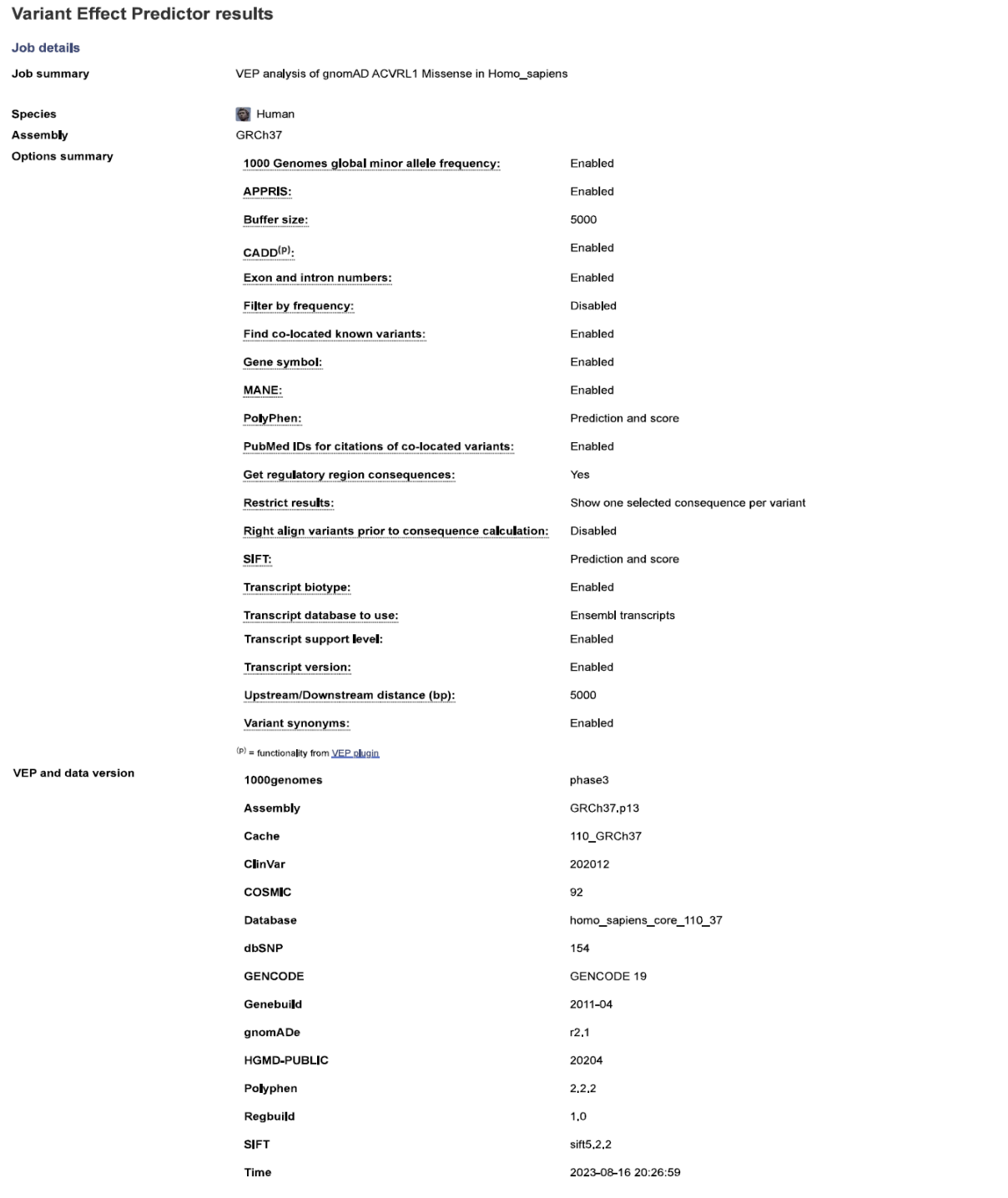


Figure 1. Methodology Workflow

Flowchart of the methodological process performed in this study. "X" represents the number of predictions that labeled the variant as pathogenic. "Y" represents the number of prediction methods that output any prediction/score. Note: some uploaded VEP results displayed mislabeled variants (e.g. called a "synonymous" variant and a "missense" prediction was given), in these cases, the variant was re-run and the prediction for the appropriate consequence was used.

6.1.2 Figure 2



Variant Effect Predictor results

Job details

Job summary	VEP analysis of gnomAD ACVRL1 Missense in Homo_sapiens	
Species	Human	
Assembly	GRCh37	
Options summary	**1000 Genomes global minor allele frequency:**	Enabled
	APPRIS:	Enabled
	Buffer size:	5000
	CADD(p):	Enabled
	Exon and intron numbers:	Enabled
	Filter by frequency:	Disabled
	Find co-located known variants:	Enabled
	Gene symbol:	Enabled
	MANE:	Enabled
	PolyPhen:	Prediction and score
	PubMed IDs for citations of co-located variants:	Enabled
	Get regulatory region consequences:	Yes
	Restrict results:	Show one selected consequence per variant
	Right align variants prior to consequence calculation:	Disabled
	SIFT:	Prediction and score
	Transcript biotype:	Enabled
	Transcript database to use:	Ensembl transcripts
	Transcript support level:	Enabled
	Transcript version:	Enabled
	Upstream/Downstream distance (bp):	5000
	Variant synonyms:	Enabled
	(p) = functionality from VEP plugin	
VEP and data version	**1000genomes**	phase3
	Assembly	GRCh37.p13
	Cache	110_GRCh37
	ClinVar	202012
	COSMIC	92
	Database	homo_sapiens_core_110_37
	dbSNP	154
	GENCODE	GENCODE 19
	Genebuild	2011-04
	gnomADe	r2.1
	HGMD-PUBLIC	20204
	Polyphen	2.2.2
	Regbuild	1.0
	SIFT	sift5.2.2
	Time	2023-08-16 20:26:59

Figure 2. VEP Settings

VEP settings used to analyze all genetic variants.

6.1.3 Figure 3

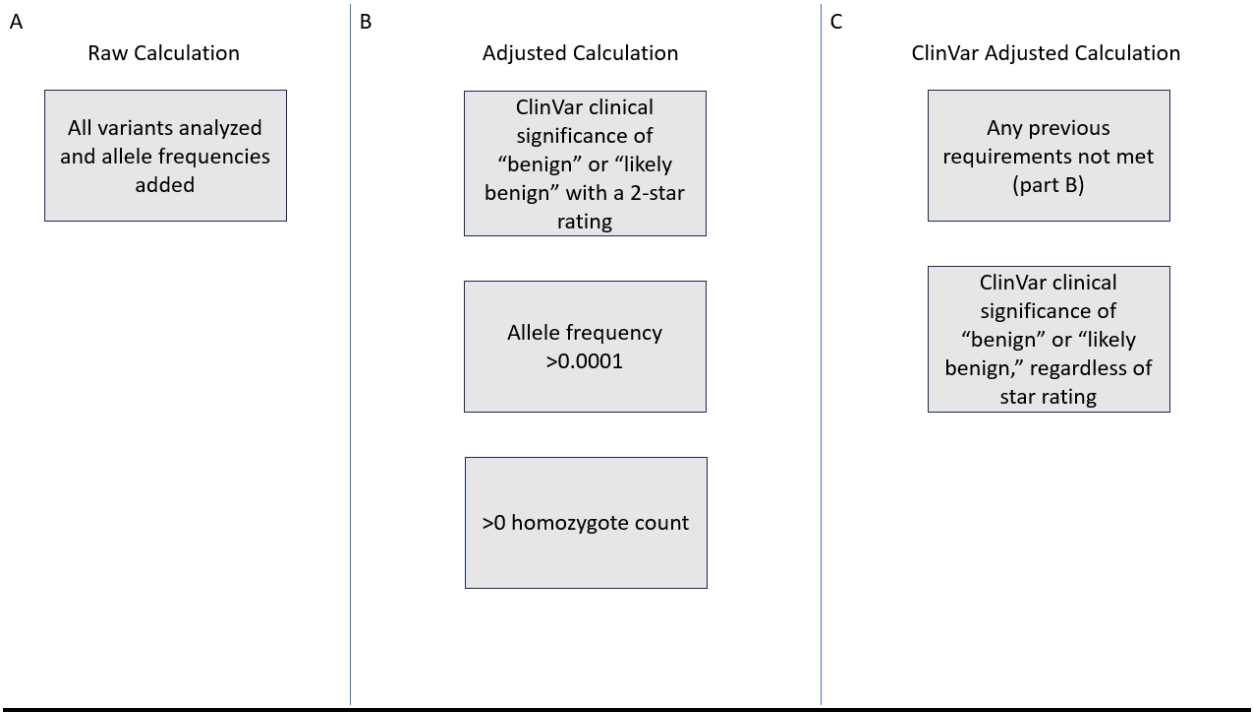


Figure 3. Variant Exclusion Criteria

Figure 3. (A) For the Raw calculation all variants with pathogenic prediction scores via Ensembl VEP were included. (B and C) Exclusion criteria for adjusted (B) and ClinVar Adjusted (C) calculations. If any statement was true for a given variant, the variant was excluded from the calculation.

6.2 Tables

6.2.1 Table 1

Table 1. Descriptive Variant Data

Table 1. Counts of the number of variants within type of variant categories.

Type of Variant	***ACVRL1***	***ENG***	**Totals**
Missense	243	365	608
StopGain/FrameShift/StartLoss	17	3	20
SpliceSite/InFrameDeletion	31	74	105
Synonymous	172	193	365
3'UTR/5'UTR	58	72	130
Totals	521	707	1228

6.2.2 Table 2

Table 2. HHT Prevalence: Raw Calculation

Table 2	*ACVRL1*			*ENG*			Combined
	Type of Variant	Number of Contributing Variants	Pathogenic Allele Frequency	Type of Variant	Number of Contributing Variants	Pathogenic Allele Frequency	**Raw Prevalence**
1 or More Pathogenic Prediction (≥1/3)	Missense	143	0.357985123	Missense	199	0.022818716	0.38520
	StopGain/FrameShift/StartLoss	13	0.000191822	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	16	0.000349184	
	Synonymous	1	4.02023E-06	Synonymous	1	3.99345E-06	
	3'UTR/5'UTR	4	0.000288832	3'UTR/5'UTR	15	0.003481433	
	Total	171	0.358546948	Total	232	0.026657998	
Majority Pathogenic Predictions (≥1/2)	Missense	119	0.004282049	Missense	189	0.022508229	0.03113
	StopGain/FrameShift/StartLoss	11	0.000127995	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	16	0.000349184	
	Synonymous	1	4.02023E-06	Synonymous	1	3.99345E-06	
	3'UTR/5'UTR	4	0.000288832	3'UTR/5'UTR	15	0.003481433	
	Total	145	0.004780047	Total	222	0.026347512	
All Pathogenic Predictions (3/3, 2/2, 1/1)	Missense	67	0.003029065	Missense	103	0.020122176	0.027489 (137/5000)
	StopGain/FrameShift/StartLoss	11	0.000127995	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	16	0.000349184	
	Synonymous	1	4.02023E-06	Synonymous	1	3.99345E-06	
	3'UTR/5'UTR	4	0.000288832	3'UTR/5'UTR	15	0.003481433	
	Total	93	0.003527063	Total	136	0.023961458	

6.2.3 Table 3

Table 3. HHT Prevalence: Adjusted Calculation

Table 3	*ACVRL1*			*ENG*			Combined
	Type of Variant	Number of Contributing Variants	Pathogenic Allele Frequency	Type of Variant	Number of Contributing Variants	Pathogenic Allele Frequency	**Adjusted Prevalence**
1 or More Pathogenic Prediction (≥1/3)	Missense	137	0.002952313	Missense	186	0.002624133	0.00633
	StopGain/FrameShift/StartLoss	13	0.000191822	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	16	0.000349184	
	Synonymous	1	4.02023E-06	Synonymous	1	3.99345E-06	
	3'UTR/5'UTR	3	1.31455E-05	3'UTR/5'UTR	10	0.000114469	
	Total	164	0.003238451	Total	214	0.003096452	
Majority Pathogenic Predictions (≥1/2)	Missense	115	0.002037692	Missense	177	0.002469336	0.00520
	StopGain/FrameShift/StartLoss	11	0.000127995	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	16	0.000349184	
	Synonymous	1	4.02023E-06	Synonymous	1	3.99345E-06	
	3'UTR/5'UTR	3	1.31455E-05	3'UTR/5'UTR	10	0.000114469	
	Total	140	0.002260003	Total	205	0.002941655	
All Pathogenic Predictions (3/3, 2/2, 1/1)	Missense	64	0.000983579	Missense	99	0.001524908	0.003203 (16/5000)
	StopGain/FrameShift/StartLoss	11	0.000127995	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	16	0.000349184	
	Synonymous	1	4.02023E-06	Synonymous	1	3.99345E-06	
	3'UTR/5'UTR	3	1.31455E-05	3'UTR/5'UTR	10	0.000114469	
	Total	89	0.00120589	Total	127	0.001997226	

6.2.4 Table 4

Table 4. HHT Prevalence: ClinVar Adjusted Calculation

Table 4	*ACVRL1*			*ENG*			Combined
	Type of Variant	Number of Contributing Variants	Pathogenic Allele frequency	Type of Variant	Number of Contributing Variants	Pathogenic Allele frequency	**ClinVar Adjusted Prevalence**
All Pathogenic Predictions (3/3, 2/2, 1/1) including ClinVar	Missense	64	0.000983579	Missense	93	0.001409977	0.002945 (15/5000)
	StopGain/FrameShift/StartLoss	11	0.000127995	StopGain/FrameShift/StartLoss	1	4.67207E-06	
	SpliceSite/InFrameDeletion	10	7.71501E-05	SpliceSite/InFrameDeletion	14	0.000210252	
	Synonymous	1	4.02023E-06	Synonymous	0	0	
	3'UTR/5'UTR	3	1.31455E-05	3'UTR/5'UTR	10	0.000114469	
	Total	89	0.001205890	Total	118	0.001739370	

Appendix A Supplementary Tables

Appendix A.1 Supplementary Table 1

Supplementary Table 1. All Variants in GnomAD and Variant Analyses

Supplementary Table 1. An excel sheet for both *ACVRL1* and *ENG* gene variant analyses as well as sheets for the complete variant lists downloaded from gnomAD at the time of study. In the analysis sections, variants are separated by analysis groupings. Variants in the analysis sections are color coded to indicate pathogenicity (red), conflicting or uncertain clinical significance (grey), benign (green), inconsistent varaint type annotations that were re-run (organge), variants that produced no VEP score (yellow), and variants excluded from the adjusted calculation (blue).

gnomADv2.1.1 ALL variants 08.09.2023

Appendix A.2 Supplementary Table 2

Supplementary Table 2. Variants Without Pathogenicity Scores

ACVRL1			*ENG*		
Variant	Variant Type	Allele Frequency	Variant	Variant Type	Allele Frequency
rs1353580616	Frameshift	3.99071E-06	rs754066649	Inframe insertion	5.70017E-05
rs745544888	Frameshift	4.06075E-06	rs1476851940	Inframe insertion	5.65796E-06
rs753624883	Inframe deletion	4.06904E-06	rs765377503	Inframe insertion	4.00404E-06
rs1184640812	Splice region	2.04424E-05	rs1221530128	Splice region	1.2107E-05
rs1282724548	Splice region	4.0084E-06	ENG:c.1687-12_1687-7delTTTCTC	Splice region	3.98061E-06
rs755419795	Splice region	4.0113E-06	rs1260888046	Splice region	3.97807E-06
rs1206786731	5'UTR	3.18715E-05	rs1389761719	3'UTR	6.46939E-06
			rs1449495191	3'UTR	5.77034E-06
			rs748184867	3'UTR	1.11809E-05

Appendix A.3 Supplementary Table 3

Supplementary Table 3. Variants Excluded from the Adjusted Calculation

ACVRL1						*ENG*					
Variant	Variant Type	Labeled Pathogenicity	ClinVar Clinical Significance (Star Rating)	Homozygote Count	Allele Frequency	Variant	Variant Type	Labeled Pathogenicity	ClinVar Clinical Significance (Star Rating)	Homozygote Count	Allele Frequency
rs199874575	missense	2/2	Benign/LB (2)	2	0.000355	rs142896669	missense	2/2	Conflicting	0	0.000495
rs139142865	missense	2/2	Benign/LB (2)	1	0.001602	rs1800956	missense	2/2	Benign/LB (2)	134	0.008921
rs746715195	missense	2/2	Likely benign (1)	1	8.87E-05	rs41322046	missense	2/2	Benign/LB (2)	11	0.009062
rs373133784	missense	1/2	VUS	0	0.000199	rs762200397	missense	2/2	Likely benign (1)	0	0.000118
rs706816	missense	1/3	Benign (2)	10,732	0.269622	rs139334561	missense	1/2	Conflicting	0	0.000278
rs2277382	missense	1/3	Benign (1)	1,054	0.083167	rs756897517	missense	1/2	Conflicting	0	0.000127
rs532801800	5' UTR	1/1		0	0.000276	rs150932144	missense	1/2	Conflicting	0	0.000124
						rs201393380	missense	1/2	Benign/LB (2)	0	0.000221
						rs200960408	missense	1/2	Likely benign (2)	0	9.91E-05
						rs762209698	missense	1/2	Likely benign (2)	0	1.89E-05
						rs752195587	missense	1/2	Benign	1	0.000174
						rs146100407	missense	1/2	Benign	0	0.0004
						rs372045549	missense	1/3	Conflicting	0	0.000156
						rs886063476	5'UTR	1/1	VUS	0	0.000319
						rs2296702	5'UTR	1/1	Benign (1)	1	0.002039
						rs562538400	5'UTR	1/1	Benign/LB (1)	0	0.000287
						rs371104611	3'UTR	1/1	VUS	0	0.000376
						rs376579767	3'UTR	1/1	Likely benign (1)	0	0.000346

Appendix A.4 Supplementary Table 4

Supplemental Table 4. Variants Excluded from the ClinVar Adjusted Calculation

ENG				
Variant	Variant Type	Labeled Pathogenicity	ClinVar Clinical Significance	Allele Frequency
rs777633247	missense	2/2	Benign	1.67735E-05
rs532649202	missense	2/2	Benign	3.97934E-05
rs748457491	missense	2/2	Likely benign	4.24304E-05
rs121918401	missense	3/3	Likely benign	7.95279E-06
rs1800956	missense	2/2	Likely benign	3.97646E-06
rs754842280	missense	2/2	Likely benign	4.00449E-06
rs202048202	splice region	1/1	Benign	9.94748E-05
rs374628465	splice region	1/1	Likely benign	3.94568E-05
rs1280679232	synonymous	1/1	Likely benign	3.99345E-06

Appendix A.5 Supplementary Table 5

Supplementary Table 5. ClinVar Clinical Signifiance Descriptive Data

	Number of Variants Labeled		
ClinVar Clinical Significance Label	***ACVRL1***	***ENG***	**Total**
Benign	7	16	23
Benign/Likely benign	15	34	49
Likely benign	62	163	225
Likely pathogenic	4	3	7
Pathogenic/Likely pathogenic	3	4	7
Pathogenic	11	0	11
Conflicting interpretations of pathogenicity	17	47	64
Uncertain significance	40	108	148
Total	159	375	534

Appendix A.6 Supplementary Table 6

Supplementary Table 6. Flagged Variants

Supplementary Table 6. A display of all flagged variants within the gnomAD databse for *ACVRL1* and *ENG* genes. Flagg abbreviations are as follows: "multinucleotide variant" (mnv), "low confidence predicted loss of function" (lc_lof), "predicted loss of function flag" (lof_flag), "low complexity region" (lcr), and "outside canonical splice site predicted loss of function" (os_lof).

ACVRL1					*ENG*				
Variant	Variant Type	Flag	Labeled Pathogenicity	Allele Frequency	Variant	Variant Type	Flag	Labeled Pathogenicity	Allele Frequency
rs755922974	Missense	mnv	1/1	7.95583E-06	rs1800956	Missense	mnv	2/2	0.008921499
rs1466116430	Missense	mnv	0/2	4.86112E-06	rs1353936644	Missense	mnv	2/2	3.18654E-05
rs1268297127	Missense	mnv	0/2	4.85559E-06	rs146100407	Missense	mnv	1/2	0.000399562
rs1060499838	Frameshift	lc_lof	1/1	3.97988E-06	rs146188464	Synonymous	mnv	0/1	0.000399559
rs759344606	Frameshift	lc_lof	0/1	2.00877E-05	rs201497772	Synonymous	mnv	0/1	7.55515E-05
rs745544888	Frameshift	lc_lof	0/0	4.06075E-06	rs370943570	Synonymous	mnv	0/1	7.28237E-06
rs1353580616	Frameshift	lc_lof	0/0	3.99071E-06	rs1367345326	Splice Site	os_lof	1/1	5.00495E-06
rs764143924	Frameshift	lof_flag	0/1	3.97934E-06					
rs778959565	Stop Gain	mnv	1/1	7.95539E-06					
rs1057517944	Stop Gain	lc_lof	1/1	3.97741E-06					
rs1457108711	Stop Gain	lc_lof	0/1	3.18979E-05					
rs576100542	Stop Gain	lc_lof	0/1	1.08462E-05					
rs746661046	Splice Site	os_lof	1/1	2.90818E-05					
rs1480142671	Splice Site	os_lof	1/1	4.05022E-06					
rs61734312	Splice Site	os_lof	1/1	3.97972E-06					
rs771035372	3'UTR	lcr	0/1	0.0003121					
rs776507354	3'UTR	lcr	0/1	0.000103529					
rs753232692	3'UTR	lcr	0/1	6.51093E-05					
rs771035372	3'UTR	lcr	0/1	3.19407E-05					

Supplementary Table 6. Flagged Variants (continued)

ACVRL1				
Variant	Variant Type	Flag	Labeled Pathogenicity	Allele Frequency
rs764500987	3'UTR	lcr	0/1	2.018E-05
rs759673510	3'UTR	lcr	0/1	1.94709E-05
rs765683960	3'UTR	lcr	0/1	1.61558E-05
rs201106717	3'UTR	lcr	0/1	1.29505E-05
rs768608853	3'UTR	lcr	0/1	1.23147E-05
rs759673510	3'UTR	lcr	0/1	1.0785E-05
rs1405608386	3'UTR	lcr	0/1	6.74645E-06
rs768608853	3'UTR	lcr	0/1	6.15733E-06

Appendix A.7 Supplementary Table 7

Supplementary Table 7. Representation of Ethnicities in GnomAD

	Percent of Allele Number Contributed to the Sample		
Ethnic Origin of Sample	***ACVRL1***	***ENG***	**Overall**
African/African-American	7.32%	7.32%	7.32%
Latino/Admixed American	13.55%	13.60%	13.58%
Ashkenazi Jewish	3.98%	3.96%	3.97%
East Asian	7.31%	7.35%	7.33%
European (Finnish)	8.31%	8.50%	8.42%
European (non-Finnish)	45.13%	44.91%	45.00%
South Asian	11.88%	11.85%	11.86%
Other	2.52%	2.52%	2.52%

Bibliography

1 Shovlin, C. L. Hereditary haemorrhagic telangiectasia: pathophysiology, diagnosis and treatment. *Blood Rev* **24**, 203-219 (2010). https://doi.org:10.1016/j.blre.2010.07.001

2 Salibe-Filho, W., Oliveira, F. R. & Terra-Filho, M. Update on pulmonary arteriovenous malformations. *J Bras Pneumol* **49**, e20220359 (2023). https://doi.org:10.36416/1806-3756/e20220359

3 de Gussem, E. M. *et al.* Life expextancy of parents with Hereditary Haemorrhagic Telangiectasia. *Orphanet J Rare Dis* **11**, 46 (2016). https://doi.org:10.1186/s13023-016-0427-x

4 McDonald, J. *et al.* Hereditary hemorrhagic telangiectasia: genetics and molecular diagnostics in a new era. *Front Genet* **6**, 1 (2015). https://doi.org:10.3389/fgene.2015.00001

5 Robert, F., Desroches-Castan, A., Bailly, S., Dupuis-Girod, S. & Feige, J. J. Future treatments for hereditary hemorrhagic telangiectasia. *Orphanet J Rare Dis* **15**, 4 (2020). https://doi.org:10.1186/s13023-019-1281-4

6 Richards-Yutz, J., Grant, K., Chao, E. C., Walther, S. E. & Ganguly, A. Update on molecular diagnosis of hereditary hemorrhagic telangiectasia. *Hum Genet* **128**, 61-77 (2010). https://doi.org:10.1007/s00439-010-0825-4

7 Sanchez-Martinez, R. *et al.* Current HHT genetic overview in Spain and its phenotypic correlation: data from RiHHTa registry. *Orphanet J Rare Dis* **15**, 138 (2020). https://doi.org:10.1186/s13023-020-01422-8

8 Oh, S. P. *et al.* Activin receptor-like kinase 1 modulates transforming growth factor-beta 1 signaling in the regulation of angiogenesis. *Proc Natl Acad Sci U S A* **97**, 2626-2631 (2000). https://doi.org:10.1073/pnas.97.6.2626

9 Heimdal, K. *et al.* Mutation analysis in Norwegian families with hereditary hemorrhagic telangiectasia: founder mutations in ACVRL1. *Clin Genet* **89**, 182-186 (2016). https://doi.org:10.1111/cge.12612

10 Lesca, G. *et al.* Distribution of ENG and ACVRL1 (ALK1) mutations in French HHT patients. *Hum Mutat* **27**, 598 (2006). https://doi.org:10.1002/humu.9421

11 Torring, P. M., Brusgaard, K., Ousager, L. B., Andersen, P. E. & Kjeldsen, A. D. National mutation study among Danish patients with hereditary haemorrhagic telangiectasia. *Clin Genet* **86**, 123-133 (2014). https://doi.org:10.1111/cge.12269

12 Ruiz-Llorente, L. *et al.* Endoglin and alk1 as therapeutic targets for hereditary hemorrhagic telangiectasia. *Expert Opin Ther Targets* **21**, 933-947 (2017). https://doi.org:10.1080/14728222.2017.1365839

13 Gallione, C. J. *et al.* A combined syndrome of juvenile polyposis and hereditary haemorrhagic telangiectasia associated with mutations in MADH4 (SMAD4). *Lancet* **363**, 852-859 (2004). https://doi.org:10.1016/S0140-6736(04)15732-2

14 Grosse, S. D., Boulet, S. L., Grant, A. M., Hulihan, M. M. & Faughnan, M. E. The use of US health insurance data for surveillance of rare disorders: hereditary hemorrhagic telangiectasia. *Genet Med* **16**, 33-39 (2014). https://doi.org:10.1038/gim.2013.66

15 Kjeldsen, A. D., Vase, P. & Green, A. Hereditary haemorrhagic telangiectasia: a population-based study of prevalence and mortality in Danish patients. *J Intern Med* **245**, 31-39 (1999). https://doi.org:10.1046/j.1365-2796.1999.00398.x

16 Dakeishi, M. *et al.* Genetic epidemiology of hereditary hemorrhagic telangiectasia in a local community in the northern part of Japan. *Hum Mutat* **19**, 140-148 (2002). https://doi.org:10.1002/humu.10026

17 Donaldson, J. W., McKeever, T. M., Hall, I. P., Hubbard, R. B. & Fogarty, A. W. The UK prevalence of hereditary haemorrhagic telangiectasia and its association with sex, socioeconomic status and region of residence: a population-based study. *Thorax* **69**, 161-167 (2014). https://doi.org:10.1136/thoraxjnl-2013-203720

18 Shovlin, C. L. *et al.* European Reference Network For Rare Vascular Diseases (VASCERN) Outcome Measures For Hereditary Haemorrhagic Telangiectasia (HHT). *Orphanet J Rare Dis* **13**, 136 (2018). https://doi.org:10.1186/s13023-018-0850-2

19 Westermann, C. J., Rosina, A. F., De Vries, V. & de Coteau, P. A. The prevalence and manifestations of hereditary hemorrhagic telangiectasia in the Afro-Caribbean population of the Netherlands Antilles: a family screening. *Am J Med Genet A* **116A**, 324-328 (2003). https://doi.org:10.1002/ajmg.a.10002

20 Bayrak-Toydemir, P. *et al.* Genotype-phenotype correlation in hereditary hemorrhagic telangiectasia: mutations and manifestations. *Am J Med Genet A* **140**, 463-470 (2006). https://doi.org:10.1002/ajmg.a.31101

21 Pierucci, P. *et al.* A long diagnostic delay in patients with Hereditary Haemorrhagic Telangiectasia: a questionnaire-based retrospective study. *Orphanet J Rare Dis* **7**, 33 (2012). https://doi.org:10.1186/1750-1172-7-33

22 Faughnan, M. E. *et al.* Second International Guidelines for the Diagnosis and Management of Hereditary Hemorrhagic Telangiectasia. *Ann Intern Med* **173**, 989-1001 (2020). https://doi.org:10.7326/M20-1443

23 Hammill, A. M., Wusik, K. & Kasthuri, R. S. Hereditary hemorrhagic telangiectasia (HHT): a practical guide to management. *Hematology Am Soc Hematol Educ Program* **2021**, 469-477 (2021). https://doi.org:10.1182/hematology.2021000281

24 Shovlin, C. L. *et al.* Diagnostic criteria for hereditary hemorrhagic telangiectasia (Rendu-Osler-Weber syndrome). *Am J Med Genet* **91**, 66-67 (2000). https://doi.org:10.1002/(sici)1096-8628(20000306)91:1

25 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-755 (2011). https://doi.org:10.1038/nrg3031

26 Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835-D844 (2020). https://doi.org:10.1093/nar/gkz972

27 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020). https://doi.org:10.1038/s41586-020-2308-7

28 Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* **43**, 1012-1030 (2022). https://doi.org:10.1002/humu.24309

29 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016). https://doi.org:10.1186/s13059-016-0974-4

30 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003). https://doi.org:10.1093/nar/gkg509

31 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009). https://doi.org:10.1038/nprot.2009.86

32 Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894-3900 (2002). https://doi.org:10.1093/nar/gkf493

33 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014). https://doi.org:10.1038/ng.2892

34 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019). https://doi.org:10.1093/nar/gky1016

35 Borges, P., Pasqualim, G., Giugliani, R., Vairo, F. & Matte, U. Estimated prevalence of mucopolysaccharidoses from population-based exomes and genomes. *Orphanet J Rare Dis* **15**, 324 (2020). https://doi.org:10.1186/s13023-020-01608-0

36 Zhao, T., Fan, S. & Sun, L. The global carrier frequency and genetic prevalence of Upshaw-Schulman syndrome. *BMC Genom Data* **22**, 50 (2021). https://doi.org:10.1186/s12863-021-01010-0

37 Gao, J., Brackley, S. & Mann, J. P. The global prevalence of Wilson disease from next-generation sequencing data. *Genet Med* **21**, 1155-1163 (2019). https://doi.org:10.1038/s41436-018-0309-9

38 de Andrade, K. C. *et al.* Variable population prevalence estimates of germline TP53 variants: A gnomAD-based analysis. *Hum Mutat* **40**, 97-105 (2019). https://doi.org:10.1002/humu.23673

39 Kaler, S. G., Ferreira, C. R. & Yam, L. S. Estimated birth prevalence of Menkes disease and ATP7A-related disorders based on the Genome Aggregation Database (gnomAD). *Mol Genet Metab Rep* **24**, 100602 (2020). https://doi.org:10.1016/j.ymgmr.2020.100602

40 Roberts, A. D. & Wadhwa, R. in *StatPearls* (2023).

41 Singh, A., Saini, N., Behl, G., Aggarwal, S. & Kolar, G. Recurrent Vein of Galen Aneurysmal Malformation as a Presentation of Hereditary Hemorrhagic Telangiectasia. *Mol Syndromol* **13**, 440-446 (2022). https://doi.org:10.1159/000522352

42 El-Harith el, H. A. *et al.* Hereditary hemorrhagic telangiectasia is caused by the Q490X mutation of the ACVRL1 gene in a large Arab family: support of homozygous lethality. *Eur J Med Genet* **49**, 323-330 (2006). https://doi.org:10.1016/j.ejmg.2005.09.002