High-throughput molecular diagnosis of von Willebrand disease by next generation sequencing methods

Irene Corrales,[1] Susana Catarino,[2] Júlia Ayats,[3] David Arteta,[2] Carmen Altisent, [3]Rafael Parra,[1,3] and Francisco Vidal[1]

[1]Unitat de Diagnòstic i Teràpia Molecular, Banc de Sang i Teixits, Barcelona; [2]Progenika Biopharma SA, Derio, Vizcaya; and [3]Unitat d'Hemofília, Hospital Universitari Vall d'Hebron, Barcelona, Spain

Citation: Corrales I, Catarino S, Ayats J, Arteta D, Altisent C, Parra R, and Vidal F. High-throughput molecular diagnosis of von Willebrand disease by next generation sequencing methods. Haematologica 2012;97(7):1003-1007. doi:10.3324/haematol.2011.055285

Online Supplementary Appendix

Preliminary pilot study

One of the most important challenges in applying NGS platforms to genetic diagnosis of monogenic disease is the capability of sequencing a large number of samples from different patients simultaneously. Several methods have been designed to address this requisite, such as physical divisions in the sequencer holder and/or indexing samples with barcode tags, that is, short sequences added to the ends of fragments during polymerase chain reaction (PCR) or after isolating the targeted sequences.[1] Before sequencing, the region of interest (a gene, group of genes, or genomic region) must be enriched.[2] Several imaginative enrichment strategies[3] have been developed and tested, including long-range PCR (LR-PCR), amplicon sequencing, array-based gene capture (e.g. Nimblegen), gene capture in solution (e.g. Agilent SureSelect), and other approaches (e.g. Raindance).

To examine the feasibility of *VWF* sequencing by NGS technology, hybridization capture methods were ruled out because their cost-effectiveness is adequate for capturing megabase genomic regions.[4] Conversely, PCR amplification of the target region of interest is the traditional method used to enrich areas of the genome (e.g. single gene studies) upstream of medical re-sequencing procedures. We performed a pilot study that tested two different strategies: i) *VWF* amplification from patient genomic DNA by a newly designed LR-PCR and normalization by densitometric quantification of the PCR bands; and ii) *VWF* amplification from patient genomic DNA by a previously described PCR protocol,[5] mixing all the amplicons to be sequenced. These strategies were analyzed in a single run using a portion of an Illumina flow cell (*Online Supplementary Table S1*) that is physically divided into 8 lanes. Samples from 3 patients in whom the molecular defects had been previously identified by traditional Sanger method were amplified by LR-PCR and sequenced in 3 independent Illumina NGS lanes. A sample from one of these patients was also amplified using the short PCR protocol and analyzed in an additional flow cell lane.

Long-range polymerase chain reaction enrichment

A new procedure for *VWF* amplification has been designed based on LR-PCR and subsequent application to NGS. An accurate analysis of the sequence allowed selection of 14 pairs of specific primers for amplification, covering all exons and a large percentage of intronic regions (*Online Supplementary Figure S1A*).

According to our results, the SequalPrep System provides a highly efficient and robust enrichment solution to meet NGS re-sequencing throughput capabilities. Only one of the 14 PCRs designed presented set-up problems: LR-PCR number 6, which includes exons 14, 15 and 16, was refractory to amplification under the conditions tested, probably due to the high GC content of this region.[5] Amplification of LR-PCR 6 was redesigned to amplify the region in two shorter fragments (6A and 6B), in an attempt to circumvent these difficulties. However, we were only able to efficiently amplify region 6B (covering exon 16 and partially covering introns 15 and 16) (*Online Supplementary Figure S1B*). The LR-PCRs were then mixed at an equimolar concentration with two conventional PCRs covering exons 14 and 15 and the corresponding flanking intronic regions. After sonication and sequencing, we were able to verify that these regions also have a significant representation in the final sequence obtained. This indicates that nebulization could also be a suitable method for fragmentation of short amplicons prior to library construction. The final protocol established can be adapted to any of the currently available NGS sequencers and allows amplification of 114 Kb, approximately 64% of the *VWF* sequence.

The sequencing run in the preliminary pilot study was properly developed, and the parameters studied (number of clusters passing filters, percentage of reads aligning to the reference, percentage of error rate after alignment), which were indicative of the quality of the experiment for the 4 samples processed plus the Illumina control library PhiX, were unproblematic and generated the expected clusters with conventional data quality.

Once the run was completed, the sequences were analyzed with Illumina software and exported into BWA, SAMtools, and the CLC Genomics Workbench for variant detection. The current results indicate that expected mutations and polymorphisms (13 to 17 SNPs) were identified with high reliability, regardless of the procedure used. Furthermore, in addition to the variations identified by traditional sequencing exons, LR-PCR detected several changes (7, 14 and 8, respectively, in each of the 3 patients) in deep intronic regions that are not described as polymorphisms in dbSNP Build 133 (*Online Supplementary Table S1*).

Previous studies have shown that coverage in the 10- to 15-fold range may suffice for re-sequencing applications, but higher coverage depths provide better alignment, assembly and accuracy.[6] In our case, both the LR-PCR and conventional PCR approaches had higher coverage than needed for a robust resequencing assay (*Online Supplementary Table S1*), allowing simultaneous analysis of many patients by pooling the patients' PCRs in an Illumina flow cell lane. Nonetheless, although all the approaches were suitable to a lesser or greater extent for simultaneous analysis of several patients, implementation of the LR-PCR was technically more complex because normalization of amplicon concentrations for the LR-PCRs represented an additional time-consuming technical complication in the overall protocol. The possibility of analyzing gene variations in deep intronic regions could be particularly relevant to those patients in whom no candidate mutation was identified in *VWF* coding sequence. Nevertheless, it is still very difficult, based on current knowledge of *VWF*, to extract conclusions from extra information discovered with the LR-PCR approach and the available *in silico* prediction tools lack the power to predict the outcome of an array of variants that would be encountered by direct sequencing of introns and to decide which of these variants should be tracked as likely to be pathogenic. Taking these data together, we chose the conventional highly optimized PCR procedure as a first approach in designing the proof-of-concept study. Nonetheless, we do not exclude the utilization of the LR-PCR approach in future experiments. In fact, validation of the variations found and investigation in their biological significance could help to clarify some basic aspects of the pathophysiology of VWD.

Design and Methods

Primer design for LR-polymerase chain reaction

Primers were designed within intronic regions of *VWF* (GenBank n. NC_000012.10; range 5928301:6104097) for *VWF* amplification in 14 LR-PCRs. Primers used to amplify the region homologous to the pseudogene (exons 23 to 34, LR regions 9 and 10) were designed taking into account the particular differences between the *VWF* sequence and the pseudogene, to ensure highly specific amplification. Furthermore, all the primers designed were aligned against the *VWF* region according to dbSNP (Build 133) to corroborate the absence of SNPs in primer binding sequences that could result in preferential or single allele amplification, and lead to missing a mutation. The primer sequences and positions in *VWF*, and the LR-PCR product size are presented in *Online Supplementary Table S2*.

VWF long-range amplification

All 14 regions were amplified with the SequalPrep Long PCR Kit with dNTPs (Invitrogen, Carlsbad, CA, USA) and the primers listed in the *Online Supplementary Table S2*. The expected range of sizes was between 4924 and 9557 bp, and all primers specifically annealed under the same salt conditions and thermocycling parameters, giving rise to optimal LR-PCR products (*Online Supplementary Figure S1B*). The LR-PCR solution contained 1X SequalPrep reaction buffer, 0.4 μL DMSO, 1X SequalPrep enhancer B, 1.8 U SequalPrep long polymerase, 0.75 μM of each primer, and 50 ng of DNA in a total volume of 20 μL. To obtain good yield in the PCR conditions established (*Online Supplementary Figure S1B*), primer concentration was adjusted to 1.25 μM for amplification of LR regions 1, 10, 11 and 14, and 2.5 μM for regions 3, 7 and 12 (*Online Supplementary Table S2*). After initial denaturation at 94°C for 2 min, 10 cycles were performed at 94°C for 10 s and 68°C for 18.5 min, followed by 20 additional cycles at 94°C for 10 s and at 68°C for 18.5 min (plus 20 s/cycle), and a final extension at 72°C for 5 min. LR-PCR products were separated on 0.5% agarose gel and visualized by ethidium bromide staining. Only one of the 14 PCRs designed presented set-up problems: LR-PCR number 6, which includes exons 14, 15 and 16, was refractory to amplification under the conditions tested, probably due to the high GC content of this region.[5] Amplification of LR-PCR 6 was redesigned to amplify the region in two shorter fragments (6A and 6B) in an attempt to circumvent these difficulties. However, we were only able to efficiently amplify region 6B (covering exon 16 and partially covering introns 15 and 16) (*Online Supplementary Figure S1B*).

Normalization and fragmentation of LR-polymerase chain reactions

Quantitative detection of ethidium bromide-stained PCR product bands by densitometry after agarose gel electrophoresis was used for product quantification. The LR-PCRs were then mixed at an equimolar concentration with two conventional PCRs covering exons 14 and 15 and the corresponding flanking intronic regions. Pools of long-range amplicons were fragmented by sonication with a Bioruptor (Diagenode, Philadelphia, PA, USA).

Massively parallel sequencing

Normalized pooled PCR products (5 μg), comprising the amplified sequences of *VWF*, were used for Illumina GA sample library preparations. The subsequent procedure steps, including ends repair, addition of adenine bases to the 3' end of DNA fragments, and adapter ligation were performed according to the Illumina protocol. Specific samples were connected to different adapters in order to allow index sequencing (these specific adapters with Tags have the following structure: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATCtagT3', where all the Tags have 5 nucleotides). After DNA fragmentation and ligation to the adaptors, the samples were amplified and processed according to Illumina GA protocols. Briefly, fragmented PCR products were purified by excising a portion of gel corresponding to a 250-500 bp region for the pilot study and 150-250 bp region for the proof-of-concept experiment. DNA fragments were enriched by PCR, and the library was validated. In order to achieve the highest quality of data on Illumina sequencing platforms, it is important to create optimum cluster densities across every lane of every flow cell. This requires accurate quantitation of enriched DNA library templates by using qPCR as described in the Illumina qPCR Quantification Guide (*http://www.illumina.com*). To verify the size of the PCR enriched fragments, the template size distribution was checked by running an aliquot of the enriched library on an Agilent 2100 Bioanalyzer. Lastly, all samples were diluted to 11 pM for the sequencing run. Cluster generation was performed in the cluster station in individual lanes of the Illumina flow cell. All reagents were from the Illumina Single Read Cluster Generation Kit, v2. A custom-designed primer (5'ACACTCTTTCCCTACACGACGCTCTTCCGATC3') for the indexed samples and the Illumina sequencing primer for other samples were hybridized to the prepared flow cell. Thirty-eight cycles of base incorporation were carried out on Illumina GA II, following the Illumina protocol and using the Genome Analyzer Sequencing Control software, v2.4. The reagents consumed in this step were from the Illumina 36 Cycle Sequencing Kit, v3.0. A PhiX library was used as a control lane to validate the quality of each run. Images taken during the sequencing reactions were processed in three stages using Illumina software (version 1.5): Firecrest for image analysis, Bustard for base-calling, and Gerald for sequence analysis. The Illumina GA pipeline output sequence file (fast format) was used as input for the analysis with CLC

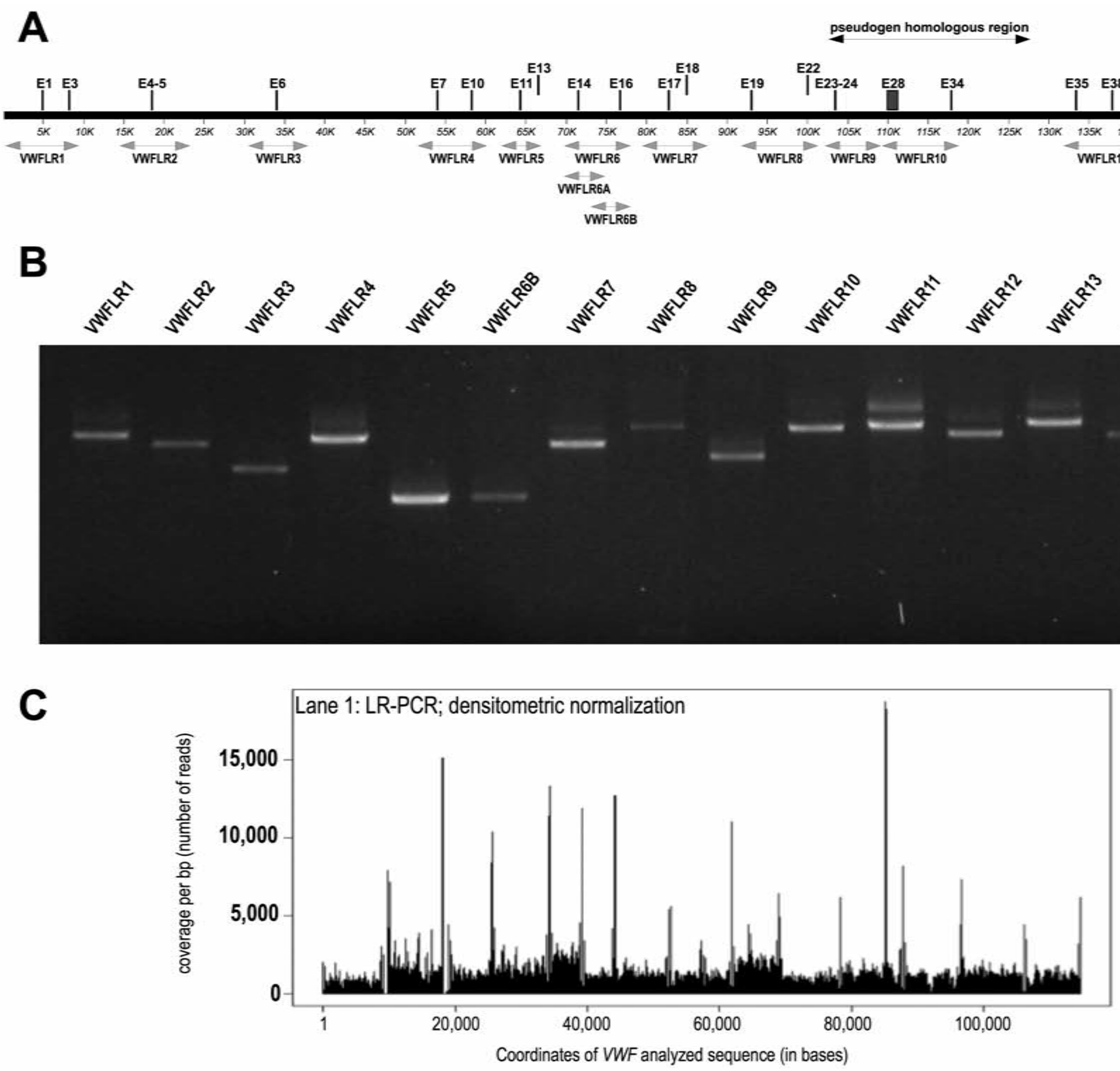
Genomic Workbench version 4.5 software (CLCbio, Aarhus, Denmark). After several adjustments, the optimal parameters used for mutation detection analysis were the following: coverage over 50X, minor allele counts over 5, and percent of variant allele over 3 and below 25. Additional analyses to complete and/or corroborate ambiguous results were performed with Burrows-Wheeler Aligner[7] (BWA), version 0.5.5 and SAMtools,[8] version 0.1.7. Concatenated sequences of exon and exon/intron boundaries covered by PCR products, the CDS sequence for VWF, and the annotated genomic sequence for *VWF* (GenBank accession n. NG_009072) were used as reference sequences to detect and assign the previously described mutations and polymorphisms. To this end, in addition to using the dbSNP Build 133 database, we reviewed the *VWF* international mutation database (*http://www.sheffield.ac.uk/vwf/index.html*) and published VWD literature to collect previously described variants clinically associated with the disease.

In silico *analysis of novel putative mutations*

Alamut 1.51 software (Interactive Biosoftware, Rouen, France) was used to investigate the predicted impact of the novel missense mutations described in VWF as a consequence of nucleotide substitution. Missense mutations evaluation include automated access to web-based variant scoring methods (PolyPhen-2, SIFT, Align GVGD) (*Online Supplementary Table S3*).[9] The novel potential splice site mutations (PSSM) impact was also performed with Alamut 1.51 that integrates data from known constitutive human splicing signals and a number of prediction methods such as the SpliceSiteFinder-like, the MaxEntScan, the GeneSplicer, the ESEFinder and the RESCUE-ESE methods. Analysis with NetGene2 software (*http://www.cbs.dtu.dk/services/NetGene2/*), not integrated in Alamut, was also performed (*Online Supplementary Table S4*).

References

1. Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. Nature Protocols. 2008;3(2):267-78.
2. Summerer D. Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. Genomics. 2009;94(6):363-8.
3. Hedges DJ, Guettouche T, Yang S, Bademci G, Diaz A, Andersen A, et al. Comparison of Three Targeted Enrichment Strategies on the SOLiD Sequencing Platform. PLoS One. 2011;6(4):e18595.
4. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111-8.
5. Corrales I, Ramirez L, Altisent C, Parra R, Vidal F. Rapid molecular diagnosis of von Willebrand disease by direct sequencing. Detection of 12 novel putative mutations in VWF gene. Thromb Haemost. 2009;101(3):570-6.
6. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 2008;18(10):1638-42.
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
9. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24(3):133-41.



Online Supplementary Figure S1. Long-range PCR approach. (A) Schematic scale representing the genomic region of VWF with location of primers used for LR-PCR. (B) Fragments amplified by LR-PCR under the same thermocycling parameters were separated on 0.5% agarose gel and visualized by ethidium bromide staining. The 14 PCRs comprising approximately 64% of the locus have a size range from 4924 to 9557 bp. (C) Sequence coverage diagram representative of the Illumina GA reads for LR-PCR samples normalized by agarose gel densitometry. The x axis represents the position on the reference sequence comprised by concatenated sequences of *VWF* amplified regions. The y axis shows the fold coverage (number of reads per bp). Regions of very high coverage represent the ends of the LR-PCR amplicons.

Online Supplementary Table S1. Mutation detection and coverage data for samples analyzed in the pilot study.

Lane	Enrichment method	Max. coverage	Median coverage[a]	Detected changes	Mutations[b]	Additional variations in VWF
1	LR-PCR (Patient A)	18,759	1166	10	p.S182S/ c.546G>A c.7082-2A>G c.8155+3G>C	7
2	LR-PCR (Patient B)	12,737	1078	16	p.V1409F/c.4225G>T p.P2063S/c.6187C>T	14
3	LR-PCR (Patient C)	7,837	609	9	p.P2063S/c.6187C>T	8
4	Short PCR (Patient A)	380,913	27,162	3	p.S182S/c.546G>A c.7082-2A>G c.8155+3G>C	0

[a] *Median coverage represents the coverage per bp. The higher coverage in short PCR method is due to the smaller sequence analyzed in comparison with LR-PCR that has a higher overall coverage of the VWF sequence.* [b] *Optional parameters used in CLC Bio Genomics Workbench for mutation detection analysis: coverage higher than 50; variant frequency lower than 30. c.6187C>T is also described as polymorphism in ISTH VWD mutation database.*

Online Supplementary Table S2. Primers used for amplification and sequencing of *VWF* by LR-PCR.

PCR FORWARD PRIMER a	PCR REVERSE PRIMER[a]	AMPLIFIED REGION	PCR SIZE
VWFLR1-1: TCCTGTCTTACATGTCATTGCTATCTGG(1) 6092689 .. 6092716	VWFLR1-2: ACAGCCCAGATCTCAGGTAAGTGTCC(1) 6083599 .. 6083624	Promotor + Exon 1 to 3	9118 bp
VWFLR2-1B: CTTGGACTTCCTGATGCTGGTCTCTCC 6078128 .. 6078154	VWFLR2-2: CCACAAAAAGACAAATGCTGTATGATTCC 6069761 .. 6069789	Exon 4 to 5	8754 bp
VWFLR3-1B: GGACTCTGGTAATATTTCAGAAACCATCC(2) 6061678 .. 6061706	VWFLR3-2C: GGTTTTAGAACTTAAAGAAACCTTAAGAACC(2) 6055143 .. 6055173	Exon 6	7437 bp
VWFLR4-1: TGGGGACCTGCTCACTATTCTAGAGG 6041535 .. 6041560	VWFLR4-2: AAGCTCTATCTGCAGTCATCACACTGG 6032941 .. 6032967	Exon 7 to 10	8620 bp
VWFLR5-1: GGCTCTGTTGCTCAGTTGTACCGAGG 6030986 .. 6031011	VWFLR5-2C: GGGAGATTGGACAGCAAACCTGCTCC 6026088 .. 6026113	Exon 11 to 13	4924 bp
VWFLR6.5-1: CACCTAGTGTTCGTTCAGCACAGAAGG 6019959 .. 6019985	VWFLR6-2: ACTGGTCTTCCCTAGTAGACTATCAGG 6015017 .. 6015043	Exon 16	4969 bp
VWFLR7-1: TTGAGTTACCATCTTGGTGAACAAATGG(2) 6013516 .. 6013543	VWFLR7-2: ATTCCTACACTTCTGTAGAACTCTCTGG(2) 6005399 .. 6005426	Exon 17 to 18	8145 bp
VWFLR8-1: CTCACATACGATTCTAGCCTGGGTCC 6001294 .. 6001319	VWFLR8-2: TTGGTTTGCTGCTGTAGGTTAAGAAACC 5991763 .. 5991790	Exon 19 to 22	9557 bp
VWFLR9-1B: GATGACATTCAGCCCACACAGATAATCC 5990913 .. 5990940	VWFLR9-2: GGGTCTCCACGGTGTCAGGCCTAG 5983798 .. 5983821	Exon 23 to 27	7143 bp
VWFLR10-1: TTCTTGGAGACACTTGTAAGAAGGCTTG(1) 5983534 .. 5983561	VWFLR10-2B: TTGATTATTACGCAAGAGTGGGTATCTAG(1) 5974260 .. 5974288	Exon 28 to 34	9302 bp
VWFLR11-1: GATGGAAGGACCTCAAGTTTTACTATACC(1) 5961159 .. 5961187	VWFLR11-2B: GACAACCCTGGCGACAGCATGCAGG(1) 5951693 .. 5951717	Exon 35 to 38	9554 bp
VWFLR12-1: GTACACACCTATAAAACCATTGACATAACC(2) 5951323 .. 5951352	VWFLR12-2: GACAGTTGCTAGAGCATCAACTCATGG(2) 5942557 .. 5942583	Exon 39 to 42	8796 bp
VWFLR13-1: GAAGGAGTGAAGTGCAGTTAAGGCAGG 5939981 .. 5940007	VWFLR13-2: ATCCAAATGTCCCGCAATAATAGCTTGG 5930465 .. 5930492	Exon 43 to 47	9543 bp
VWFLR14-1: TTGCATGTGCTATACTTTTACATGACTGG(1) 5919606 .. 5919634	VWFLR14-2: GCTGCTTTCATCACTTGTAAAGGAGTGG(1) 5911047 .. 5911074	Exon 48 to 52	8588 bp

[a] *Primer postions referred to GenBank accession NT_009759.15. The final PCR primer concentration was 0.75 M except for: (1) 1.25 μM; (2) 2.5 μM. Within black frame: specific primers used to amplify the VWF region homologous to the pseudogene (exons 23 to 34). Exons 14 and 15 were amplified with short PCR.*[4]

Online Supplementary Table S3. Summary of the *in silico* analysis of the new mutations.

Missense mutations

Nucleotide change	Amino acid change	PolyPhen prediction (score[1])	SIFT prediction (score[2])	Align GVGD (score[3])
c.1109G>A	p.C370Y	Probably damaging (3.185)	Deleterious (0.00)	Highly conserved (1)
c.3788C>T	p.S1263L	Benign (0.093)	Deleterious (0.01)	Weakly conserved (0.00)
c.5311G>A	p.G1771R	Probably damaging (2.172)	Deleterious (0.00)	Highly conserved (1)
c.6890C>T	p.P2297L	Probably damaging (2.105)	Deleterious (0.05)	Highly conserved (0.98)
c.7150C>T	p.R2384W	Probably damaging (2.255)	Deleterious (0.00)	Weakly conserved (0.32)

[1] Large values of PolyPhen Score indicate that the substitution is rarely or never observed in the protein family, suggesting likelihood that the amino acid replacement will be deleterious. [2] SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. Substitutions with less than 0.05 score are deleterious. [3] Align GVGD combines the biophysical characteristics of amino acids and protein multiple sequence alignments to predict where missense substitutions in genes of interest fall in a spectrum from enriched delterious (score of 1) to enriched neutral (score of 0).

Online Supplementary Table S4.

Potential splice site mutations (PSSM)

Nucleotide change	Amino acid change	Affected splice site	NetGene2 score native-mutated	HSF score native-mutated	GeneSplicer score native-mutated	NNSPLICE score native-mutated	ESE finder score native-mutated
c.126C>T	p.(=)	-	No difference	62.9-63.0	No difference	No difference	SRp55 ESE element disrupted (3.84-0)
c.874+8G>A	-	DSS intron 7	0.97-0.95	43.86-72.8 New potential ASS	8.09-7.38	0.80-0.63	-
c.4917G>A	p.(=)	-	0.49-0.46	39.85-68.8 New potential ASS	No difference	No difference	New SRp40 ESE element (0-2.74)
c.4923G>A	p.(=)	-	0.49-0.51	67.74-64.61	No difference	No difference	SF2/ASF ESE element disrupted (4.15-0)
c.5455+1G>A	-	DSS intron 31	0.71-Native DSS destroyed	81.38-54.55 Broken splice site	5.56-Native DSS destroyed	0.89-Native DSS destroyed	-

PSSM: potential splice site mutation; ASS: aceptor splice site; DSS: donor splice site; ESE: exonic splicing enhancer. Synonymous sequence changes are included in the list of candidate mutations when not described as polymorphism in the general population (SNPdb Build 133) and some of the in silico tools predicts some effect.