Non-canonical regulation of Endoglin by rare and common variants: new molecular and clinical perspectives for Hereditary Hemorrhagic Telangiectasia and beyond

Omar Soukarieh[1,2#*], Gaëlle Munsch[1#], Clémence Deiber[1], Caroline Meguerditchian[1], Carole Proust[1], Ilana Caro[1], Maud Tusseau[3], Alexandre Guilhem[4,5], Shirine Mohamed[6], INVENT consortium, Béatrice Jaspard-Vinassa[2], Aurélie Goyenvalle[7], Stéphanie Debette[1,8], Sophie Dupuis-Girod[3,9], David-Alexandre Trégouët[1*]

[1] Univ. Bordeaux, INSERM, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France
[2] Univ. Bordeaux, INSERM, Biology of Cardiovascular Diseases, U1034, F-33600 Pessac, France
[3] Hospices Civils de Lyon, French National HHT Reference Center and Genetics department, Hôpital Femme-Mère-Enfant, 69677, Bron, France
[4] Service de Médecine Interne et Immunologie Clinique, Centre de compétence maladie de Rendu-Osler, Centre Hospitalo-Universitaire Dijon Bourgogne, Dijon, France
[5] Université de Bourgogne, INSERM, EFS BFC, UMR1098, RIGHT Interactions Greffon-Hôte-Tumeur/Ingénierie Cellulaire et Génique, Dijon, France
[6] Département de Médecine interne et Immunologie Clinique, CHRU BRABOIS, Vandoeuvre-lès-Nancy, France
[7] Université Paris-Saclay, UVSQ, Inserm, END-ICAP, Versailles, France
[8] Department of Neurology, Institute for Neurodegenerative Diseases, Bordeaux University Hospital, France
[9] Univ. Grenoble Alpes, Inserm, CEA, Laboratory Biology of Cancer and Infection, F-38000 Grenoble, France.

[#] These authors have contributed equally to this work
[*] Joint corresponding authors: omar.soukarieh@inserm.fr; david-alexandre.tregouet@u-bordeaux.fr

Abstract

Endoglin, encoded by the *ENG* gene, is a transmembrane glycoprotein with a major implication in angiogenesis. Loss-of function *ENG* variants are responsible for Hereditary Hemorrhagic Telangiectasia (HHT), a rare vascular disease, characterized with a large inter-individual clinical heterogeneity. But, Endoglin and its soluble form have also been reported to be involved in other pathologic conditions including cancer and thrombosis. Thus, dissecting the genetic regulation of Endoglin holds the potential to deepen our understanding of the pathophysiology underlying HHT and other human diseases.

To follow-up our latest study in which we characterized 5 rare HHT-causing variations in the 5'UTR of ENG, all creating overlapping upstream Open Reading Frames (upORFs) initiated with upstream AUG, we here performed an exhaustive *in silico* analysis of all possible single nucleotide variants (n=328) predicted to create or modify any type of upORF in the 5'UTR of ENG. We demonstrated that 85% (11/13) of variants creating uAUGs in frame with the same stop codon located at position c.125, decrease the Endoglin levels *in vitro.* We identified the moderate effect on ENG of a rare uCUG-creating variant found in HHT patients. Our obtained experimental results were in partial correlation with bioinformatics predictions based on Kozak sequence and PreTIS scores.

In parallel, we leveraged results from large scale plasma proteogenomics resources and identified 8 loci (*ABO, ASGR1, B3GNT8, ENG, HBS1L, NCOA6, PLAUR,* and *TIRAP*), presenting common polymorphisms, significantly associated with Endoglin levels. The ABO locus, coding for the ABO blood groups, explain ~5% of the inter-individual variability of ENG plasma levels. Overall, these loci candidates could contribute to explain the incomplete penetrance of known pathogenic mutations and/or the clinical heterogeneity of HHT patients. Of note, 4 of these loci are also associated with venous thrombosis in the latest INVENT Genome Wide Association Study initiative.

This project brings new insights on the interpretation of *ENG* non-coding variants and on molecular mechanisms participating to the regulation of Endoglin. It also exemplifies how the incorporation of genotype data on common polymorphisms could enhance the management of rare diseases.

Introduction

Endoglin, also known as CD105, is a transmembrane glycoprotein playing a major role in endothelial cells and highly involved in angiogenesis[1,2]. Endoglin has mainly been known as an auxiliary receptor of ALK1 (encoded by the *ACVRL1* gene) in response to bone morphogenetic proteins (BMPs), mainly BMP9 and BMP10, belonging to the transforming growth factor β (TGF-β) superfamily[3–5]. In that context, the *ENG* gene coding for Endoglin is, together with *ACVRL1* and *SMAD4* genes, one of the 3 major genes for Hereditary Hemorrhagic Telangiectasia (HHT), a multiorgan rare vascular disease[6]. More recent works have highlighted new functions for Endoglin[7], in particular in leukocyte adhesion and transmigration under inflammatory conditions[8], but also in platelets biology[9] and cancer-associated fibroblasts[10]. There is then accumulating evidence that Endoglin and its soluble form (S-Endoglin) are involved in many human diseases[11] such as thrombosis[9], coronary atherosclerosis[12], preeclampsia[13], hypertension[14], autoimmune disease[15], some cancers[16], and beyond. Notably, elevated levels of S-Endoglin have been detected in specific conditions such as hypertension and coronary atherosclerosis [14,17,18]. As a consequence, dissecting the genetic regulation of Endoglin holds the potential not only to enhance molecular diagnosis and clinical management for HHT patients but also to deepen our understanding of the pathophysiology underlying more common complex diseases associated with ENG.

The *ENG* gene is located on chromosome 9q34.11 and loss-of function *ENG* variants are responsible for HHT, also known as Osler-Rendu-Weber disease. HHT-causing ENG variants have been shown to be associated with mild to severe decrease of Endoglin levels in functional assays[19,20]. HHT is characterized with very heterogeneous clinical manifestations between patients, thus requiring a multidisciplinary approach for the clinical diagnosis and treatment[21–23]. The clinical diagnosis of HHT is based on Curaçao criteria corresponding to epistaxis, multiple telangiectasias at characteristic sites (lips, oral cavity, fingers and nose), visceral vascular lesions including gastrointestinal telangiectasia and/or arteriovenous malformations (AVMs), and family history[24]. Frequent and repeated nose bleeding

and/or digestive telangiectasias can lead to anemia and visceral AVMs lead to vascular shunts, strokes and cerebral abscess secondary to pulmonary AVMs and high cardiac output secondary to liver AVMs[24].

For a long time, the search for pathogenic *ENG* mutations in HHT patients was mainly restricted to exonic/flanking intronic regions[6,19,20,25]. However, we and others have recently reported 5 pathogenic variations in the 5'UTR of *ENG* identified in HHT patients[26–31] and associated with a decrease of Endoglin levels and of the response to BMP9[26]. All these variations create upstream AUGs (uAUGs) in frame with the same stop codon in position c.125 (uStop-c.125), generating overlapping upstream Open Reading Frames (uoORFs). These data then suggested that *ENG* could be enriched in HHT-causing uAUG-creating variants and that the created uAUGs are able to initiate the translation. Moreover, it is now well known that non-canonical translation initiation site (TIS) differing of one nucleotide from AUG codon could also be recognized by the ribosomes[32–34]. As such, upstream ORFs (upORFs) starting with non-canonical TIS could also contribute to regulate the translation of the coding sequence (CDS). Note that upORFs naturally exist in coding transcripts and are main regulators of translation[35–37]. As a consequence, any genetic variations that create new stop codons or delete existing ones in existing natural upORFs could impact CDS translation and associate with genetic diseases[38,39].

In order to provide a comprehensive overview of genetic variations altering upORFs in the 5'UTR of *ENG*, we here performed an exhaustive *in silico* analysis of all possible single nucleotide variations (SNVs) in the 5'UTR of ENG (ENST00000373203.9; NM_001114753.3) that could create canonical and non-canonical upstream TIS (uTIS), create new stop codons, and/or delete existing upstream stop codons (uStop) by using an updated version of the bioinformatics tool MORFEE[40]. Such upORF-altering variants were then looked for in a national collection of patients with unresolved molecular diagnosis for HHT. Finally, the functional impact on Endoglin levels of the detected variants as well as of all additional variants creating uAUGs in frame with the stop codon at position c.125 was assessed. Again, there is a large inter-individual variability in the clinical manifestations and complications of HHT, even in HHT patients from the same family and/or carrying the same pathogenic ENG mutations[22,23].

Hypothesizing that such variability could be partially due to (common/rare) variants outside *ENG*, we also leveraged in this work results from large scale plasma proteogenomics resources where Endoglin has been measured to identify new candidate variants/genes regulating plasma Endoglin levels. Such candidates could then contribute to explain the incomplete penetrance of known pathogenic mutations and/or the clinical heterogeneity of HHT patients.

Materials and Methods

Nomenclature

DNA sequence variant nomenclature follows current recommendations of the HGVS[41].

In silico mutational saturation of the 5'UTR of ENG and search for upORF-altering variants

We *in silico* mutated each position in the 5'UTR of the main transcript of ENG (MANE select, ENST00000373203.9) reported in the latest version of Ensembl database (GRCh38.p14) with the 3 alternative nucleotides to generate a vcf file containing all possible single nucleotide variations (SNVs) between positions c.-303 (i.e., first nucleotide of the 5'UTR) and c.-1 (Supplementary Table 1). The generated vcf file was then annotated using an updated version of the MORFEE bioinformatics tool[26,42] now available on https://github.com/CarolineMeg/MORFEE. Compared to its initial version that abled to annotate variants creating upstream AUGs or deleting upstream stop codons, the current version can annotate variations predicted to (i) create canonical and non-canonical TIS; (ii) create new stop codons (TAA, TAG and TGA); and/or (iii) delete existing stop codons along a given transcript. The resulting list of all possible SNVs creating or altering upORFs in the 5'UTR of ENG is provided in Supplementary Table 2.

To identify variations that could alter naturally existing upORFs in the 5'UTR of ENG, we extracted known upORFs from public databases reporting small ORFs that have been identified through ribosome profiling and/or mass spectrometry in human cells. Assessed databases included: sORFs repository (www.sORFs.org)[43] ; metamORF database (https://metamorf.hb.univ-amu.fr/)[44] ; vUTR

interface (https://vutr.rarediseasegenomics.org/) ; and smoRFs browser (https://smorfs.ddnetbio.com)[45].

Selection of *ENG* variants for experimental validation

Five rare variants creating uAUGs in frame with the stop codon at position c.125 have been identified in HHT patients and have shown drastic effects on ENG protein levels[26]. To get a more general view of the possible impact on Endoglin of this specific kind of variation in the 5'UTR of ENG, we selected all additional 8 variations identified by MORFEE to create uAUGs in frame with the stop codon at position c.125 for experimental validation (Supplementary Table 2).

We also selected for experimental validation the c.-76C>T variation. This variation was the only one from Supplementary Table 2 creating a non-canonical uTIS in frame with the c.125 stop codon that was further identified in a collection of HHT patients with unresolved molecular diagnosis.

Plasmid constructs

Except for the c.-287C>A variant located at the beginning of *ENG* transcript close to BamHI restriction site, preparation of pcDNA3.1-L-ENG constructions was performed by directed mutagenesis on pcDNA3.1-L-ENG-WT construct[26] using the 2-step overlap extension PCR method[46] and primers listed in Supplementary Table 3. BamHI and SacII or BamHI and BlpI were used as cloning sites, depending on the inserted variant (Supplementary Table 3). All new constructs were verified by sanger sequencing of the insert and cloning sites.

Functional analysis of ENG 5'UTR variants

Transfection of HeLa cells, RNA and protein extractions as well as western blot and RT-qPCR analyses were carried out as described in Soukarieh *et al*., 2023[26]. Briefly, HeLa cells were cultured in 6 well plates and transfected with 1µg of each of the mutant constructions, in parallel with the wild-type (WT) construct and an empty vector as a negative control. Forty-eight hours after transfection, proteins and RNA were extracted from the same well and analyzed by western blot and RT-qPCR, respectively. Antibodies against the myc tag detecting the exogenous ENG and β-actin were used during the western

blot analysis and ENG levels were normalized on β-actin levels. RT-qPCR was performed with specific ENG and α-tubulin primers (Supplementary Table 3) and relative amounts of ENG to α-tubulin were quantified following the $2^{-\Delta\Delta Ct}$ method.

Kozak sequence interpretation and bioinformatics predictions

Different bioinformatics tools have been developed to predict the translation efficiency of a given uTIS using in particular the strength of the Kozak sequencing encompassing the uTIS. We here used the obtained results in our functional assays on *ENG* variants (n = 14, Table 1) to evaluate the predictive power of the strength of the Kozak sequence as well as of predictive scores provided by TIS-predictor[47] and PreTIS[48] tools.

The Kozak sequence is defined as the genomic sequence surrounding a TIS with nucleotides at positions -3 and +4 regarding the TIS being the most important for translation initiation. The optimal Kozak sequence is [**A/G**]CCATG**G**, underlined nucleotides corresponding to the TIS and bolded nucleotides to the most conserved positions. We have here considered a given Kozak sequence as (i) strong when it contains a purine at position -3 AND a guanine at position +4; (ii) moderate when it contains a purine at position -3 OR a guanine at position +4, and; (iii) weak when it does not contain a purine at position -3 nor a guanine at position +4. We extended our evaluation of the importance of the Kozak sequence by using TIS-predictor[47], a recently developed bioinformatics tool that computes a Kozak Similarity Score (KSS) to evaluate the strength of Kozak sequence based on ± 10 nucleotides surrounding TIS. KSS ranges from 0 and 1. Canonical and non-canonical TIS with KSS higher than 0.64 and 0.61, respectively, have been proposed to have high potential to initiate the translation [47].

Finally, we also computed PreTIS scores from https://service.bioinformatik.uni-saarland.de/pretis/ for annotated variants located > 99 nucleotides from the beginning of ENG 5'UTR. PreTIS is another recent tool predicting the translation efficiency of a given TIS in the 5'UTR of coding genes[48]. It is based on ribosome profiling data and takes into account 64 criteria (e.g., conservation, upORF size, secondary structure) to calculate scores for translation efficiency.

Statistical analysis of *in vitro* data

Differential protein and RNA levels according to tested variants were assessed using analysis of variance followed by Tukey's multiple comparison test. A threshold of $p < 0.05$ was used to declare statistical significance.

Plasma proteogenomics investigations for common polymorphisms associated with Endoglin levels

To identify common polymorphisms associated with plasma Endoglin levels, we meta-analyzed Genome Wide Association Study (GWAS) summary statistics from 2 large scale proteogenomics resources where Endoglin have been measured. These resources include 10,708 participants from the Fenland study[49] and 35,559 Icelander participants of the Decode project[50] with Endoglin plasma levels measured using the Somalogic platform. The meta-analysis of these GWAS results was performed using the METAL software implementing the Z-score fixed-effect model[51]. The heterogeneity of genetic associations across studies was assessed using the Cochran-Mantel-Haenszel statistical test and its magnitude was expressed in terms of I^2 [52]. Only associations with moderate heterogeneity $I^2 < 30\%$ were considered.

Capitalizing on the results of this meta-analysis as well as those of the GWAS analysis on VT risk performed by the INVENT consortium[53] , we deployed colocalization[54] and two-sample Mendelian Randomization (MR)[55,56] analyses to assess a possible causal role of plasma endoglin on VT risk. Using the Coloc R package, we applied colocalization at each locus significantly associated ($p < 5\ 10^{-8}$) with Endoglin plasma levels using all single nucleotide polymorphisms (SNPs) located at $\pm$ 100 kb from the lead variant. Using the lead SNPs at each locus associated with Endoglin level as instrument variables, we deployed several MR methodologies (Inverse Variance Weighted[57], Weighted Median[58], Egger[59] methods) implemented in the TwoSampleMR R package (verion 0.5.7).

To fine map genomic findings obtained from this meta-analysis, we used individual data from an additional plasma proteogenomic resource consisting of a sample of 1,056 population-based participants of the 3C-Dijon study[60] with GWAS data and profiled on the Olink Explore 3072 panel. Olink

proteomic profiling was conducted on EDTA plasma tube of more than 500 µL not thawed/refrozen using the PEA technology following the manufacturer's protocol[61]. Profiling was performed at the McGill Genome Center (Montreal, Canada). Pre-processing of the proteomic data included plate-based normalization and QC checks based on appropriate Olink protocols. Data were transformed and normalized to Olink's Normalized Protein eXpression (NPX) values, relative protein quantification unit in a logarithmic base 2 scale. Twelve samples were removed after principal component analysis of all proteins because they are found to deviate of more than 5 standard deviations from the mean. Besides, three proteins were removed as more than 50% of NPX values were below the protein's detection limit. After exclusion of 3 participants presenting with extremely low outlier Endoglin values, 966 3C-Dijon participants with GWAS data remained for fine mapping analysis.

In 3C-Dijon, association of common polymorphisms with Endoglin levels was conducted on centered and standardized NPX values adjusted for age and sex.

To estimate the genetic correlation between plasma levels of Endoglin and of the PLAUR protein identified from the aforementioned plasma proteogenomics investigations, we used the Linkage Disequilibrium Score regression approach[62,63] implemented in the LDSC package (https://github.com/bulik/ldsc). This method was applied to Endoglin and PLAUR protein summary GWAS statistics separately from the Fenland and Decode studies. The Fisher z-transformation[64] was then applied to the obtained genetic correlation estimates. The two resulting z coefficients were then meta-analyzed using the fixed effect Mantel-Haenszel methodology[52] and the combined z coefficient was then transformed back to obtain a combined estimate of the genetic correlation between plasma levels of Endoglin and PLAUR protein.

Results

Identification of all possible upORF-altering variations in in the 5'UTR of ENG

The *in silico* mutational saturation analysis of the 5'UTR of ENG (ENST00000373203.9; NM_001114753.3) generated a total of 909 possible SNVs (Supplementary Table 1). Among these

variants, 328 were predicted by MORFEE to create any type of uTIS, uStop, and/or to delete existing uStop. In total, 360 different upORFs could be generated by these 328 variations (Figure 1a, Supplementary Table 2). More precisely, 255 variants have been predicted to create new uTIS, 30 to create new uStop and 12 to delete existing uStop (Figure 1a). The remaining 31 SNVs show multiple consequences (Figure 1a; Supplementary Table 4): 28 of them are predicted to create uTIS and new stop codons at the same time, 2 variants create uTIS and delete existing uStop, and one variant seem to simultaneously create a uTIS, delete a uStop and create a new one.

At least 8 upORFs naturally existing in the 5'UTR of ENG and resulting from experimental studies conducted in human cells have been reported in databases. While 7 of these upORFs are initiated with a CUG start codon, one is initiated with a GUG start codon (c.-68). Five of these upORFs end with the stop codon located at c.-166, 2 with the stop codon at position c.-33 and 1 is overlapping ending at the uStops-c.125 (Supplementary Figure 1a). Interestingly, MORFEE annotated 8 variations that could delete the uStop codon located at position c.-166, and 7 that delete the one at position c.-34, thus elongating these existing upORFs into either longer fully upstream ORFs (uORFs) ending at positions c.-34 or into uoORFs ending at new stop codon c.90, respectively (Supplementary Figure 1b). By contrast, MORFEE identified 17 variations that could create new stop codons then shortening existing upORFs reported in databases (i.e., starting with uTIS at positions c.-280, c.-268, c.-265, c.-262, c.-256, c.-139, c.-109 or c.-68) and 42 that could create new stop codons at the origin of new upORFs (Supplementary Figure 1c). None of the annotated uStop-deleting or new stop-creating variants in the 5'UTR of ENG are reported in ClinVar. However, 2 of those generating new upORFs by the creation of new stop codons (c.-253C>A and c.-87C>T) have been reported as rare in GnomAD V4 0.0 database (https://gnomad.broadinstitute.org/) with the c.-87C>T showing double consequence (Supplementary Table 4).

Finally, 286 SNVs creating canonical or non-canonical uTIS could be at the origin of 294 upORFs, with the understanding that a given SNV may create different uTIS (Figure 1b; Supplementary Table 5).

Among these upORFs, 86 are fully located in the 5'UTR (uORFs) ending at 2 different stop codons (c.-166 and c.-34), 122 are overlapping with the CDS (uoORFs) ending with stop codons at position c.90 or c.125, and 86 correspond to elongated CDS (eCDS) ending at the main stop codon (Figure 1c). Interestingly, 18 of these uTIS-creating variants are reported in ClinVar as associated with HHT, and 13 of them are classified as variants of unknown significance or with conflicting interpretations (Supplementary Table 6). In addition, 14 additional uTIS-creating variants have been reported in GnomAD V4 0.0 database with allele frequency lower than 0.01% and without any evidence of association with HHT (Supplementary Table 7).

A high proportion of *ENG* SNVs creating uAUG in frame with the uStop-c.125 drastically alter the protein levels

Recently, we have demonstrated that 5 *ENG* 5'UTR variants (c.-142A>T, c.-127C>T, c.-79C>T, c.-68G>A and c.-10C>T) identified in HHT patients and creating uAUGs leading to uoORFs of various length all ending at the same stop codon at position c.125, were responsible of decreased ENG protein levels and activity[26] revealing their pathogenicity. In order to extrapolate whether these deleterious effects also hold for any other uAUG-creating SNVs at the origin of uoORFs ending at the uStop-c.125, we conducted the same experimental work[26] on the remaining 8 variations (c.-287C>A; c.-271G>T, c.-249C>G, c.-182C>A, c.-167C>A, c.-37G>T, c.-33A>G and c.-31G>T) predicted by the MORFEE *in silico* analysis (Figure 2a; Table 1).

For two variations (c.-287C>A and c.-37G>T), no decrease of protein levels was observed (Fig. 2b-c). While the c.-287C>A variant showed similar protein levels comparing to the wild-type, the c.-37G>T tended to associate with an increase of ENG levels in our assay (Figure 2b-c; Supp. Figure 2a). Of note, no significant difference of ENG transcript levels was observed in our RT-qPCR experiments between mutants and wild-type for these 2 variants (Supp. Figure 2b). Interestingly, the c.-37G>T variant is predicted to simultaneously create a uAUG-initiated uoORF and to shorten an existing upORF (Supplementary Table 4).

All other 6 variations were associated with a protein level lower than 40% compared to the wild-type construct. More precisely, c.-271G>T, c.-249C>G, c.-182C>A and c.-33A>G led to a drastic decrease of the protein levels (remaining amount ≤ 20%), while c.-167C>A and c.-31G>T showed a less drastic effect (remaining amount ≤ 30-40%). No significant difference was observed in our RT-qPCR experiments between mutants and WT (Supplementary Figure 2b).

In total, 85% (11/13) of uAUG-creating variants at the origin of uoORFs ending at position c.125 (those studied in the current work together with those previously investigated[26]) decrease the ENG protein levels *in vitro*.

Bioinformatics predictions partially correlate with the observed functional effects

Among the 13 5'UTR variants we experimentally studied, all leading to uAUGs at the origin of uoORFs ending at stop codon c.125, 6 are characterized by a moderate Kozak sequence (Table 1). These 6 variants (c.-271G>T, c.-249C>G, c.-142A>T, c.-127C>T, c.-33A>G and c.-31G>T) were all associated with decreased protein levels in our assay with 5/6 variants showing the most drastic effects. The 7 remaining variants were associated with a weak Kozak sequence and 5 of them (c.-182C>A, c.-167C>A, c.-79C>T, c.-68G>A and c.-10C>T) were associated with a decrease of the protein levels. Finally, the 2 variants with no decrease of Endoglin protein levels (c.-287C>A and c.-37G>T) are associated with weak Kozak sequences. Our analysis suggests that the sole information on the Kozak sequence of uTIS cannot be used to predict the potential effect of uAUG-creating variants in ENG.

We extended the predictions of the Kozak sequence by applying predictions from TIS-predictor, which takes into account the 10 nucleotides surrounding a given canonical or non-canonical TIS to generate KSS scores reflecting the strength of Kozak sequences. The authors defined a threshold of 0.64 for translation initiation by uAUGs. We extracted KSS scores for our 13 uAUG-creating variants of interest (Table 1; Supplementary Table 2) and found that 6/13 uAUGs have scores higher than 0.64. These 6 variants (c.-271G>T, c.-249C>G, c.-142A>T, c.-127C>T, c.-79C>T and c.-33A>G) were observed to decrease the protein levels of Endoglin in our experiments. Interestingly, these variants are those

showing the most drastic effects in our assays. However, 5/11 variants decreasing the protein levels (c.-182C>A, c.-167C>A, c.-68G>A, c.-31G>T and c.-10C>T) presented with KSS scores below the recommended threshold of 0.64 as the two variants with no decrease of protein levels.

Furthermore, we evaluated the efficiency of PreTIS scores to predict uTIS that can alter the protein levels. These scores predict the efficiency of a given uTIS to initiate the translation. We hypothesized that the highest PreTIS scores predicting the most efficient uTIS would be associated with the most drastic effect on the protein levels. We collected PreTIS scores for the 10/13 uAUGs created by 5'UTR variants in ENG. The remaining 3 variants being located within the first 99 nucleotides of the 5'UTR (c.-287C>A, c.-271G>T and c.-249C>G), they cannot be predicted with PreTIS. The totality of the predicted uAUGs (10/10) has PreTIS scores higher than the predefined threshold for translation efficiency (0.54) (Table 1). Importantly, 9 are associated with a decrease of the protein levels in our assays but the remaining one (c.-37G>T) increases Endoglin levels. These observations suggest that while PreTIS could be efficient to predict which uTIS creating variant could have a functional impact, it cannot predict the direction of its molecular effect.

New uoORF-creating variants identified in HHT patients: contribution to molecular diagnosis

By crossing the results of our *in silico* mutational saturation of the 5'UTR of ENG with the list of variants identified in HHT patients from the French National reference center for HHT with unresolved molecular diagnosis, we identified 2 uTIS-creating variants in HHT patients. The first one was the aforementioned c.-33A>G variant creating a uAUG predicted to generate an uoORF ending at the c.125 codon. This variant, never reported in public databases, was identified in a patient with definite HHT according to Curaçao criteria and the experimental study described above (Figure 2b-c) provided strong argument for its pathogenicity. The second variant was the uCUG-creating c.-76C>T variant also predicted to generate an uoORF ending at the c.125 codon (Figure 3a). This variant (rs943786398) was detected in 2 unrelated patients with suspected HHT (Supplementary Table 8) and has been classified as variant of unknown significance in ClinVar with no evidence about its potential functional effect on

Endoglin. The proband in the first family had an atypical presentation for HHT with stroke and deep vein thrombosis associated with few telangiectasias. In the second family, the proband presented with pulmonary AVM and the father was an asymptomatic carrier of the variant (normal Ct scan and no clinical signs at over 70 years of age). Following the same experimental workflow as above, we observed that the c.-76C>T variant was associated with decreased ENG levels of more than 25% in comparison with the WT (Figure 3b-c). No significant difference of ENG transcript amounts was observed between c.-76C>T and WT by RT-qPCR (Supplementary Figure 2c). Very interestingly, the uCUG created by *ENG* c.-76C>T, associated with a moderate decrease of ENG levels, is encompassed by a strong Kozak sequence (**A**CGCTG**G**) and carries very high scores with TIS-predictor and PreTIS (0.87 and 0.93, respectively; Table 1). Five additional variants creating non-canonical uTIS in frame with the stop codon at position c.125 have been reported in ClinVar (Supplementary Table 6).

Common polymorphisms associate with Endoglin plasma levels

By meta-analyzing results from two GWAS on Endoglin plasma levels in up to 46,091 healthy individuals, we identified 8 loci significantly ($p < 5.10^{-8}$) associated with Endoglin levels (Figure 4, Supplementary Table 9). Among them, *ABO* on chr9q34.2 was the locus most significantly associated with Endoglin plasma levels ($p = 4.25.10^{-262}$, $I^2 = 0\%$), with rs558240 being the lead SNP. The second associated locus mapped to 19q13.31 with the *PLAUR* non synonymous rs4760 (p.Leu371Pro) being the lead SNP ($p = 1.14.10^{-158}$, $I^2=10.1\%$). The structural *ENG* gene on chr9q34.11 ranked in third position with the intronic rs10987756 as lead SNP ($p = 2.35.10^{-43}$, $I^2=0\%$). The remaining five loci associated with Endoglin plasma levels were *TIRAP* on chr11q24.2 (rs8177398, intronic, $p=8.82.10^{-14}$, $I^2=18\%$), *NCOA6* on chr20q11.22 (rs73106997, intronic, $p= 1.12.10^{-12}$, $I^2=0\%$), *B3GNT8* on chr19q13.2 (rs284660, p.Ala200Ala, $p=8.46.10^{-10}$, $I^2=0\%$), *ASGR1* on 17p13.1 (rs67143157, downstream, $p=1.83.10^{-8}$, $I^2=20.3\%$) and *HBS1L* on chr6q23.2 (rs1547247, intronic, $p =1.98.10^{-8}$, $I^2= 19.1\%$).

The *ABO* locus codes for the ABO blood groups whose main groups can be genetically characterized by rs2519093 (A1), rs1053878 (A2), rs8176743 (B), rs8176719 (O1) and rs41302905 (O2)[65]. To further

clarify the association observed at the *ABO* locus, we investigated the association of ABO blood groups with Endoglin plasma levels in 966 heathy individuals of the 3C study[60] in which Endoglin was measured using the Olink panel technology[61]. The pattern of associations is shown in Figure 5. To summarize, in this healthy population, assuming additive allele effects, the ABO B group was associated with increased Endoglin levels (β = + 0.50 ± 0.088, p = $1.8.10^{-8}$) compared to the O1 blood group whereas the A1 group was associated with decreased levels (β = - 0.23 ± 0.056, p = $5.12.10^{-5}$). Altogether, ABO blood groups explained 5.8% of the interindividual variability in Endoglin plasma levels. An additional 1.8% was explained by the 7 lead SNPs at the other genome-wide significant loci.

It is worth noting that the aforementioned *ASGRL1* rs67143157 we observed to associate with Endoglin levels also associated with plasma levels of Plasminogen Activator, Urokinase Receptor encoded by *PLAUR*, both in the Decode study (p=$5.64.10^{-16}$) and in Fenland (p=$4.22.10^{-5}$). As PLAUR was also available on the Olink panel used in 3C-Dijon, we assessed the biological correlation between PLAUR and ENG plasma levels. After adjusting for age and sex, both protein plasma levels strongly correlated with each other (ρ=+ 0.37, p=$2.18.10^{-33}$). Consistently, the genetic correlation estimated between these two proteins as derived from summary GWAS statistics was ρ=0.24 (p <10^{-16}) (ρ = 0.54 and ρ = 0.15 in Fenland and Decode, respectively).

As mentioned in the Introduction, there is accumulating evidence that Endoglin participates to thrombus formation in a platelet-dependent manner[7,9]. It is then relevant to note that 4 of the 8 genome wide significant loci found to associate with plasma Endoglin levels exhibited suggestive statistical evidence for association with venous thrombosis (VT) in the latest INVENT GWAS initiative[53]. In addition to well established VT-associated ABO locus, these loci include *TIRAP* rs8177398 (p=$2.1.10^{-7}$), *NCOA6* rs73106997 (p=$6.3.10^{-6}$) and *ENG* rs10987756 (p=0.0051) (Supplementary Table 10). For *TIRAP* only, colocalization analysis (PP4>0.90) suggests the variant influencing Endoglin levels is the one associated with the risk of VT (Supplementary Table 10). Consistently, MR analyses were not supportive for a causal association between increased endoglin levels and VTE (Supplementary Table 11).

Discussion

For several years, the search for pathogenic variations in *ENG* was mainly restricted to those located within coding exons or splice sites and implicated in HHT[6,19]. However, our group and others shed light on 5'UTR *ENG* variants identified in HHT patients and acting as loss-of-function variants leading to Endoglin deficiency[26–31]. All these 5'UTR variants are predicted to create uAUGs at the origin of overlapping upORFs terminating at the same position in the CDS. This is in line with the bioinformatics analysis of Whiffin and collaborators revealing the ENG between genes with high-impact upORF-altering variants (i.e., uAUG-creating and uStop-deleting variants) that are more likely to be deleterious[39]. Based on these data, we here extended the bioinformatics analysis of 5'UTR variants to those also creating non-canonical TIS or new stop codons or deleting stop codons in the ENG. This was achieved by cataloguing all possible, already reported or yet unreported artificial ones, SNVs creating or modifying upORFs in the 5'UTR of ENG with the aim at facilitating the identification of pathogenic variants. Interestingly, among *ENG* variants reported in ClinVar, 43 are located in the 5'UTR (~ 14%) with 18 (~42%) that have been annotated with MORFEE as creating uTIS. 2022Inserm1034!

We here further demonstrate that 6 out of 8 additional variations creating uAUG-initiated uoORFs ending with the Stop codon located at position c.125 also associate with decreased Endoglin levels *in vitro*. One of these variations, c.-33A>G, was identified in a French patient with definite HHT. In addition, we experimentally demonstrate that the rare c.-76C>T variant identified in 2 unrelated suspected HHT patients was associated with a moderate decrease in Endoglin levels. This variant creates a uCUG-initiated uoORF also ending at stop codon c.125 suggesting that rare variants predicted to create non canonical uTIS in frame with the stop codon at position c.125 should be considered as candidates for causing HHT. For instance, 5 remaining variants (c.-249C>A, c.-188G>A, c.-98G>A, c.-91C>T and c.-70C>T) associated with HHT and creating the same type of upORFs are reported in ClinVar and should be considered (Supplementary Table 6). Unexpectedly, one of the tested uAUG-creating

variants, c.-37G>T, is associated with higher levels of Endoglin in our assay. Interestingly, in addition to the creation of a uAUG, this variant is predicted to simultaneously create a stop codon (uUGA). The predicted stop codon is in frame with several non-canonical uTIS (a uAGG at position c.-142 and 4 uCUGs at positions c.-139, c.-109, c.-52 and c.-46) (Supplementary Table 4). Noteworthy, 2 of these in frame uTIS (c.-139 and c.-109 CUGs) have been reported in databases to initiate natural uORFs (Supplementary Figure 1a). Consequently, the c.-37G>T could shorten these two natural uORFs. Of note, the one at position c.-139 is encompassed with strong Kozak sequence and has high KSS and PreTIS scores (Supplementary Table 10). Whether this uORF modification could explain the increase of protein levels (i.e., by increasing downstream translation re-initiation) we observed in our *in vitro* models still need to be explored. Finally, even if rare 5'UTR ENG variants responsible for increased Endoglin levels cannot explain HHT, they should not be neglected in other vascular diseases (i.e., coronary atherosclerosis)[12,18] and solid and hematological cancers[66,67] in which increased levels of Endoglin have been observed. Noteworthy, the bioinformatics predictions and experimental validation of a given variant could be challenging for variants with multiple consequences on upORFs.

In total, we experimentally analyzed 13 uAUG-creating and 1 uCUG-creating variant in the 5'UTR of ENG. We used this modest number of variants to evaluate the predictive potential of Kozak score strength and of the 2 uTIS predictive scores, KSS and PreTIS. In summary, none of the three bioinformatics metrics we evaluated were able to perfectly discriminate the observed impacts (increase, decrease or null effect) on protein levels of the tested variants. However, using at least two positive predictions (moderate/strong Kozak sequence or high KSS or PreTIS score > 0.54) allowed to identify 8 out of 12 (~67%) variants associated with decreased Endoglin levels without any false positive. But this observation needs to be extended on a larger number of variants with various molecular effects. Once optimized, such predictions applied to *ENG* variants identified by MORFEE (Supplementary Table 2) could be used to prioritize pathogenic candidates in ENG for functional validation.

HHT is well known to exhibit a large heterogeneity in its clinical manifestations including in patients carrying the same variant[22]. The source of such heterogeneity has been proposed to be attributable, at least partially, to either a second hit variant in *ENG* or to some modifier genes[6]. Our present work on rare variants modifying upORFs in the 5'UTR of the ENG suggests that such variations, with strong or moderate effects, could indeed be good candidates to look at in the context of molecular diagnosis. However, our plasma proteogenomic work suggests that more frequent variants should also be taken into account in the molecular diagnosis of HHT. Indeed, we here demonstrated that at least 8 loci (*ABO, ASGR1, B3GNT8, ENG, HBS1L, NCOA6, PLAUR,* and *TIRAP*) present common polymorphisms (allele frequency > 1% in the general population) influencing plasma Endoglin levels. In particular, we observed that the *ABO* locus explains ~5% of inter-individual variability with the highest and lowest plasma Endoglin levels observed in individuals carrying B and A1 blood groups, respectively. This suggests that ABO blood groups should be considered in the context of clinical diagnosis for HHT, especially for explaining incomplete penetrance of identified (likely) pathogenic *ENG* variations or those associated with moderate effect as we here observed for the c.-76C>T variation associated with incomplete penetrance in one HHT family. Our hypothesis is that patients carrying an *ENG* variation responsible for a moderate decrease of Endoglin may be more prone to definite HHT when they are of ABO A1 group compared to non-A1 group. The remaining 7 loci jointly explain less than 2% of the variability of Endoglin plasma levels. Before envisaging any translation of these results in the clinical management of HHT patients, efforts are needed to characterize the exact molecular mechanisms/variants underlying the observed associations. Nevertheless, our results pinpoint to a new role of PLAUR in the regulation of Endoglin. PLAUR codes for the cellular receptor of the urokinase-type plasminogen activator (uPAR), a key component of the plasminogen activation system, known to play a major role in thrombosis but also in inflammation, cell migration and adhesion and vascularization[68–70]. These results not only provide additional support for the involvement of Endoglin in hemostasis but also open novel research avenues to dissect the exact biological and molecular links between uPAR and Endoglin.

To follow up on these observations, international efforts are needed to perform genome wide association studies on HHT and other ENG-related diseases and to integrate their results within the framework of Mendelian Randomization studies in order to facilitate the identification of causal biomarkers between Endoglin levels and HHT as well as the repositioning of existing drugs. These analyses could have inputs to find new therapeutic perspectives. For instance, PLAUR has been reported as a pharmaceutical target for Ruxolitinib (a family member of JAK inhibitor) on the context of myelofibrosis and essential thrombocythemia in the Therapeutic Target Database[71]. Similarly, *ASGR1* is currently studied as a pharmaceutical target for AMG 529 in the context of diseases of the circulatory system including essential hypertension.

To conclude, our work provides new insights on the interpretation of ENG non-coding variants and related frequent ones. It has direct implication in HHT and will contribute to better understand the implication of ENG in other vascular diseases and cancers. Finally, this project could be used as a proof-of-principle for other HHT genes (i.e., *ACVRL1*, *SMAD4*) and for any gene implicated in rare diseases.

Acknowledgments

This project was carried out in the framework of the French National Research Agency (ANR) ANR-23-CE17-0042-01 program as part of the ENDOMORF project and of the INSERM GOLD Cross-Cutting program (D-A.T.). O.S was financially supported by a grant of the Lefoulon-Delalande Foundation. The 3C proteomics project was supported by a grant overseen by the ANR as part of the “Investment for the Future Program” ANR-18-RHUS-0002 and by the Precision and Global Vascular Brain Health Institute (VBHI) funded by the France 2030 IHU3 initiative.

Statistical analyses benefited from the CBiB computing centre of the University of Bordeaux.

Data availability

ENG constructs generated during this study are available upon request by email from the corresponding author (omar.soukarieh@inserm.fr).

Code Availability

The used version of MORFEE tool is available at https://github.com/CarolineMeg/MORFEE.

Author contributions

OS and DAT conceived the project. OS and CP designed the experiments. OS, CP and CD performed the experiments. OS and CP analyzed the data. BJV and AG provided technical support and suggestions on the project and the experiments. SDG and MT were in charge of clinical management of HHT patients. CM performed the mutational saturation and variant annotation with MORFEE. GM performed analyzed proteogenomics data. IC performed bioinformatics analysis on the 3C study under the supervision of SD. OS and DAT drafted the paper that was further shared to co-authors who read/corrected/ and approved the final manuscript.

Competing interests

The authors declare that they have no known competing financial or non-financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Rossi, E., Bernabeu, C. & Smadja, D. M. Endoglin as an Adhesion Molecule in Mature and Progenitor Endothelial Cells: A Function Beyond TGF-β. *Front Med (Lausanne)* **6**, 10 (2019).
2. López-Novoa, J. M. & Bernabeu, C. The physiological role of endoglin in the cardiovascular system. *Am J Physiol Heart Circ Physiol* **299**, H959-974 (2010).
3. Nolan-Stevaux, O. *et al.* Endoglin requirement for BMP9 signaling in endothelial cells reveals new mechanism of action for selective anti-endoglin antibodies. *PLoS One* **7**, e50920 (2012).
4. Blanco, F. J. *et al.* Interaction and functional interplay between endoglin and ALK-1, two components of the endothelial transforming growth factor-beta receptor complex. *J Cell Physiol* **204**, 574–584 (2005).

5. Lebrin, F. *et al.* Endoglin promotes endothelial cell proliferation and TGF-beta/ALK1 signal transduction. *EMBO J* **23**, 4018–4028 (2004).
6. Shovlin, C. L. *et al.* Mutational and phenotypic characterization of hereditary hemorrhagic telangiectasia. *Blood* **136**, 1907–1918 (2020).
7. Rossi, E. & Bernabeu, C. Novel vascular roles of human endoglin in pathophysiology. *J Thromb Haemost* **21**, 2327–2338 (2023).
8. Rossi, E. *et al.* Endothelial endoglin is involved in inflammation: role in leukocyte adhesion and transmigration. *Blood* **121**, 403–415 (2013).
9. Rossi, E. *et al.* Soluble endoglin reduces thrombus formation and platelet aggregation via interaction with αIIbβ3 integrin. *J Thromb Haemost* **21**, 1943–1956 (2023).
10. Paauwe, M. *et al.* Endoglin Expression on Cancer-Associated Fibroblasts Regulates Invasion and Stimulates Colorectal Cancer Metastasis. *Clin Cancer Res* **24**, 6331–6344 (2018).
11. Bernabeu, C., Olivieri, C. & Rossi, E. Editorial: Role of membrane-bound and circulating endoglin in disease. *Front Med (Lausanne)* **10**, 1271756 (2023).
12. Chen, H. *et al.* Negative correlation between endoglin levels and coronary atherosclerosis. *Lipids Health Dis* **20**, 127 (2021).
13. Margioula-Siarkou, G. *et al.* The role of endoglin and its soluble form in pathogenesis of preeclampsia. *Mol Cell Biochem* **477**, 479–491 (2022).
14. Blázquez-Medela, A. M. *et al.* Increased plasma soluble endoglin levels as an indicator of cardiovascular alterations in hypertensive and diabetic patients. *BMC Med* **8**, 86 (2010).
15. Grignaschi, S. *et al.* Endoglin and Systemic Sclerosis: A PRISMA-driven systematic review. *Front Med (Lausanne)* **9**, 964526 (2022).
16. Hakuno, S. K. *et al.* Endoglin and squamous cell carcinomas. *Front Med (Lausanne)* **10**, 1112573 (2023).
17. Cruz-Gonzalez, I. *et al.* Identification of serum endoglin as a novel prognostic marker after acute myocardial infarction. *J Cell Mol Med* **12**, 955–961 (2008).

18. Plasma endoglin as a marker to predict cardiovascular events in patients with chronic coronary artery diseases - PubMed. https://pubmed-ncbi-nlm-nih-gov.proxy.insermbiblio.inist.fr/21667051/.
19. Mallet, C. *et al.* Functional analysis of endoglin mutations from hereditary hemorrhagic telangiectasia type 1 patients reveals different mechanisms for endoglin loss of function. *Hum Mol Genet* **24**, 1142–1154 (2015).
20. Ali, B. R. *et al.* Endoplasmic reticulum quality control is involved in the mechanism of endoglin-mediated hereditary haemorrhagic telangiectasia. *PLoS One* **6**, e26206 (2011).
21. Kritharis, A., Al-Samkari, H. & Kuter, D. J. Hereditary hemorrhagic telangiectasia: diagnosis and management from the hematologist's perspective. *Haematologica* **103**, 1433–1443 (2018).
22. Faughnan, M. E. *et al.* International guidelines for the diagnosis and management of hereditary haemorrhagic telangiectasia. *J Med Genet* **48**, 73–87 (2011).
23. Ola, R. *et al.* Executive summary of the 14th HHT international scientific conference. *Angiogenesis* **26**, 27–37 (2023).
24. Faughnan, M. E. *et al.* Second International Guidelines for the Diagnosis and Management of Hereditary Hemorrhagic Telangiectasia. *Ann Intern Med* **173**, 989–1001 (2020).
25. Sánchez-Martínez, R. *et al.* Current HHT genetic overview in Spain and its phenotypic correlation: data from RiHHTa registry. *Orphanet J Rare Dis* **15**, 138 (2020).
26. Soukarieh, O. *et al.* uAUG creating variants in the 5'UTR of ENG causing Hereditary Hemorrhagic Telangiectasia. *NPJ Genom Med* **8**, 32 (2023).
27. Bossler, A. D., Richards, J., George, C., Godmilow, L. & Ganguly, A. Novel mutations in ENG and ACVRL1 identified in a series of 200 individuals undergoing clinical genetic testing for hereditary hemorrhagic telangiectasia (HHT): correlation of genotype with phenotype. *Hum Mutat* **27**, 667–675 (2006).
28. Damjanovich, K. *et al.* 5'UTR mutations of ENG cause hereditary hemorrhagic telangiectasia. *Orphanet J Rare Dis* **6**, 85 (2011).

29. Kim, M.-J. *et al.* Clinical and genetic analyses of three Korean families with hereditary hemorrhagic telangiectasia. *BMC Med Genet* **12**, 130 (2011).
30. Albiñana, V. *et al.* Mutation affecting the proximal promoter of Endoglin as the origin of hereditary hemorrhagic telangiectasia type 1. *BMC Med Genet* **18**, 20 (2017).
31. Ruiz-Llorente, L. *et al.* Characterization of a family mutation in the 5' untranslated region of the endoglin gene causative of hereditary hemorrhagic telangiectasia. *J Hum Genet* **64**, 333–339 (2019).
32. Ivanov, I. P., Firth, A. E., Michel, A. M., Atkins, J. F. & Baranov, P. V. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res* **39**, 4220–4234 (2011).
33. Kearse, M. G. & Wilusz, J. E. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev* **31**, 1717–1731 (2017).
34. Cao, X. & Slavoff, S. A. Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. *Exp Cell Res* **391**, 111973 (2020).
35. Araujo, P. R. *et al.* Before It Gets Started: Regulating Translation at the 5' UTR. *Comp Funct Genomics* **2012**, 475731 (2012).
36. Dever, T. E., Ivanov, I. P. & Hinnebusch, A. G. Translational regulation by uORFs and start codon selection stringency. *Genes Dev* **37**, 474–489 (2023).
37. Brito Querido, J., Díaz-López, I. & Ramakrishnan, V. The molecular basis of translation initiation and its regulation in eukaryotes. *Nat Rev Mol Cell Biol* (2023) doi:10.1038/s41580-023-00624-9.
38. Lee, D. S. M. *et al.* Disrupting upstream translation in mRNAs is associated with human disease. *Nat Commun* **12**, 1515 (2021).
39. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat Commun* **11**, 2523 (2020).

40. Aïssi, D. *et al.* MORFEE: a new tool for detecting and annotating single nucleotide variants creating premature ATG codons from VCF files. *bioRxiv* 2020.03.29.012054 (2020) doi:10.1101/2020.03.29.012054.
41. den Dunnen, J. T. *et al.* HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat* **37**, 564–569 (2016).
42. Soukarieh, O. *et al.* Common and Rare 5'UTR Variants Altering Upstream Open Reading Frames in Cardiovascular Genomics. *Front Cardiovasc Med* **9**, 841032 (2022).
43. Olexiouk, V. *et al.* sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* **44**, D324–D329 (2016).
44. Choteau, S. A., Wagner, A., Pierre, P., Spinelli, L. & Brun, C. MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. *Database (Oxford)* **2021**, baab032 (2021).
45. Chothani, S. P. *et al.* A high-resolution map of human RNA translation. *Mol Cell* **82**, 2885-2899.e8 (2022).
46. Soukarieh, O. *et al.* Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. *PLoS Genet* **12**, e1005756 (2016).
47. Gleason, A. C., Ghadge, G., Chen, J., Sonobe, Y. & Roos, R. P. Machine learning predicts translation initiation sites in neurologic diseases with nucleotide repeat expansions. *PLoS One* **17**, e0256411 (2022).
48. Reuter, K., Biehl, A., Koch, L. & Helms, V. PreTIS: A Tool to Predict Non-canonical 5' UTR Translational Initiation Sites in Human and Mouse. *PLoS Comput Biol* **12**, e1005170 (2016).
49. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
50. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* **53**, 1712–1721 (2021).

51. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
52. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* **22**, 719–748 (1959).
53. Thibord, F. *et al.* Cross-Ancestry Investigation of Venous Thromboembolism Genomic Predictors. *Circulation* **146**, 1225–1242 (2022).
54. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
55. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, e1007081 (2017).
56. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).
57. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658–665 (2013).
58. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* **46**, 1985–1998 (2017).
59. Spring, B., Lemon, M., Weinstein, L. & Haskell, A. Distractibility in schizophrenia: state and trait aspects. *Br J Psychiatry Suppl* 63–68 (1989).
60. Godin, O. *et al.* White matter lesions as a predictor of depression in the elderly: the 3C-Dijon study. *Biol Psychiatry* **63**, 663–669 (2008).
61. Lind, L. *et al.* Use of a proximity extension assay proteomics chip to discover new biomarkers for human atherosclerosis. *Atherosclerosis* **242**, 205–210 (2015).
62. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
63. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236–1241 (2015).

64. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* **10**, 507–521 (1915).
65. Goumidi, L. *et al.* Association between ABO haplotypes and the risk of venous thrombosis: impact on disease risk estimation. *Blood* **137**, 2394–2402 (2021).
66. Minhajat, R. *et al.* Organ-specific endoglin (CD105) expression in the angiogenesis of human cancers. *Pathol Int* **56**, 717–723 (2006).
67. Andersson-Rusch, C. *et al.* High concentrations of soluble endoglin can inhibit BMP9 signaling in non-endothelial cells. *Sci Rep* **13**, 6639 (2023).
68. Blasi, F. & Carmeliet, P. uPAR: a versatile signalling orchestrator. *Nat Rev Mol Cell Biol* **3**, 932–943 (2002).
69. Singh, I. *et al.* Failure of thrombus to resolve in urokinase-type plasminogen activator gene-knockout mice: rescue by normal bone marrow-derived cells. *Circulation* **107**, 869–875 (2003).
70. Alfano, D., Franco, P. & Stoppelli, M. P. Modulation of Cellular Function by the Urokinase Receptor Signalling: A Mechanistic View. *Front Cell Dev Biol* **10**, 818616 (2022).
71. Zhou, Y. *et al.* TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res* gkad751 (2023) doi:10.1093/nar/gkad751.

Non-canonical regulation of Endoglin by rare and common variants: new molecular and clinical perspectives for Hereditary Hemorrhagic Telangiectasia and beyond

Omar Soukarieh[1,2#*], Gaëlle Munsch[1#], Clémence Deiber[1], Caroline Meguerditchian[1], Carole Proust[1], Ilana Caro[1], Maud Tusseau[3], Alexandre Guilhem[4,5], Shirine Mohamed[6], INVENT consortium, Béatrice Jaspard-Vinassa[2], Aurélie Goyenvalle[7], Stéphanie Debette[1,8], Sophie Dupuis-Girod[3,9], David-Alexandre Trégouët[1*]

[1] Univ. Bordeaux, INSERM, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France
[2] Univ. Bordeaux, INSERM, Biology of Cardiovascular Diseases, U1034, F-33600 Pessac, France
[3] Hospices Civils de Lyon, French National HHT Reference Center and Genetics department, Hôpital Femme-Mère-Enfant, 69677, Bron, France
[4] Service de Médecine Interne et Immunologie Clinique, Centre de compétence maladie de Rendu-Osler, Centre Hospitalo-Universitaire Dijon Bourgogne, Dijon, France
[5] Université de Bourgogne, INSERM, EFS BFC, UMR1098, RIGHT Interactions Greffon-Hôte-Tumeur/Ingénierie Cellulaire et Génique, Dijon, France
[6] Département de Médecine interne et Immunologie Clinique, CHRU BRABOIS, Vandoeuvre-lès-Nancy, France
[7] Université Paris-Saclay, UVSQ, Inserm, END-ICAP, Versailles, France
[8] Department of Neurology, Institute for Neurodegenerative Diseases, Bordeaux University Hospital, France
[9] Univ. Grenoble Alpes, Inserm, CEA, Laboratory Biology of Cancer and Infection, F-38000 Grenoble, France.

[#] These authors have contributed equally to this work
[*] Joint corresponding authors: omar.soukarieh@inserm.fr; david-alexandre.tregouet@u-bordeaux.fr

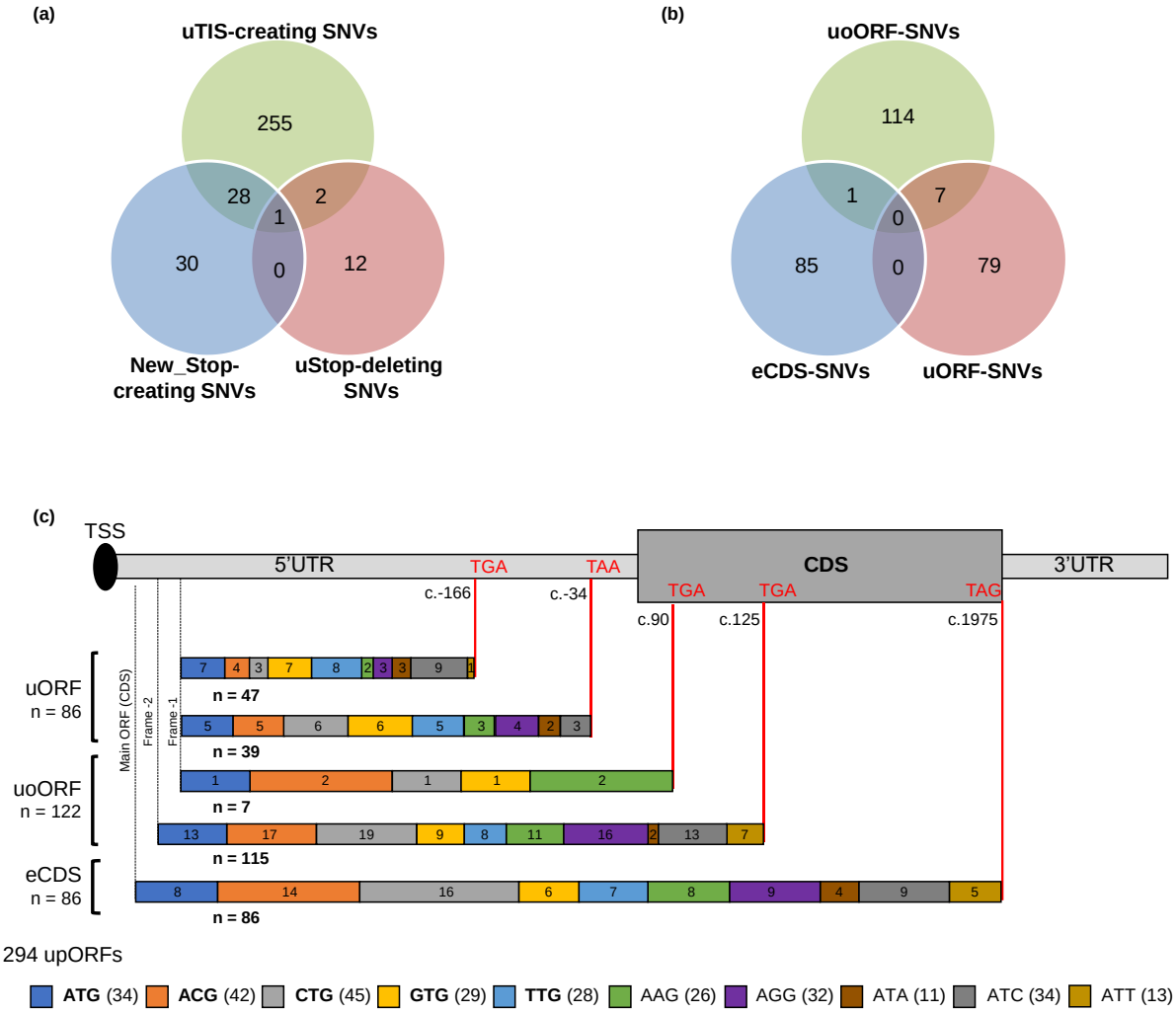

Figure 1. Identification of all possible single nucleotide variants (SNVs) modifying or creating upstream open reading frames (upORFs) in the 5'UTR of ENG. (a) MORFEE annotated SNVs that modify existing upORFs or create new ones (New_Stop-creating, uStop-deleting SNVs and uTIS-creating SNVs). (b) Three types of upORFs can result from the creation of upstream translation initiation sites (uTIS) in the 5'UTR of ENG. uoORF-SNVs, variants at the origin of upORFs overlapping the coding sequence (CDS); eCDS-SNVs, variants creating elongated CDS; uORF-SNVs, variants creating fully upstream upORFs. (c) Detailed illustration of upORFs generated by uTIS-creating SNVs. The type of uTIS and position of stop codons associated with the generated upORFs, as well as the number of each type of upORF and of uTIS are indicated. TSS, translation start site.

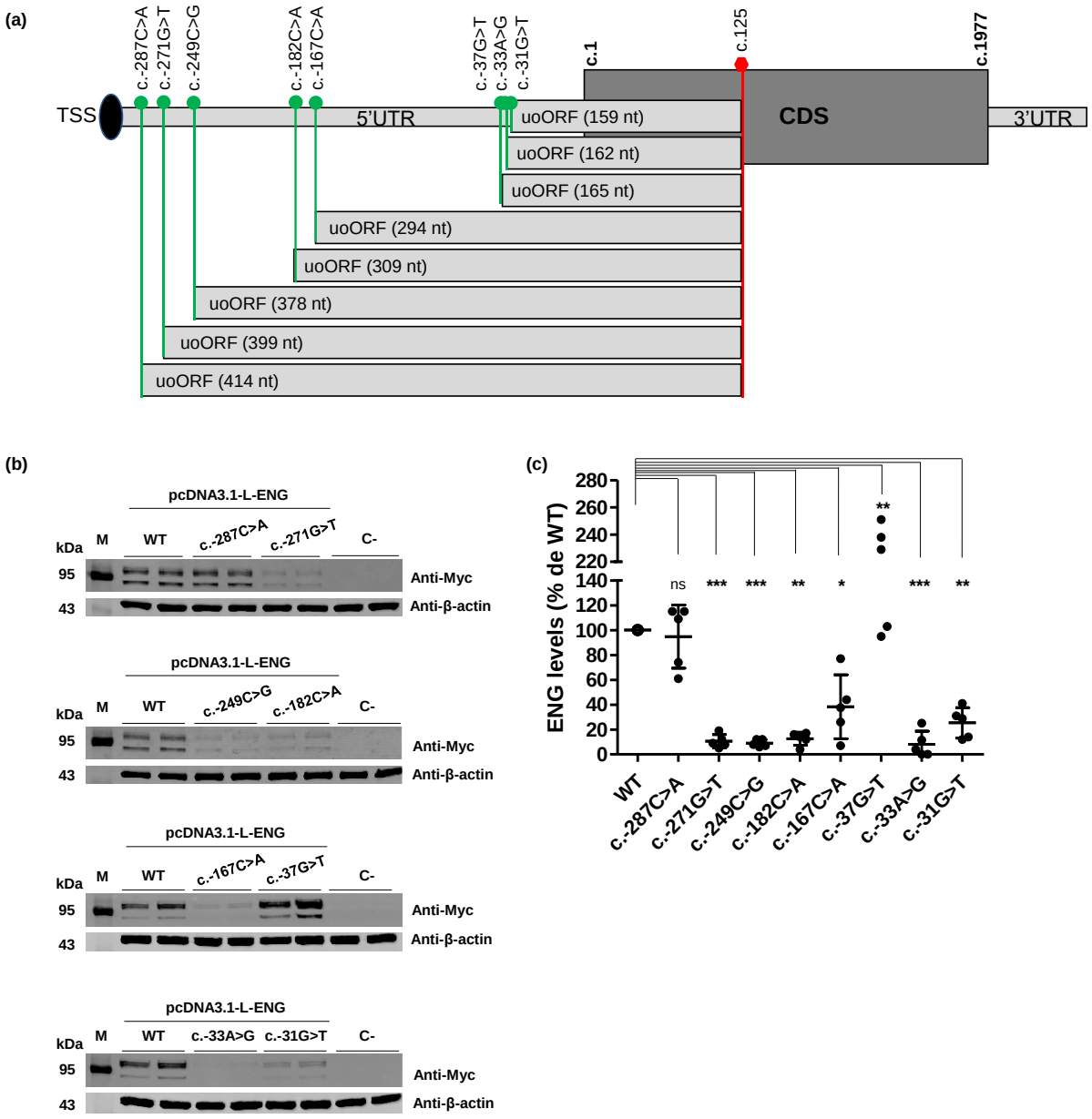

Figure 2. *ENG* variants at the origin of uAUGs in frame with the same stop codon at position c.125. (a) Eight variants in the 5'UTR create uAUG-initiated upstream overlapping open reading frames (uoORFs) with different sizes but ending at the same position. Position and identity of these variants and of created uAUG, size of the uoORFs, associated stop codon located at position c.125 and first and last positions of the CoDing Sequence (CDS) (c.1 and c.1977, respectively) are indicated. TSS, translation start site; CDS, CoDing Sequence. (b) Western blot results on total proteins extracted from transfected HeLa cells with 1 µg of pcDNA3.1-L-ENG constructs. Two bands of different molecular weights are observed for endoglin likely corresponding to more glycosylated (upper band) and less/non glycosylated (lower band) ENG monomers. Anti-Myc and anti-ENG correspond to the used antibodies for the target protein from HeLa and HUVECs, respectively, and anti-β-actin corresponds to the antibody used against the reference protein. kDa kilodalton, M protein ladder, WT wild type, C- negative control corresponding to pcDNA3.1- empty vector. Shown results are representative of 5 independent experiments. All blots were processed in parallel and derive from the same experiments. (c) Quantification of ENG steady-state levels in HeLa cells from (b). For

quantification, the average of each duplicate has been calculated from the quantified values and ENG levels for each sample have been normalized to the corresponding β-actin levels then to the WT (%). The two bands obtained for the Endoglin, corresponding to the more glycosylated (upper band) and less/non glycosylated (lower band) ENG monomers 1, were taken together for the quantification. Graphs with standard error of the mean are representative of 5 independent experiments. ***, p-value < 10^{-3}; **, p < 10^{-2}; *, p < 5.10^{-2}, ns, non-significant (two-factor ANOVA followed by Tukey's multiple comparison test of variants versus WT).

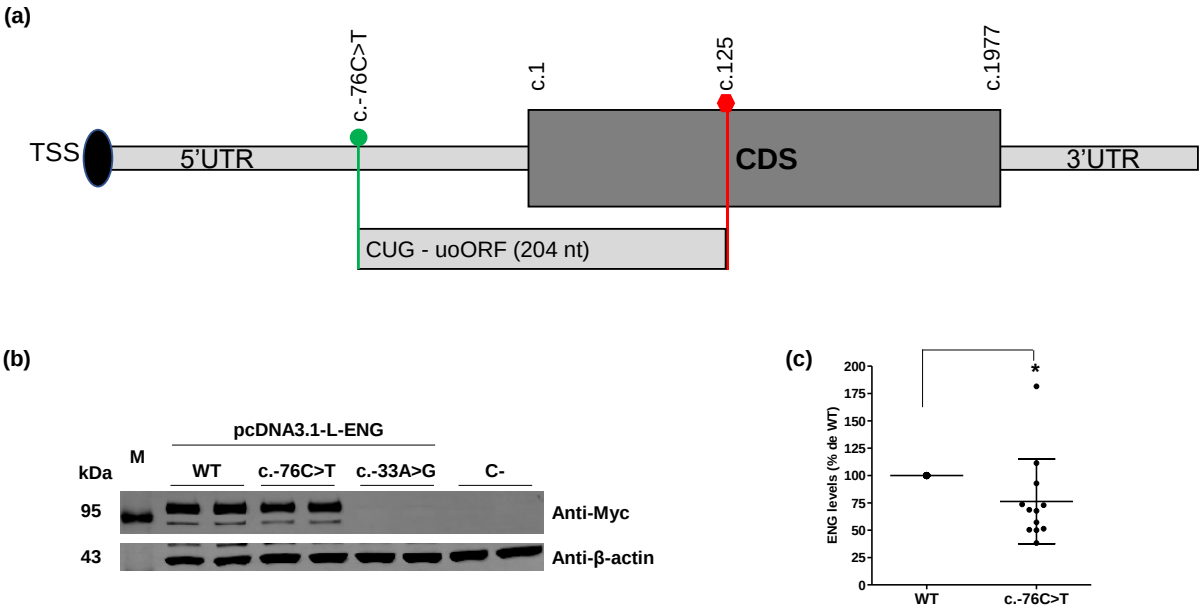

Figure 3. uoORF-creating variants identified in French HHT patient. (a) ENG c.-76C>T creating an upstream CUG in frame with the stop codon at position c.125 at the origin of an upstream Overlapping Open Reading Frame (uoORF) of 204 nucleotides (nt). Position and identity of the variant and of created uCUG, and first and last positions of the CoDing Sequence (CDS) (c.1 and c.1977, respectively) are indicated. (b) Western blot results on total proteins extracted from transfected HeLa cells with 1 µg of pcDNA3.1-L-ENG constructs. Two bands of different molecular weights are observed for endoglin likely corresponding to more glycosylated (upper band) and less/non glycosylated (lower band) ENG monomers. Anti-Myc and anti-ENG correspond to the used antibodies for the target protein from HeLa and HUVECs, respectively, and anti-β-actin corresponds to the antibody used against the reference protein. kDa kilodalton, M protein ladder, WT wild type, C- negative control corresponding to pcDNA3.1- empty vector. Shown results are representative of 5 independent experiments. All blots were processed in parallel and derive from the same experiments. (c) Quantification of ENG steady-state levels in HeLa cells from (b). For quantification, the average of each duplicate has been calculated from the quantified values and ENG levels for c.-76C>T sample have been normalized to the corresponding β-actin levels then to the WT (%). The two bands obtained for the Endoglin, corresponding to the more glycosylated (upper band) and less/non glycosylated (lower band) ENG monomers 1, were taken together for the quantification. Graphs with standard error of the mean are representative of 12 independent experiments. *, $p < 5.10^{-2}$ (two-factor ANOVA followed by Tukey's multiple comparison test of variants versus WT).

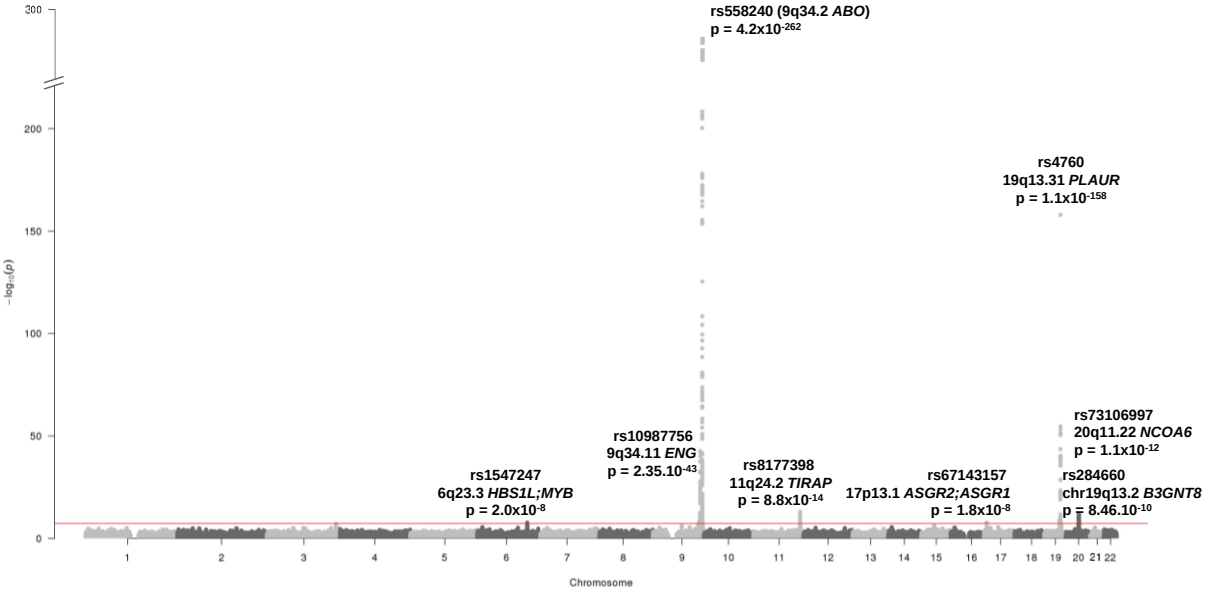

Figure 4. Manhattan plot summarizing the meta-analysis of two GWAS datasets on plasma Endoglin levels in 46,091 individuals. The identity, location and p-value of the 8 loci significantly ($p < 5.10^{-8}$) associated with endoglin levels are indicated.

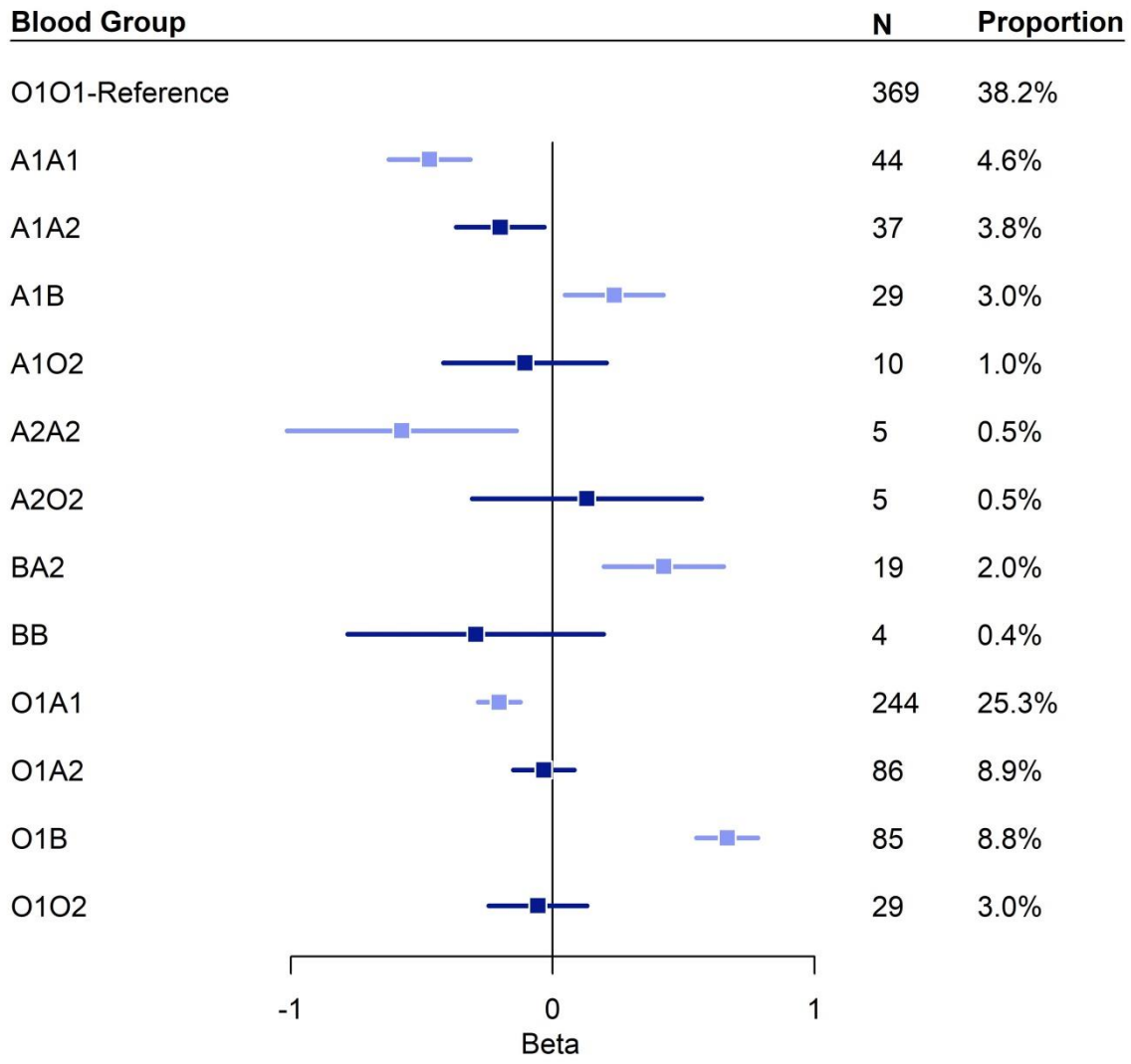

Figure 5. Forest plot summarizing the association of ABO diplotypes with Endoglin plasma levels. Diplotype effects were assessed with the O1O1 diplotype as the reference and were estimated based on standardized Endoglin values measured in 966 participants from the 3C study, where Endoglin levels were determined using Olink technology. N, number of carriers.

Table 1. Characteristics of the 13 uAUGs and the studied uCUG created by variants in the 5'UTR of ENG. uTIS position in the L-ENG; NM_001114753.3 transcript (c.1 corresponds to the A of the main AUG) is indicated. Nucleotides at positions -3 and +4 relative to the predicted uTIS are in bold when corresponding to the most conserved nucleotide. The uTIS are underlined. Bolded scores are those higher than the predefined thresholds.

uTIS-SNV	rsID	ClinVar	Protein levels *in vitro*	uoORF size (nt)	Kozak sequence	PreTIS score	TIS-Predictor score
c.-287C>A*	NA	No	~WT	414	CATATGC	na	0.42
c.-271G>T*	NA	No	< 20%	399	**G**CCATGA	na	**0.78**
c.-249C>G	NA	No	< 20%	378	**G**CCATGC	na	**0.67**
c.-182C>A	NA	No	< 20%	309	TCCATGT	**0.91**	0.57
c.-167C>A	NA	No	< 40%	294	CTCATGA	**0.87**	0.55
c.-142A>T	NA	No	< 20%	270	CAGATG**G**	**0.84**	**0.66**
c.-127C>T	rs1060501408	Yes (P/LP)	< 20%	255	**G**GGATGC	**0.84**	**0.69**
c.-79C>T	rs1564466502	Yes (VUS)	< 20%	207	CCCATGC	**0.67**	**0.65**
c.-68G>A	NA	No	< 50%	195	CCGATGC	**0.89**	0.61
c.-37G>T*	NA	No	≥ WT	165	CACATGA	**1**	0.57
c.-33A>G	NA	No	< 20%	162	**A**GGATGA	**1**	**0.73**
c.-31G>T	NA	No	< 30%	159	**A**TAATGC	**0.96**	0,63
c.-10C>T	rs756994701	Yes (Conflicting)	< 40%	138	CCCATGT	**0.85**	0.61
c.-76C>T	rs943786398	Yes (VUS)	< 75%	204	**A**CGCTG**G**	**0.93**	**0.87**

*, in addition of the uAUG-creation, c.-287C>A and c.-37G>T are predicted to create new stop codons shortening existing upORFs, and c.-271G>T is predicted to simultaneously create a uAUA at the origin of a fully upstream upORF. nt, nucleotide; P, pathogenic; LP, Likely-Pathogenic; VUS, variant of unknown significance; na, non-applicable.