The EAHAD Blood Coagulation Factor VII Variant Database

Muriel Giansily-Blaizot1, Pavithra M Rallapalli2, Stephen J Perkins2, Geoffrey Kemball-Cook3, Daniel J Hampshire4, Keith Gomez3, Christopher A Ludlam5 and John H McVey6

1Hematologie Biologique, CHU Montpellier, University of Montpellier, Montpellier, France.
2Department of Structural & Molecular Biology, University College London, UK.
3Katherine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, UK.
4Department of Biomedical Sciences, University of Hull, Hull, UK.
5University of Edinburgh, Edinburgh, UK.
6School of Biosciences & Medicine, University of Surrey, Guildford, UK.

Correspondence to:
John H. McVey
Department of Biochemical Sciences
School of Biosciences & Medicine
University of Surrey
Guildford GU2 7XH
UK
+44 (0)1483 686390
j.mcvey@surrey.ac.uk
https://orcid.org/0000-0002-7416-533X

Running title: Factor VII Variant Database

ORCID ID:
Muriel Giansily-Blaizot: 0000-0002-3705-6177
Pavithra M Rallapalli: 0000-0002-8261-6209
Stephen J Perkins: 0000-0001-9218-9805
Geoffrey Kemball-Cook: 0000-0002-2649-9503
Daniel J Hampshire: 0000-0002-1387-8926
Keith Gomez: 0000-0002-8934-0700
Christopher A Ludlam
John H McVey: 0000-0002-7416-533X
Abstract

Hereditary blood coagulation Factor VII (FVII) deficiency is a rare autosomal recessive bleeding disorder resulting from variants in the gene encoding FVII (F7). Integration of genetic variation with functional consequences on protein function is essential for the interpretation of the pathogenicity of novel variants. Here, we describe an integration of previous locus specific databases for F7 into a single curated database with enhanced features. The database provides access to in silico analyses that may be useful in the prediction of variant pathogenicity as well as cross-species sequence alignments, structural information and functional and clinical severity described for each variant where appropriate. The variant data is shared with the F7 LOVD. The updated database now includes 221 unique variants, representing gene variants identified in 728 individuals. Single nucleotide variants are the most common type (88%) with missense representing 74% of these variants. A number of variants are found with relatively high minor allele frequencies that are not pathogenic but contribute significantly to the likely pathogenicity of co-inherited variants due to their effect on FVII plasma levels. This comprehensive collection of curated information significantly aids the assessment of pathogenicity.

Key words

Factor VII deficiency, genetic variation, LSDB, haemostasis, blood coagulation disorders
Introduction

The initiation of blood coagulation and subsequent wound repair is a fundamental defence mechanism conserved in all vertebrates. Exposure of blood coagulation factor (F) VII/VIIa to cells expressing its cellular receptor and cofactor tissue factor (TF) is both necessary and sufficient to initiate blood coagulation in vivo, leading to the generation of thrombin and a fibrin clot (Figure 1A).

FVII is a zymogen of a vitamin K-dependent serine protease that is synthesized in the liver and circulates in plasma as a single-chain molecule (416 amino acids) at a concentration of approximately 0.5µg/mL (10nmol/L). In common with the other serine proteases of the coagulation network (FIX, FX, prothrombin and protein C) as well as protein S and protein Z, FVII has an N-terminal domain that contains 10 glutamic acid residues that are post-translationally modified by the addition of a carboxyl group to the ϑ-carbon by a vitamin K-dependent carboxylase. This ϑ-carboxyglutamic acid (GLA) domain confers affinity to negatively charged phospholipid membranes such as those of activated platelets, promoting the assembly of functional multiprotein complexes on these surfaces. The primary translation product of FVII (466 amino acids) contains a pre-pro leader sequence consisting of a secretory leader (signal) sequence and a highly conserved (pro) sequence found in other vitamin K-dependent proteins that directs the ϑ-carboxylation. The GLA domain is followed by two epidermal growth factor (EGF)-like domains, the connecting or activation peptide, and the serine protease domain (Figure 1B).

FVII is converted to its activated form FVIIa as the result of a single proteolytic cleavage between Arg212 and Ile213 (all residues are numbered according to Goodeve et al (Goodeve, Reitsma, & McVey, 2011) using Human Genome Variation Society (HGVS) nomenclature, numbering the initiation methionine of the reference protein sequence (NP_000122.1) as +1), and may differ from legacy numbering +60), producing a disulfide linked two-chain molecule. In blood, 4% of the total circulating FVII is in the form of FVIIa that has little functional activity in the absence of its cofactor TF. Unlike other members of the trypsin superfamily, the neo N-terminus generated upon activation of FVII fails to insert into the activation pocket leading to a non-optimal alignment of the catalytic machinery, rendering the FVIIa ‘zymogen like’ with significantly reduced catalytic activity. Binding of FVIIa to TF allosterically corrects this defect, transforming FVIIa into a catalytically competent enzyme. In addition, TF ensures optimal orientation and positioning of the FVIIa catalytic domain above the membrane for optimal interaction with its substrates, thereby
enhancing the proteolytic activity by $10^6$-fold. The substrates of the TF-FVIIa complex are blood coagulation FIX and FX. A schematic of TF-FVIIa initiated thrombin generation is shown in Figure 1A.

The FVII gene ($F7$) is located on chromosome 13 (13q34) and spans 14909 base pairs. $F7$ lies adjacent to the factor X gene ($F10$), 2118bp apart, suggesting gene duplication during evolution. Three alternative mRNA transcripts have been described. The longest reference transcript (NM_000131.4) has 9 exons and encodes the longest isoform composed of 466 amino acid residues including the secretory leader (1-20) and propeptide (21-60) required for appropriate $\gamma$-carboxylation of glutamic acid residues in the N-terminal GLA domain. Transcript NM_019616.4 lacks exon 2, which encodes residues 22-43 of the propeptide but encodes an identical mature FVII protein. NM_019616.4 is the most abundant and physiologically relevant mRNA transcript. Transcript NM_001267554.1 is the shortest transcript lacking exons 2, 3 & 4, thus encoding a protein lacking the critically functional GLA and EGF1 domains.

Deficiency of FVII is an autosomal recessive bleeding disorder with a highly variable phenotype that results from variants in $F7$. The clinical heterogeneity ranges from lethal to mild or even asymptomatic forms. In many cases, especially from Southern Europe and African countries, the identification of variants associated with an asymptomatic clinical phenotype arises from the use of non-human TF as the trigger for functional coagulation assays. The use of TF from animal sources or recombinant versions thereof may result in a discrepant in vitro result that is inconsistent with FVII activity (FVII:C) assays performed using human TF and the bleeding phenotype of the individual concerned (Bolton-Maggs, Hay, Shanks, Mitchell, & McVey, 2007). Whereas in the more common inherited bleeding disorders haemophilia A and B there is a good correlation between residual clotting activity and the severity of any associated bleeding, in FVII deficiency this relationship is less clear. This might be explained by the lack of sensitivity in assays in differentiating between a total absence of FVII:C activity from extremely low levels. Most individuals experience mild mucous membrane bleeding, menorrhagia and post-surgical bleeding but more significant events such as life threatening central nervous system bleeds are well recognized. Neonatal central nervous system bleeds that are often fatal are characteristic of severely affected cases and arise in individuals with variants that result in extremely low or undetectable FVII levels. This mirrors the phenotype observed in mice made null for $F7$ by homologous recombination who die either from fatal intra-abdominal haemorrhage in the peri-partum period or intracranial haemorrhage before the age of 24 days (Rosen et al., 1997).
The first web-based FVII variant database was established in 2001 (McVey, Boswell, Mumford, Kemball-Cook, & Tuddenham, 2001), however availability of time and funding impacted the ability to maintain the database as well as continuity of access to a stable URL. A locus specific database (LSDB) loses its relevance once it is not maintained and updated on a regular basis. Although a replacement F7 variant database (umd.be/F7) (Beroud et al., 2005) was developed, the European Association for Haemophilia and Allied Disorders (EAHAD) initiated a Coagulation Factor Variant Database Project with the aim of gathering together single gene variant databases involved in clinical bleeding disorders that would provide a single web portal (dbs.eahad.org) to LSDBs for genes in haemostasis, mirroring the data to Leiden Open Variation Databases (LOVD) thus addressing the concerns of ‘Yet another database?’ (den Dunnen, 2018). The first interactive coagulation factor web database was developed for F11 variants (Saunders, O’Connell, Lee, Perry, & Perkins, 2005) and the EAHAD databases build on this structure. The LSDBs would share a common architecture making navigation of the database(s) easier as well as providing greater support for maintenance of the sites. New databases for FVIII (f8-db.eahad.org) and FIX (f9-db.eahad.org) (Rallapalli, Kemball-Cook, Tuddenham, Gomez, & Perkins, 2013) gene variants have evolved from previously developed single gene variant coagulation database projects (Giannelli et al., 1990; Kemball-Cook, Tuddenham, & Wacey, 1998), incorporating new data, new analysis tools and a new common database architecture with new interfaces and filters. The project aims to improve the quality and quantity of information available to the haemostasis research and clinical communities, thereby enabling accurate classification of disease severity in order to make assessments of likely pathogenicity. The databases are curated by international experts in the field who are contributing to the ClinGen (clinicalgenome.org) initiative. We now report the development of the FVII gene (F7) Variant Database (f7-db.eahad.org).

**Materials and Methods**

**Database Structure**

The database was built on a common architecture developed for blood coagulation variant databases, using a MySQL platform and HTML, CSS, JavaScript, Perl and PHP interface. The first EAHAD database using this architecture was for F9 variants (Rallapalli et al., 2013). The database is available at f7-db.eahad.org. The variant data in the
databases is shared with LOVD; databases.lovd.nl/shared/genes/F7, which is a freely available gene centred collection of DNA variant data and is part of the GEN2PHEN and Human Variome projects (Fokkema et al., 2011).

**Identification of variants**

Data was initially imported from the original MRC FVII mutation database (McVey et al., 2001) and the UMD-F7 mutation database (umd.be/F7)(Beroud et al., 2005). Subsequently additional variants were identified in the published literature. All variants incorporated into the new database were verified for accuracy and HGVS nomenclature was generated and checked with Mutalyzer (mutalyzer.nl/). All data referring to individual cases with variants in the F7 database is pseudo-anonymized and no information is provided on the site that identifies individuals.

**Nomenclature**

It is particularly important in molecular genetic analysis that there is no confusion resulting from differences in variant nomenclature between laboratories/publications. Many coagulation genes were cloned and initially sequenced during the 1980s, prior to the introduction of standardised nomenclature. As a result, genes and proteins have their own idiosyncrasies of naming and numbering. This can lead to confusion in the laboratory, literature and diagnostic setting. To reduce confusion, an ISTH working group published recommendations that full adoption of standard gene names and of DNA and protein sequence variant numbering according to HGVS guidelines should be adopted for all genes/proteins in haemostasis (Goodeve et al., 2011). All variants described in the Factor VII gene (F7) variant database therefore conform to HGVS guidelines and are reported in relation to reference sequences. The longest transcript NM_000131.4 is used to describe DNA variants with the A of the ATG initiator methionine utilized as the sequence start point +1 and similarly the first methionine is numbered +1 in the protein reference NP_000121.1. This differs from legacy numbering of FVII protein amino acid variants where the signal peptide and propeptide were numbered negatively, and amino acid numbering started from the beginning of the mature protein. However, to aid comparison with previous variant descriptions, both HGVS and legacy numbering is provided for each variant. Variants are described using cDNA and protein reference sequences, however in future releases variants will also be described using a locus genomic reference (LRG) sequence. This description is currently available
from the Leiden Open Variation Database (LOVD; lovd.nl) to which we mirror all variants described in the EAHAD database.

**In silico analyses**

The database provides access to *in silico* analyses that may be useful in the prediction of variant pathogenicity, using open access software packages (Align-GVGD, PolyPhen-2, SIFT & PROVEAN). The minor allele frequency (MAF) from the Genome Aggregation Database (gnomAD; gnomad.broadinstitute.org) is also provided. Where a MAF is not presented this indicates that this variant has not been identified in the data set and is a rare variant or in a region that is not read to an adequate depth in the control datasets. The position of a missense variant can also be visualized if the variant appears in the molecular model based on the structure of active site inhibited FVIIa in complex with human soluble TF (Protein Data Bank: 2A2Q). Finally, the evolutionary conservation of the amino acid sequence of the protein at the variant residue in closely related species: chimpanzee, gorilla, gibbon, bushbaby and marmoset can be inspected. Further multiple sequence alignments from more distantly related species and multiple alignment of human vitamin K-dependent coagulation factor protease domains are also available from the AA Alignments tab.

**Database content**

A total of 728 individual cases with plasma FVII:C levels outwith the normal range, identified by *in vitro* functional coagulation assays, and variants identified in their F7 gene(s) are compiled in the FVII gene (F7) variant database (f7-db.eahad.org). 221 unique variants have been identified in the cohort. The database allows simple searches of variants based on nucleotide or amino acid numbering (HGVS and legacy), variant type, or location within the gene. In addition, advanced search options allow further refinement of searches based on variant effect, protein domain and severity (where known). A search returns the unique variants at the selected position in the gene or protein as a list with links where appropriate to in-depth analysis of the variation and anonymized patient information.

**Mutational spectrum**

The majority of the F7 variants reported are small lesions including deletions (7.7%), duplications (1.8%), insertions (0.5%), indel rearrangements (1.8%) all smaller than 20 nucleotides with the exception of two large deletions; and
single nucleotide substitutions (88.2%) with the majority of these being missense variants (74.3%) (Figure 2A; f7-db.eahad.org/statistics.html.php). F7 is located 2.8kb upstream of F10 and therefore large rearrangements often involve both genes leading to combined FVII and FX deficiency (Pavlova et al., 2015).

Variants have been identified throughout the gene with 14 identified in the 5' untranslated and flanking region, 184 in exons and 23 in introns. Similarly, variants have been identified in the coding sequence for all the protein domains: signal peptide (9), propeptide (9), GLA (18), EGF1 (13), EGF2 (17), activation peptide (7) and protease (116) domains (Figure 3). The 66bp exon 2 is an alternatively used exon present in the longest reference transcript (NM_000131.4) but absent from the most abundant and physiologically relevant mRNA transcript (NM_019616.3). It encodes additional amino acids in the propeptide sequence. Only 2 variants are described in this exon, the synonymous c.66C>T p.(Gly22=) and the c.130G>A p.(Val44Ile) variant. The increased number identified in the protease domain reflects the relative size of the coding sequence and not a higher variant rate.

Some sites however have multiple variants in a single codon (29 codons with 2 variants and 8 codons with 3 or more variants), for example p.Ala251 has 4 different amino acid substitutions associated with variants in both c.751G and c.752C.

There are 195 single nucleotide substitutions: 144 missense, 15 nonsense, 4 silent (synonymous) within the coding sequence and an additional 10 within the 5’ flanking region located within transcription factor binding sites, 2 in the 5’ untranslated region and 20 intronic. Tools for assessing the pathogenicity of these variants are discussed below but many missense variants are predicted to result in a variety of consequences including functional and misfolding defects. The variants in transcription factor binding sites have been extensively studied and shown to result in either increased or decreased transcription. The intronic variants are predicted to result in splice variants as a consequence of various mechanisms: 10 are located in the canonical “-1, +1, -2, +2” splice nucleotides, 4 in the critical intronic “+5” position, 2 are located at less conserved positions namely -11 and -12. In addition, a small 4 nucleotide-deletion within intron 8 (c.805+3_805+6del) and a synonymous variant located at the last nucleotide of exon 3 (c.291G>C or p.(Thr97=)) are also predicted to result in aberrant splicing (Millar et al., 2000; Pinotti et al., 1998).

Impact of common variants known to modulate FVII levels
A number of common variants have been identified in F7 (Table 1). Some of these variants, which are in strong disequilibrium with each other, have been extensively studied in relation to FVII:C and antigen (FVII:Ag) levels. The F7 promoter haplotype [c.-401G>T, c.-325_-324insCCTATATCCT, c.-122T>C], and the missense variant c.1238G>A p.(Arg413Gln) are associated with decreased levels of FVII:C and FVII:Ag of approximately 30% and 23% respectively (Bernardi et al., 1996) in patients and confirmed by in vitro studies (Hunault, Arbini, Lopaciuk, Carew, & Bauer, 1997; Pollak, Hung, Godin, Overton, & High, 1996). In contrast, the rare c.-402A allele is associated with increased transcriptional activity (van ’t Hooft et al., 1999) resulting in significantly higher FVII:C and FVII:Ag levels than individuals homozygous for the common c.-402G allele (Lindman, Pedersen, Arnesen, Hjerkinn, & Seljeflot, 2005).

Intron 8 is the location of a VNTR repeat of 37bp with its own modulating effect. A quantitative analysis of transcripts indicated a parallel decrease of the VNTR repeat number and mRNA relative expression. Both c.795_805+26[8] and c.795_805+26[7] repeats showed higher values than the reference c. 795_805+26[6], whereas c.795_805+26[5] repeats showed lower values (Pinotti et al., 2000). The in vitro results were confirmed in genetic association studies for the c.795_805+26[7] allele contributing to a 17% variance of FVII:C levels (Bernardi et al., 1996).

The impact of the common variants on the combined expression of function FVII:C activity is complex as they may be either inherited in cis or trans with the rare variant being analyzed. Often it will not be possible to assign the linkage because family studies are usually not reported. Nevertheless, it may be possible to draw conclusions of the impact of the common variants when they are present in a homozygous state. To illustrate this point, 8 patients heterozygous for the p.(Ala304Val) missense variant but with various combinations of the p.(Arg413Gln) common variant were compared to each other. Four of them were homozygous for the p.(Arg413Gln) allele resulting in the p.[(Ala304Val); Arg413Gln];[(Arg413Gln)] genotype whereas the remaining four were classified into the other possible combinations of genotypes: p.[(Ala304Val;Arg413Gln)];[(Arg413=)], p.[(Ala403Val;Arg413=)];[(Arg413=)] or p.[(Ala403Val;Arg413=)];[(Arg413Gln)] according to the provided pedigree (Alshinawi, Scerri, Galdies, Aquilina, & Felice, 1998). Interestingly, both p.[(Ala304Val;Arg413Gln)];[(Arg413Gln)] and p.[(Ala403Val;Arg413=)];[(Arg413Gln)] patients with no wild type allele remaining, displayed the lowest FVII:C levels ranging from 25 to 32% (Tamary et al., 1996) whereas the p.[(Ala304Val;Arg413Gln)];[(Arg413=)] and p.[(Ala403Val;Arg413=)];[(Arg413=)] had the highest
FVII:C levels of 38-45% and 46% respectively. The location in cis or in trans of the modulating common variants will therefore be provided when available for future submissions to this database.

**Individual data**

In common with the other EAHAD coagulation factor databases, the F7 database stores and displays all cases reported regardless of how many individual reports of the identical genetic variant have been reported, allowing users to survey phenotypic variability among cases with the same genetic variant. Where appropriate additional variants within an individual case are reported, including the common F7 variants allowing analysis of combined genotypes of a rare variant and a common variant. It should be noted that in some cases the frequency of reporting variants is biased by extensive family studies or founder effects in certain ethnic groups, for example p.(Ala354Val) in North and North-East Europe (99 individuals) and p.(Ala304Val) in Jewish populations (38 individuals). It is possible to visualize/export the data either as a list of unique variants at the selected position, or as a multiple patient list showing all individuals with a particular variant at that position, allowing between-case comparisons to assess variant frequency and variability of presentation. Thus, despite the wide heterogeneity, 17 variants are highly frequent. Four missense variants are found more than 50 times, namely p.(Gln160Arg), p.(Ala304Val), p.(Ala354Val) combined with p.(Pro464Hisfs*32) and p.(Arg364Gln) (also known as FVII Padua). Five missense variants are found between 20 and 50 times and nine variants between 10 and 19 times (Figure 3).

**Impact of thromboplastin source for FVII:C measurement**

It is well-documented that some FVII protein variants display variations in the FVII:C measurement according to the species of the thromboplastin reagent (tissue factor; TF) that has been used to trigger the in vitro measurement. Historically bovine and rabbit sources were used before the introduction of recombinant human thromboplastin. The FVII:C value has therefore been recorded in the database with the species of the corresponding thromboplastin reagent where this information was available in the original report. Thirty-nine individuals were reported with FVII:C levels determined with thromboplastin of both human and rabbit origin. Seventeen displayed significant variations between FVII:C levels when measured with the different thromboplastins. The corresponding variants are: p.(Gly391Asp), p.(Arg364Gln), p.(Arg364Trp), p.(Arg337His), p.(Arg139Gln) (Matsushita, Kojima, Emi, Takahashi, & Saito, 1994; Mourey et al., 2014; O'Brien et al., 1991; Takamiya & Takeuchi, 1998; Zheng, Shurafa, & James, 1996).
Importantly, the impact of the residual FVII:C levels measured *in vitro* on the potential bleeding phenotype of the individual should only be considered when assayed with human thromboplastin.

**Clinical phenotype and genotype**

The database presents statistics and graphics on all the variants in the database by specific type of variant, by protein domain and by disease severity available from the *Variants* tab allowing users to analyze relationships between clinical phenotype and genotype. Mariani *et al* (Mariani *et al.*, 2005) proposed a standard classification of clinical severity for FVII deficiency which we would encourage all reporting of FVII deficiency to use. However, the lack of a standard classification of the observed clinical phenotype prior to 2005 requires users to refer back to the original publication to ascertain the scheme used if the phenotypic severity is to be used in the interpretation of the pathogenicity of the variant.

The 728 patients present a wide range of clinical phenotypes, from asymptomatic (34.6%; often identified in pre-operative coagulation screening) to individuals suffering severe bleeds (21.2%) (Figure 2B). The large majority of the 154 patients (92%) with severe bleeding tendency displayed FVII:C levels below 5%. Of those determined with human thromboplastin the majority displayed FVII:C levels below 2.9%. Genotypically, severe patients are either homozygous or compound heterozygous for nonsense, canonical splice site, frameshift, missense variants resulting in substitution of critical residues for the FVII function (namely, arginine residue at the proteolytic activation site (Arg212), the catalytic triad (Asp302), or residues involved in the disulfide bond (Cys195), between the light and heavy chains of the activated form of FVII), variants at transcription factor binding sites within the *F7* promoter or variants predicted to dramatically alter the protein folding namely p.(Gln160Arg) or p.(Thr419Met). However, the converse was not always true. Of the 65 individuals homozygous for variants predicted to be pathogenic (Table 2) there were 27 reported not to have a severe bleeding tendency. Two asymptomatic patients were homozygous for the p.(Arg462*) nonsense variant occurring only 4 residues before the natural termination signal. The 4 residue-carboxy-terminal-truncated protein has been reported to have increased specific activity and gain-of-function features (Branchini *et al.*, 2012) explaining the observed asymptomatic phenotype. Conversely, it was unexpected that an individual homozygous for the consensus splice site c.572-1G>A was reported with only epistaxis at the age of three (Kwon, Yoo, Lee, Kim, & Kim, 2011) or an individual homozygous for the critical variant p.(Arg212Gln)
reportedly with a mild bleeding phenotype despite FVII:C levels of <1% and 4 other cases homozygous for this variant reported as having severe haemorrhagic phenotypes. As for other clotting factors deficiencies, homozygotes for frameshift variants displayed the largest clinical heterogeneity. One explanation could be a partial correction by ribosomal slippage and DNA replication/RNA transcription errors as suggested for F8 (Young et al., 1997).

Assessing pathogenicity of F7 variants

Variant classification is central to the utility of molecular genetic diagnostics in clinical practice, however, predicting whether gene variants are likely to be pathogenic may not be straightforward. The F7 variant database currently does not assign pathogenicity to an individual variant but rather provides access to a number of tools to allow assessment of the variant according to published guidelines that establish a framework for variant classification (Nykamp et al., 2017; Richards et al., 2015). In future releases the database will link directly to ClinVar (ncbi.nlm.nih.gov/clinvar) which is an open access database that reports curated information and likely pathogenicity on gene variants. EAHAD curators are working as co-chairs or members of relevant curation panels that input into ClinVar.

Evidence for pathogenicity can be obtained from evaluation of variant frequencies from large population datasets; variant type; clinical observations; experimental studies and computational analyses. Each variant in the F7 database is assigned a variant ID and the variant ID view (Figure 4A) displays the cDNA change, the amino acid change, the type of variant (deletion, duplication, indel, insertion or point), the effect (frameshift, in-frame, intronic, missense, nonsense, promoter or silent), the location within the gene (untranslated region, flanking sequence, exon or intron) as well as within the protein (domain) and finally, the minor allele frequency (MAF) is obtained from the gnomAD database that calculates the frequency from 125,748 exomes and 15,708 genomes.

The MAF value can provide strong evidence that a variant is benign or that it is sufficiently rare to be considered a candidate pathogenic variant. However, MAF values can sometimes be misleading. For example, two variants with high MAFs of 0.1419 and 0.1341 respectively are p.(His175=) and p.(Arg413Gln). The high frequency and variant type would lead to the assignment of p.(His175=) as non-pathogenic. Whereas p.(Arg413Gln) is a frequent missense variant associated with decreased levels of FVII and although not directly pathogenic it can contribute to the consequence of co-inherited variants.
In order to assess the clinical observations each variant is associated with the number of individuals reported, laboratory measurements of FVII:C (TF: unknown, human or rabbit) and FVII:Ag plasma levels, reported clinical severity as well as whether the variant is homozygous or heterozygous. Additional variants identified in each individual are also collated. For each missense variant computational analysis using \textit{in silico} tools (Align-GVGD, Polyphen-2, SIFT and PROVEAN) to predict the likely impact on the resultant protein are provided. It is also possible to visualize the potential structural implications by viewing the variant in a molecular model based on the X-ray crystal structure of the TF-FVIIa complex or in multiple sequence alignments of the FVII amino acid sequence across primates or the serine protease domains of the human vitamin K dependent coagulation factors (FVII, FIX, FX, Protein C) (Figure 4B).

\textbf{Conclusions}

DNA sequence analysis (either classical Sanger DNA sequencing or more commonly exome/complete genome sequencing) is now standard in the characterisation of inherited disease. Most variants identified are single nucleotide variants and predicting the pathogenicity of these variants may be difficult. Deficiency of FVII is an autosomal recessive bleeding disorder with a highly variable phenotype, ranging from lethal to mild to even asymptomatic. The correlation between residual clotting activity and the clinical severity of associated bleeding is less clear than in the more common inherited bleeding disorders haemophilia A and B. This is further complicated by the use of non-human TF as the trigger for functional coagulation assays resulting in discrepant \textit{in vitro} results that are inconsistent with FVII:C assays performed using human TF and the bleeding phenotype of the individual concerned. This has often resulted in the identification of individuals with variants in their $F7$ gene who are asymptomatic. The co-inheritance of common variants that can affect the expression levels of plasma FVII:Ag further complicates the analysis of potential pathogenicity.

An integrated comprehensive analysis is therefore key to understanding the consequence of a variant. In the $F7$ variant database for any variant it is possible to obtain the description at cDNA and amino acid (both HGVS and legacy) levels, the type of variant, and the location of the variant within the gene and protein. Each individual reported to carry the variant is displayed along with the reported clinical severity, whether the individual is hetero-
or homozygous for the variant and any other variants identified in each individual. The generation and maintenance of an online open access database by a team of specialised curators that incorporates all described variants in F7 and their associated properties and functional activities in addition to in depth computational analysis of the consequence of the variant significantly aids the assessment of pathogenicity.

Data availability statement

The database is available at f7-db.eahad.org

Funding

The authors acknowledge the support of the EAHAD for this project.

Acknowledgments

The authors would like to acknowledge the support of the UMD database (Beroud et al., 2005) in providing F7 variants data from the UMD-F7 database (umd.be/F7/). The authors have no interests which might be perceived as posing a conflict or bias.

References


Figure Legends

Figure 1: Factor VII structure and function

A. Schematic of TF-FVIIa initiated thrombin generation. The network includes elements of coagulation initiation via TF-FVIIa interaction, amplification via FIXa-FVIIIa and FXa-FVa complexes leading to thrombin generation and formation of a fibrin clot. Positive feedback loops are represented by dashed lines coloured red and inhibition of the TF-FVIIa-FXa complex by tissue factor pathway inhibitor by dashed blue line.

B. Molecular model of Factor VIIa derived from the complex between active site inhibited FVIIa and the extracellular domain of tissue factor (Protein Databank: 1DAN). FVIIa protein chain is represented by a ribbon. The active site inhibitor is shown in ball and stick. Calcium ions are represented by green spheres. EGF, epidermal growth factor; GLA, γ-carboxylated.

Figure 2: Pie charts indicating the variant type and clinical severity

A. Variant type of the 221 unique variants in the database

B. Classification of the 728 individual cases in the database according to clinical severity

Figure 3: Frequency and location of the 221 unique variants

The number of individuals reported with variants in the amino acid sequence of the protein is plotted. The correspondence of the primary amino acid sequence with the domain structure of the protein is shown below with the number and percentage of unique variants reported in each domain.

Figure 4: Representative screen shots of the EAHAD Factor VII Gene (F7) variant database

A. An example of a Variant ID page; shown for F7:p.(Arg139Gln). The molecular details of the variant are listed. All cases reported to carry the variant on at least one allele with links to the individual case details and reference.

B. An example of an In depth variant analysis view; shown for F7:p.(Gln160Arg). Substitution analysis scores are provided for Grantham (Align-GVGD), PolyPhen-2, SIFT and PROVEAN. The location of the missense variant is indicated by a red sphere in a molecular model derived for the X-ray structure of the TF-FVIIa complex.
<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Effect</th>
<th>Location</th>
<th>cDNA change</th>
<th>Amino acid change</th>
<th>MAF</th>
<th>Legacy Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs510317</td>
<td>Point</td>
<td>Promoter Flanking (5')</td>
<td>c.-402G&gt;A</td>
<td></td>
<td></td>
<td>0.2326</td>
<td>NA</td>
</tr>
<tr>
<td>rs510335</td>
<td>Point</td>
<td>Promoter Flanking (5')</td>
<td>c.-401G&gt;T</td>
<td></td>
<td></td>
<td>0.2041</td>
<td>NA</td>
</tr>
<tr>
<td>rs5742910</td>
<td>Insertion</td>
<td>Promoter Flanking (5')</td>
<td>c.-325_-324insCCTATATCCT</td>
<td></td>
<td></td>
<td>0.23</td>
<td>WT=A1: Var=A2</td>
</tr>
<tr>
<td>rs61241</td>
<td>Point</td>
<td>Promoter Flanking (5')</td>
<td>c.-122T&gt;C</td>
<td></td>
<td></td>
<td>0.1436</td>
<td>WT=P1: Var=P2</td>
</tr>
<tr>
<td>rs6039</td>
<td>Point</td>
<td>Intrinsic Intron (1)</td>
<td>c.6A&gt;RG&gt;A</td>
<td></td>
<td></td>
<td>0.1403992</td>
<td>G73A (DNA numbering)</td>
</tr>
<tr>
<td>rs6042</td>
<td>Point</td>
<td>Silent  Exon (6)</td>
<td>c.525C&gt;T</td>
<td>p.(His175=) (H115=)</td>
<td></td>
<td>0.1345885</td>
<td>WT=H1 (or C1): Var=H2 (or C2)</td>
</tr>
<tr>
<td>VNTR[7]</td>
<td>Indel</td>
<td>Intrinsic Exon (8)</td>
<td>c.795_805+26[7]</td>
<td></td>
<td></td>
<td>0.31</td>
<td>Var=a (7) or h7</td>
</tr>
<tr>
<td>rs6041</td>
<td>Point</td>
<td>Intrinsic Intron (8)</td>
<td>c.806-206G&gt;A</td>
<td></td>
<td></td>
<td>0.1274271</td>
<td>WT=I1: Var=I2</td>
</tr>
<tr>
<td>rs6046</td>
<td>Point</td>
<td>Missense Exon (9)</td>
<td>c.1238G&gt;A</td>
<td>p.(Arg413Gln) (R353Q)</td>
<td></td>
<td>0.1265056</td>
<td>WT=M1: Var=M2</td>
</tr>
</tbody>
</table>
Table 2: Characteristics of the 65 cases homozygous for a pathogenic variant

<table>
<thead>
<tr>
<th>Variant*</th>
<th>Phenotype</th>
<th>Number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>cDNA numbering</td>
<td>Protein numbering</td>
</tr>
<tr>
<td>Nonsense</td>
<td>c.335C&gt;G</td>
<td>p.(Ser112*)</td>
</tr>
<tr>
<td></td>
<td>c.396C&gt;A</td>
<td>p.(Cys132*)</td>
</tr>
<tr>
<td></td>
<td>c.859C&gt;T</td>
<td>p.(Gln287*)</td>
</tr>
<tr>
<td></td>
<td>c.1324C&gt;T</td>
<td>p.(Gln442*)</td>
</tr>
<tr>
<td></td>
<td>c.1384C&gt;T</td>
<td>p.(Arg462*)</td>
</tr>
<tr>
<td>Consensus splice site</td>
<td>c.291G&gt;C</td>
<td>p.(Thr97=)</td>
</tr>
<tr>
<td></td>
<td>c.291+1G&gt;C</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.292-2A&gt;G</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.430+1G&gt;A</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.572-1G&gt;A</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.681+1G&gt;T</td>
<td>NA</td>
</tr>
<tr>
<td>Frameshift</td>
<td>c.16delC</td>
<td>p.(Leu6Serfs*41)</td>
</tr>
<tr>
<td></td>
<td>c.27_28delCT</td>
<td>p.(Cys10Profs*16)</td>
</tr>
<tr>
<td></td>
<td>c.1061C&gt;T</td>
<td>p.(Ala354Val) + p.(Pro464Hisfs*32)</td>
</tr>
<tr>
<td>Missense variants resulting in substitution of critical residues</td>
<td>c.583T&gt;C</td>
<td>p.(Cys195Arg)</td>
</tr>
<tr>
<td></td>
<td>c.635G&gt;A</td>
<td>p.(Arg212Gln)</td>
</tr>
<tr>
<td></td>
<td>c.635G&gt;A + c.416G&gt;A</td>
<td>p.(Arg212Gln) + p.(Arg139Gln)</td>
</tr>
<tr>
<td></td>
<td>c.904G&gt;A</td>
<td>p.(Asp302Glu)</td>
</tr>
<tr>
<td>Promoter transcription factor binding site</td>
<td>c.-94C&gt;G</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.-65G&gt;C</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.-63T&gt;G</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.-60-59delTT</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>c.-55C&gt;T</td>
<td>NA</td>
</tr>
</tbody>
</table>

*The variants classified as pathogenic are: nonsense, canonical splice site, frameshift, missense variants resulting in substitution of critical residues for FVII function namely, residues at the Arg212-Ile213 bond for proteolytic activation, at the catalytic site (His253, Asp302, Ser404), or residues involved in the unique disulfide bond between the light and heavy chains of the activated form of FVII (Cys195, Cys322), or variants at transcription factor binding sites within the F7 promoter (HNF4 and Sp1).
Figure 1
Figure 2

A

B

- deletion
- duplication
- Indel
- insertion
- point

- frameshift
- in frame
- intronic
- misense
- nonsense
- promoter
- silent
Figure 3

Number of Unique variants

<table>
<thead>
<tr>
<th>Component</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-pro leader</td>
<td>18</td>
<td>9.2</td>
</tr>
<tr>
<td>Gla</td>
<td>18</td>
<td>9.1</td>
</tr>
<tr>
<td>EGF1</td>
<td>13</td>
<td>6.6</td>
</tr>
<tr>
<td>EGF2</td>
<td>17</td>
<td>8.6</td>
</tr>
<tr>
<td>Activation peptide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serine protease</td>
<td>116</td>
<td>58.9</td>
</tr>
</tbody>
</table>