Genetics in Medicine

Targeted re-sequencing of FECH locus reveals that a novel deep intronic pathogenic variant and eQTLs may cause erythropoietic protoporphyria (EPP) through a methylation-dependent mechanism.

--Manuscript Draft--

Manuscript Number:	GIM-D-19-00032R3
Article Type:	Article
Section/Category:	Clinical Genetics and Genomics
Keywords:	erythropoietic protoporphyria; deep intronic pathogenic variant; eQTLs; CpG sites; pre-mRNA splicing pattern
Corresponding Author:	Elena Di Pierro, Ph.D. Ospedale Maggiore Policlinico Milano, ITALY
First Author:	Matteo Chiara
Order of Authors:	Matteo Chiara
	Ilaria Primon
	Letizia Tarantini
	Luca Agnelli
	Valentina Brancaleoni
	Francesca Granata
	Valentina Bollati
	Elena Di Pierro, Ph.D.
Manuscript Region of Origin:	ITALY
Abstract:	Purpose Existing data do not explain the reason why some individuals homozygous for the hypomorphic FECH allele develop erythropoietic protoporphyria (EPP) while the majority are completely asymptomatic. This study aims to identify novel possible genetic variants contributing to this variable phenotype. Methods High throughput re-sequencing of the FECH gene, qualitative analysis of RNA and quantitative DNA methylation examination were performed on a cohort of 72 subjects. Results A novel deep intronic variant was found in four homozygous carriers developing a clinically overt disease. We demonstrate that this genetic variant leads to the insertion of a pseudoexon containing a stop codon in the mature FECH transcript by the abolition of an exonic splicing silencer site and the concurrent institution of a new methylated CpG di-nucleotide. Moreover, we show that the hypomorphic FECH allele is linked to a single haplotype of about 20kb in size that encompasses three non-coding variants that were previously associated with expression quantitative trait loci (eQTLs). Conclusion This study confirms that intronic variants could explain the variability in the clinical manifestations of EPP. Moreover, it supports the hypothesis that the control of the FECH gene expression can be mediated through a methylation-dependent modulation of the pre-mRNA splicing pattern.

FONDAZIONE IRCCS CA' GRANDA
OSPEDALE MAGGIORE POLICLINICO
DIPARTIMENTO MEDICINA INTERNA
U.O. MEDICINA INTERNA 1A
DIRETTORE MARIA DOMENICA CAPPELLINI

Milan, 4th June 2019

To Genetics in Medicine
Editorial Office

Dear Editors,

thank you very much for accepting the revision of our manuscript entitled "Targeted re-sequencing of FECH locus reveals that a novel deep intronic pathogenic variant and eQTLs may cause erythropoietic protoporphyria (EPP) through a methylation-dependent mechanism". Higher-quality figure files have been provided and the requested changes have been introduced into the text. Hoping that now the paper could be suitable for publication

On behalf of all authors,

best regards,

Elena Di Pierro
Fondazione IRCCS "Cà-Granda" Ospedale Maggiore Policlinico
Dipartimento di Medicina Interna
Tel:+390255036155
Fax:+390250320296
e-mail: elena.dipierro@unimi.it

Dear reviewers,

Thank you very much for accepting the revision of our manuscript. The changes have been introduced into the text according to suggestions. Detailed point-by-point answers to the additional comments are given below. Hoping that now the paper could be suitable for publication

On behalf of all authors,

Best regards

Elena Di Pierro

Reviewer Comments:

Reviewer #2: Overall, this manuscript is much improved. The introduction now adequately frames the study to reflect the goals and in general the data is present in a much more concise way. This is an interesting study that shows new genetic variants that can modulate the EPP phenotype, and interesting mechanism that may regulate FECH expression. Overall, the authors have addressed all my previous queries and I have only the following additional comment to my previous points.

13. the authors have addressed this query by adding the following text "Interestingly we notice that 94.27% of the low coverage..... reflecting the difficulty in mapping short Illumina reads in highly repetitive regions of the genome". However, this only occurred in 20 samples, not all - therefore the possibility that there may be a more complex structural variant or STR expansion in this region cannot be ruled out, this should be acknowledged/discussed.

We agree with the referee on this point. Therefore, in the text we now explicitly address the possibility that the decrease in coverage observed at highly repetitive regions in some patients, might be explained by the presence of complex structural rearrangements like for example copy number alterations or large indels.

30. This discussion is much improved, however, poison exons or NMD exons are by definition, alternatively spliced in cell-types where the expression of the gene is suppressed (PMID: 27565344, 30526861, 26829591). Is this the case with this putative 'poison exon' or is this merely deletion of an ESS and activation of a cryptic splice site. There are multiple RNA datasets available to differentiate between these two possibilities.

The exon identified in the 4 carriers of the c.464–1169A>C allele has not been observed nor described before, in any large scale study of the expression pattern of the FECH gene. Therefore it can not be described as a “poison exon”. The text has now been corrected accordingly. We now use the term “pseudoexon” or “cryptic exon”, which in our opinion are more correct.

In addition, I have the following comments having re-reviewed the paper in the context of all the changes from the previous version.

Major comments:

Figure 4B - why do the haplotype frequencies of the EPP and 1000 genomes data not sum to 1 in each cohort? Moreover, the authors state: "Only a single individual with clinically overt disease carries the ATC" - how is the haplotype frequency then 0.25 in figure 4B? in the discussion "Considerations on the relative frequency of the GTT and GTC haplotypes (GTT 22% and GTC 11%) in the healthy population", different values to Figure 4B.

We apologize with the referee on this point. We have mistakenly uploaded the wrong figure. The correct figure is now provided, and as you can see from this table all the frequency values now sum to unity.

	TSI (1000G) %	EPP cohort %
ACT	65	53.5
GTC	11	31.9
GTT	19	11.1
ACC	2	0
GCT	1	2.8
ATC	1	0.7
ATT	1	0

Allele frequency estimates of GTT and GTC haplotypes were obtained from the 1000G dataset. The slight differences in haplotype frequencies between figure 4b and the values reported in the discussion are due to that, in the table only individuals of Italian ancestry (TSI=Toscani in Italia) population are considered, while in the discussion we refer to the average frequencies estimated on all the 26 human populations included in 1000G.

Minor comments:

1. "Clinical manifestation of EPP is typically associated with a compound heterozygous genetic background at FECH locus", the word 'background' here is incorrect, should read "Clinical manifestation of EPP is typically associated with compound heterozygous variants at the FECH locus inherited in an autosomal recessive manner".

Since only 4% of patients inherit two pathogenic variants, the sentence has been corrected as suggested removing the last part regarding the inheritance. "Clinical manifestation of EPP is typically associated with compound heterozygous variants at the FECH locus".

2. "It is widely accepted that the aberrant alternatively spliced mRNA is degraded by a nonsense mediated decay (NMD) mechanism, leading, in the context of the null allele, to sufficient FECH enzyme deficiency". This sentence does not make sense, if the mRNA is degraded and the other allele is null - how is there functional FECH protein? I think what the authors mean to say is that only a subset of transcripts from the C allele are aberrantly spliced? Thus, there is still sufficient FECH protein to prevent EPP?

Actually, the use of cryptic alternative acceptor splice site is also observed in individuals carrying the T allele. Carriers of the C allele show an increased utilization of this site resulting in a higher proportion of aberrantly spliced transcripts. This leads to an additional FECH enzyme deficiency, which is necessary for protoporphyrin overproduction and clinical symptoms. The sentence has been changed for clarity.

3. I would recommend figure S1 be included in the main manuscript to aid in interpretation of results. For instance, within figure1 as the first 'A' panel. Figure S1 needs a legend, what are the # and other symbols. Also it would be helpful if the resulting FECH protein annotation was added to this figure e.g. c.464-1169A>C (p.Ala155GlyfsTer22)

The figure S1 has been incorporated in the figure 1 as suggested. The legend was already included in the text as well as the explanations of the symbols in the figure itself.

4. (p.(Ala155GlyfsTer22)) should be (p.Ala155GlyfsTer22)

No, the HGVS general recommendations for sequence variant nomenclature (http://varnomen.hgvs.org/recommendations/general/) suggest to use of (parentheses) to indicate uncertainties and predicted consequences; NC_000023.9:g.(123456_234567)_(345678_456789)del, p.(Ser123Arg). Also in our case the consequences at protein level are predicted, thus we indicated this fact as p.(Ala155GlyfsTer22). In the text there are other 2 parentheses because of the sentence construction.

5. "A novel deep intronic variant (c.464-1169A>C) and an annotated low frequency genetic variant (c.804+659G>A; rs754770993, frequency in 1000 genomes 0; Kaviar 0,000115)" Frequency in bravo/topmed? What was the data source referenced in Kaviar?

Frequencies in TopMed, gnomAD and ExAC have now been included in the text. Kaviar is a large database that incorporates data of human genetic variation from more than 30 independent studies. A complete list of the datasets that are currently incorporated in Kaviar can be found at: http://db.systemsbiology.net/kaviar/cgi-pub/Kaviar.pl?show=sources . Details concerning the version of Kaviar used in this study are provided in the materials and methods section.

6. "For PCR amplicon, covering exons 4 to 8" amplicon should be amplification. In general the manuscript needs to be thoroughly proofed for grammatical and style errors such as this.

The manuscript has now been subjected to a further round of proofing, which was performed by a native speaker (see acknowledgments).

7. While the analysis of the sequences around the known c.68-23 A>T and c.315-48 T>C eQTLs indicated formation of new ESS sites ((TT)TCATGT(GAG) and (G)GCTG(CTAA) respectively). This section would be more relevant to include in the next section focusing on the GTC haplotype.

The sentence has been moved as required

Reviewer #3: With the changes that the authors have made to the text, it reads clearly, and has sufficient analysis details.

Thanks

Targeted re-sequencing of *FECH* locus reveals that a novel deep intronic pathogenic variant and eQTLs may cause erythropoietic protoporphyria (EPP) through a methylation-dependent mechanism.

Matteo Chiara PhD[1#], Ilaria Primon BSc [2], Letizia Tarantini BSc[3], Luca Agnelli PhD[4], Valentina Brancaleoni MSc[2], Francesca Granata MSc[2], Valentina Bollati PhD[3], Elena Di Pierro PhD[2# *]

[1] Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy.

[2] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, UOC Medicina Generale, Milan, Italy.

[3] EPIGET - Epidemiology, Epigenetics and Toxicology Lab - Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy.

[4] Department of Oncology and Hemato-oncology, University of Milan; and Hematology 1, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

# These authors contributed equally to this work. The paper must be cited as Chiara, Di Pierro et al.

* Corresponding author

Elena Di Pierro, PhD

Fondazione IRCCS Ca' Granda,

Ospedale Maggiore Policlinico

Via F. Sforza 35, 20122 Milano, Italy

+390255036155

elena.dipierro@policlinico.mi.it

elena.dipierro@unimi.it

ABSTRACT

Purpose

Existing data do not explain the reason why some individuals homozygous for the hypomorphic *FECH* allele develop erythropoietic protoporphyria (EPP) while the majority are completely asymptomatic. This study aims to identify novel possible genetic variants contributing to this variable phenotype.

Methods

High throughput re-sequencing of the *FECH* gene, qualitative analysis of RNA and quantitative DNA methylation examination were performed on a cohort of 72 subjects.

Results

A novel deep intronic variant was found in four homozygous carriers developing a clinically overt disease. We demonstrate that this genetic variant leads to the insertion of a pseudoexon containing a stop codon in the mature *FECH* transcript by the abolition of an exonic splicing silencer site and the concurrent institution of a new methylated CpG di-nucleotide. Moreover, we show that the hypomorphic *FECH* allele is linked to a single haplotype of about 20kb in size that encompasses three non-coding variants that were previously associated with expression quantitative trait loci (eQTLs).

Conclusion

This study confirms that intronic variants could explain the variability in the clinical manifestations of EPP. Moreover, it supports the hypothesis that the control of the *FECH* gene expression can be mediated through a methylation-dependent modulation of the pre-mRNA splicing pattern.

KEYWORDS

erythropoietic protoporphyria; deep intronic pathogenic variant; eQTLs; CpG sites; pre-mRNA splicing pattern

INTRODUCTION

Erythropoietic protoporphyria (EPP, MIM#177000) is a heritable metabolic disorder resulting from a reduction, to less than 35% of normal levels, of ferrochelatase (FECH, EC 4.99.1.1) activity.[1] FECH is the last enzyme of the heme biosynthetic pathway and reduced activity leads to significantly elevated metal-free protoporphyrin (PPIX) levels mainly in erythrocytes and subsequently in skin and liver, causing clinical manifestations of the disease.[2] EPP patients experience severe cutaneous phototoxic reactions in sun-exposed areas since their early childhood. Besides this, the accumulation of PPIX in the liver may lead to mild hepatic injury and, in approximately 2% of cases, to severe cholestatic damage progressing to liver failure and requiring consequent liver transplantation.[3]

Clinical manifestation of EPP is typically associated with compound heterozygous variants at *FECH* locus, mapped to chromosome 18q21.3. Currently about 200 pathogenic variants are reported in the Human Gene Mutation Database (HGMD professional v2018.3 http://www.hgmd.cf.ac.uk/ac/index.php). Although a limited number of patients, carrying two null *FECH* alleles has been reported,[4] in the majority of the affected individuals, a rare null *FECH* allele is co-inherited in *trans* with a common hypomorphic allele that is associated with decreased levels of *FECH* gene expression.[5, 6] This hypomorphic *FECH* allele is characterized by the presence of a common variant in intron 3 (c.315–48T>**C**, **C** variant, rs2272783), which modulates the usage of a cryptic alternative acceptor splice site, 63 bp upstream of the constitutive site (Figure 1A). It is widely accepted that the aberrant alternatively spliced mRNA is degraded by a nonsense mediated decay (NMD) mechanism, leading, in the context of the null allele, to additional deficiency of ferrochelatase, which is necessary for protoporphyrin accumulation and clinical symptoms.[7, 8] Recent independent studies have reported a few cases with a mild EPP phenotype in the presence of the **C** variant in homozygosis[9, 10] while others have described a late-onset of the EPP phenotype, secondary to a myelodysplastic syndrome.[11] The **C** variant is normally present in healthy human populations, with frequencies ranging from 1% - 5% in Africa and Europe to 32% - 37% in East Asia and in

America. Increased frequencies of the **C** variant are not associated with an increase in the prevalence of EPP, even in populations where a high percentage of homozygous subjects (19–22%) is observed.[12] Several independent studies have reported that two other common variants, c.1–252A>**G** in the promoter (**G** variant, rs17063905) and c.68–23C>**T** in intron 1 (**T** variant, rs2269219) are consistently found in association with the c.315-48T>**C** variant in EPP patients, forming a so-called **GTC** haplotype.[13, 14] Interestingly all three variants of the **GTC** haplotype are associated with expression quantitative trait loci (eQTL) that reduce the expression of the *FECH* gene according to the GTEX study (GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)), suggesting a possible role for the **GTC** haplotype as a whole in the pathogenesis of EPP. However, at present it is still unclear whether a homozygous **GTC** state in isolation is sufficient to provoke EPP.[15] In this study, we have performed a high throughput targeted re-sequencing of the *FECH* locus in a cohort of 72 individuals belonging to 24 Italian unrelated EPP families. Notably, five subjects out of 72 were homozygous for the hypomorphic **GTC** allele and showed a variable phenotype where one was completely asymptomatic, while others developed a clinically overt disease from childhood. By comparing hypomorphic **GTC** alleles between patients and unaffected carriers, the study aimed to identify possible functional variants responsible for the variable outcome of EPP.

METHODS

Study subjects

All patients were diagnosed based on the clinical history of photosensitivity in the presence of plasma porphyrin peak at 635nm and high levels of protoporphyrin in the erythrocytes and feces. The sequence of the *FECH* gene was also determined by Sanger sequencing. Both parents when available and healthy relatives were recruited in order to facilitate the reconstruction of the hypomorphic **GTC** haplotype. The study was approved by the Fondazione IRCCS Ca' Granda ethics committee (n°2952, 12-18-2015) and all the subjects signed informed consent prior to their inclusion in the study (Table S1).

Targeted *FECH* re-sequencing

A custom enrichment panel was designed by the means of the Agilent SureDesign[tm] software to capture 90.2 Kb of genomic DNA, from 40Kb upstream to 10kb downstream of the FECH gene, including all exons and introns (chr18 55202704-55292856 on the hg19 human genome assembly). Genomic DNA was extracted from peripheral blood using the Maxwell®16 Automated System (Promega Corporation, Madison, USA) and spectrofluorometrically quantified using QuantiFluor® One dsDNA kit on GloMax Discover® instrument (Promega Corporation, Madison, USA) according to manufacturer's instructions. Standard Haloplex Target Enrichment system procedure (Agilent Technologies, Santa Clara, USA) was applied for library preparation and 150 bp paired–end reads were generated using a MiSeq sequencer (Illumina, San Diego, USA). Several coverage metrics were recorded in order to define the accuracy and possible limitations of the enrichment panel.

Variant calling, phasing and association tests

Variant calling was performed by the CoVaCS pipeline as described by Chiara et al. 2018.[16] The Annovar software was employed for variants annotation.[17] The following annotation resources were considered for the estimation of allele frequencies: ExAC (version 1.0 updated 02-27-2017)[18], 1000G (phase3)[19], gnomAD (version 2.1, updated 12-10-2018)[18], dbSNP (build 151)[20], Kaviar (version 160204-Public)[21] and TopMed (freeze5, accessed on 02-28-2019, nhlbiwgs.org). Refseq (release 106)[22] was considered for genes and transcript annotations, Clinvar (version 1.55, updated 12-26-2018)[23] and HGMD-Pro 2018.3[24] for the annotation of disease–causing variants and the dbNSPF (v4.0b1, updated 12-30-2018)[25] database for the evaluation of non-synonymous substitutions effect. Nucleotides are numbered based on the FECH NM_000140.3 - Human Refseq transcript, with the A of the ATG initiation codon as '+1'.

Adapted functions of the R *alleHap* package (https://cran.r-project.org/web/packages/alleHap/index.html) were used to obtain the most likely genotype combination and haplotype phasing for trios. Single marker association statistics and visualization of local linkage disequilibrium (LD) were obtained using the Haploview software (https://www.broadinstitute.org/haploview/haploview).[26]

FECH gene expression analysis using public data

Vcf files containing the genetic profiles of the individuals included in the GTEx[27] and 1000G studies were retrieved from the dbGaP database[28] (dbGaP Study Accession: phs000424.v7.p2) and the 1000G data repository (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) respectively. The expression profiles of the *FECH* gene were obtained directly from the GTEx portal (https://storage.googleapis.com/gtex_analysis_v7/rna_seq_data/GTEx_Analysis_2016–01–15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz). A custom Perl script was used to extract haplotypes for the *FECH* gene and cross-reference genetic with expression data. A total of 136 individuals, for which both genotypic and gene expression data were available, were considered.

Qualitative RNA analysis by RT-PCR

Total RNA was isolated from peripheral blood of patients, using the LEV simplyRNA Blood Kit for Maxwell®16 (Promega Corporation, Madison, USA), according to the protocol described in Fiorentino et al. 2016.[29] 400 ng of total RNA were reverse transcribed using the Superscript IV VILO Master Mix (Thermo Fisher, San Francisco, USA) following the protocol supplied with the kit. 50 ng of cDNA was amplified using 10 pmol/µL of each primer, in the presence of 1X buffer, 1.5mM Mg2+, 0.2 mM dNTPs, and 1.25 U of Taq polymerase. The reaction was performed under the following conditions: an initial step at 94 °C for 5 min followed by 35 cycles of denaturation at 94 °C for 30 s; amplification at 60 °C for 30 s and 72 °C for 1 min; and a final extension at 72 °C for 10 min. The region spanning from exons 1 to 8 of the *FECH* gene was amplified using different sets of primer pairs and directly sequenced (Table S2). In order to increase the signal of abnormal bands during sequencing, cDNA product were re-amplified with original and nested primers. Splicing motifs analysis was carried out through Human Splicing Finder v.3 software (http://www.umd.be/HSF3/index.html).[30]

DNA methylation analysis

For DNA methylation analysis 500 ng of DNA was treated with bisulfite using the EZ–96 DNA methylation-Gold kit (Zymo Research, CA, USA) in a final elution volume of 200 µL according to

manufacturer's instructions. Bisulfite-treated DNA was amplified with PCR for each region of interest: a PCR reaction in 50 μL volume was carried out with 25 μL of Hot Start GoTaq Green Master mix (Promega), 1 pmol of forward primer, 1 pmol of reverse primer and 25 ng of bisulfite-treated genomic DNA. Biotin-labeled primers (forward or reverse, depending on the assay) were used to purify the PCR products with Sepharose beads. PCR products were bound to a Streptavidin Sepharose HP (Amersham Biosciences, Uppsala, Sweden), purified, washed, denatured with 0.2 M NaOH and washed again with the Pyrosequencing Vacuum Prep Tool (Pyrosequencing, Inc., Westborough, MA, USA). Then, pyrosequencing primer (0.3 μm) was annealed to the purified single-stranded PCR product, and methylation analysis was performed by PyroMark MD Q96 (Pyrosequencing, Inc. Westborough, MA, USA). PCR cycling conditions and primer sequences are shown in Table S3.

RESULTS

Targeted resequencing of the *FECH* gene

Observed coverage levels, which are reported in Table S4, were well in line with the recommendations for the usage of NGS based resequencing assays in diagnostics: 92.5% of the target regions were covered at 10x or more and 87.8% at 20x or more. However, a consistent proportion of the targeted regions, corresponding to 14.5% of total target was covered by 10x or less reads in more than 20 samples, suggesting systematic biases in the coverage profiles. Accordingly, these regions were excluded from subsequent analyses. Interestingly we notice that 94.27% of the low coverage regions correspond with RepeatMasker[31] annotated repeats in the hg19 human genome assembly, reflecting either a reduced rate of mapping of short Illumina reads in highly repetitive regions of the genome or the possible presence of large structural rearrangements or repeat copy number alterations. A markedly increased coverage (no single region below 20x in all the samples) was observed for the exonic and non-repetitive regions (Table S4).

Genotyping and haplotype phasing

A total of 510 variable sequence positions at the target *FECH* region were identified in the 72 individuals included in this study (Table S5). Of these 109 were completely novel and had never been

reported in any of the publicly available repositories of human genetic variations considered in this study (i.e., 1000G, gnomAD, ExAC, TopMed and Kaviar). Importantly, all the EPP–causing variants, previously identified by the Sanger sequencing, were recovered also by the NGS based assay. In fact, re-sequencing confirmed that among the 24 patients, 18 co-inherited a null allele in *trans* to the hypomorphyic allele, 2 carried only one hypomorphic allele and no pathogenic variants in *trans*, while 4 carried two hypomorphic alleles. Notably the asymptomatic mother of a patient with classical genotype also carried two hypomorphic alleles. The most likely genotypes were obtained by haplotype phasing for all the 72 sequenced subjects in order to compare single alleles.

Two novel deep intronic variants are associated with EPP

A novel deep intronic variant (NG_008175.1(FECH_v002):c.464–1169A>C) and an annotated low frequency genetic variant (c.804+659G>A; rs754770993, frequency in 1000G and ExAC 0; Kaviar and gnomAD 0,0001; TopMed 0.0002), were found in four patients carrying the hypomorphic allele in homozygosis and displaying a clinically overt disease. These variants, which were located, respectively, in introns 4 and 7 of the *FECH* gene, were inherited in *cis* with one of the two hypomorphic alleles from one of the two parents. More importantly, they were not observed in the homozygous hypomorphic subject without symptoms of EPP and in all the other 20 patients carrying the hypomorphic allele.

c.464–1169A>C variant activates a pseudo–exon in intron 4

RT-PCR was performed in order to investigate possible effects of the c.464–1169A>C deep intronic variant on the *FECH* splicing pattern. For amplification, covering exons 4 to 8 an additional longer band was amplified only in four symptomatic **GTC** homozygous patients (Figure 1B). Re-amplification of the exon 4-8 cDNA fragments with the original and nested primers revealed that the longer band is caused by an insertion between exons 4 and 6 and not between exons 7 and 8 (Figure 1C). Direct sequencing showed an insertion in the transcript of a pseudo–exon of 156bp containing a stop codon sequence (Figure 2). This exon, located in intron 4, also encompassed the c.464–1169A>C deep intronic variant and presented common consensus sequences of splicing.

The c.464–1169A>C variant is associated with the disruption of an exonic splicing silencer (ESS)

Functional annotation of the deep intronic variants by the means of the human splicing finder program (HSF) was performed in order to identify possible functional effects on the splicing pattern. Three different algorithms predicted the disruption of an exonic splicing silencer site (ESS) in intron 4, by the c.464–1169A>C substitution. The ESS wild type sequence ((GT)T**A**GGAG) was also recognized as a core binding site for the hnRNP A1 protein; in the presence of the ESS mutated sequence ((GT)T**C**GGAG) this binding site was disrupted. No relevant alterations were reported for the c.804+659G>A variant.

A single **GTC** haplotype is linked to the hypomorphic allele

Using simple Mendelian inheritance rules, 446 variants were assigned to distinct haplotypes of *FECH* gene. Of these, 176 were found to be consistently shared among all the individuals carrying the hypomorphic *FECH* allele. Segregation analysis was carried out to establish the parental origin of each variant and the most likely genotype combination in phasing with the c.315–48 **C** was obtained for each trio. The comparisons of the selected hypomorphic alleles showed that 23 out of 24 unrelated patients shared an identical haplotype of *FECH* gene of about 20 kb in size. The same presumed haplotype was also identified in both copies of the *FECH* locus in the five individuals homozygous for hypomorphic allele. The haplotype spans from 3.7 kb upstream to the transcription start site (rs75861770) to 1.7 kb in the intron 4 (rs11874117) and contains 47 annotated SNPs including the c.1–252 **G** and the c.68–23 **T** variants (Figure 3). Interestingly only a single recombination event of the proposed haplotype is observed, that is one patient, where only the portion from intron 1 (rs32166686) to intron 4 is retained.

The **GT**T and A**TC** haplotypes are associated with reduced *FECH* mRNA levels

Among the 47 SNPs included in the observed haplotype only the known c.315–48 **C** variant was never observed in *trans* to a mutated *FECH* allele in asymptomatic carriers. Notably only two parents presenting very light accumulation of protoporphyrins (6.3 and 5.2 vs n.v <3 mcg/gHb.), with no clinical manifestations of EPP, inherited in *trans* to the mutated allele the c.1–252 A>**G** and c.68–

23C>**T** variants; these other two known substitutions are part of **GTC** haplotype. A chi-squared test, as implemented by the Haploview software, was used to evaluate the level of association for single markers. The analysis confirmed that the c.315–48 **C** variant has the strongest association with the disease (p-value: 3.29 E–7) while a reduced but still significant association is observed for the c.68–23**T** (p-value: 7.03E–6) and c.1–252 **G** (p-value: 2.45E–5) variants (Table S6). An extensive analysis of the publicly available genotypic and gene expression data (GTEX) of 136 individuals included in the GTEX study evidenced a significant decrease in the expression level of the *FECH* gene both in the carriers of GTT and GTC haplotypes. Consistent with previous observations this decrease is more pronounced in the individuals carrying the GTC haplotype (Figure 4A).

Notably, while the GTC haplotype is over-represented in our cohort of patients with respect to a population of healthy individuals, we observe no, or very weak, evidence of recombination within the GTC haplotype in the healthy population (Figure 4B). Interestingly, among the patients included in the present study, only one individual with a clinically overt disease carried a "modified" A**TC** haplotype. At the same time, no EPP patient carried the AC**C** haplotype, which is as frequent as the A**TC** (Figure 4B). Of note, a LOD score of 12.65 was found between the **T** and **C** variants where LOD > 2 indicates significant linkage disequilibrium (LD) (Figure S1). Moreover, the HSF analysis of the sequences around the c.68–23 A>**T** and c.315–48 T>**C** eQTLs indicated formation of new ESS sites ((TT)TCATG**T**(GAG) and (G)G**C**TG(CTAA) respectively).

Differential methylation around *FECH* variants are associated with altered splicing patterns

Considering the emerging role of intragenic methylation in the regulation of the alternative splicing, different CpG sites along the gene were analyzed by bisulfite pyrosequencing. As expected for an expressed gene, the methylation in the promoter region was minimal and no measurable difference was noted between patients. Surprisingly, the patients carrying the c.464–1169A>C deep intronic pathogenic variant presented a new methylated CpG site which was not observed in other patients with a classical GTC haplotype or in unaffected subjects. Conversely, the c.68–23**T** variant in intron 1 and the variant in intron 7 abolished commonly methylated CpG sites. No alteration was detected

in the region encompassing the c.315-48 **C** variant of intron3 (Table 1). In order to establish whether the c.68-23**T** variant also affects the modulation of the *FECH* splicing, a forward primer encompassing the exons 1 and 3 and a reverse primer located in the region of alternative splicing of intron 3 were used for RT-PCR (Table S2). The direct sequencing of the PCR product confirmed the presence of an isoform of splicing showing a complete skipping of the constitutive exon 2 and the insertion of the 63bps of intron 3 (Figure S2). The identified sequence is reported in Ensembl as a non-coding processed transcript (ENST00000585699.1).

DISCUSSION

In this study, by using deep targeted resequencing of the *FECH* gene we report the first evidence of a deep intronic variant causing erythropoietic protoporphyria (EPP). We demonstrate that the c.464–1169A>C intronic substitution (p.(Ala155GlyfsTer22)) disrupts, likely through the institution of new methylated CpG site, an exonic splicing silencer (ESS) site causing the insertion of a "cryptic exon" containing a stop codon, in the mature *FECH* transcript. It is now clear that constitutive and alternative splicing events in higher eukaryotes are finely regulated through the concerted recognition of multiple well-defined and weak cis-acting elements by trans-acting factors. Depending on the effect they exert, these weak cis-acting elements are generally referred to as either enhancers or silencers.[32] Several lines of evidence suggest that silencers have a fundamental role in preventing pseudoexon inclusion in mature transcripts and in defining constitutive exons by suppressing nearby decoy splice sites.[33] Additionally, DNA methylation is emerging as an important factor in exon selection by the splicing machinery and also in the regulation of alternative splicing.[34] In particular, the increase of DNA methylation has been reported to promote the inclusion of alternative exons.[35] All these considerations are highly consistent with the proposed significance of the c.464–1169A>C variant.

Our data also provide independent confirmation that clinically overt EPP is strongly associated with inheritance of the c.315–48 **C** variant and that, clinical expression of disease typically occur when this hypomorphic allele exists in *trans* to a null *FECH* allele. These results also confirm that the

c.315–48 **C** variant in isolation is necessary but not sufficient to cause an overt disease even when inherited in homozygosis. It was recently shown that abnormal splicing events are twice as frequent in the presence of the c.314–48 **C** variant in heterozygosis. At the same time, this figure does not increase further in homozygous **C** EPP patients, which show *FECH* mRNA levels comparable to those of EPP patients with a classical genotype.[15] Moreover, FECH activity in Japanese healthy controls, homozygous for the **C** variant, was reported to be <50% of that reported for non-carriers, but anyway increased by 40% with respect to that of EPP patients.[36] Therefore the presence of the deep intronic pathogenic variant identified in this study, in *cis* with the c.315–48 **C** variant in one of the two hypomorphic alleles, could explain, at least in part, the variable outcome of EPP in homozygous **C** individuals worldwide.

According to our findings in the majority of our unrelated patients (23 of 24), the c.315–48 **C** variant is linked to a single haplotype encompassing the first 20Kb of the *FECH* gene. Among all the variants included in this haplotype, however the Haploview association analysis, recovered a significant association with EPP only for two other known variants: c.1 –252 **G** and c.68–23**T.** Both these single variants were functionally evaluated by in vitro analyses. The c.1–252A>**G** substitution in the promoter region has been reported to result in a slight decrease in *FECH* transcriptional activity.[37] While, the c.68–23C>**T** substitution in intron 1, was found to alter the pre-mRNA structure leading to exon 2 skipping during the splicing process.[38] Notably in our cohort, both these single variants were inherited in *trans* to a null allele in two subjects presenting very mild accumulation of protoporphyrins without any apparent clinical symptoms. This evidence suggests that both variants in *trans* to a null *FECH* allele can result in a slight decrease in the FECH activity and in a mild PPIX accumulation but are not sufficient to consistently cause clinical expression of the disease. Consistent with this hypothesis, extensive publicly available gene expression data from the GTEX study provide evidence for a decrease in the expression level of the *FECH* transcript also in individuals carrying the **GTT** haplotype. Taken together these observations are consistent with the idea that the hypomorphic allele is prevalently inherited in the form of the **GTC** haplotype. Considerations on the relative

frequency of the **GT**T and **GTC** haplotypes (**GT**T approx. 22% and **GTC** approx**.** 11%) in the healthy population suggest that **GTC** is a derived form of the **GT**T haplotype and that it is associated with a more marked decrease in the expression levels of the *FECH* gene.

The observation that in our cohort, one symptomatic patient was not carrying the c.1–252 **G** variant in the promoter suggests that this might not be required to induce an overt disease. On the contrary, all patients showed extended levels of linkage disequilibrium between the c.68–23**T** and c.315–48 **C** variants, located in introns 1 and 3 respectively, supporting the hypothesis that both are necessary for the lower steady state level of *FECH* mRNA resulting in protoporphyrin overproduction and photosensitivity. Importantly both these variants are associated with the creation of new exonic splicing silencer sites (ESS) according to the HSF tool. Moreover, we demonstrate that the c.68–23**T** variant alters the DNA methylation pattern by abolishing a methylated CpG site, with an opposite effect with respect to the c.464–1169A>C deep intronic pathogenic variant. Additionally, the identification of a non-coding splicing isoform, with a complete skipping of the constitutive exon 2 and the insertion of the 63bps of intron 3 strongly supports the conclusion that both the variants are required for clinically relevant down-regulation of the *FECH* gene. Considerations regarding the relatively low frequency of the A**TC** and AC**C** with regards to the **GTC** haplotype, suggest that the functional characterization of the **T** variants warrants further investigation and probably requires the study of a larger cohort of EPP patients worldwide.

In conclusion, our findings suggest that although the majority of EPP causing variants has been shown to have a "radical" effect on the coding sequence of the *FECH* gene, the presence of non-coding variants in pathogenic process should consistently be evaluated especially in EPP patients carrying only one hypomorphic allele. Moreover, we believe that this study supports the recent findings that methylation-dependent modulation of the pre-mRNA splicing patterns may function directly to control gene expression levels through the incorporation of "poison exons" leading to NMD.[39, 40] All in all the findings of this study confirm the validity of the hypothesis that "hidden" sources of

variability that are not normally considered in clinical genetic screenings might explain at least in part the variability in the clinical manifestations of diseases with incomplete penetrance.

ACKNOWLEDGEMENTS

The authors are grateful to all patients and their families who donated samples for this study. We thank Dr. Pasquale Missineo for biochemical analysis, D.ssa Valeria Fiorentino for support in the panel design, Archt. Luca Turchet for graphic assistance and Prof. David Horner for English editing and comments that greatly improved the manuscript. We also thank Prof.sa MD Cappellini for her constant and continuous support in the research activity.

CONFLICT OF INTEREST DISCLOSURE

This research was supported by grants from the Italian Ministry of Health (GR–2011–02347129 to Dr. Di Pierro) and in part from Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (RC2019).

AUTHORS' CONTRIBUTION

MC performed bioinformatic analyses and contributed to write the paper, IP prepared libraries for targeted resequencing experiments, LT and VBollati performed pyrosequencing analysis, LA and VBrancaleoni provided technical support for allehap and haploview bioinformatic tools, FG and VBrancaleoni performed the diagnostic genetic screening. EDP designed the study, performed the qualitative RNA experiments, supervised research and wrote the paper. All authors read and agreed to the final version of the manuscript, revised it critically and they have no relevant conflict of interest to disclose.

References

1. Lecha M, Puy H, and Deybach JC (2009) Erythropoietic protoporphyria. Orphanet J Rare Dis 4:19

2. Balwani M, Naik H, Anderson KE, Bissell DM, Bloomer J, Bonkovsky HL, Phillips JD, Overbey JR, Wang B, Singal AK, Liu LU, and Desnick RJ (2017) Clinical, Biochemical, and Genetic Characterization of North American Patients With Erythropoietic Protoporphyria and X-linked Protoporphyria. JAMA Dermatol 153 (8):789-796

3. Wahlin S, Stal P, Adam R, Karam V, Porte R, Seehofer D, Gunson BK, Hillingso J, Klempnauer JL, Schmidt J, Alexander G, O'Grady J, Clavien PA, Salizzoni M, Paul A, Rolles K, Ericzon BG, and Harper P (2011) Liver transplantation for erythropoietic protoporphyria in Europe. Liver Transpl 17 (9):1021-1026

4. Whatley SD, Mason NG, Khan M, Zamiri M, Badminton MN, Missaoui WN, Dailey TA, Dailey HA, Douglas WS, Wainwright NJ, and Elder GH (2004) Autosomal recessive erythropoietic protoporphyria in the United Kingdom: prevalence and relationship to liver disease. J Med Genet 41 (8):e105

5. Whatley SD, Mason NG, Holme SA, Anstey AV, Elder GH, and Badminton MN (2010) Molecular epidemiology of erythropoietic protoporphyria in the U.K. Br J Dermatol 162 (3):642-646

6. Balwani M, Doheny D, Bishop DF, Nazarenko I, Yasuda M, Dailey HA, Anderson KE, Bissell DM, Bloomer J, Bonkovsky HL, Phillips JD, Liu L, and Desnick RJ (2013) Loss-of-function ferrochelatase and gain-of-function erythroid-specific 5-aminolevulinate synthase mutations causing erythropoietic protoporphyria and x-linked protoporphyria in North American patients reveal novel mutations and a high prevalence of X-linked protoporphyria. Mol Med 19:26-35

7. Gouya L, Puy H, Robreau AM, Bourgeois M, Lamoril J, Da S, V, Grandchamp B, and Deybach JC (2002) The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype FECH. Nat Genet 30 (1):27-28

8. Barman-Aksozen J, Beguin C, Dogar AM, Schneider-Yin X, and Minder EI (2013) Iron availability modulates aberrant splicing of ferrochelatase through the iron- and 2-oxoglutarate dependent dioxygenase Jmjd6 and U2AF(65.). Blood Cells Mol Dis 51 (3):151-161

9. Schneider-Yin X, Mamet R, Minder EI, and Schoenfeld N (2008) Biochemical and molecular diagnosis of erythropoietic protoporphyria in an Ashkenazi Jewish family. J Inherit Metab Dis 31 Suppl 2:S363-S367

10. Mizawa M, Makino T, Nakano H, Sawamura D, and Shimizu T (2016) Incomplete erythropoietic protoporphyria caused by a splice site modulator homozygous IVS3-48C polymorphism in the ferrochelatase gene. Br J Dermatol 174 (1):172-175

11. Suzuki H, Kikuchi K, Fukuhara N, Nakano H, and Aiba S (2017) Case of late-onset erythropoietic protoporphyria with myelodysplastic syndrome who has homozygous IVS3-48C polymorphism in the ferrochelatase gene. J Dermatol 44 (6):651-655

12. Nakano H, Nakano A, Toyomaki Y, Ohashi S, Harada K, Moritsugu R, Takeda H, Kawada A, Mitsuhashi Y, and Hanada K (2006) Novel ferrochelatase mutations in Japanese patients with erythropoietic protoporphyria: high frequency of the splice site modulator IVS3-48C polymorphism in the Japanese population. J Invest Dermatol 126 (12):2717-2719

13. Li C, Di Pierro E, Brancaleoni V, Cappellini MD, and Steensma DP (2009) A novel large deletion and three polymorphisms in the FECH gene associated with erythropoietic protoporphyria. Clin Chem Lab Med 47 (1):44-46

14. Colombo FP, Rossetti MV, Mendez M, Martinez JE, Enriquez de SR, del CBA, and Parera VE (2013) Functional associations of genetic variants involved in the clinical manifestation of erythropoietic protoporphyria in the Argentinean population. J Eur Acad Dermatol Venereol 27 (6):754-762

15. Brancaleoni V, Granata F, Missineo P, Fustinoni S, Graziadei G, and Di Pierro E (2018) Digital PCR (dPCR) analysis reveals that the homozygous c.315-48T>C variant in the FECH gene might cause erythropoietic protoporphyria (EPP). Mol Genet Metab 124 (4):287-296

16. Chiara M, Gioiosa S, Chillemi G, D'Antonio M, Flati T, Picardi E, Zambelli F, Horner DS, Pesole G, and Castrignano T (2018) CoVaCS: a consensus variant calling system. BMC Genomics 19 (1):120

17. Wang K, Li M, and Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38 (16):e164

18. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536 (7616):285-291

19. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, and Abecasis GR (2015) A global reference for human genetic variation. Nature 526 (7571):68-74

20. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29 (1):308-311

21. Glusman G, Caballero J, Mauldin DE, Hood L, and Roach JC (2011) Kaviar: an accessible system for testing SNV novelty. Bioinformatics 27 (22):3216-3217

22. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44 (D1):D733-D745

23. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, and Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42 (Database issue):D980-D985

24. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, and Cooper DN (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136 (6):665-677

25. Liu X, Wu C, Li C, and Boerwinkle E (2016) dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat 37 (3):235-241

26. Barrett JC, Fry B, Maller J, and Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21 (2):263-265

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

28. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, and Feolo M (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res 42 (Database issue):D975-D979

29. Fiorentino V, Brancaleoni V, Granata F, Graziadei G, and Di Pierro E (2016) The assessment of noncoding variant of PPOX gene in variegate porphyria reveals post-transcriptional role of the 5' untranslated exon 1. Blood Cells Mol Dis 61:48-53

30. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, and Beroud C (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37 (9):e67

31. Smith, AF, Hubley, R, and Green, P. RepeatMasker Open-3.0. http://www.repeatmasker.org. 2010 Ref Type: Electronic Citation

32. Chasin LA (2007) Searching for splicing motifs. Adv Exp Med Biol 623:85-106

33. Sironi M, Menozzi G, Riva L, Cagliani R, Comi GP, Bresolin N, Giorda R, and Pozzoli U (2004) Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res 32 (5):1783-1791

34. Lev MG, Yearim A, and Ast G (2015) The alternative role of DNA methylation in splicing regulation. Trends Genet 31 (5):274-280

35. Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, Nissim-Rafinia M, Cohen AH, Rippe K, Meshorer E, and Ast G (2015) HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. Cell Rep 10 (7):1122-1134

36. Tahara T, Yamamoto M, Akagi R, Harigae H, and Taketani S (2010) The low expression allele (IVS3-48C) of the ferrochelatase gene leads to low enzyme activity associated with erythropoietic protoporphyria. Int J Hematol 92 (5):769-771

37. Di Pierro E, Cappellini MD, Mazzucchelli R, Moriondo V, Mologni D, Zanone PB, and Riva A (2005) A point mutation affecting an SP1 binding site in the promoter of the ferrochelatase gene impairs gene transcription and causes erythropoietic protoporphyria. Exp Hematol 33 (5):584-591

38. Nakahashi Y, Fujita H, Taketani S, Ishida N, Kappas A, and Sassa S (1992) The molecular defect of ferrochelatase in a patient with erythropoietic protoporphyria. Proc Natl Acad Sci U S A 89 (1):281-285

39. Carvill GL, Engel KL, Ramamurthy A, Cochran JN, Roovers J, Stamberger H, Lim N, Schneider AL, Hollingsworth G, Holder DH, Regan BM, Lawlor J, Lagae L, Ceulemans B, Bebin EM, Nguyen J, Barsh GS, Weckhuysen S, Meisler M, Berkovic SF, De JP, Scheffer IE, Myers RM, Cooper GM, and Mefford HC (2018) Aberrant Inclusion of a Poison Exon Causes Dravet Syndrome and Related SCN1A-Associated Genetic Epilepsies. Am J Hum Genet 103 (6):1022-1029

40. Yan Q, Weyn-Vanhentenryck SM, Wu J, Sloan SA, Zhang Y, Chen K, Wu JQ, Barres BA, and Zhang C (2015) Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. Proc Natl Acad Sci U S A 112 (11):3445-3450

FIGURE LEGENDS:

Figure 1: Scheme of the transcriptional isoforms of the *FECH* gene analyzed in this study and RT-PCR experiments. In Panel A, gray stars are used to indicate the variants that form the GTC haplotype, the black triangle indicates the deep intronic variant in intron 4, identified in this study and the gray point indicates the common variant in intron 7. The asterisk marks the presence of a stop codon in the mature transcript. Primers used for the RT-PCR and nested RT-PCR are also displayed. Panel B shows the PCR product spanning exons 4 to 8. Lane M is the molecular weight marker. The GTC homozygous symptomatic patients carrying the deep intronic pathogenic variant are indicated as Pt1 to Pt4. The following individuals were used as controls: an asymptomatic GTC homozygous carrier, an EPP patient with the common GTC haplotype trans to a pathogenic variant and an healthy subject. Panel C shows the re-amplification of the PCR fragment with primers encompassing exons 4–8, 7–8, and 4–6, respectively.

Figure 2: Sanger sequencing of the PCR product encompassing exons 4-6. The sequence shows the insertion of a fragment corresponding to a portion of intron 4 into the *FECH* transcript. The upper panel shows the beginning of the inserted sequence, the middle panel highlights the presence of a stop codon and the bottom panel shows the junction with exon 4.

Figure 3: Sequence Logo of the conserved haplotype associated with the hypomorphic allele. Common SNPs that form the haplotype are designated by their respective dbSNP rs code (X-axis). A red rectangle is used to illustrate the polymorphic positions that are conserved between all the carriers of the haplotype. The G (rs17063095), T (rs2269219), and C (rs2272783) variants are marked by a purple square.

Figure 4: Comparison of *FECH* gene expression in publicly available data. A Boxplots of fold change of expression levels of *FECH* in 136 individuals from the GTEx study carrying the GTC, GTT and ATC haplotypes. Median expression of *FECH* across all the 48 tissues considered in GTEx is used as the baseline for the calculation of the fold change. Fold changes are expressed using a base 2 logarithmic scale. Values lower than 0 indicate down-regulation. Values higher than 0 indicate up regulation. **B** Barplot of haplotype frequencies in a healthy population and in our cohort EPP patients. Haplotypes are indicated based on the G (rs17063095), T (rs2269219), and C (rs2272783) variants.

Table 1: DNA methylation analysis.

Methylation levels are reported in the form of dinucleotides percentage of methylation. Relevant differences between the genotype are underlined. A gray scale is used to indicate low, medium and high levels of methylation.

Conflict of Interest Statement

FONDAZIONE IRCCS CA' GRANDA
OSPEDALE MAGGIORE POLICLINICO
DIPARTIMENTO MEDICINA INTERNA
U.O. MEDICINA INTERNA 1A
DIRETTORE MARIA DOMENICA CAPPELLINI

Milan, 4th June 2019

To Genetics in Medicine

Editorial Office

This research was supported by grants from the Italian Ministry of Health (GR–2011–02347129 to Dr. Di Pierro) and in part from Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (RC2019).

Best regards

Elena Di Pierro
Fondazione IRCCS "Cà-Granda" Ospedale Maggiore Policlinico
Dipartimento di Medicina Interna
Tel:+390255036155
Fax:+390250320296
e-mail: elena.dipierro@unimi.it

Figure 1

Click here to access/download;Figure;1.tif

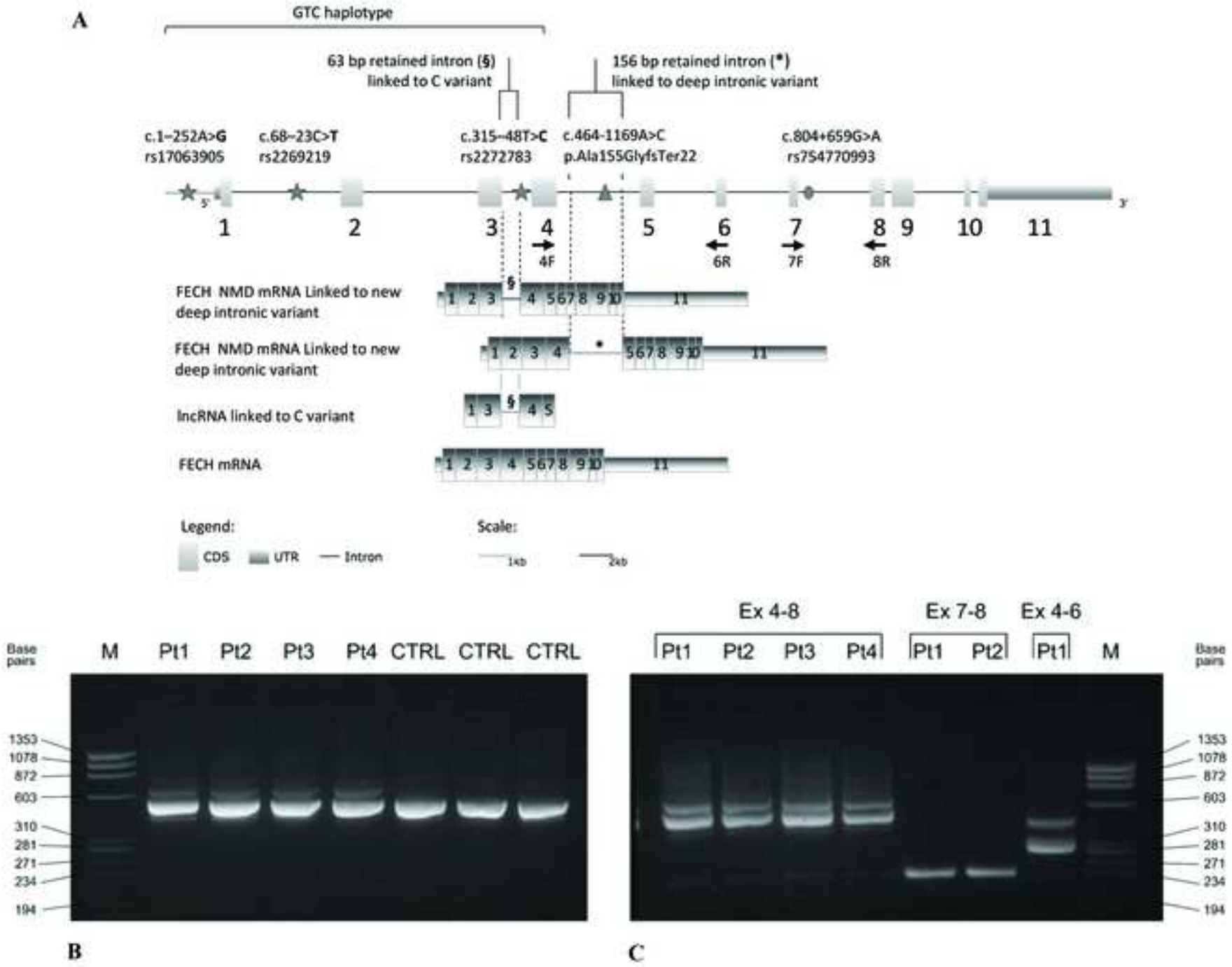

Figure 2

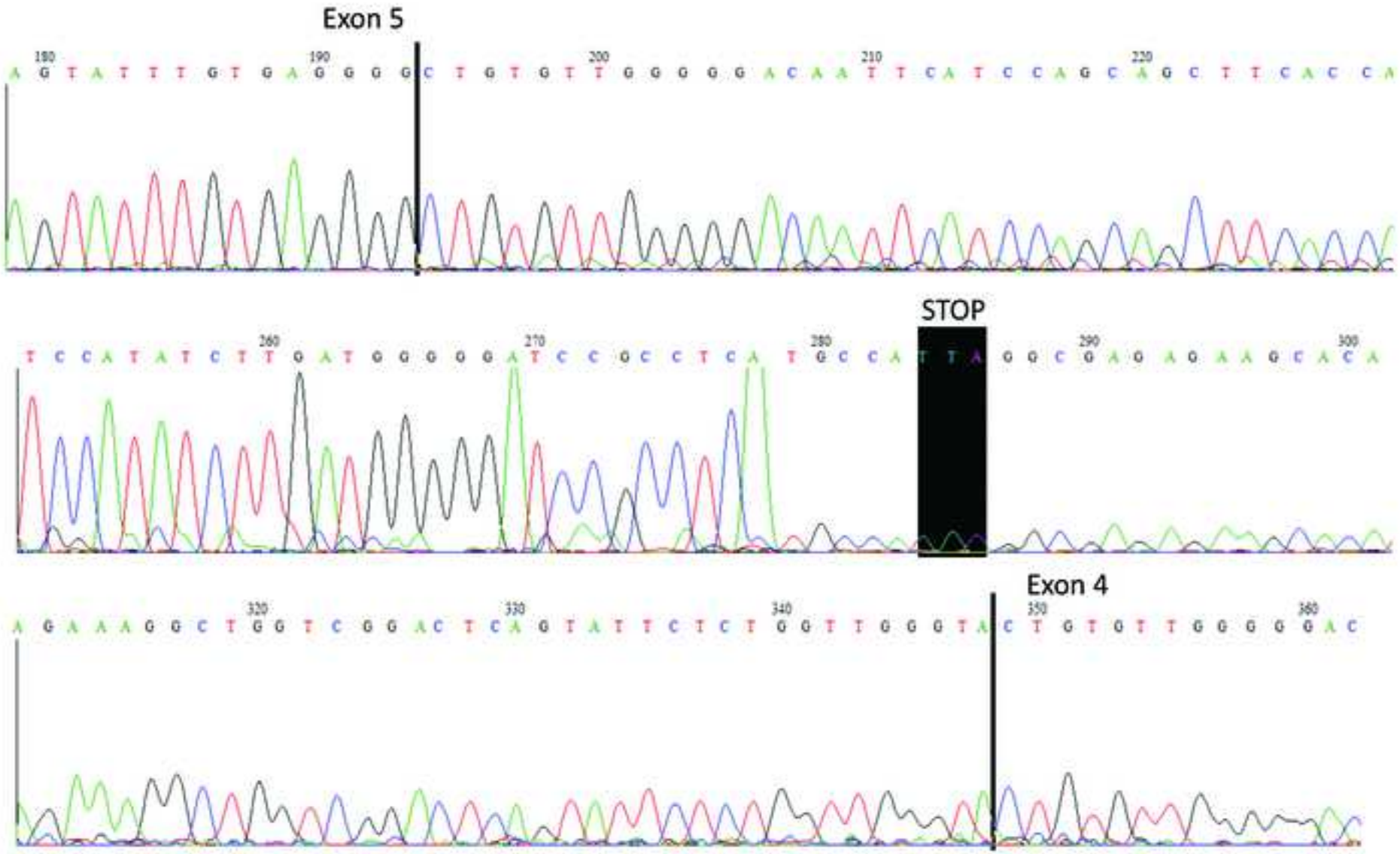

Figure 3

Click here to access/download;Figure;3.tif

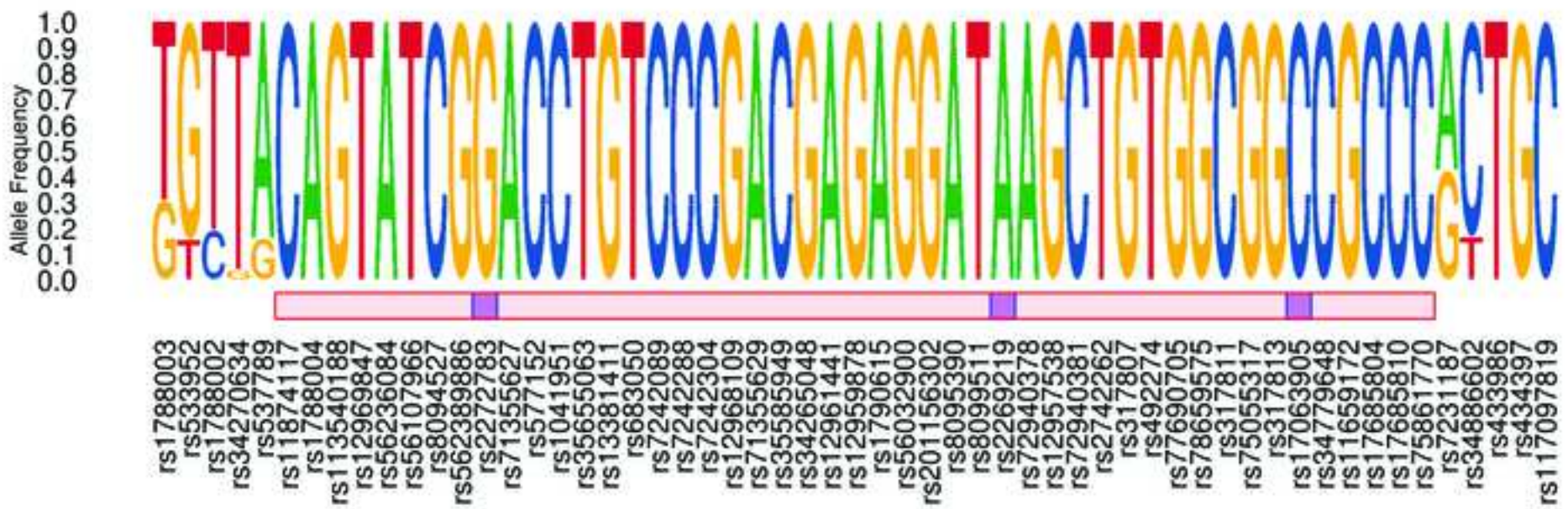

Figure 4

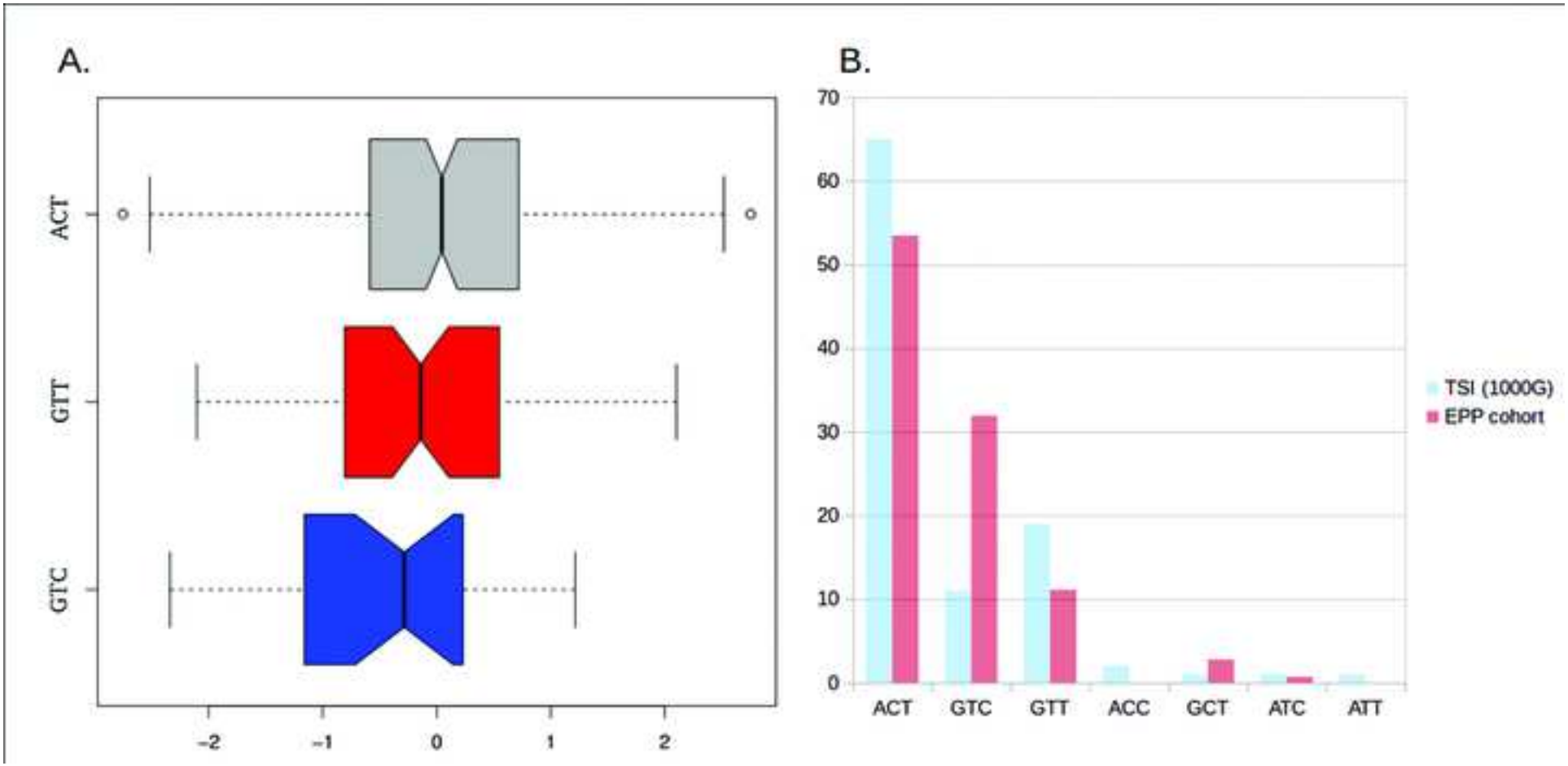

Figure S1

Click here to access/download;Figure;Slide1.tiff

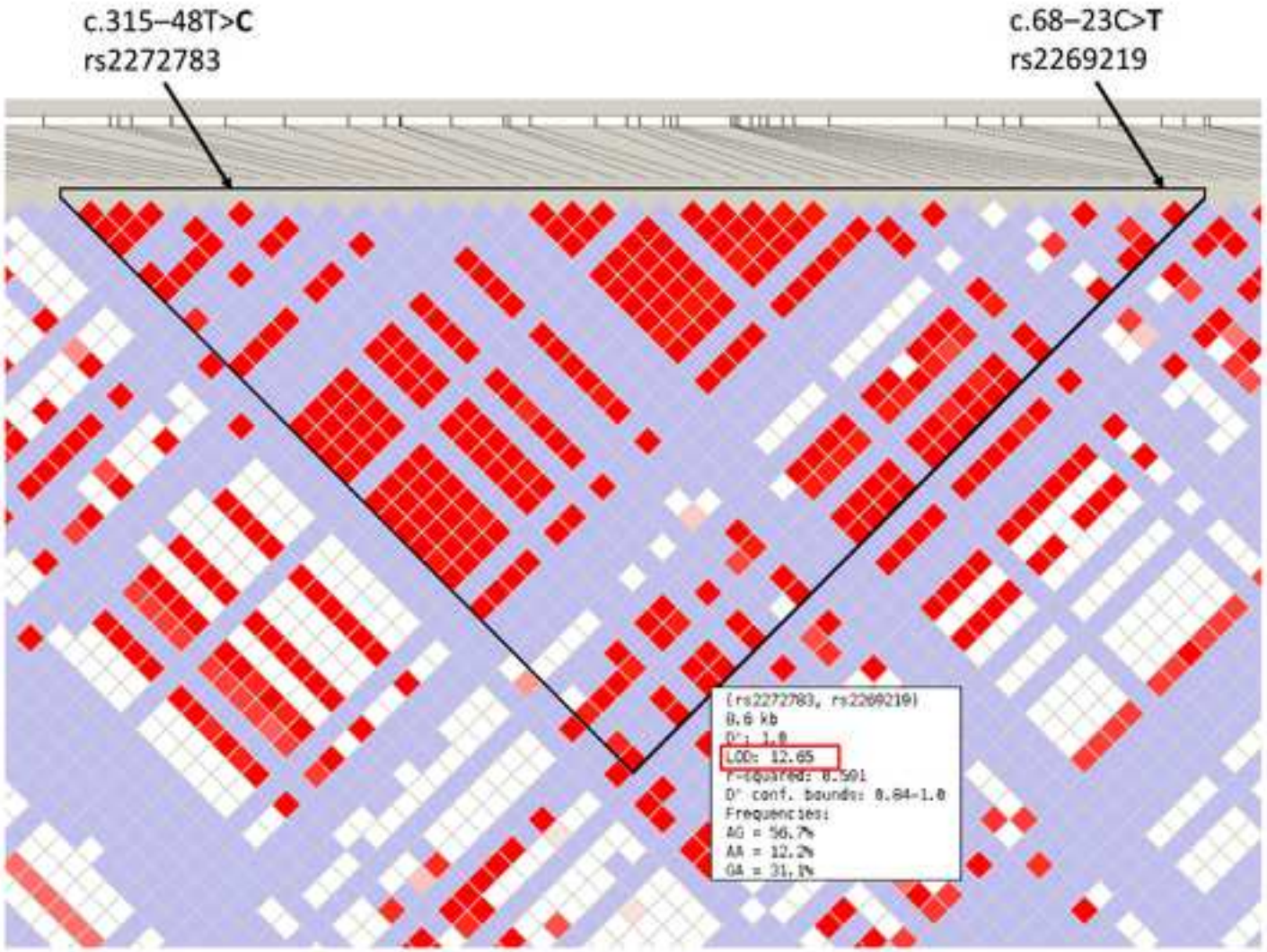

Figure S2

Click here to access/download;Figure;6.tif

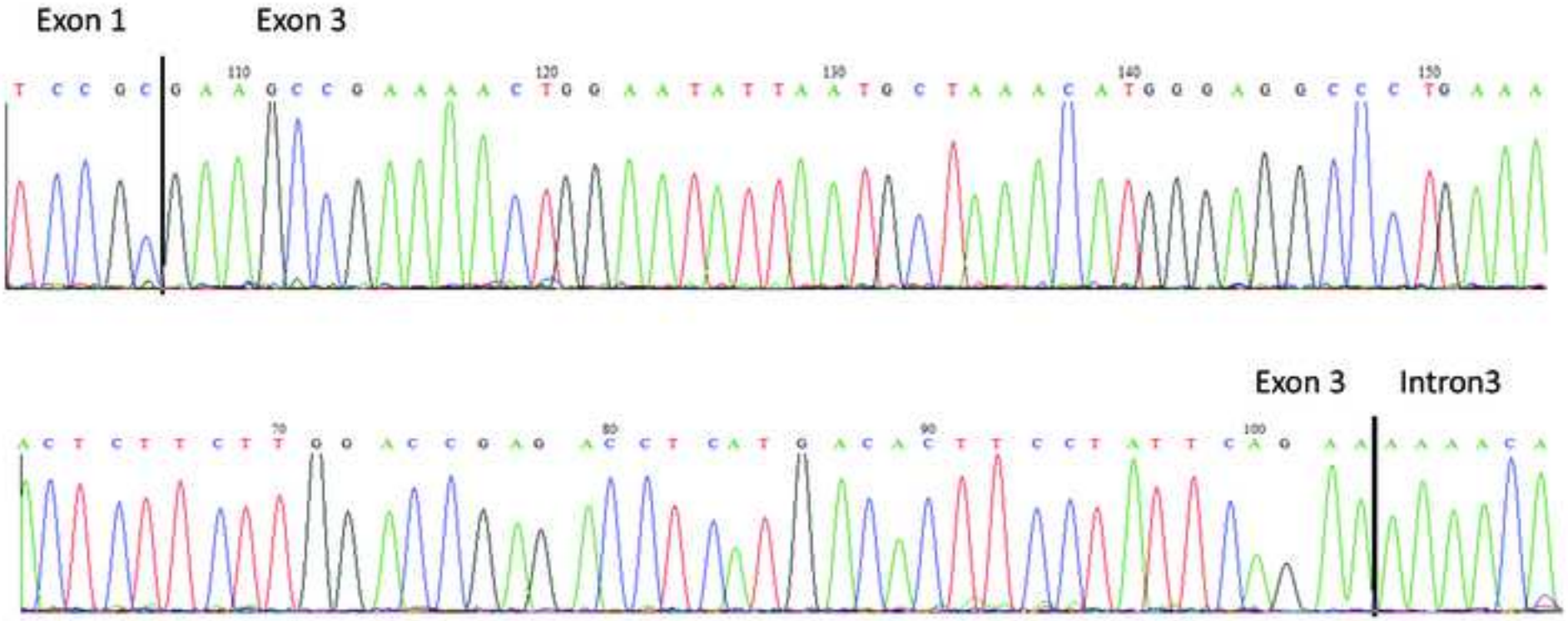

Table 1: The results of DNA methylation analysis.
The numbers indicate the percentage of methylation of each analyzed dinucleotides. The relevant differences according to the genotype are underlined. The gray scale indicates low, medium and high levels of percentage of methylation.

		Analized patient	CTRL1	CTRL2	CTRL3	1561	1553	1574	1611	1655	1704	717	1535	1185	713	1525	1526	1169	1262	1382	1683	1448
		Genotype	ACT/ACT	ACT/ATT	ATT/ATT	ACT/ACT	ACT/ACT	ACT/ACT	ACT/ACT	ACT/ACT	ACT/GTT	ACT/GTT	ACT/GTT	ACT/GTC	ACT/GTC	ACT/ATC	ACT/ATC	GTC/GTC	GTC/GTC	GTC/GTC	GTC/GTC	GTC/GTC
																		c.464–1169A>C	c.464–1169A>C	c.464–1169A>C	c.464–1169A>C	
Analyzed region	CpG dinucleotides																	c.804+659G>A	c.804+659G>A	c.804+659G>A	c.804+659G>A	
Promoter	CG1		0,6	1,8	2,0	2,3	0,5	0,0	0,5	0,0	nd	0,0	0,7	8,8	1,0	0,5	3,1	0,9	3,0	0,5	1,2	0,6
Sequencing A	CG2		0,0	0,6	1,1	1,1	0,0	0,0	1,7	1,4	nd	0,0	0,0	1,2	0,4	0,0	8,3	1,9	1,1	1,2	0,0	1,0
	CG3		4,0	2,6	3,8	4,0	3,7	5,3	3,6	4,6	nd	4,6	5,1	6,5	2,8	1,9	4,5	3,2	3,1	3,3	6,0	2,2
	CG4		0,0	0,0	1,1	0,7	0,0	3,4	0,7	0,0	nd	0,0	0,0	1,7	0,0	0,0	0,0	0,0	0,7	0,7	0,0	0,8
	CG5		0,7	0,6	0,0	1,1	2,8	0,0	1,2	1,7	nd	0,0	0,0	0,5	1,1	0,0	0,0	0,0	0,6	0,5	0,0	0,0
	CG6		0,0	0,0	0,8	0,6	0,0	0,0	0,0	0,0	nd	0,0	0,0	0,5	0,0	1,1	4,0	0,0	0,0	0,6	2,8	0,0
	CG7		0,0	0,8	1,9	0,9	0,0	0,0	2,4	1,0	nd	0,0	6,3	1,9	2,1	2,7	3,8	0,0	1,3	1,8	0,0	1,7
	CG8		2,9	2,7	2,5	4,0	5,0	4,3	3,5	4,4	nd	11,7	5,0	3,3	4,0	5,2	5,2	3,5	3,2	3,3	6,7	3,3
	CG9		1,6	1,9	1,5	1,9	2,3	2,8	4,8	2,6	nd	4,6	3,5	2,7	2,0	2,2	2,3	2,2	1,7	1,5	2,8	1,7
	CG10		3,2	3,1	3,0	3,4	5,1	6,2	4,7	6,7	nd	10,0	6,4	3,7	3,0	4,3	6,8	4,2	3,4	3,2	7,1	3,0
	CG11		1,6	1,8	1,5	2,0	2,1	2,2	2,3	2,9	nd	1,9	3,5	2,2	1,6	2,5	3,2	1,5	1,6	1,8	4,0	1,9
Promoter	CG1		0,0	0,0	1,3	0,0	1,4	1,3	0,0	0,0	0,0	0,0	0,0	0,7	1,6	0,0	0,0	0,0	1,3	0,0	0,0	0,0
Sequencing B	CG2		0,9	2,2	1,0	0,0	1,7	0,9	1,0	0,0	0,0	1,5	0,0	1,9	2,0	1,5	0,0	0,0	1,9	2,0	1,8	2,4
	CG3		9,8	9,9	9,5	10,9	9,0	8,5	9,8	9,5	16,0	13,7	10,6	6,7	14,2	13,8	14,8	8,3	9,9	8,6	16,8	11,6
	CG4		1,0	2,2	1,0	0,0	1,2	1,9	0,0	0,0	0,0	0,0	0,0	1,3	1,6	0,0	2,0	0,0	0,0	0,0	1,6	0,0
	CG5		1,7	2,9	2,4	3,9	2,5	2,5	3,5	3,6	0,0	0,0	1,9	2,5	4,4	0,0	2,6	2,3	3,3	3,0	1,7	3,8
	CG6		0,0	0,0	0,0	0,0	2,2	0,0	1,1	0,0	0,0	0,0	0,0	1,5	0,0	0,0	0,0	0,0	0,9	0,0	0,0	1,0
	CG7		0,0	1,1	0,0	1,2	0,0	2,6	1,2	0,0	0,0	1,5	0,0	0,8	0,0	0,0	0,0	0,0	1,1	2,3	0,0	0,0
Intron 1	CG1		92,7	93,3	93,6	95,1	95,0	92,2	96,3	94,5	96,4	93,3	94,2	97,1	96,7	93,8	93,6	94,3	97,6	95,3	95,9	94,7
	CG2		94,5	96,6	96,4	96,0	96,2	96,7	97,1	95,7	96,6	96,4	96,2	97,1	98,4	96,6	96,6	97,0	96,6	95,6	96,8	96,8
	CG3 pol	c.68-23C>T	72,0	38,4	3,6	73,8	66,9	71,3	75,6	74,9	38,2	39,4	32,0	41,6	42,7	38,9	34,6	2,2	2,2	3,8	2,5	3,2
	CG4		82,8	74,3	75,4	80,7	72,8	82,6	82,9	80,1	84,5	82,2	73,9	84,5	80,6	77,8	75,6	69,0	84,3	71,1	83,5	79,9
	CG5		73,4	66,5	67,8	72,9	64,9	70,0	74,5	72,2	78,6	73,4	66,3	76,2	65,1	72,1	70,0	60,7	78,7	63,4	78,1	72,7
	CG6		83,2	76,1	78,2	79,5	68,1	76,9	78,2	79,7	82,6	77,6	68,7	83,6	77,7	79,6	76,5	73,3	84,4	65,9	81,9	77,1
Intron3	CG1		95,6	92,2	93,3	88,6	91,4	91,1	91,3	89,9	92,1	92,5	90,5	91,0	85,3	92,9	92,7	90,5	93,5	92,3	92,0	88,7
	CG2		85,1	81,0	81,0	86,0	83,3	80,3	80,9	79,6	82,6	84,5	82,8	84,3	83,5	84,7	82,3	83,9	80,7	81,4	80,5	85,1
	CG3		80,1	89,6	93,3	82,0	89,2	93,8	88,8	87,5	93,0	85,9	89,5	82,1	67,2	88,8	87,2	82,2	87,7	92,2	94,3	81,2
	CG4		99,1	95,4	95,0	95,0	98,0	92,1	95,4	93,0	95,7	94,6	97,9	98,6	96,3	94,1	95,1	93,3	97,9	96,6	96,6	96,8
	CG5		98,0	96,6	94,6	96,9	95,0	94,2	93,0	94,9	94,6	95,4	94,3	95,5	94,3	94,5	95,1	95,8	95,7	96,0	94,7	95,7
Intron4	CG1		95,4	99,6	99,7	nd	nd	nd	nd	nd	nd	nd	nd	98,7	99,3	96,2	98,3	98,6	99,9	100,0	98,7	98,9
	CG2 mut	c.464–1169A>C	8,1	6,2	5,6	nd	nd	nd	nd	nd	nd	nd	nd	6,9	7,5	11,9	8,7	54,3	53,3	52,2	52,6	7,3
Intron7	CG1 pol	c.804+659G>A	97,3	96,6	97,0	nd	nd	nd	nd	nd	nd	nd	nd	95,4	96,8	96,2	97,2	55,7	56,0	57,5	57,1	95,2

Table S1: list of 72 sequenced subjects; the ID number in bold indicates proband positive at biochemical tests; in gray are underlined individuals homozygous for the GTC hypomorphyic allele among them the **c.=** in bold indicates the presence of new deep intronic variant.

Family Number	ID number	sanger sequencing results		
		pathogenetic variant	*Cis*	*Trans*
1	235	c.215dupT	ACT	ACT
	236	c.=	ACT	GTC
	237	**c.215dupT**	ACT	**GTC**
2	689	c.=	GTT	GTC
	690	c.215dupT	ACT	ACT
	691	**c.215dupT**	ACT	**GTC**
3	712	c.=	GTT	GTC
	713	**c.843delC**	ACT	**GTC**
	717	c.843delC	ACT	GTT
4-5	**760**	**c.400delA**	ACT	**GTC**
5	761	c.843delC	ACT	ACT
	762	**c.843delC**	ACT	**GTC**
6	**912**	**c.782C>T**	ACT	**GTC**
	922	c.782C>T	ACT	ACT
	923	c.=	GTT	GTC
7	**955**	**c.67+5G>A**	ACT	**GTC**
	956	c.=	ACT	GTC
	957	c.67+5G>A	ACT	ACT
8	**1025**	**c.215dupT**	ACT	**GTC**
	1611	c.215dupT	ACT	ACT
	1613	c.=	ACT	GTC
9	**1154**	**c.215dupT**	ACT	**GTC**
	1655	c.215dupT	ACT	ACT
	1656	c.=	ACT	GTC
10	**1185**	**c.215dupT**	ACT	**GTC**
	1704	c.215dupT	ACT	GTT
	1705	c.=	ACT	GTC
11	**1284**	**c.215dupT**	ACT	**GTC**
	1573	c.=	ACT	GTC
	1574	c.215dupT	ACT	ACT
12	**1352**	c.215dupT	ACT	**GTC**
	758	c.=	ACT	GTC
	759	c.215dupT	ACT	ACT
13	**1391**	c.343C>T	GTT	**GTC**
	1447	c.343C>T	GTT	ACT
	1448	c.=	**GTC**	**GTC**
	1604	c.=	ACT	GTC
	1605	c.=	ACT	GTC
	1603	c.=	ACT	ACT
14	**1526**	c.67+5G>A	ACT	**ATC**

	1535	c.67+5G>A	ACT	GTT
	1536	c.=	ACT	ATC
	1525	c.67+5G>A	ACT	**ATC**
	1537	c.=	ACT	GTT
	1749	c.67+5G>A	ACT	ACT
	1750	c.=	ACT	GTT
15	**1532**	c.706-3C>G	ACT	**GTC**
	1561	c.706-3C>G	ACT	ACT
	1562	c.=	ACT	GTC
16	**1550**	c.544delC	ACT	**GTC**
	1552	c.=	ACT	GTC
	1553	c.544delC	ACT	ACT
17	**1754**	c.215dupT	ACT	**GTC**
	1765	c.=	ACT	GTC
	1766	c.215dupT	ACT	ACT
18	**1691**	c.315-15T>A	ACT	**GTC**
19	**882**	c.=	GTT	**GTC**
20	**1061**	c.=	GTT	**GTC**
21	**1169**	**c.=**	**GTC**	**GTC**
	1545	c.=	ACT	GTC
22	**1262**	**c.=**	**GTC**	**GTC**
23	**1382**	**c.=**	**GTC**	**GTC**
	1534	c.=	ACT	GTC
24	**1683**	**c.=**	**GTC**	**GTC**
CTRL family 1	249	c.=	ACT	ATC
CTRL	250	c.=	GTT	ATC
CTRL	251	c.=	ACT	ACT
CTRL family 2	1071	c.=	GTT	GTC
CTRL	1072	c.=	GTT	GCT
CTRL	1073	c.=	ACT	GTC
CTRL family 3	847	c.=	ACT	GTC
CTRL family 4	1728	c.=	ACT	GTC

Table S2: Primers used for RNA Analysis

Analyzed Region	Primer name	Sequence
Exons 4–8	FECH Q4F	5’GAGGCGGATCCCCCATCAAGATATG3’
	FECH Q8R	5’ATTGCCACACCAGTCGGTAG3’
Exons 4–6	FECH Q4F	5’GAGGCGGATCCCCCATCAAGATATG3’
	FECH Q6R	5’GCCACCTGTCAATAGTGCTCCACT3’
Exons 7–8	FECH Q7F	5’GATCATATTCTAAAGGAACTGGACCA3’
	FECH Q8R	5’ATTGCCACACCAGTCGGTAG3’
Exon 1–3-ins	FECH Q1–3F	5’TCCGCGAAGCCGAAAACT3’
	FECH Q1–3R	5’GTTTTTTCTGAATAGGAAGTGTCATGA3’

Table S3: Assays conditions used for DNA methylation analysis

Region	Primer name	Sequences	Analyzed Sequence
Promoter	Forward	5'BIO-GTTGAGTTATGGTTGAGGATTTTG3'	5'C**A/G**C**A/G**CCCTCCC**A/G**AC**A/G**C**A/G**CCC**A/G**CTCATTC**A/G**CTACC**A/G**AACC**A/G**AAAC**A/G**AAC**A/G**AAACC3'
	Reverse	5'TCCCCAACCCCCTAACCT3'	
	Sequencing A	5'TCCCCAACCCCCTAACCT3'	
	Sequencing B	5'AAACCAAACAAAAACAC3'	5'**A/G**CCCCTC**A/G**C**A/G**AAAAACCC**A/G**CCCCAACTAAAC**A/G**CC**A/G**CTCCCTAC**A/G**TA3'
Intron 1	Forward	5'AGGAGGTGTGTGTAGTTTTTAAAATG3'	5'TTTTTTTGTAGGTTTTTAT**C/T**GGTTATTTTAGGGGAG**C/T**GATTTTTTATTTT3'
	Reverse	5'BIO-AACTTCCACCTCCATAACTAACAAAC3'	
	Sequencing A	5'AGGAGGTGTGTGTAGTTTTTAAAATG3'	
	Sequencing B	5'AATTATTTTTTAGTTAGATT3'	5'TTATG**C/T**GAGTATTTTAATTTT3'
	Forward	5'BIO-AGTTATGGAGGTGGAAGTTAGGTG3'	5'AC**A/G**ACC**A/G**CTACAACTACACCTAACTT CC3'
	Reverse	5'CACCTTTCCTCCCAAACAACTT3'	
	Sequencing A	5'ACTAAACTATTTCTATAATA3'	
	Sequencing B	5'CACCTTTCCTCCCAAACAACTT3'	5'ATATCTATTAATATCTACATC**A/G**ATAAAAAAAAA3'
Intron 3	Forward	5'TAGTTGGGTATTTTTTAGAGAGGG3'	5'T**C/T**GTTAAA**C/T**GT**C/T**GAATTTTTAAGATTTAAG3'
	Reverse	5'BIO-CAATTCATCCAACAACTTCACC3'	
	Sequencing A	5'TTTAGTAAGTTGGTATTATTTA3'	
	Sequencing B	5'TTTAAGATTTAAGAGTAGTA3'	5'T**C/T**GTAGGATTGGAGG**C/T**GGATTTTTTATTAAGA3'
Intron 4	Forward	5'BIO-TGGTTTTTAGTTTTTTGGGTTAAG3'	5'AAAAC**A/G**ACATTCCTC**CG/A/T**AA3'
	Reverse	5'ACCACAAAAAACATTTAAATTAA3'	
	Sequencing	5'TTTATACCTAAATATATACA3'	
Intron 7	Forward	5'GTGGGATAGTTAAGAGAAGGTTGT3'	5'TAAG**C/T**GTTTTTAA3'
	Reverse	5'BIO-TCCCAACTTAATCCTCTATATTCA3'	
	Sequencing	5'GGGGGGTTTTTATTTAGA3'	

SUPPLEMENTAL LEGENDS:

Figure S1: Haploview analysis. The figure shows a large block of the linkage disequilibrium (LD) between intron 1 and intron 4 of the *FECH* gene. The LD block was identified by Haploview using the method described in Gabriel et al. 2002. A LOD score of 12.65 was obtained for the c.68–23**T** and c.315–48 **C** variants, where a LOD > 2 indicates significant LD.

Figure S2: Sanger sequencing analysis. The figure shows the sequence of the long non-coding transcript. The black line indicates the point of junction between the exons.

Table S1: List of the 72 subjects included in this study; probands positive to the biochemical tests are indicated in bold; individuals homozygous for the GTC hypomorphyic allele are underlined in gray. The **c.=** symbol in bold is used to indicate the presence of new deep intronic variant.

Table S2: Primers used for RNA Analysis

Table S3: Assays conditions used for DNA methylation analysis

Table S4: Detailed coverage statistics of the target region. Base by base coverage of the targeted genomic locus (chr18 55202704-55292856). For every targeted position, the median coverage (Med), upper quartile (75th percentile, UQ) and 90th percentile (90th_p) of the coverage distribution, calculated on the 72 individuals included in this study are reported. A boolean value (1 yes, 0 no) is used to indicate regions showing reduced levels of coverage, that is a upper quartile of the coverage distribution lower than 10 (L10) or lower than 20 (L20). Due to limitations in the number of rows that are allowed by MS-excels, data are reported in two blocks of columns. A complete table with full coverage data for all the individuals included in the study can be retrieved from:

https://github.com/matteo14c/supplementary_files_chiara_dipierro_et_al_ajhg. To avoid the aforementioned limitations on the number of rows, the full table is provided in csv (comma separated values) and ods (open office spreadsheet) format.

Table S5: List of genetic variants identified in this study. List of the 510 genetic variants identified in the targeted genomic locus in the 72 individuals included in the study. Variants are reported by genomic coordinates on the hg19 human genome reference assembly. For every polymorphic position the following information is also provided: Variant Quality Score (as calculated by the GATK haplotype caller), median coverage (Median_Cov), total depth of coverage (DP), allele frequency in the cohort (AF), presence/absence (Yes/No) in the collection of publicly available database of human genetic variation used in this study.

Table S6: Case-control chi square test results from Haploview analysis

LICENCE TO PUBLISH

SPRINGER NATURE

Manuscript Number: GIM-D-19-00032

Journal Name: Genetics in Medicine (the "Journal")

Proposed Title of the Article: Targeted re-sequencing of FECH locus reveals that a novel deep intronic pathogenic variant and eQTLs may cause erythropoietic protoporphyria (EPP) through a methylation-dependent mechanism. (the "Article")

Author(s) [Please list all authors, continuing on a separate sheet if necessary]: Matteo Chiara PhD, Ilaria Primon BSc, Letizia Tarantini BSc, Luca Agnelli PhD, Valentina Brancaleoni MSc, Francesca Granata MSc, Valentina Bollati PhD, Elena Di Pierro PhD (the "Author(s)")

Miscellaneous [for office use only]:

1. American College of Medical Genetics and Genomics (the 'Licensee') will consider publishing this article, including any supplementary information and graphic elements therein (e.g. illustrations, charts, moving images) (the 'Article'). Headings are for convenience only.

2. Grant of Rights
In consideration of the Licensee evaluating the Article for publication, the Author(s) grant the Licensee the exclusive (except as set out in clauses 3, 4 and 5a) iv) and sub-licensable right, unlimited in time and territory, to copy-edit, reproduce, publish, distribute, transmit, make available and store the Article, including abstracts thereof, in all forms of media of expression now known or developed in the future, including pre- and reprints, translations, photographic reproductions and extensions. Furthermore, to enable additional publishing services, such as promotion of the Article, the Author(s) grant the Licensee the right to use the Article (including the use of any graphic elements on a stand-alone basis) in whole or in part in electronic form, such as for display in databases or data networks (e.g. the Internet), or for print or download to stationary or portable devices. This includes interactive and multimedia use as well as posting the Article in full or in part or its abstract on social media, and the right to alter the Article to the extent necessary for such use. The Licensee may also let third parties share the Article in full or in part or its abstract on social media and may in this context sub-license the Article and its abstract to social media users. Author(s) grant to Licensee the right to re-license Article metadata without restriction (including but not limited to author name, title, abstract, citation, references, keywords and any additional information as determined by Licensee).

3. Self-Archiving
Author(s) are permitted to self-archive a pre-print and an author's accepted manuscript version of their Article.
a) A pre-print is the Author's version of the Article before peer-review has taken place ("Pre-Print"). Prior to acceptance for publication, Author(s) retain the right to make a Pre-Print of their Article available on any of the following: their own personal, self-maintained website; a legally compliant, non-commercial pre-print server such as but not limited to arXiv and bioRxiv. Once the Article has been published, the Author (s) should update the acknowledgement and provide a link to the definitive version on the publisher's website: "This is a pre-print of an article published in [insert journal title]. The final authenticated version is available online at: https://doi.org/[insert DOI]"
b) An Author's Accepted Manuscript ("AAM") is the version accepted for publication in a journal following peer review but prior to copyediting and typesetting that can be made available under the following conditions:
i) Author(s) retain the right to make an AAM of their Article available on their own personal, self-maintained website immediately on acceptance,
ii) Author(s) retain the right to make an AAM of their Article available for public release on any of the following 6 months after first publication ("Embargo Period"): their employer's internal website; their institutional and/or funder repositories. AAMs may also be deposited in such repositories immediately on acceptance, provided that they are not made publicly available until after the Embargo Period.
An acknowledgement in the following form should be included, together with a link to the published version on the publisher's website: "This is a post-peer-review, pre-copyedit version of an article published in [insert journal title]. The final authenticated version is available online at: http://dx.doi.org/[insert DOI]"

4. Author's Retained Rights
Author(s) retain the following non-exclusive rights for the published version provided that, when reproducing the Article or extracts from it, the Author(s) acknowledge and reference first publication in the Journal:
a) to reuse graphic elements created by the Author(s) and contained in the Article, in presentations and other works created by them;
b) they and any academic institution where they work at the time may reproduce the Article for the purpose of course teaching (but not for inclusion in course pack material for onward sale by libraries and institutions); and
c) to reproduce, or to allow a third party Licensee to reproduce the Article in whole or in part in any printed volume (book or thesis) written by the Author(s).

5. Warranties
The Author(s) warrant and represent that:
a) (i) the Author(s) are the sole copyright owners or have been authorised by any additional copyright owner to assign the rights defined in clause 2, (ii) the Article does not infringe any intellectual property rights (including without limitation copyright, database rights or trade mark rights) or other third party rights and no licence from or payments to a third party are required to publish the Article, (iii) the Article has not been previously published or licensed, (iv) if the Article contains materials from other sources (e.g. illustrations, tables, text quotations), Author(s) have obtained written permissions to the extent necessary from the copyright holder(s), to license to the Licensee the same rights as set out in clause 2 but on a non-exclusive basis and without the right to use any graphic elements on a stand-alone basis and have cited any such materials correctly;
b) all of the facts contained in the Article are according to the current body of science true and accurate;
c) nothing in the Article is obscene, defamatory, violates any right of privacy or publicity, infringes any other human, personal or other rights of any person or entity or is otherwise unlawful and that informed consent to publish has been obtained for all research participants;
d) nothing in the Article infringes any duty of confidentiality which any of the Author(s) might owe to anyone else or violates any contract, express or implied, of any of the Author(s). All of the institutions in which work recorded in the Article was created or carried out have authorised and approved such research and publication; and
e) the signatory (the Author or the employer) who has signed this agreement has full right, power and authority to enter into this agreement on behalf of all of the Author(s).

6. Cooperation
a) The Author(s) shall cooperate fully with the Licensee in relation to any legal action that might arise from the publication of the Article and the Author(s) shall give the Licensee access at reasonable times to any relevant accounts, documents and records within the power or control of the Author(s). The Author(s) agree that the distributing entity is intended to have the benefit of and shall have the right to enforce the terms of this agreement.
b) The Author(s) authorise the Licensee to take such steps as it considers necessary at its own expense in the Author(s)' name and on their behalf if the Licensee believes that a third party is infringing or is likely to infringe copyright in the Article including but not limited to initiating legal proceedings.

7. Author List
After signing, changes of authorship or the order of the authors listed will not be accepted unless formally approved in writing by the Licensee.

8. Edits & Corrections
The Author(s) agree(s) that the Licensee may retract the Article or publish a correction or other notice in relation to the Article if the Licensee considers in its reasonable opinion that such actions are appropriate from a legal, editorial or research integrity perspective.

9. Governing Law
This agreement shall be governed by, and shall be construed in accordance with, the laws of the State of New York. The courts of New York, New York shall have the exclusive jurisdiction.

Signed for and on behalf of the Author(s): Elena Di Pierro
(PLEASE NOTE, ONLY HANDWRITTEN SIGNATURES ARE ACCEPTED)

Print Name: ELENA DI PIERRO

Date: 04 April 1919

Address: Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico Via F. Sforza 35, 20122 Milano

Corresponding Author Name: Dr. Elena Di Pierro
Manuscript Number: GIM-D-19-00032

Reporting Checklist

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. **Please respond completely to all questions relevant to your manuscript. For sections that are not applicable please fill in NA.** For more information, please read the journal's Guide to Authors.

⬓ Check here to confirm that the following information is available in the Material & Methods section:

- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represent **technical or biological replicates** (including how many animals, litters, culture, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**: (For small sample sizes (n<5) descriptive statistics are not appropriate, instead plot individual data points)
  - very common tests, such as *t*-test, simple χ^2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - **statistical test results**, e.g., ***P* values**;
  - definition of **'center values'** as **median or mean**;
  - definition of **error bars as s.d. or s.e.m. or c.i.**

Please ensure that the answers to the following questions are reported both **in the manuscript itself and in the space below.** We encourage you to include a specific subsection in the methods section each for statistics, reagents and animal models. Below, provide the text as it appears in the manuscript as well as the page number.

Statistics and general methods	**Text AND page number from manuscript**
1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? (Give text and page #)	not applicable
For animal studies, include a statement about sample size estimate even if no statistical methods were used.	not applicable
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established? (Give text and page #)	not applicable
3. If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it. (Give text and page #)	not applicable
For animal studies, include a statement about randomization even if no randomization was used.	not applicable

4. If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding. (Give text and page #)	not applicable
For animal studies, include a statement about blinding even if no blinding was done.	not applicable
5. For every figure, are statistical tests justified as appropriate?	not applicable
Do the data meet the assumptions of the tests (e.g., normal distribution)?	not applicable
Is there an estimate of variation within each group of data?	not applicable
Is the variance similar between the groups that are being statistically compared? (Give text and page #)	not applicable

Reagents	**Text AND page number from manuscript**
6. Report the source of antibodies (vendor and catalog number)	not applicable
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination	not applicable

Animal Models	**Text AND page number from manuscript**
8. Report species, strain, sex and age of animals	not applicable
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	not applicable

10. We recommend consulting the ARRIVE guidelines (*PLoS Biol.* **8**(6), e1000412,2010) to ensure that other relevant aspects of animal studies are adequately reported.

Human subjects	**Text AND page number from manuscript**
11. Identify the committee(s) approving the study protocol.	Fondazione IRCCS "Ca' Granda" ethics commitee. Methods, "Study subjects " section, pag. 4
12. Include a statement confirming that informed consent was obtained from all subjects.	Methods, "Study subjects " section, pag. 4
13. For publication of patient photos, include a statement confirming that consent to publish was obtained. For more information, please see http://www.icmje.org/recommendations/browse/roles-and-responsibilities/protection-of-research-participants.html.	not applicable
14. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent).	not applicable

15. For phase II and III randomized controlled trials, please refer to the CONSORT statement and submit the CONSORT checklist with your submission.

16. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines.

Data deposition	**Text AND page number from manuscript**
17. Provide accession codes for deposited data. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Microarray data	Submission of sequencing (fastq) data and genotypes (vcf) at the European Genome Phenome Archive is currently ongoing (EGA #335272). However accession numbers were not received before the deadline for the resubmission of the paper. Genetic variants identified in the study are provided in the form of pseudo vcf file with allele frequencies (table S5). Accession codes will be updated upon acceptance of the

Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available in the Guide to Authors. We encourage the provision of other source data in supplementary information or in unstructured repositories such as Figshare and Dryad. We encourage publication of Data Descriptors (see Scientific Data) to maximize data reuse.

18. If computer code was used to generate results that are central to the paper's conclusions, include a statement in the Methods section under "**Code availability"** to indicate whether and how the code can be accessed. Include version information as necessary and any restrictions on availability.	not applicable

Table S4

Click here to access/download

Large Excel File

Table_S4_small.xls

Table S5

Click here to access/download

Large Excel File

Table_S5.xls

Table S6

Click here to access/download

Large Excel File

Table S6.xls

COLOR ARTWORK FORM

This form must be completed for all papers. We will be unable to process your paper through to production until we receive instructions concerning color files. Please upload this form with your revised files.

JOURNAL: ***Genetics in Medicine***

ARTICLE TITLE TARGET RE-SEQUENCING OF FBCR LOCUS REVEALS THAT A NOVEL ....

MANUSCRIPT NUMBER GIM-D-19-00032

CORRESPONDING AUTHOR NAME DR.SSA ELENA Di PIERRO

Genetics in Medicine has charges for figures printed in color: $500 for the first color figure and $250 for each subsequent figure. Current ACMG members (excluding Student Members) who are first or senior/corresponding authors are exempt,* as are authors who have opted for Open Access.*

Please note that figures can appear online in color in the HTML version of your manuscript, and in black and white in the PDF/print version of the manuscript. Color figures will also be at the discretion of the editorial office.

Please check:

☒ Yes, my manuscript contains material that must be printed in color. I agree to pay the color charges in full and hereby authorize Nature Publishing Group to invoice me for the cost of reproducing color artwork in print.

☐ Yes, my manuscript contains material that should be in color in the online HTML version, but in black and white in the PDF/print version of the manuscript. No charges incurred.

☐ Yes, my manuscript contains material that must be printed in color, but I have chosen Open Access for my article or am an ACMG member, so am exempt.

☐ No, my manuscript does not contain material that must be printed in color.

Which figures should be printed in color? (e.g. Figures 1a, 2, 3b)

FIGURES 2 | 3

College Member's Name: _____

Membership will be verified by the ACMG, and false claims will be liable for the full color charge rate. Signed completion of this form constitutes a full and total acceptance of the terms listed. College membership has no bearing on the review process at *Genetics in Medicine*, and papers from members and nonmembers alike will be given equal consideration.

Color figures will be set close to the citation and in the best possible position.

20% VAT will be added to the total charge amount upon invoicing. This applies to all EU authors who do not provide a valid VAT number upon returning this form. Customers outside of the EU will not be charged VAT but local taxes will be added where applicable.

Signature: Elena Di Pierro VAT no. (if applicable): 04724150968

Print Name: ELENA Di PIERRO Date: 4 JUNE 2019

Updated 8/16/16

nature publishing group npg