Linkage analysis combined with whole-exome sequencing identifies a novel prothrombin (*F2*) gene mutation in a Dutch Caucasian family with unexplained thrombosis

René Mulder,[1] Ton Lisman,[2] Joost C.M. Meijers,[3] James A. Huntington,[4] André B. Mulder[1] and Karina Meijer[5]

[1]Department of Laboratory Medicine, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; [2]Surgical Research Laboratory and Section of Hepatobiliary Surgery and Liver Transplantation, Department of Surgery, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; [3]Department of Molecular and Cellular Hemostasis, Sanquin Research, Amsterdam and Amsterdam UMC, University of Amsterdam, Department of Experimental Vascular Medicine, Amsterdam Cardiovascular Sciences, Amsterdam, the Netherlands; [4]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK and [5]Department of Hematology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

Correspondence: RENÉ MULDER - r.mulder01@umcg.nl.

doi:10.3324/haematol.2019.232504

Supplementary Appendix

Linkage analysis combined with whole exome sequencing identifies a novel prothrombin (*F2*) gene mutation in a Dutch Caucasian family with unexplained thrombosis

René Mulder[1], Ton Lisman[2], Joost C.M. Meijers[3], James A. Huntington[4], André B. Mulder[1], and Karina Meijer[5]

[1]Department of Laboratory Medicine, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; [2]Surgical Research Laboratory and Section of Hepatobiliary Surgery and Liver Transplantation, Department of Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [3]Department of Molecular and Cellular Hemostasis, Sanquin Research, Amsterdam, the Netherlands; Amsterdam UMC, University of Amsterdam, Department of Experimental Vascular Medicine, Amsterdam Cardiovascular Sciences, Amsterdam, the Netherlands; [4]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK; [5]Department of Hematology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands;

Correspondence: Dr. René Mulder at the Department of Laboratory Medicine, University Medical Center Groningen, Hanzeplein 1, 9700 RB, Groningen, the Netherlands, r.mulder01@umcg.nl.

Linkage analysis combined with whole exome sequencing identifies a novel prothrombin (*F2*) gene mutation in a Dutch Caucasian family with unexplained thrombosis

René Mulder[1], Ton Lisman[2], Joost C.M. Meijers[3], James A. Huntington[4], André B. Mulder[1], and Karina Meijer[5]

[1]Department of Laboratory Medicine, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[2]Surgical Research Laboratory and Section of Hepatobiliary Surgery and Liver Transplantation, Department of Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[3]Department of Molecular and Cellular Hemostasis, Sanquin Research, Amsterdam, the Netherlands; Amsterdam UMC, University of Amsterdam, Department of Experimental Vascular Medicine, Amsterdam Cardiovascular Sciences, Amsterdam, the Netherlands

[4]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK

[5]Department of Hematology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

Correspondence: Dr. René Mulder at the Department of Laboratory Medicine, University Medical Center Groningen, Hanzeplein 1, 9700 RB, Groningen, the Netherlands, r.mulder01@umcg.nl.

Genomic DNA

Genomic DNA was obtained from citrated and EDTA samples using the Qiacube system.

Linkage analysis

Genome-wide SNP genotyping was performed on 19 family members (5 affected and 14 unaffected; Figure 1A) using Illumina Infinium Core Exome-24v1.1 bead chip. The bead chip included 551,839 SNP markers distributed evenly across the genome. The mean and median intervals between markers were 5.26 kb and 1.82 kb, respectively.

For linkage analysis, we first excluded SNPs with genotype call rate less than 95%, MAF less than 0.001 and HWE test p-value less than 0.0001. Next, frequency and genotyping pruning was performed. From the subset, SNPs were checked for strong LD, since high LD within dense SNP regions can result in false positive results. We applied prune SNPs within strong LD using –indep 50 5 10 option in PLINK v1.07. Strong LD pruning can be defined as removing SNPs within a 50 SNP window, with 5 SNPs to shift the window at each step that had r^2>0.9 (corresponding to a variance inflation factor, VIF, greater than 10) with all other SNPs in the window. We then checked the SNPs with Mendelian inheritance error using Pedstats program[1] and masked the genotypes for the SNPs with error using Pedwipe program. For pruned and masked SNP data, non-parametric linkage analyses will be performed using the Merlin 1.1.2.[2] We performed non-parametric analyses to get LOD scores and statistics. A LOD score of more than 0 would be suggestive for evidence of linkage.

Whole exome sequencing (WES)

The SureSelect Human All Exon V5 was used for WES on 5 family members (4 affected: 26101, 26103, 26105, 26206; 1 unaffected: 26312; Figure 1A).

Read mapping

Paired-end sequences produced by HiSeq Instrument were first mapped to the human genome, with the reference sequence as UCSC assembly hg19 (original GRCh37 from NCBI, Feb. 2009), and without unordered sequences or alternate haplotypes. 'BWA' (version 0.7.12) was used as the mapping tool, which generates the mapping result file in BAM format using 'BWA-MEM'. Then, programs within Picard-tools (ver.1.130) were used in order to remove PCR duplicates, reducing those reads to identically match to a position at start into a single one, using MarkDuplicates.jar, which requires reads to be sorted. The local realignment process were performed to consume BAM files and to locally realign reads such that the number of mismatching bases is minimized across all reads. Base quality score recalibration (BQSR) and local realignment around indels were performed using Genome Analysis Toolkit to locally realign reads such that the number of mismatching bases was minimized.

Variant calling and annotation

Based on the BAM file previously generated, variant genotyping for each sample was performed with Haplotype Caller of GATK (v3.4.0). In this stage SNPs and short indels candidates were detected at the nucleotide resolution. These variants were annotated by a program called SnpEff v4.1g and converted to vcf file format. They were filtered with dbSNP version 142, and SNPs from the 1000

genome project. Then, in-house program and SnpEff were applied to filter additional databases, including ESP6500, ClinVar, dbNSFP2.9.

Family analysis

To identify the genetic variants underlying rare diseases, variants in all family members were joint-called with GATK (v3.4.0) and filtered with an in-house developed script. Firstly, exon-, exonic-splicing, and splicing regions are selected to merge variant data. Synonymous single nucleotide variants (SNVs) within the remaining variants were then removed. Furthermore, dbSNP135 common (freq>=1%) and 1000 genome freq >0.01 variants were eliminated. Next, variants uncommon among affected family members were filtered out. In addition, only variants at chromosomes with linkage peaks were further filtered based on their distance (~ 3 Mb) to the linkage peak. Remaining variants with variant ids were then filtered out. Finally, variants in genes known to be associated with coagulation were selected and further filtered based on the in silico prediction where only variants that scored deleterious with MetaSVM and MetaLR remained.[3]

Molecular dynamics

The starting model for the molecular dynamics simulation was thrombin from 1JOU (chains A and B), including surrounding water molecules and sodium ion, modified by mutation of Ala195 back to Ser and Arg173 to Trp. The simulation was conducted using the program YASARA within a hydrated box surrounding the thrombin molecule.[4, 5] The YAMBER forcefield was used with simulated annealing at 335K. The parent molecule (p.Ala195Ser) was run for 18 ps to obtain the minimized starting structure which did not result in significant changes. Arg173 was then mutated to Trp and the simulation was restarted. In order to bury the side chain of Trp173 it was selected and pulled toward the S4 pocket, followed by 26 ps of energy minimization with simulated annealing.

References

1. Wigginton JE, Abecasis GR. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. Bioinformatics. 2005;21(16):3445-3447.

2. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30(1):97-101.

3. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-2137.

4. Krieger E, Vriend G. New ways to boost molecular dynamics simulations. J Comput Chem. 2015;36(13):996-1007.

5. Krieger E, Darden T, Nabuurs SB, Finkelstein A, Vriend G. Making optimal use of empirical energy functions: force-field parameterization in crystal space. Proteins. 2004;57(4):678-683.