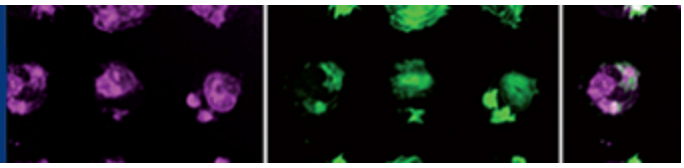
Thrombosis and Haemostasis
ISSN 0340-6245
International Journal for Vascular Biology and Medicine



Long-read sequencing identifies the first retrotransposon insertion and resolves structural variants causing antithrombin deficiency.

Journal:	*Thrombosis and Haemostasis*
Manuscript ID	TH-21-09-0539.R1
Manuscript Type:	Original Article: New Technologies, Diagnostic Tools and Drugs
Category:	Clinical Studies
Date Submitted by the Author:	n/a
Complete List of Authors:	de la Morena-Barrio, Belén; Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Arrixaca, CIBERER Stephens, Jonathan ; Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK; NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK. de la Morena-Barrio, Maria Eugenia; Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Arrixaca, CIBERER, Luca, Stephanucci; Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK; National Health Service Blood and Transplant (NHSBT), Cambridge Biomedical Campus, Cambridge, CB2 0PT, UK.; BHF Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK Padilla, Jose; Centro Regional de Hemodonación, Miñano, Antonia; Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Arrixaca Gleadall, Nicholas ; Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK; NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK. García, Juan Luis; Servicio de Hematología, Hospital Universitario de Salamanca, Salamanca, Spain. Lopez Fernandez, M Fernanda; C.Hospitalario Universitario A Coruña, Hematology Morange, Pierre; Inserm, Hematology Puurunen, Marja; Boston University School of Medicine, Framingham Heart Study Undas, Anetta; Jagiellonian University School of Medicine, Department of Medicine Vidal, Francisco; Blood and Tissue Bank (BST), Congenital Coagulopathies Laboratory; Vall d'Hebron University Hospital, Barcelona, Spain, Haemophilia Unit

	Raymond, F Lucy; NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK.; Department of Medical Genetics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK Vincente, Vincente; Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Arrixaca, CIBERER Ouwehand, Willem; Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK; NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK. Corral, Javier; Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Arrixaca, CIBERER Sanchis-Juan, Alba; Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK; NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK.
Keywords:	Long-read sequencing, Antithrombin deficiency, Structural variants, SVA retrotransposon
Abstract:	The identification of inherited antithrombin deficiency (ATD) is critical to prevent potentially life-threatening thrombotic events. Causal variants in SERPINC1 are identified for up to 70% of cases, the majority being single-nucleotide variants and indels. The detection and characterization of structural variants (SVs) in ATD remain challenging due to the high number of repetitive elements in SERPINC1. Here, we performed long-read whole-genome sequencing on 10 familial and 9 singleton cases with type I ATD proven by functional and antigen assays, that were selected from a cohort of 340 patients with this rare disorder because genetic analyses were either negative, ambiguous, or not fully characterized. We developed an analysis workflow to identify disease-associated SVs. This approach resolved, independently of its size or type, all 8 SVs detected by MLPA, and identified for the first time a complex rearrangement previously misclassified as a deletion. Remarkably, we identified the mechanism explaining ATD in 2 out of 11 cases with previous unknown defect: the insertion of a novel 2.4Kb SINE-VNTR-Alu retroelement, which was characterized by de novo assembly and verified by specific PCR amplification and sequencing in the probands and affected relatives. The nucleotide-level resolution achieved for all SVs allowed breakpoint analysis, which revealed repetitive elements and microhomologies supporting a common replication-based mechanism for all the SVs. Our study underscores the utility of long-read sequencing technology as a complementary method to identify, characterize and unveil the molecular mechanism of disease-causing SVs involved in ATD, and enlarges the catalogue of genetic disorders caused by retrotransposon insertions.

Title

Long-read sequencing identifies the first retrotransposon insertion and resolves structural variants causing antithrombin deficiency.

Running title

Long-read sequencing to resolve structural variants in *SERPINC1*

Author list

Belén de la Morena-Barrio, [1] Jonathan Stephens, [2,3] María Eugenia de la Morena-Barrio, [1] Luca Stefanucci, [2,4,5] José Padilla, [1] Antonia Miñano, [1] Nicholas Gleadall, [2,3] Juan Luis García, [6] María Fernanda López-Fernández, [7] Pierre-Emmanuel Morange, [8] Marja K Puurunen, [9] Anetta Undas, [10] Francisco Vidal, [11] NIHR BioResource, [3] F Lucy Raymond, [3,12] Vicente Vicente García, [1] Willem H Ouwehand, [2,3] Javier Corral, [1] Alba Sanchis-Juan [2,3]

Affiliations

1. Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Murcia, Spain.
2. Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK.
3. NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK.

4. National Health Service Blood and Transplant (NHSBT), Cambridge Biomedical Campus, Cambridge, CB2 0PT, UK.
5. BHF Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK.
6. Servicio de Hematología, Hospital Universitario de Salamanca, Salamanca, Spain.
7. Servicio de Hematología, Complexo Hospitalario Universitario de A Coruña, A Coruña, Spain.
8. Laboratory of Haematology, La Timone Hospital, Marseille, France; C2VN, INRAE, INSERM, Aix-Marseille Université, Marseille, France.
9. National Heart, Lung and Blood Institute's. The Framingham Heart Study, Framingham, MA, US.
10. Institute of Cardiology, Jagiellonian University Medical College and John Paul II Hospital, 80 Prądnicka St, Kraków, Poland.
11. Banc de Sang i Teixits, Barcelona, Spain; Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona (VHIR-UAB), Barcelona, Spain; CIBER de Enfermedades Cardiovasculares, Madrid, Spain.
12. Department of Medical Genetics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

Correspondence

Alba Sanchis-Juan, University of Cambridge, Department of Haematology, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK. Email: as2635@cam.ac.uk. Phone number: +44(0)1223588035.

Javier Corral, University of Murcia, Centro Regional de Hemodonación, Calle Ronda de Garay s/n, Murcia 30003, Spain. Email: javiercorraldelacalle@gmail.com. Phone number: +34968341990.

Authorship Contributions

BMB, WHO, JC and ASJ designed the study.

MMB, LS, JP, AM, NG, FLR and VV helped with study design.

BMB, MMB, JP, AM performed laboratory experiments and analyzed the experimental data.

JS performed sample preparation and executed long-read sequencing.

ASJ developed the analysis workflow for long-read sequencing, applied this to data processing and performed the computational and statistical analyses.

BMB performed computational analyses and variant validation.

JM, FV, provided valuable insight into microarray and NGS data analysis.

AU, MF, MP and PM recruited participants and collected the clinical data and samples.

BMB, WHO, JC and ASJ wrote the manuscript.

All authors read and approved the final manuscript.

Data and Code Availability

The workflow developed for the detection of structural variants is publicly available at http://github.com/who-blackbird/magpie.

Funding

This work was supported by the National Institute for Health Research England (NIHR) for the NIHR BioResource project (grant numbers RG65966 and RG94028), the

PI18/00598, PI21/00174 and PMP21/00052 projects (ISCIII & FEDER) and the 19873/GERM/15 project (Fundación Séneca).

Declaration of Interests

The authors declare that they have no conflicts of interest.

Patient consent statement

All included subjects gave their written informed consent to enter the study.

Ethics

This study was approved by the Ethics Committee of Morales Meseguer Hospital and the East of England Cambridge South national institutional review board (13/EE/0325). The research conforms with the principles of the Declaration of Helsinki and their later amendments.

Keywords

Long-read sequencing; Antithrombin deficiency; Structural variants; SVA retrotransposon.

Abstract

The identification of inherited antithrombin deficiency (ATD) is critical to prevent potentially life-threatening thrombotic events. Causal variants in *SERPINC1* are identified for up to 70% of cases, the majority being single-nucleotide variants and indels. The detection and characterization of structural variants (SVs) in ATD remain challenging due to the high number of repetitive elements in *SERPINC1*. Here, we performed long-read whole-genome sequencing on 10 familial and 9 singleton cases

with type I ATD proven by functional and antigen assays, that were selected from a cohort of 340 patients with this rare disorder because genetic analyses were either negative, ambiguous, or not fully characterized. We developed an analysis workflow to identify disease-associated SVs. This approach resolved, independently of its size or type, all 8 SVs detected by MLPA, and identified for the first time a complex rearrangement previously misclassified as a deletion. Remarkably, we identified the mechanism explaining ATD in 2 out of 11 cases with previous unknown defect: the insertion of a novel 2.4Kb SINE-VNTR-Alu retroelement, which was characterized by *de novo* assembly and verified by specific PCR amplification and sequencing in the probands and affected relatives. The nucleotide-level resolution achieved for all SVs allowed breakpoint analysis, which revealed repetitive elements and microhomologies supporting a common replication-based mechanism for all the SVs. Our study underscores the utility of long-read sequencing technology as a complementary method to identify, characterize and unveil the molecular mechanism of disease-causing SVs involved in ATD, and enlarges the catalogue of genetic disorders caused by retrotransposon insertions.

WC: 250

Introduction

Antithrombin deficiency is the most severe congenital thrombophilia firstly identified in 1965 by O Egeberg[1]. The key hemostatic role of this anticoagulant serpin explains the high risk of thrombosis associated to congenital antithrombin deficiency (OR: 20-30), which is mainly caused by haploinsufficiency of *SERPINC1*, the coding gene.[2] Accurate genetic diagnosis of antithrombin deficiency facilitates the management of both symptomatic and asymptomatic carriers[3,4], and increases the antithrombotic arsenal of

carriers with antithrombin concentrates.[5] Routine investigation of antithrombin deficiency combines functional assays, antigen quantification and genetic analyses to determine the molecular base. However, most studies do not reach a molecular characterization, despite it could contribute to a better definition of the thrombotic risk.[2]

In genetic diagnostic centers, causal Single Nucleotide Variants (SNVs) and small insertions or deletions (indels) are routinely identified in *SERPINC1* by Sanger sequencing, and copy number changes are investigated by multiple ligation-dependent probe amplification (MLPA).[2] Only few cases with gross gene defects have been analyzed by microarray to determine the extent of the variants. These methods identify causal mutations in *SERPINC1* for 70% of cases, whilst 5% of patients harbor defects in other genes and 25% remain without a genetic diagnosis.[2] To date, 441 causal variants in *SERPINC1* have been reported,[6] and these adhere to the typical spectrum observed in disorders with a dominant inheritance, being 63% SNVs, 28% indels and 9% structural variants (SVs).[7,8]

However, there are important limitations to these techniques, including that neither MLPA nor microarray consider the full spectrum of SVs and do not provide nucleotide-level resolution, which is important for confirming causality and reveal insights into SVs formation.[7,9,10] These limitations may now be addressed by long-reads, that can span repetitive or other problematic regions, allowing identification and characterization of SVs.[10–14] This is particularly advantageous for antithrombin deficiency due to the high number of repetitive elements in and around *SERPINC1* (where 35% of sequence are interspersed repeats)[15], that hinders SVs identification by other methods.

Here, we report on the results of long-read whole-genome sequencing (LR-WGS) on 19 unrelated cases with antithrombin deficiency, selected from one of the largest cohort of patients with this disorder based on negative or ambiguous results, as well as not fully characterized SVs provided by routine molecular tests. Our aim was to identify new causal variants, resolve ambiguous ones and investigate the most likely mechanism of formation of SVs involved in this severe thrombophilia.

Methods

Cohort

Nineteen unrelated individuals with antithrombin deficiency were selected from our cohort of 340 cases, recruited between 1994 and 2019 and largely characterized by functional, biochemical, and molecular analyses. Selection was done based on negative results from multiple genetic studies evaluating *SERPINC1* gene, including Sanger sequencing followed by Next-Generation Sequencing (NGS) and MLPA, as well as negative glycosylation analysis (N= 11). Additionally, individuals with SVs that could not be characterized or that were identified by MLPA but had ambiguous results from other approaches (such as microarray and/or long-range PCR) were also selected (N= 8) (Table 1). Detailed information of these procedures is shown in Supplemental Methods. Measurement of antithrombin levels and function were performed for all participants as previously described.[16,17]

Long-read whole-genome sequencing

Long-read whole-genome sequencing (LR-WGS) of DNAs purified from peripheral blood leukocytes using Gentra Puregene Qiagen kit, used to reduce the fragmentation of DNA, was done using the PromethION platform (Oxford Nanopore Technologies).

Samples were prepared using the 1D ligation library prep kit (SQK-LSK109) and genomic libraries were sequenced on R9 flow cell. Read sequences were extracted from base-called FAST5 files by Guppy (versions 3.0.4 to 3.2.8; 3.0.4+e7dbc23 to 3.2.8+bd67289) to generate FASTQ files, that were then merged per sample.

Data processing and SV identification

We used the Snakemake library to develop an *in-house* multi-modal analysis workflow for the sensitive detection of SVs,[18] which is publicly available at https://github.com/who-blackbird/magpie. An overview of the workflow is shown in Figure 1A. Detailed information is provided in Supplemental Methods.

De novo assembly of the SINE-VNTR-Alu retroelement

Local *de novo* assembly was performed to characterize the SINE-VNTR-Alu retroelement insertion in P9. Reads within the region [GRCh38/hg38] Chr1:173,840,000-174,820,000 were extracted from the alignment of this individual and converted to a FASTQ file using Samtools.[19] *De novo* assembly was performed with wtdbg2 v2.5, using the parameters `-x ont -g 980k -X 10 -e 3`.[20] The *de novo* contig was then aligned to the reference genome using minimap2[21] with default parameters for nanopore reads. The genomic sequence containing the SINE-VNTR-Alu retroelement was then extracted from the alignment and analyzed with RepeatMasker (http://www.repeatmasker.org) to characterize the type of SINE-VNTR-Alu and its subelements.

Validations and breakpoint flanking sequence analysis

All candidate SV junctions were confirmed by PCR amplification and Sanger sequencing to verify all variant configurations at nucleotide level resolution. We then manually identified the presence of microhomology, insertions and deletions at the breakpoints as previously described.[22] The percentage of repetitive sequence was also calculated for each junction (+/− 150 bps) by intersecting these regions with the human genomic repeat library (hg38) from RepeatMasker version open-4.0.5 using bedtools.[23]

Results

Long-read sequencing identifies SVs involving SERPINC1

Nanopore sequencing in 21 runs produced reads with an average length of 4,499bp and median genome coverage of 16x (Figure 1B). After a detailed quality control analysis (Figure S1), 83,486 SVs were identified, consistent with previous reports using LR-WGS (Figure S2).[11] Focusing on rare variants (allele count <= 10 in gnomAD v3, NIHR BioResource and NGC project)[11,24,25] in *SERPINC1* and flanking regions, 10 candidate heterozygous SVs were observed in 9 individuals (Figure 1C). Visual inspection of read alignments identified an additional heterozygous SV in a region of low coverage involving *SERPINC1* in an additional patient (Table 1).

Resolution of causal SVs: identification of the first complex SV

Nanopore sequencing resolved the precise configuration of all SVs previously identified by MLPA in 8 individuals (P1-P8). Structural variants were identified independently of their size (from 7Kb to 968Kb, restricted to *SERPINC1* or involving neighboring genes) and their type (six deletions, one tandem duplication and one complex SV) (Figure 2,

Table 1). In all the cases the extension of the variants was determined, and nucleotide level resolution of breakpoints was achieved by the long reads (Table 1). Importantly, nanopore sequencing facilitated the resolution of the SVs identified in two patients (P2 and P6) that presented inconsistent or ambiguous results from MLPA and long-range PCR and NGS results (Table 1).

For the first case (P2) MLPA detected a deletion of exon 1, but long-range PCR followed by NGS suggested a deletion of exons 1 and 2. The discordant results were explained by nanopore sequencing, as this method revealed a complex SV in *SERPINC1* resulting in a dispersed duplication of exons 2 and 3 and a deletion spanning exons 1 and 2, both in the same allele (Figure 3). Specific PCR amplification and Sanger sequencing validated this complex structural variant in the proband and his affected daughter also with antithrombin deficiency.

For the second case (P6) MLPA detected a duplication of exons 2, 3 and 5 and a deletion of exon 6. Here, our sequencing approach identified a tandem duplication of exons 1 to 5, which was confirmed by long-range PCR (Figure 4). The tandem duplication of exons 1-5 was observed to be present in the affected son of P6, also with antithrombin deficiency.

A SINE-VNTR-Alu *retroelement insertion is identified in two previously unresolved cases and characterized by de novo assembly*

We aimed to identify new disease-causing variants in the remaining 11 participants with negative results using current molecular methods. Remarkably, two cases (P9 and P10)

presented an insertion of 2,440bp in intron 6. Blast analysis of the inserted sequence revealed a new SINE-VNTR-Alu retroelement (Figure 2, Table 1). Local *de novo* assembly using the data from P9 revealed an antisense-oriented SINE-VNTR-Alu element flanked by a target site duplication (TSD) of 14bp (Figure 2C), consistent with a target-primed reverse transcription mechanism of insertion into the genome.[26,27] Interestingly, the TSD in both individuals was also the same. The inserted sequence was aligned to the canonical SINE-VNTR-Alu A-F sequences (Figure S3A) and it was observed to be closest to the SINE-VNTR-Alu E in the phylogenetic tree (Figure S3B). Moreover, the VNTR sub-element harbored 1,449bp, which was longer than the typical ~520bp-long VNTR in the canonical sequences. Multiple PCRs covering the retroelement were attempted to validate this insertion, but all PCRs using flanking primers failed due to the highly repetitive sequence of this element, specially the VNTR sub-element, which is longer in this new SINE-VNTR-Alu. Only one specific PCR using an internal SINE-VNTR-Alu primer, which designed was facilitated by the Nanopore data, was able to amplify the breakpoint (Figure S4). This method was used to confirm the insertion in P9, P10 and to confirm the Mendelian inheritance of this SINE-VNTR-Alu, as it was also present in two affected relatives, both with antithrombin deficiency (Figure S4).

Breakpoint analysis supports a replication-based mechanism for the majority of SVs

Breakpoint analysis was performed to investigate the mechanism underlying the formation of these SVs involving *SERPINC1*. Nanopore sequencing facilitated primer design to perform Sanger sequencing confirmations for all the new formed junctions, demonstrating a 100% accuracy in 7/10 (70%) SVs called. Repetitive elements (RE)

were detected in all the SVs, with Alu elements being the most frequent (16/24, 67%) (Table S1). Additionally, breakpoint analysis identified microhomologies (7/11, 64%) and insertions, deletions or duplications (7/11, 64%) (Figure S5 and Table S2). Importantly, we observed a non-random formation driven by the presence of REs in some of the SVs. We point out an Alu element in intron 5, involved in SVs of P6, P7 and P8 (Figure 2B, Table S1).

Discussion

In this study we aimed to resolve the precise configuration of SVs involved in antithrombin deficiency using nanopore, to identify new candidate variants in previously unresolved cases and to investigate the possible mechanisms of formation of these SVs by breakpoint analysis. We have characterized disease-causing SVs in eight individuals with previous positive findings from MLPA and other methods but with unresolved variants, in two cases with previous contradictory results. Additionally, we reported a new causal SINE-VNTR-Alu retroelement insertion in two unrelated individuals that we characterized by local *de novo* assembly. Finally, we presented evidence for a replication-based mechanism of formation for most of the SVs causing this severe thrombophilia.

We show new evidence of how LR-WGS can be used to identify SVs causal of a genetic disease, in this case antithrombin deficiency, independently of its length or type. LR-WGS also gives information for the exact extension of the event involved and resolves conflictive data obtained by other methods. Additionally, we show how this approach is particularly powerful to investigate complex SVs, which are genomic rearrangements typically composed of three or more breakpoint junctions. Since these

are particularly challenging to detect and interpret by other methods, complex SVs are typically missed or misclassified in research and clinical diagnostic pipelines, although they have been reported as associated with multiple Mendelian diseases.[10] Here we show for the first time a complex SV in a patient with antithrombin deficiency, expanding the landscape of SV types involved in this disorder. Further investigations will be required to elucidate the exact mechanism of formation, since it remains unclear if this event occurred by one or multiple mutational events.

Additionally, we identified an intronic SINE-VNTR-Alu retroelement insertion in 2/11 (18%) previously unresolved individuals (P9 and P10). SINE-VNTR-Alu retroelements, along with other retrotransposons, are a source of regulatory variation in the human genome, but can also cause disease.[28] Although the number of pathogenic retroelements has increased during the last years with the use of WGS technologies, [25,29–31] these are usually missed by routine diagnostic methods. With LR-WGS we have not only identified the causal mutation in two previously unresolved families, but also performed local *de novo* assembly to characterize the exact sequence and length of its sub-elements, which might be relevant for future studies to investigate their possible role in severity and age of disease onset as other studies have shown.[32]

Furthermore, the genomic heterogeneity observed between the causal SINE-VNTR-Alu retroelement and the canonical sequences highlights the diverse genomic landscape of these retroelements and underscores the importance of their characterization in order to obtain a reliable catalogue of novel mobile elements to identify and interpret this type of causal variants in other patients and other disorders where retrotransposon insertions might also be involved.[27,33,34] This characterization has been historically challenging by

the application of classic technologies, but here we show that it can be achieved by *de novo* assembly of long-reads.

The decreased levels of antithrombin in plasma of P9 and P10 might be consistent with transcriptional interference of *SERPINC1* induced by the SINE-VNTR-Alu retroelement, as reported for other cases with pathogenic SINE-VNTR-Alu insertions.[28] Besides, the 2.4Kb insertion of a retroelement in intron 6 could introduce splicing signals affecting the normal splicing of *SERPINC1* RNA. However, the specific hepatic expression of *SERPINC1* hinders investigation of the exact mechanism, but the co-segregation of this variant with antithrombin deficiency observed in family studies of both probands supports the pathogenic consequences of this insertion. The identification of the same retrotransposon in two unrelated families from different regions of Spain (570 km far from each other) with the same TSD, does not only support the germline transmission of this SV, but also suggests a shared mechanism of formation or a founder effect, which must be confirmed by further studies.

In antithrombin deficiency, the detection and characterization of SVs remain particularly challenging due to the high number of repetitive elements in and around *SERPINC1* (35% of sequence in these gene are interspersed repeats). Specific mutational signatures can yield insights into the mechanisms by which the SVs are formed. Our breakpoint analysis suggested for most of the cases (P1-P8) a replication-based mechanism (such as BIR/MMBIR/FoSTeS),[35] consistent with previous studies done in antithrombin deficiency,[36,37] but importantly, we observed a non-random formation in some instances given the recurrent involvement of specific REs such as *Alu* elements in intron 5 of *SERPINC1*. It has been suggested that RE may provide

larger tracks of microhomologies, also termed 'microhomology islands', that could assist strand transfer or stimulate template switching during repair by a replication-based mechanism.[35] These microhomology islands were present in the SVs of 3 cases (P6, P7, P8), highlighting the important role that RE play in the formation of non-recurrent, but non-random, SVs. These results highlight that *SERPINC1* might be a hot-spot for SVs given the high number of REs in this gene and shows how LR-WGS can be used to investigate and resolve events occurring in repetitive genes and regions.

In total, 9 cases in this cohort remain yet unresolved, three of whom reported to have familial disease. An explanation may be that the causal variant was missed due to low coverage, or alternatively the variant is located in an unidentified transacting gene or in a regulatory element for *SERPINC1*, as we have recently reported for other genes.[13] The observation that the antithrombin deficiency in patients without causal SVs have significantly higher anti-FXa activity than those with SVs (Figure 1D) is supportive of the notion that causal variants may regulate gene expression, which must be analyzed in future studies.

Altogether this study provides insight into the molecular mechanism of SVs causing antithrombin deficiency and highlights the importance of identifying a new class of causal variants to improve diagnostic rates, lead to new therapeutic opportunities, and provide accurate family counselling, as decisions about long term anticoagulant prophylaxis are complex and carry significant morbidity and mortality risks. Moreover, our study suggests that SVs, which are often overlooked or misclassified by conventional methods, may be more common than anticipated as a genetic mechanism of antithrombin deficiency.

Description of Supplemental Data

Supplemental Material includes additional methods information, five figures and two tables.

Acknowledgments

We thank the participants involved in this study and their families. We thank NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centers, NHS Trusts and staff for their contribution. We thank the National Institute for Health Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

1. EGEBERG O. THROMBOPHILIA CAUSED BY INHERITABLE DEFICIENCY OF BLOOD ANTITHROMBIN. *Scand J Clin Lab Invest*. 1965;17(1):92. ed. & tran. Accessed January 21, 2021.
2. Corral J, de la Morena-Barrio ME, Vicente V. The genetics of antithrombin. *Thromb Res*. 2018;169:23-29. ed. & tran. Accessed March 29, 2021.
3. Lijfering WM, Brouwer JLP, Veeger NJGM, et al. Selective testing for thrombophilia in patients with first venous thrombosis: Results from a retrospective family cohort study on absolute thrombotic risk for currently known thrombophilic defects in 2479 relatives. *Blood*. 2009;113(21):5314-5322.

ed. & tran. Accessed February 4, 2021.

4. Mahmoodi BK, Brouwer J-LP, Ten Kate MK, et al. A prospective cohort study on the absolute risks of venous thromboembolism and predictive value of screening asymptomatic relatives of patients with hereditary deficiencies of protein S, protein C or antithrombin. *J Thromb Haemost*. 2010;8(6):1193-1200. ed. & tran. Accessed August 2, 2018.

5. Bravo-Pérez C, Vicente V, Corral J. Management of antithrombin deficiency: an update for clinicians. *Expert Rev Hematol*. 2019;12(6):397-405. ed. & tran. Accessed January 22, 2021.

6. Stenson PD, Ball E V., Howells K, Phillips AD, Mort M, Cooper DN. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics*. 2009;4(2):69-72.

7. Ordulu Z, Kammin T, Brand H, et al. Structural Chromosomal Rearrangements Require Nucleotide-Level Resolution: Lessons from Next-Generation Sequencing in Prenatal Diagnosis. *Am J Hum Genet*. 2016;99(5):1015-1033. ed. & tran. Accessed March 29, 2021.

8. Beauchamp NJ, Makris M, Preston FE, Peake IR, Daly ME. Major structural defects in the antithrombin gene in four families with type I antithrombin deficiency. Partial/complete deletions and rearrangement of the antithrombin gene. *Thromb Haemost*. 2000;83(5):715-721. ed. & tran. Accessed February 25, 2021.

9. Lam HYK, Mu XJ, Stütz AM, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol*. 2010;28(1):47-

55. ed. & tran. Accessed March 29, 2021.

10. Sanchis-Juan A, Stephens J, French CE, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med*. 2018;10(1):95. ed. & tran. Accessed March 18, 2020.

11. Beyter D, Ingimundardottir H, Eggertsson HP, et al. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. *bioRxiv*. November 2019:848366. ed. & tran. Accessed January 30, 2021.

12. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018;15(6):461-468.

13. Cretu Stancu M, Van Roosmalen MJ, Renkens I, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun*. 2017;8(1):1-13. ed. & tran. Accessed January 27, 2021.

14. French CE, Delon I, Dolling H, et al. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. *Intensive Care Med*. 2019;45(5):627-636. ed. & tran. Accessed March 1, 2021.

15. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet*. 2011;7(12).

16. de la Morena-Barrio M, Sandoval E, Llamas P, et al. High levels of latent antithrombin in plasma from patients with antithrombin deficiency. *Thromb Haemost*. 2017;117(5):880-888.

17. de la Morena-Barrio ME, Martínez-Martínez I, de Cos C, et al. Hypoglycosylation

is a common finding in antithrombin deficiency in the absence of a SERPINC1 gene defect. *J Thromb Haemost*. 2016;14(8):1549-1560.

18. De Coster W, De Rijk P, De Roeck A, et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res*. 2019;29(7):1178-1187. ed. & tran. Accessed June 2, 2021.
19. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. ed. & tran. Accessed February 9, 2021.
20. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 2020;17(2):155-158. ed. & tran. Accessed March 29, 2021.
21. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100. ed. & tran. Accessed January 27, 2021.
22. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61(1):437-455. ed. & tran. Accessed November 28, 2018.
23. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. ed. & tran. Accessed February 9, 2021.
24. Köster J, Rahmann S. Snakemake - A scalable bioinformatics workflow engine. *Bioinformatics*. 2018;34(20):3600. ed. & tran. Accessed February 9, 2021.
25. Turro E, Astle WJ, Megy K, et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*. 2020;583(7814):96-102. ed. & tran. Accessed February 14, 2021.
26. Vogt J, Bengesser K, Claes KBM, et al. SVA retrotransposon insertion-associated

deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. *Genome Biol*. 2014;15(6):15:R80. ed. & tran. Accessed March 29, 2021.

27. Payer LM, Burns KH. Transposable elements in human genetic disease. *Nat Rev Genet*. 2019;20(12):760-772. ed. & tran. Accessed February 22, 2021.

28. Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet*. 2012;46:651-675. ed. & tran. Accessed February 22, 2021.

29. Nakamura Y, Murata M, Takagi Y, et al. SVA retrotransposition in exon 6 of the coagulation factor IX gene causing severe hemophilia B. *Int J Hematol*. 2015;102(1):134-139. ed. & tran. Accessed May 31, 2021.

30. Van der Klift HM, Tops CM, Hes FJ, Devilee P, Wijnen JT. Insertion of an SVA element, a nonautonomous retrotransposon, in PMS2 intron 7 as a novel cause of lynch syndrome. *Hum Mutat*. 2012;33(7):1051-1055. ed. & tran. Accessed May 31, 2021.

31. Aneichyk T, Hendriks WT, Yadav R, et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell*. 2018;172(5):897-909.e21. ed. & tran. Accessed May 31, 2021.

32. Bragg DC, Mangkalaphiban K, Vaine CA, et al. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proc Natl Acad Sci U S A*. 2017;114(51):E11020-E11028. ed. & tran. Accessed May 31, 2021.

33. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA*. 2016;7(1).

34. Kazazian HH, Moran J V. Mobile DNA in Health and Disease. *N Engl J Med*.

2017;377(4):361-370. ed. & tran. Accessed January 26, 2021.

35. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*. 2016;17(4):224-238.

36. Kato I, Takagi Y, Ando Y, et al. A complex genomic abnormality found in a patient with antithrombin deficiency and autoimmune disease-like symptoms. *Int J Hematol*. 2014;100(2):200-205. ed. & tran. Accessed May 31, 2021.

37. Picard V, Chen J-M, Tardy B, et al. Detection and characterisation of large SERPINC1 deletions in type I inherited antithrombin deficiency. *Hum Genet*. 2010;127(1):45-53. ed. & tran. Accessed August 2, 2018.

Figure Titles and Legends

Figure 1. **Long-read sequencing workflow and results**. A) Overview of the general stages of the SVs discovery workflow. Algorithms used are depicted in yellow boxes. B) Nanopore sequencing results. i) Sequence length template distribution. Average read length was 4,499 bp (sd ± 4,268); the maximum read length observed was 2.5Mb. ii) Genome median coverage per participant. The average across all samples was 16x (sd ± 7.7). C) Filtering approach and number of SVs obtained per step. *SERPINC1* + promoter region corresponds to [GRCh38/hg38] Chr1:173,903,500-173,931,500. D) Anti-FXa percentage levels for the participants with a variant identified (P1-P10), cases without a candidate variant (P11-P19) and 300 controls from our internal database. The statistical significance is denoted by asterisks (*), where ***P<0.001, ****P≤ 0.0001. p-values calculated by one-way ANOVA with Tukey's post-hoc test for repeated measures. ATD=Antithrombin Deficiency; ONT=Oxford Nanopore Technologies; SV=Structural Variant.

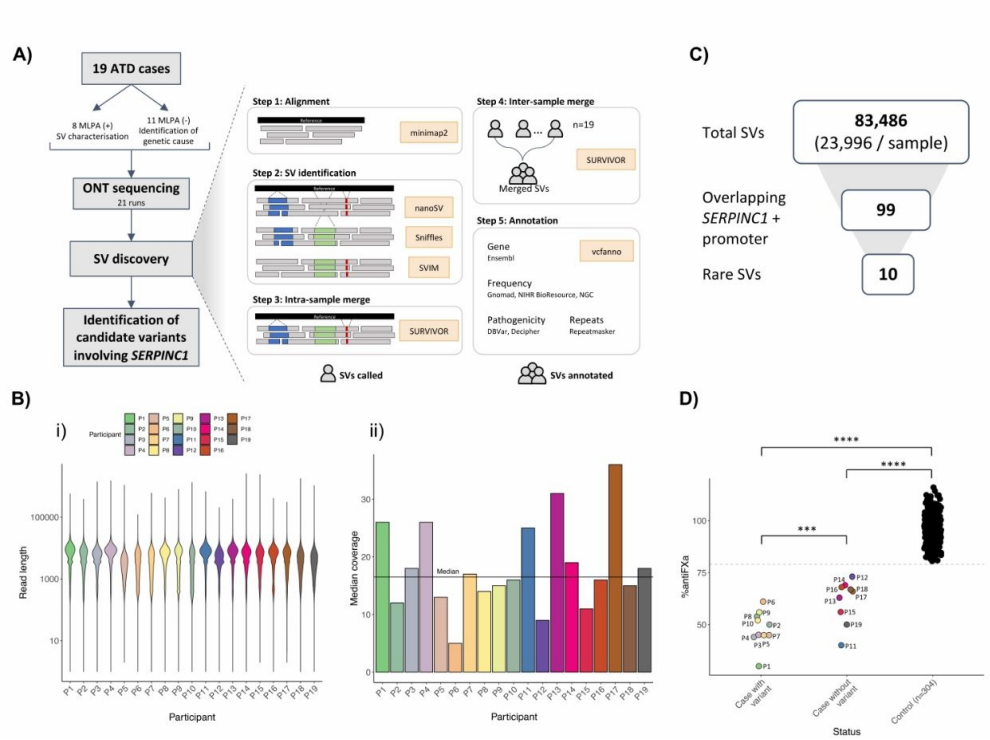
Figure 2. Candidate SVs identified by long-read sequencing. A) Schematic of chromosome 1 followed by protein coding genes falling in the zoomed region (1q25.1). SVs for each participant (P) are colored in red (deletions) and blue (duplications). The insertion identified in P9 and P10 is shown with a black line. B) Schematic of *SERPINC1* gene (NM_000488) followed by repetitive elements (RE) in the region. SINEs and LINEs are colored in light and dark grey respectively. Asterisks are present where the corresponding breakpoint falls within a RE. C) Characteristics of the antisense-oriented SINE-VNTR-Alu (SVA) retroelement (respect to the canonical sequence) observed in P9. Lengths of the fragments are subject to errors from Nanopore sequencing. TSD=Target site duplication.

Figure 3. Resolution of a complex SV. Schematic representation of genetic diagnostic methods used to characterize the SVs in participant P2. Results from MLPA, LR-PCR and Nanopore are shown in white boxes. Primers used for both LR-PCR and Sanger validation experiments are shown representing the genetic location of each one with orange and green arrows respectively. *SERPINC1* gene in the IGV screenshot is represented in blue and exons are indicated. J1 and J2 correspond to the new formed junctions described in Figure S5. J=New junction; M1k=1 kb Molecular weight marker; M=100bp molecular weight marker; P=patient; C=control; B=Blank.

Figure 4. Schematic representation of genetic diagnostic methods used to characterize the SVs in participant P6. Results from MLPA, LR-PCR and Nanopore are shown in white boxes. Primers used for both LR-PCR and Sanger validation experiments are shown representing the genetic location of each one with orange and green arrows respectively. *SERPINC1* gene in the IGV screenshot is represented in blue and exons are indicated. J1 corresponds to the new formed junctions described in Figure S5. J=New junction; M=Molecular weight marker 1Kb or 100b; P=patient; C=control; B=Blank. For the LR-PCR results, C1 and P1 correspond to PCR 1 (done with Primer F + Primer R), and C2 and P2 correspond to PCR2 (done with Primer F + Primer R2).

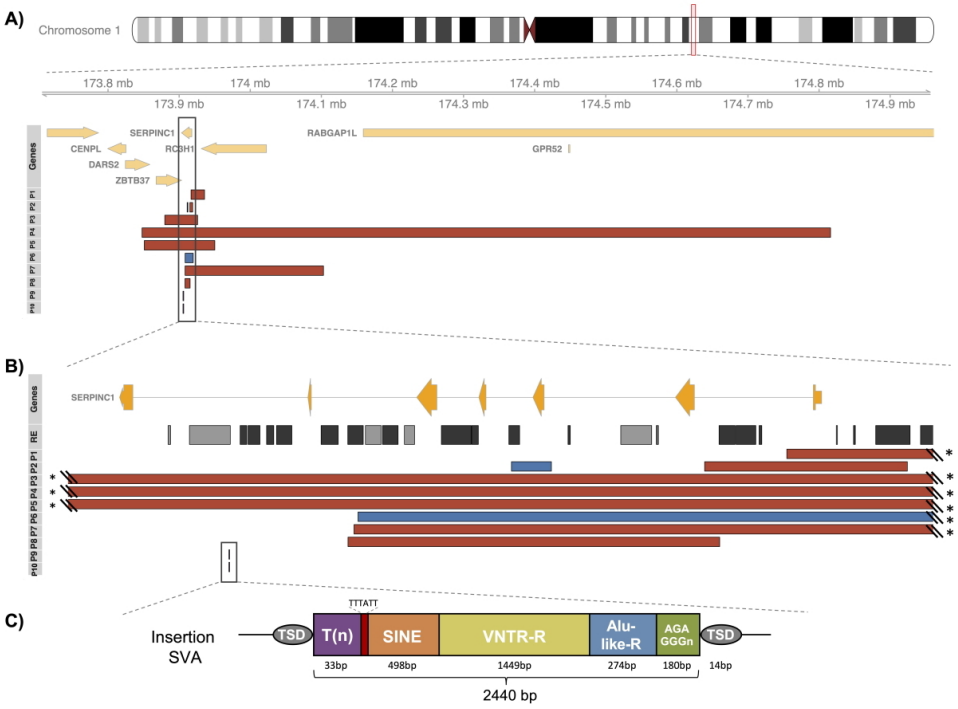
Table 1. Cohort of individuals included in this study, demographic, antithrombin values and genetic results. *SERPINC1* gene driven tests include MLPA, PGM sequencing (Ion Torrent) and long range PCR (LR-PCR) amplification and Myseq sequencing (Illumina). Genome wide tests are CGHa and whole genome sequencing (WGS) using nanopore technology (ONT). Coordinates have been confirmed by Sanger sequencing. Length refers to the extension of the structural variants. Het= Heterozygous; Ag= Antigen; bp= Base pair.

Participant	Antithrombin		Family history	Gender	MLPA *SERPINC1*	PGM	CGHa	LR-PCR & Illumina sequencing	WGS ONT	Algorithm	Genotype	Coordinates	Length (bp)
	antiFXa%	Ag (%)											
P1	30	30	Yes	M	Deletion exon 1	-	Negative	Deletion exon 1	Deletion exon 1	nanosv;sniffles;svim	Het	1:173916704-173935703	18999
P2	54	41	Yes	M	Deletion exon 1	-	Negative	Deletion exons 1,2	CxSV(Deletion exon 1;Duplication exon 3)	nanosv;sniffles	Het;Het	1:173911379-173915115; 1:173912151-173919034	3737;6884
P3	44	41	Yes	F	Complete deletion	-	Deletion 2 genes	-	Deletion 2 genes	nanosv;sniffles	Het	1:173879820-173925989	46169
P4	45	38	No	M	Complete deletion	-	Deletion 20 genes	-	Deletion 20 gene	nanosv;sniffles	Het	1:173847847-174816147	968005
P5	36	50	Yes	F	Complete deletion	-	-	-	Deletion 5 genes	nanosv	Het	1:173850996-173950174	99178
P6	61	46	Yes	M	Duplication exons 1,2 and 4; Deletion exon 6	-	Negative	Tandem duplication exons 1-5	Tandem Duplication exons 1-5	nanosv	Het	1:173908412-173919816	11404
P7	45	38	No	M	Deletion exons 1-5	-	Deletion exons 1-5 + 1 gene	-	Deletion 2 genes	nanosv;sniffles	Het	1:173908334-174103015	194389
P8	52	37	Yes	F	Deletion exons 2-5	-	-	-	Deletion exons 2-5	nanosv;sniffles	Het	1:173908218-173915405	7187
P9	56	61	Yes	F	Negative	Negative	-	Negative	Insertion SVA	nanosv	Het	1:173905922	2440
P10	50	46	Yes	F	Negative	Negative	-	Negative	Insertion SVA	visual inspection	Het	1:173905922	2440
P11	40	41	Yes	F	Negative	Negative	-	Negative	Negative				
P12	73	62	No	F	Negative	Negative	-	Negative	Negative				
P13	63	58	No	M	Negative	-	-	Negative	Negative				
P14	69	NA	No	F	Negative	Negative	-	Negative	Negative				
P15	56	45	Yes	F	Negative	-	-	-	Negative				
P16	68	54	No	M	Negative	Negative	-	Negative	Negative				
P17	66	67	No	M	Negative	Negative	-	Negative	Negative				
P18	67	70	No	F	Negative	-	-	-	Negative				
P19	50	70	Yes	M	Negative	-	-	-	Negative				



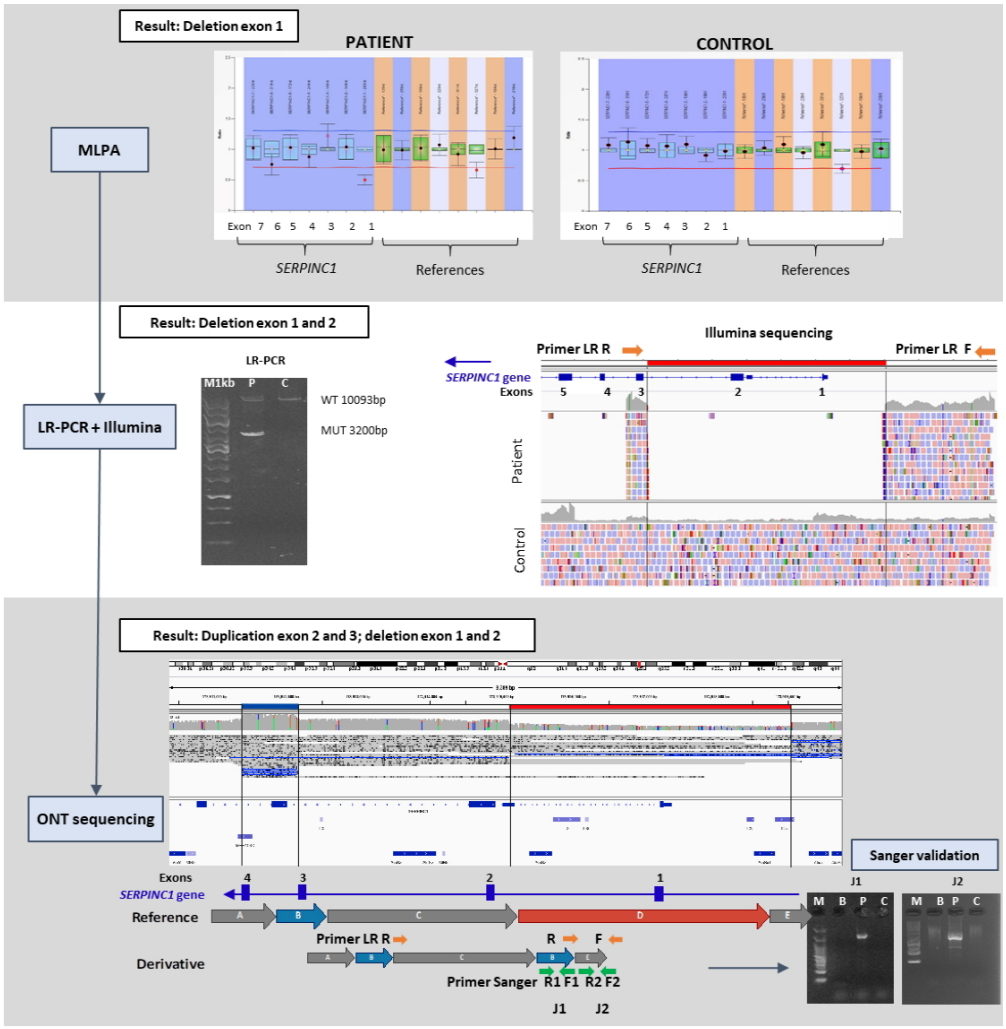
Long-read sequencing workflow and results. A) Overview of the general stages of the SVs discovery workflow. Algorithms used are depicted in yellow boxes. B) Nanopore sequencing results. i) Sequence length template distribution. Average read length was 4,499 bp (sd ± 4,268); the maximum read length observed was 2.5Mb. ii) Genome median coverage per participant. The average across all samples was 16x (sd ± 7.7). C) Filtering approach and number of SVs obtained per step. SERPINC1 + promoter region corresponds to [GRCh38/hg38] Chr1:173,903,500-173,931,500. D) Anti-FXa percentage levels for the participants with a variant identified (P1-P10), cases without a candidate variant (P11-P19) and 300 controls from our internal database. The statistical significance is denoted by asterisks (*), where ***P<0.001, ****P≤ 0.0001. p-values calculated by one-way ANOVA with Tukey's post-hoc test for repeated measures. ATD=Antithrombin Deficiency; ONT=Oxford Nanopore Technologies; SV=Structural Variant.

331x245mm (96 x 96 DPI)



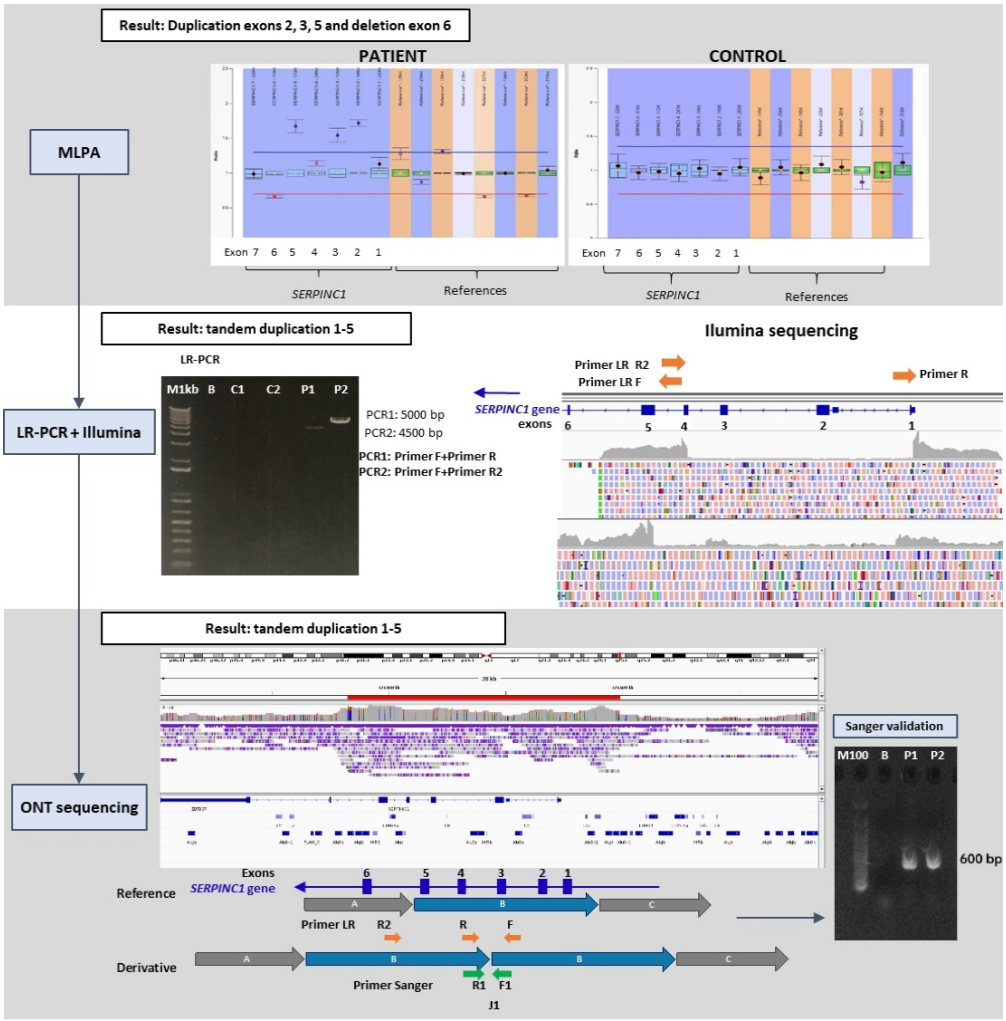
Candidate SVs identified by long-read sequencing. A) Schematic of chromosome 1 followed by protein coding genes falling in the zoomed region (1q25.1). SVs for each participant (P) are colored in red (deletions) and blue (duplications). The insertion identified in P9 and P10 is shown with a black line. B) Schematic of SERPINC1 gene (NM_000488) followed by repetitive elements (RE) in the region. SINEs and LINEs are colored in light and dark grey respectively. Asterisks are present where the corresponding breakpoint falls within a RE. C) Characteristics of the antisense-oriented SINE-VNTR-Alu (SVA) retroelement (respect to the canonical sequence) observed in P9. Lengths of the fragments are subject to errors from Nanopore sequencing. TSD=Target site duplication.

331x243mm (96 x 96 DPI)



Resolution of a complex SV. Schematic representation of genetic diagnostic methods used to characterize the SVs in participant P2. Results from MLPA, LR-PCR and Nanopore are shown in white boxes. Primers used for both LR-PCR and Sanger validation experiments are shown representing the genetic location of each one with orange and green arrows respectively. SERPINC1 gene in the IGV screenshot is represented in blue and exons are indicated. J1 and J2 correspond to the new formed junctions described in Figure S5. J=New junction; M1k=1 kb Molecular weight marker; M=100bp molecular weight marker; P=patient; C=control; B=Blank.

295x301mm (96 x 96 DPI)



Schematic representation of genetic diagnostic methods used to characterize the SVs in participant P6. Results from MLPA, LR-PCR and Nanopore are shown in white boxes. Primers used for both LR-PCR and Sanger validation experiments are shown representing the genetic location of each one with orange and green arrows respectively. SERPINC1 gene in the IGV screenshot is represented in blue and exons are indicated. J1 corresponds to the new formed junctions described in Figure S5. J=New junction; M=Molecular weight marker 1Kb or 100b; P=patient; C=control; B=Blank. For the LR-PCR results, C1 and P1 correspond to PCR 1 (done with Primer F + Primer R), and C2 and P2 correspond to PCR2 (done with Primer F + Primer R2).

292x297mm (96 x 96 DPI)

SUPPLEMENTAL MATERIAL

Supplemental methods

Genetic diagnostic methods

Genetic diagnostic methods used to evaluate *SERPINC1* gene included: i) Sanger sequencing of exons and flanking regions, ii) Multiplex Ligation-dependent Probe Amplification (MLPA) covering the 7 exons of this gene (SALSA MLPA Kit P227 SerpinC1; MRC Holland, Amsterdam, The Netherlands), ii) whole gene sequencing by Ion Torrent technology (PGM; Thermo Fisher Scientific, Waltham, MA, USA) , iv) long-range PCR (LR-PCR) amplification of the whole gene followed by Next Generation Sequencing (NGS) MiSeq platform (Illumina, San Diego, CA, USA) and / or v) Comparative Genomic Hybridization array (CGHa; CytoScan® HD Array; Thermo Fisher Scientific). These methods were performed as previously described. [1,2]

SV identification

Reads were aligned against the GRCh38/hg38 human reference genome using minimap2 (2.17-r941)[3] with default parameters for nanopore data ('`-ax map-ont`' parameter). SV discovery was done using a combination of three different algorithms:

- Sniffles v1.0.11[4] was executed with a supporting read evidence of 4 ('-s 4' parameter) due to coverage variability.
- NanoSV v1.2.4[5] was executed with default parameters. It was run on each independent chromosome in parallel to optimize compute time, with the limitation that inter-chromosomal variants were not detected.
- SVIM v1.2.0[6] was executed with default parameters. Resulting SV calls with a quality score of less than 10 were filtered out of the dataset and not used for analysis. SV calls were merged at two different levels: intra-sample merge, to merge calls within individuals that had been identified by all three algorithms, and inter-sample merge, to merge SV calls across individuals.

Intra-sample merge. For each of the 19 samples, SV calls from all three different algorithms were merged using SURVIVOR v.1.0.7.[7] VCF files were concatenated into one using bcftools[8] and SURVIVOR was run using the command '`SURVIVOR merge in.fofn 500 1 -1 -1 -1 -1 out.vcf`', requiring a maximum distance of 500bp between breakpoints. Additionally, intra-sample merge was done independently of the SV type, since different SV detection algorithms determine the type in different ways. For example, Sniffles determines canonical (DEL, DUP, INS, INV, TRA) and some complex SV types, while NanoSV calls only breakends (BND). The following options were turned-off: take the strands of SVs into account, estimate distance based on the size of SV, minimum size of SVs to be taken into account. After running SURVIVOR,

an *in-house* script was used to select the most common SV type. If there was no common type, the order of selection was NanoSV (if the SV type was not a BND) > Sniffles > SVIM type.

Inter-sample merge. Then, all the 19 samples were merged using SURVIVOR, taking the SV type into account. The command run was: '`SURVIVOR merge in.fofn 500 1 1 -1 -1 -1 out.vcf`'.

Identification of candidate SVs in SERPINC1

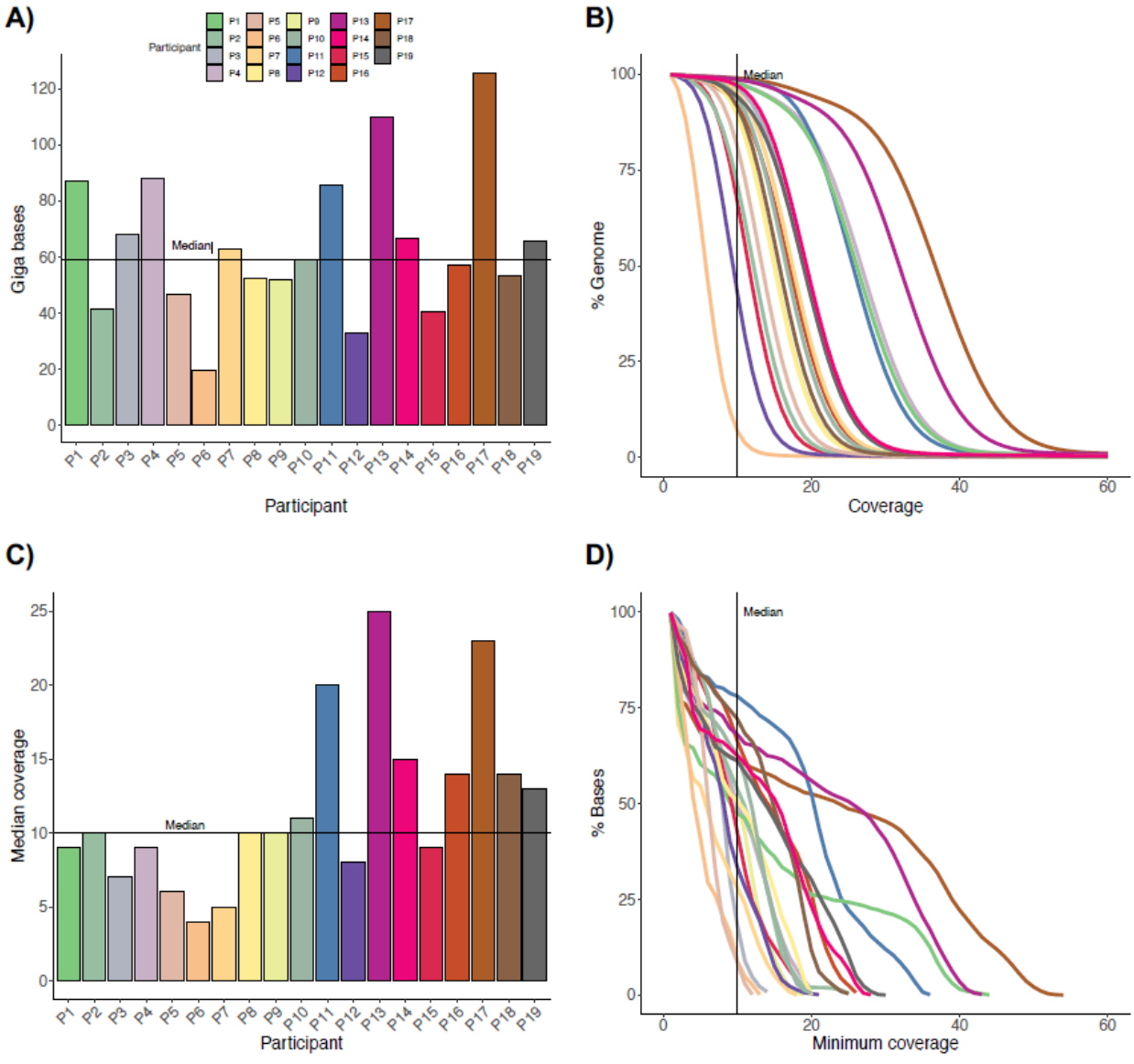
After filtering out variants that overlapped bad mapping quality regions (obtained from [9]), a total number of 83,486 SVs were identified across all participants, with a median of 23,996 (sd ± 3,431) SVs per sample. In order to identify disease-causing SVs associated with antithrombin deficiency, we filtered for SVs overlapping the region [GRCh38/hg38] chr1:173,903,500-173,931,500, which includes *SERPINC1* gene and its promoter region. A total number of 99 SVs were observed, of which 10 were absent in gnomAD, NGC and the NIHR BioResource (5-7). These 10 SVs were observed in 9 samples: 6 were deletions, 1 was a tandem duplication, 1 was a complex SV formed by 1 duplication and 1 deletion, and 1 was a SINE-VNTR-Alu (SVA)-type retrotransposon insertion.

Manual inspection of SVs

The alignments for all the cases without a candidate variant were manually inspected at the above locus of interest using IGV. [10] An additional SVA insertion was observed in P10, at the exact same position than P9. There were only two reads supporting the alternate, hence explaining why it had not been called by any variant caller. Running Sniffles with a read evidence of 2 ('-s 2' parameter) on the P10 data resulted in the SVA insertion being called.

Supplemental figures

Figure S1. **Sequencing results colored by participant. (A)** Giga bases sequenced **(B)** Percentage of bases of the genome covered at a specific minimum coverage **(C)** Median coverage in *SERPINC1* + promoter region **(D)** Coverage distribution in *SERPINC1* + promoter region **(E)** Percentage of reads with a minimum Q score **(F)** Read N50, which refers to a value where half of the data is contained within reads with alignable lengths greater than this.



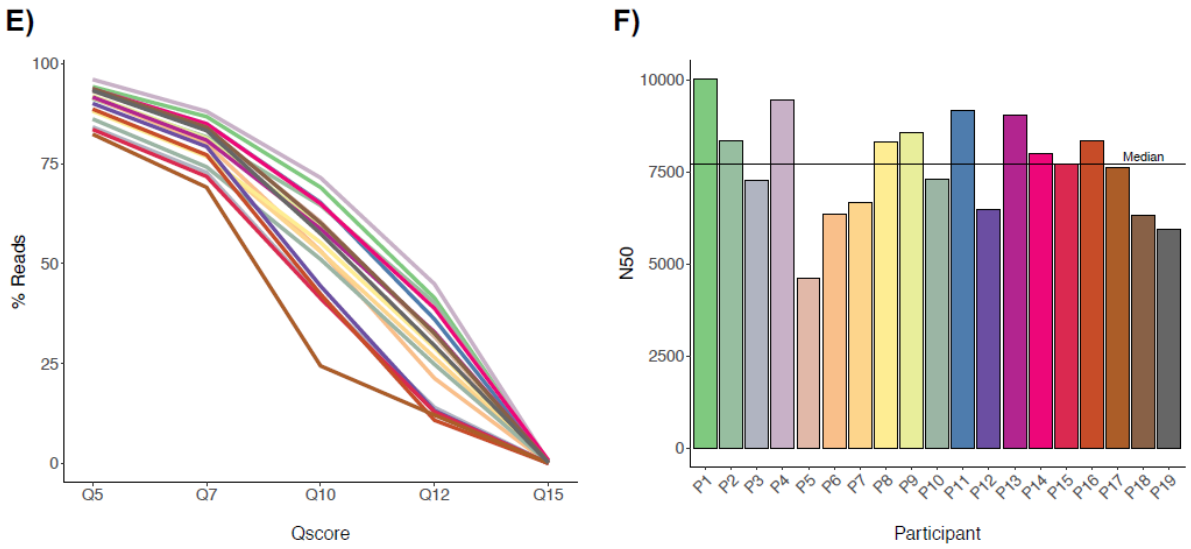


Figure S2. Structural variants metrics. (A) Number of SVs identified by type and participant **(B)** Number of SVs by SV size **(C)** Fraction of SVs per allele count in our internal cohort of 62 individuals with long-read sequencing data **(D)** Number of SVs by median coverage and participant.

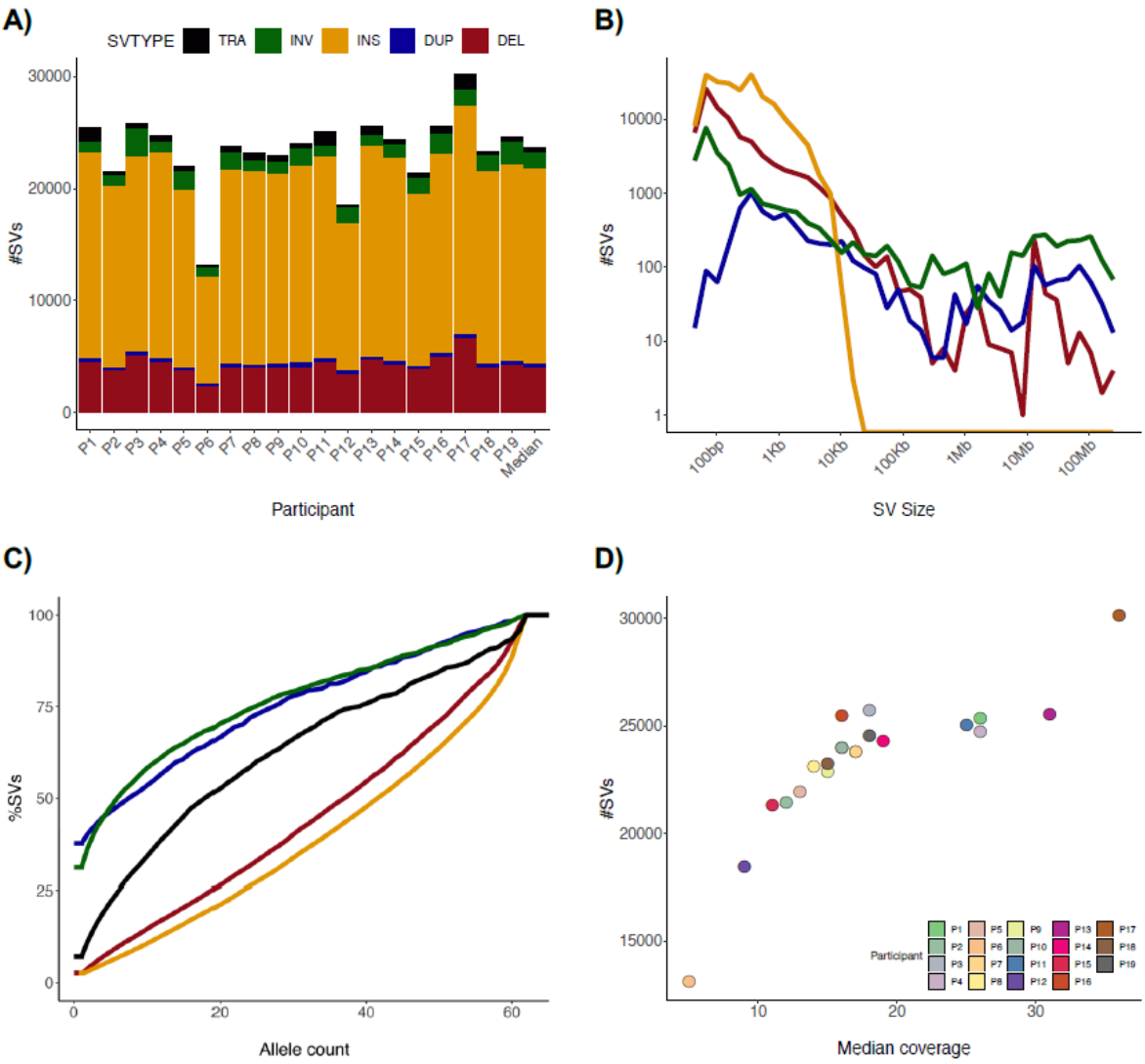
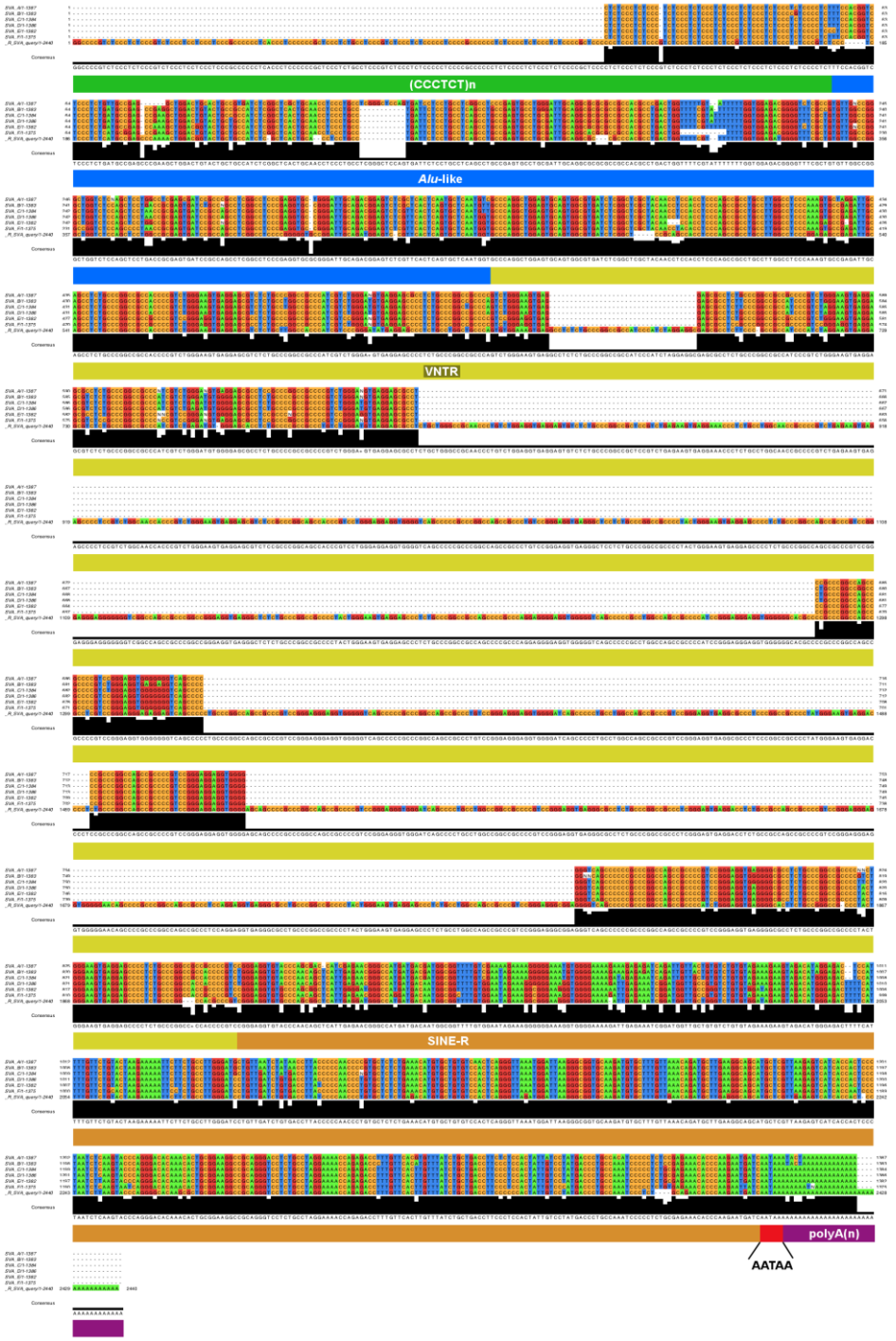


Figure S3. **SVA sequence alignments. (A)** The consensus sequences of SVA-A, -B, -C, -D, -E, and -F were taken from RepeatMasker (http://www.repeatmasker.org), then aligned using MAFFT[11] with default parameters. SVA_query corresponds to the SVA insertion in P9. Alignments were visually inspected and coloured by nucleotide using JalView. [12] Sub-elements of the SVAs are indicated underneath the consensus sequence matching colours in Figure 2C. **(B)** Phylogenetic tree was constructed with the Neighbour-Joining (NJ) algorithm using the Jukes-Cantor substitution model and visualized with iTol [13]. The SVA insertion in P9 was observed to be closest to the SVA E in the phylogenetic tree.

(A)



(B)

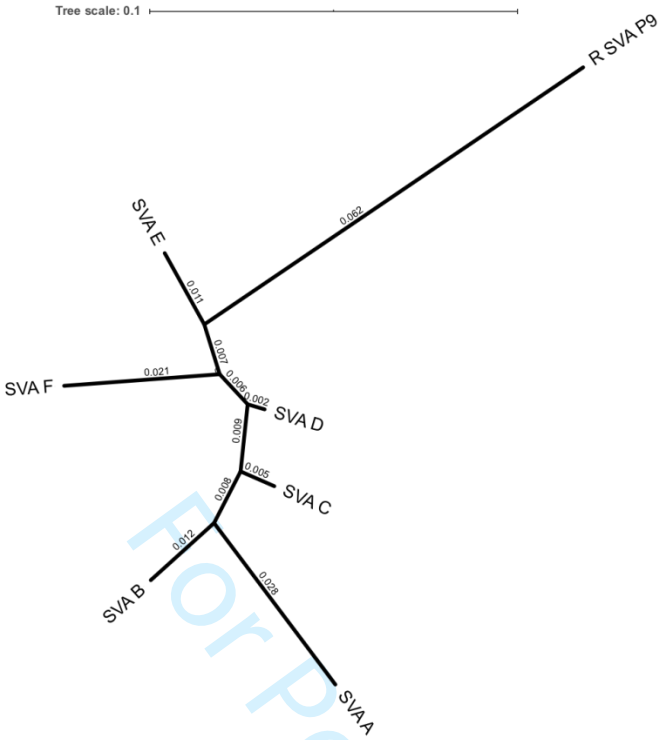


Figure S4. PCR validation of the SVA insertion in P9 and P10. (A) Schematic of *SERPINC1* gene (NM_000488) with zoom to intron 6 showing the SVA structure. Primers used in the long- and short-range PCRs are shown in orange and green respectively. Primer 8* was specifically designed within the inserted SVA sequence. Briefly, four reads of the retrotransposon present in the nanopore data for P9 and P10 were aligned to identify regions without any mismatch in order to select a 20 nucleotide sequence to be used as primer. That sequence was also checked to be present in *de novo* assembly alignment. **(B)** Primer combinations for PCR amplifications and expected sizes for wild type and mutated alleles are shown in the table. PCRs 1-4 were tested under different experimental conditions, and although in all cases the wild type allele was always amplified, no amplification of the mutated allele containing the SVA was obtained in P9 or P10. **(C)** The amplification of PCR 4 in agarose gel is shown. Only the 800 base pairs (bp) of the wild type allele was amplified in P9, P10 but also a healthy control. Only PCR 5, using the primer specific of the SVA rendered positive results and a specific 550bp band was obtained in P9, P10 and two relatives. **(D)** Family pedigrees of P9 and P10, including clinical information, the diagnosis of antithrombin deficiency (semi-filled symbols) and the anti-FX activity (as % of a reference plasma). B=Blank; M=Molecular Weight Marker; DVT=Deep vein thrombosis; PE=Pulmonary embolism.

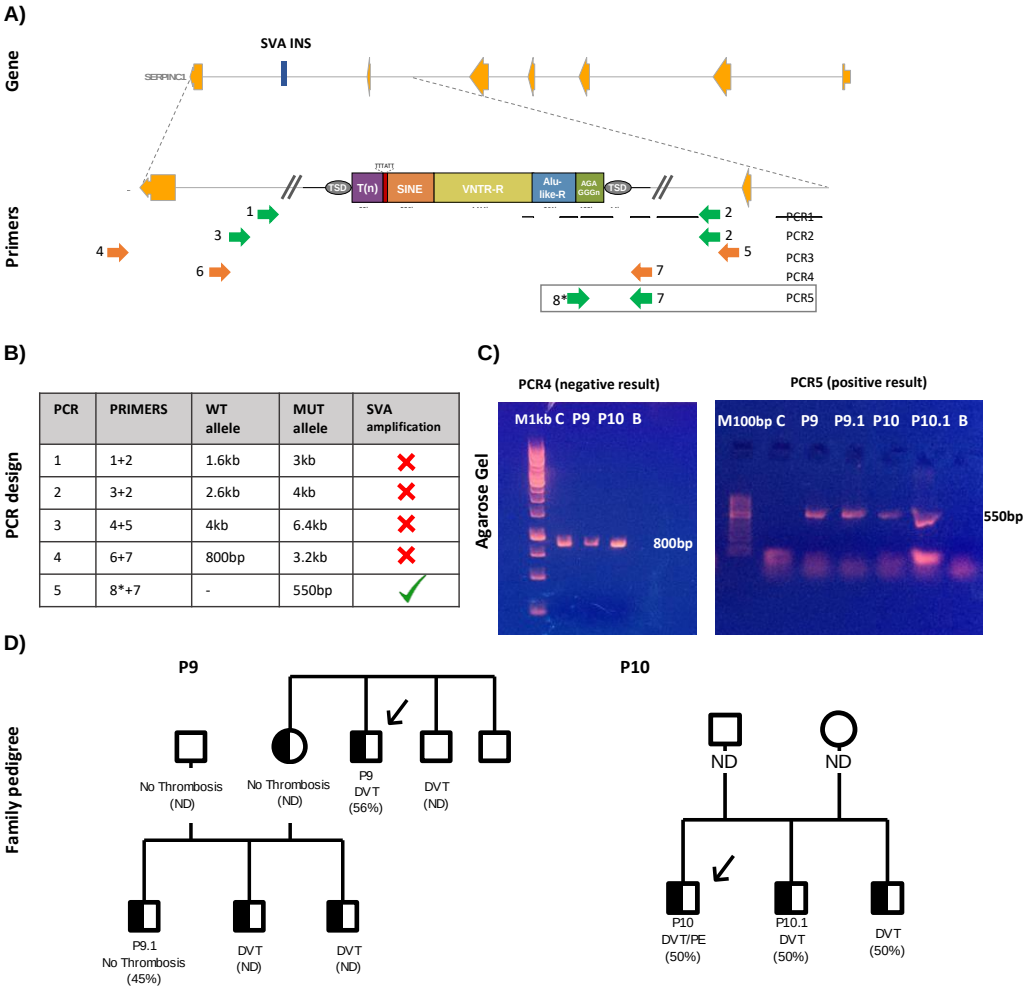
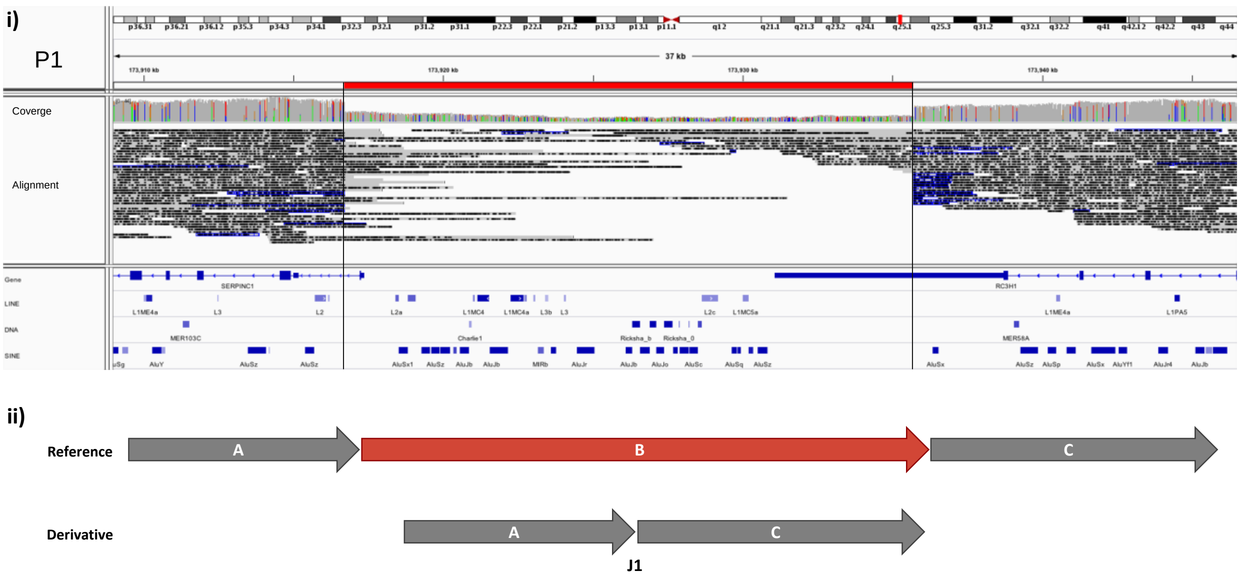


Figure S5. Nucleotide level characterization of the candidate SVs. Breakpoint junction sequence is aligned to the proximal and distal genomic reference sequence or sequence of breakpoint junction in hypothetical intermediate, as shown. Alignment is only shown for novel breakpoint junctions in the derivative chromosome. Microhomology at the breakpoint is indicated in red. Sequence in blue indicates inserted sequences at the breakpoint junction. Underline indicates repetitive elements in the reference, specified in Italic. J=Junction. **(A-H)** P1-8, **(I)** P9 and P10.

(A) P1

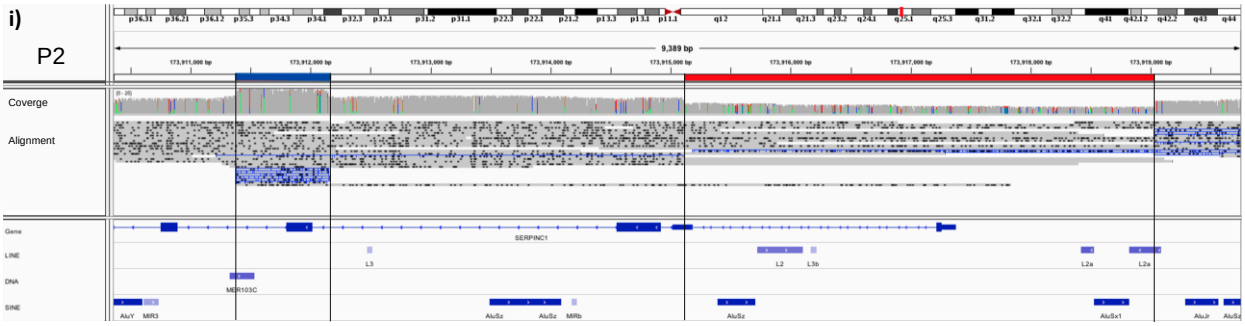


iii)

```
1-173916654   CCAAGACAGTTGTGCCACCACCCCGTTTCTAGCCTCCTGTGACAACTCA*GATGAAGAAGGGAGTGTGTGTGGCATTGAGGGAGGGACAACTCTCATTGTG   1739167 54
J1-1          CCAAGACAGTTGTGCCACCACCCCGTTTCTAGCCTCCTGTGACAACTCA* GGAGCAAGGAACAAGGAAAAAAAGCAAAAAACAAAAAACAAATTTTGGGGG   100
1-173935659   AAATATGCCATTTGTCCTTACAGGCTCATGATATCACAGAGAAAAGGAA* ------------AAGGAAAAAAAGCAAAAAACAAAAAACAAATTTTGGGGG   173935759
```

A-rich

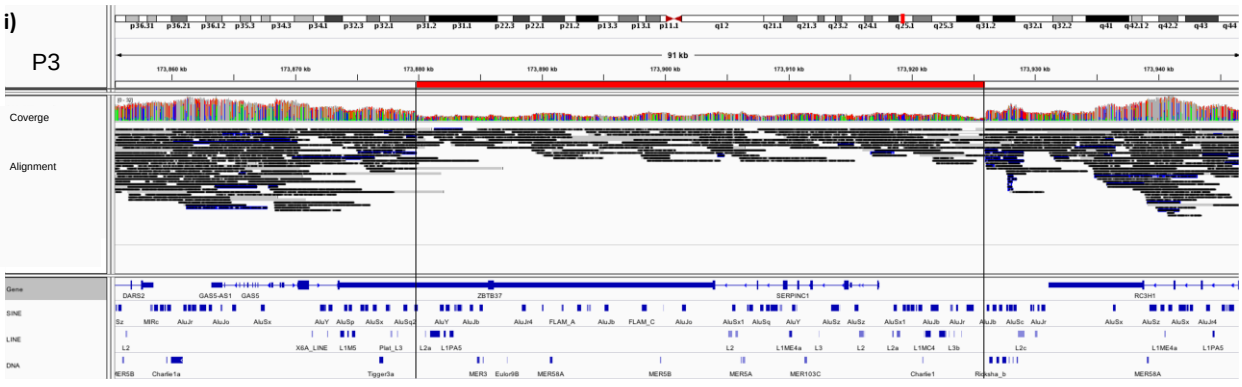
(B) P2



iii)

```
1-173915066   CCGGGACAGGTTCAGTCCTAGACTTCTTGCCAGGGGACAGTTCA GTTGC*CTGGACGTGGTCATGTTGGAATTAGAATCATCTGATGGTTATTTGGAGTA   173915165
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
J1-1          CCGGGACAGGTTCAGTCCTAGACTTCTTGCCAGGGGACAGTTCA GTTGC*-----AGAGTTTCTCAGAACCTTCATAGGCCCATGTGCTTCATGAATCTC   94
              |  |  ||  ||  |  |  |  |  |  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
1-173911329   TCAAATACATTTGAGGAATGCCGATTGAATAAAGCCAACAGATTTCTTT *GTTGCAGAGTTTCTCAGAACCTTCATAGGCCCATGTGCTTCATGAATCTC   173911428
                                                MER103C                                    MER103C
1-173912101   AATAGCCACCTCAGTTGTTAACTCCTTCAAGCACAGTACCTGGGAC AGC*ATAAATGCTCAACTCTTGTGTTCTGATGAATAAAAGGATAAAAGAAGAATG   173912201
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
J2-1          AATAGCCACCTCAGTTGTTAACTCCTTCAAGCACAGTACCTGGGAC AGC*---TGTCCAAAGAGAAATCAGAAATACAAAGACCTTGAGGTTTCAGGAACA   97
              |  |  |  |  ||  ||  |  |  |  |  ||  ||  |  |  |||||||||||||||||||||||||||||||||||||||||||||||||||||
1-173918984   GAAACCAGAATGAAATGATGGAGGGGGCCATATGAATATCTTTAGGAAG* AGCTGTCCAAAGAGAAATCAGAAATACAAAGACCTTGAGGTTTCAGGAACA   173919084
                                                L2a                                        L2a
```

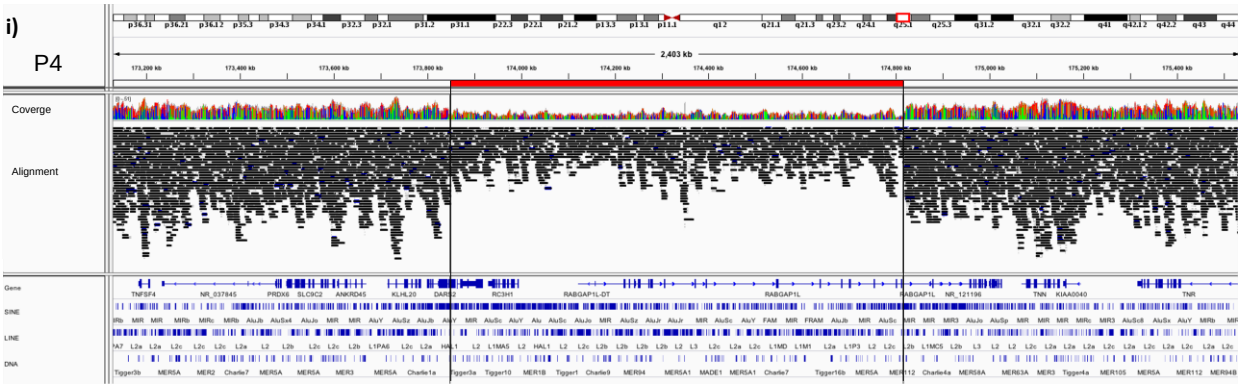
(C) P3



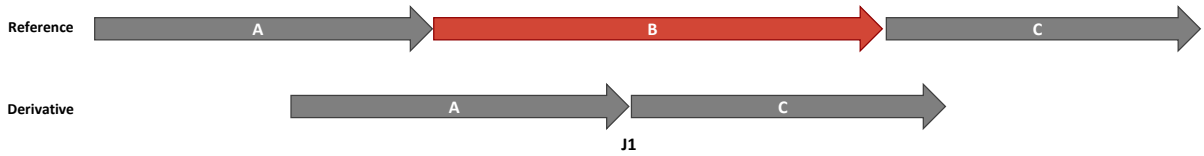
iii)

```
                                AluSx1                                                    AluSx1
1-173879770   GTGAAACCCCATCTCTACTAAATATAAAACATTAGCCAGGCATGG TGGT*GGGCACCTGTAATCTCAGCTACTCGGGAGGCTGCAACAGGAGAATCACTTG   173879870
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
J1-1          GTGAAACCCCATCTCTACTAAATATAAAACATTAGCCAGGCATGG TGGT*GCCTGGACAAGGCTATTCTTTGGAGGCTGAGCAAGAGAGAGTTGTGTGGGG   100
              |  ||  |  |  |  |  |  ||  ||  |  |  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
1-173925939   AAATGCTTACTCATCCAGAAGAATTGACACCTTGTATGAACAGAAACAT* ---TGGACAAGGCTATTCTTTGGAGGCTGAGCAAGAGAGAGTTGTGTGGGG   173926039
```

(D) P4



ii)



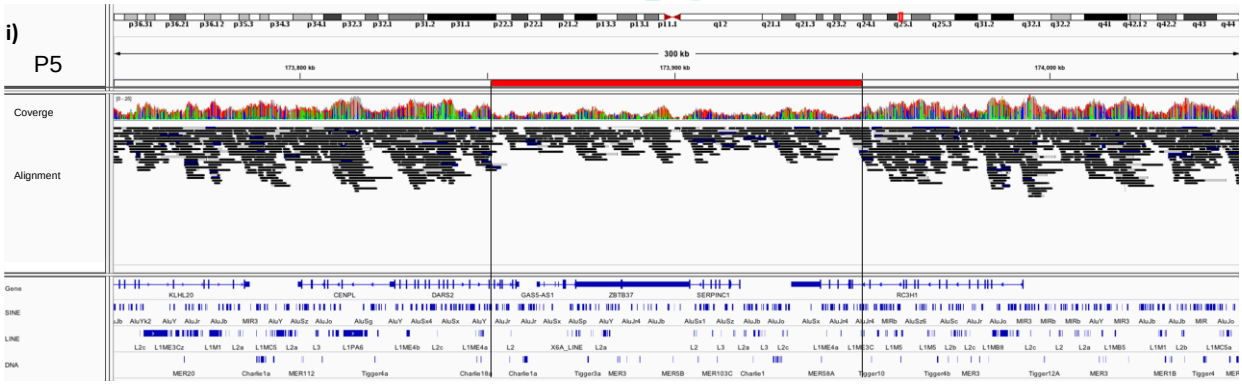
iii)

HAL1 *AluY*

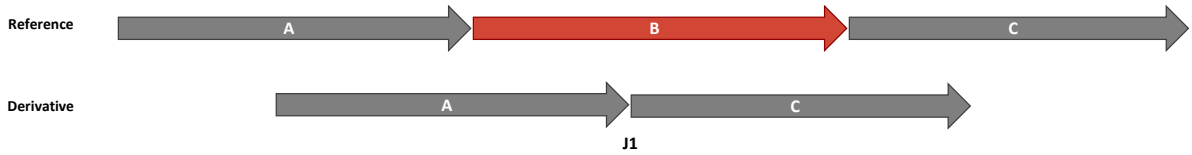
```
1-173847797   CCTCACCTAGTTACTCTGTACCTTCTGCACAGCTGTCTTTCTGAACCTTT*TTTTTTTTTTTTTTTTTTTTTTGAGACGGGAGTCTCACTCTGTCGCCCAG   173847897
              ||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||| | | | ||| |||||||
J1-1          CCTCACCTAGTTACTCTGTACCTTCTGCACAGCTGTCTTTCTGAACC---*TTTTTTTTTTTTTTTTTTTTTTGAGATAGAGTTTGCTCTTGTTGCCCAG   97
                    | | ||          | | |          | |          ||||||||||||||||||||||||||||||||||||||||||||||||
1-174816097   TTGTATGAACATATGAGCATTTTGTTTTTCTTAGTAACATAATACATAGC*TTTTTTTTTTTTTTTTTTTTTTGAGATAGTTTGCTCTTGTTGCCCAG   174816197
```

AluSp

(E) P5



ii)



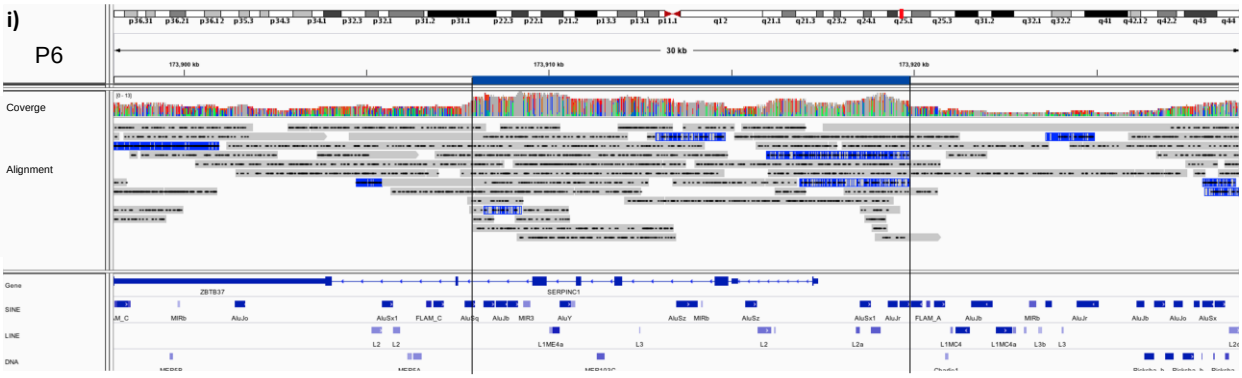
iii)

AluSx1 *AluSx1*

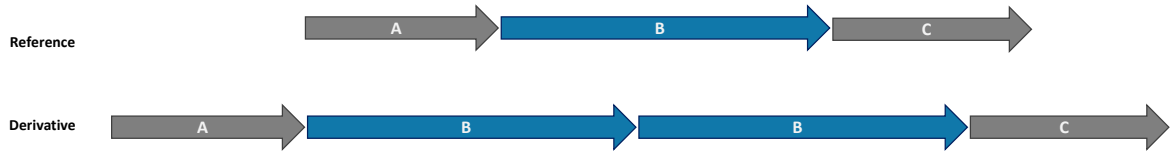
```
1-173850946   TTAAGGGTTGGCCGGGCATGGTGGCTAACGCCTGTAATCCCAGCACTTT*GGGAGGCCGAGGTGGGTGGATCGCCTGAGGTCAGGAGTTTGAGACCAGCCT   173851046
              |||||||||||||||||||||||||||||||||||||||||||||||| |||| | | | || || | | || ||
J1-1          TTAAGGGTTGGCCGGGCATGGTGGCTAACGCCTGTAATCCCAGCACTTT*-AAGTACACATCTTAAAAGCAATGCTAAGTGAAAACATGAAAATGCAAGCT   99
              | | || ||| | | | | | ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
1-173950124   TGCCACTGCACTCCAGCCTGGGTACAGAGTGGAACTCCGTCTCAAAAAA*AAA GTACACATCTTAAAAGCAATGCTAAGTGAAAACATGAAAATGCAAGCT   173950224
```

AluSx *L1ME3C*

(F) P6



ii)



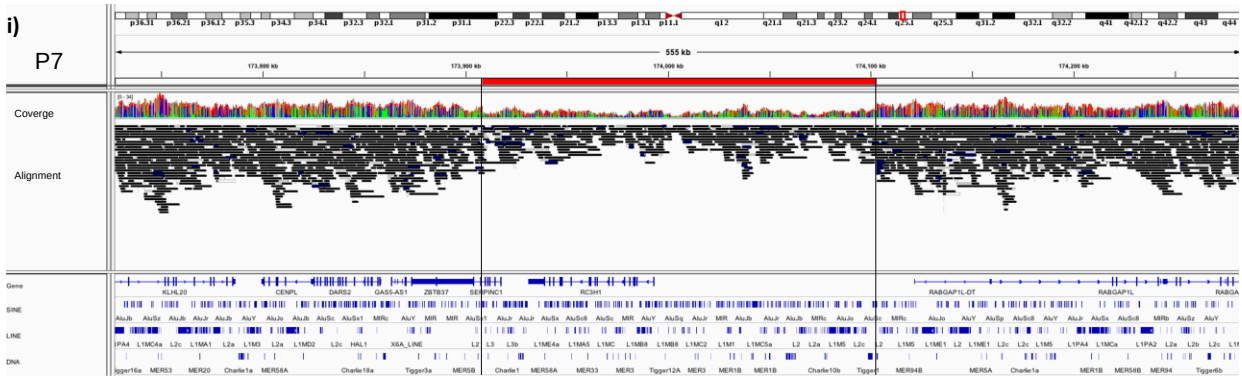
iii)

AluSz *AluSz*

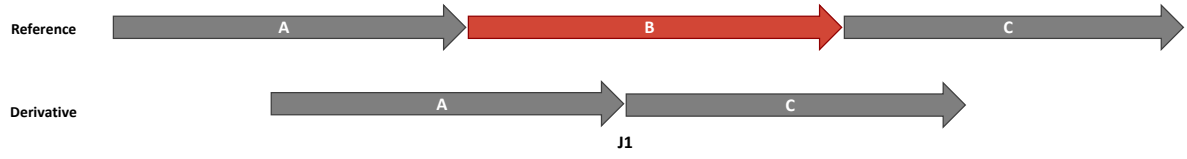
```
1-173919740   ACCTGTAATCTCAGCTACTCAGGAGGCTGAGGCAGGAGAATCACTTGAA*CCCGGGAGGTGGAGGCTGCAGTGAGTCGAGATTACGCCACTGCACTCTAGC   173919840
              ||||||||||||||||||||||||||||||||||||||||||||||||||  ||| |||||||| |||||||| | ||||| | |||||||||| |||
J1-1          ACCTGTAATCTCAGCTACTCAGGAGGCTGAGGCAGGAGAATCACTTGAA*TCTGGGGGGTGGAGGTTGCAGTGAGCCAAGATTGCACCACTGCACTCCAGC   100
              ||||||||  ||||  ||  ||||||||||||||||||||||||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||
1-173908362   GCCTGTAATCCTAGCTCCTTGGGAGGCTGAGGCAGGAGAATCACTTGAA*TCTGGGGGGTGGAGGTTGCAGTGAGCCAAGATTGCACCACTGCACTCCAGC   173908462
```

AluSx1 *AluSx1*

(G) P7



ii)



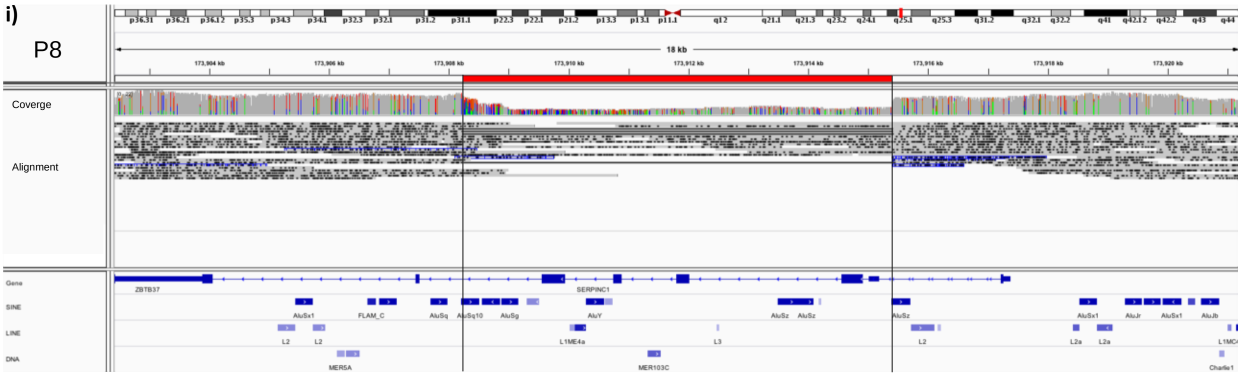
iii)

AluSx1 *AluSx1*

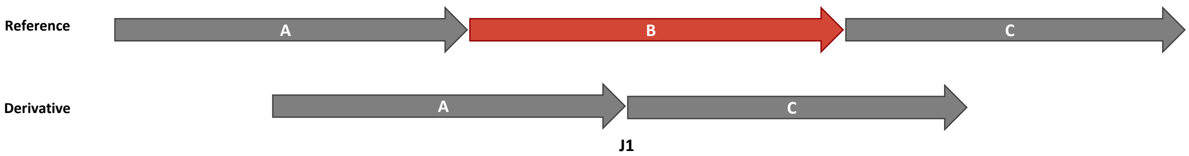
```
1-173908284   TTCAAGACCAGCCTGGCCAACATAGTGAAACCCCGTTTCTACTAAAAAT*GCAAAAATTAGCCGGGCGTGGTGGCAGGCGCCTGTAATCCTAGCTCCTTGG   173908384
              ||||||||||||||||||||||||||||||||||||||||||||||||||  |||||||||||| ||| ||||||  |||||||||||| |||| ||
J1-1          TTCAAGACCAGCCTGGCCAACATAGTGAAACCCCGTTTCTACTAAAAAT*ACAAAAATTAGCCAGGCATGGTGGTGAGCGCCTGTAATCCCAGCTACTCCA   100
              ||||||||||||||  ||||||  |||||  ||||  ||||||||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||
1-174102965   GTCAAGACCAGCCTGATCAACATGGAGAAACACCGTCTCTACTAAAAAT*ACAAAAATTAGCCAGGCATGGTGGTGAGCGCCTGTAATCCCAGCTACTCCA   174103065
```

(H) P8

i)



ii)

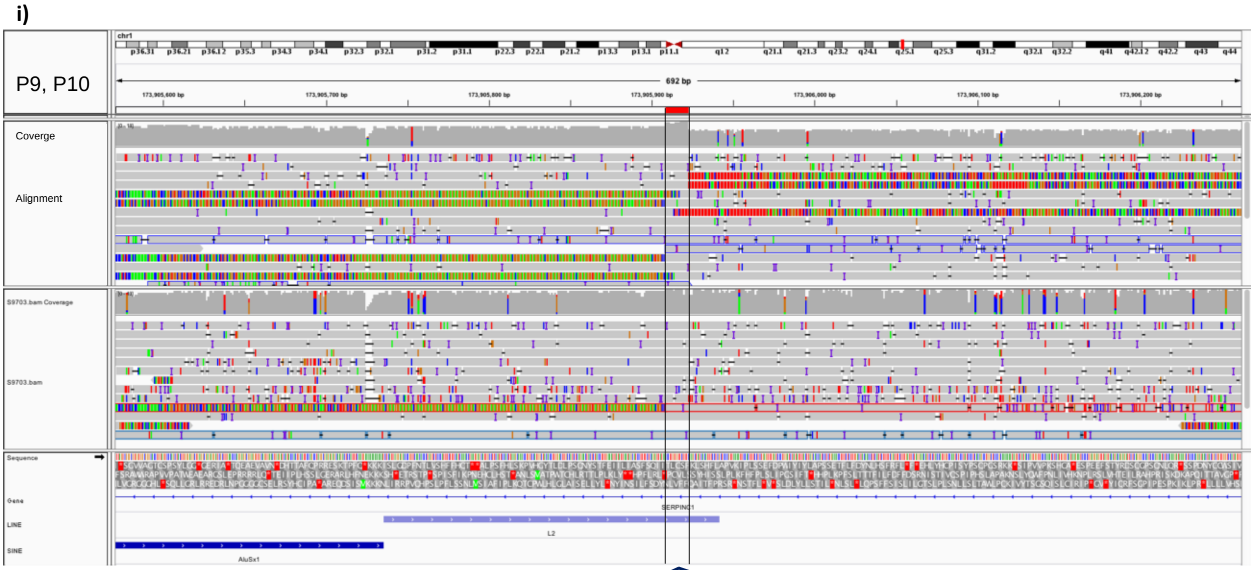


iii)

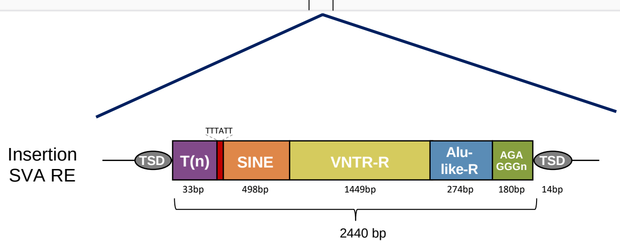
```
                                                                   AluSx1
1-173908168   GGCAACATTTTTAGTTTTCAAAACTAGCTTAAAGACAGCCCAAT AGCAC*AGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGC CAAGGCAGGTAGAT   173908268
J1-1          GGCAACATTTTTAGTTTTCAAA-CTAGCTTAAAGACAGCCCAATAGCAC* AGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGC TGAGGCATGCAGAT   99
1-173915355   AGAAGTAAGAAGAAATGGATGGAGCAAAGGCAATATC TGGCCGGGTGCG*-GTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGC TGAGGCATGCAGAT   173915454
                                                                   AluSp
```

(I) P9 and P10

i)



ii)



iii)

Participant 9

P9 39-52: TSD
P9 53-84: (T)n
P9 85-90: TTTATT sequence
P9 91-589: SINE
P9 590-2038: VNTR-R
P9 2039-2312: Alu-like-R
P9 2313-2492: (AGAGGG)n
P9 2493-2506: TSD

```
1-173905870   TTGAAATTATATTAATAGCATCCTTTTCTCAGATTATAACCTTGTGTTCTTT---------------------------------------------- 173905926
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
P9-1          TTGAAATTATATTAATAGCATCCTTTTCTCAGATTATAACCTTGTGTTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATTGATCATTCT 100

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-101        TGGGTGTTCTGCAGAGGGATTTGGCAGGGTCATAGGACAATAGTGGGGGGAAGGTCAGCAGATAAACAAGTGAACAAAGGTCTCTGGTTTTCCTAGGCAG 200

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-201        AGGACCCTGCGGCCTTCCGCAGCGCTTGTGCCCCTGGGTACTTAAGATTAGGAGTGGTGATGACTCTCAACGAGCATGCTGCCCTCAAGCATCTGTTCAA 300

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-301        CAAAGCACATCTTGCACCGCCCTTAATCCATCTAACCCTGAGTGGACACAGCACATGTCTCAGAGAGCACAGGGTTGGGGATAAGGTCACAGATCAACAG 400

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-401        GATCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACTTCTATCCACACAGACCCAGCAACCATCCGATTTCTCAA 500

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-501        TTTTTCCCCACCCTTCCCGCCTTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCATCCTCAATGAGCCGCTGGGCACACCTCCCGACGGGCGTGGC 600

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-601        CGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGCGGCCCGGCAGAAGTGCCCCTCACCTCCCAGATGGGGCGGCTGGCCGGGCGGGGGCTGACCCCTCCGC 700

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-701        CCTCCCGGACGGGCGGCTGGCCAGGCAGAGGGCTCCTCACTTCCCAGTAGGGCGGCCCGGGCAGGCGCCCTCACCTCCTGGAGGGCGGCTGGCCGGGCGG 800

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-801        GGCTGTTCCCCCACCTCCCTCCCGGACGGGGCGGCTGGCGGCAGAGGTCCTCACTCCCGAGGCCGGCCGGGCAGAGGCGCCCTCACCTCCCGGACGGGGC 900

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-901        GGCCGGCCAGGCAGGGGCTGATCCCACCCTCCCGGACGGGGCGGCTGGCCGGGCGGGGCTGCTCCCCACCTCCTCCCGGACGGGGCGGCTGGCCGGGCAG 1000

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1001       AGGGGTCCTCACTTCCCATAGGGGCGGCCGGGAGGGCGCCTCACCTCCCGGACGGGCGGCTGGCCAGGCAGGGGGCTGATCCCCACCTCCCTCCCGGACA 1100

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1101       GGGCGGCTGGCCGGGCGGGGGCTGACCCCCACCTCCCTCCCGGACGGGCGGCTGGCCGGGCAGGGGGCTGACTCCTCTCCCTCCCGGACGAGGCGGCTGG 1200

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1201       CCGGGCGGGGCGTGCCCCCCACCTCCCTCCCGGATGGGGCGGCTGGCCAGGCGGGGGCTGACCCCCACCTCCCCTCCTGGGCGGGGCTGGCGGCCGGGCA 1300

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1301       GAGGGCTCCTCACTTCCCAGTAGGGGCGGCCGGGCAGAGAGCCCTCACCTCCCGGCCGGGCGGCTGGCCGACCCCCCCTCCCTCCCGGACGGGCGGCTGG 1400

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1401       CCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCCGGGCAGAGGAGCCCTCACCTCCCGGACAGGGCGGCTGGCCGGGCGGGGGGCTGACCCCACCT 1500

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1501       CCTCCCAGGACGGGTGGCTGCCGGGCGGAGACGCTCCTCACTTCCCAGACGGGTGGTTGCCAGACGGAGGGGCTCTCACTTCTCAGACGGGGCGGTTGCC 1600

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926

P9-1601       AGGCAGAGGGTTTCCTCACTTCTCAGACGGAGCGGCCGGGCAGAGACACTCCTCACCTCCAGACAGGGTTGCGGCCCAGCAGAGGCGCTCCTCACATCCC 1700

1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-1701       AGACAGGGCGGCGGGGCAGAGGTGCTCCCACATCTCAGACGATGGGCGGCCGGGCAGAGACGCTCCTCACTTCCTAGATGGGATGGCGGCGGGAAGAGGC 1800
```

```
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-1801       GCTCGCCTCCTAGATGGGATGGCGGCCGGGCAGAGAGGCCTCACTTTCCACACTGGGCAGCCAGGCAGAGGGCTCCTCATATCCCGGGACGATGGGTGGC 1900
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-1901       CAAGCAGAGACGCTCCTCACTTCCCAGACGGGGTGGCGGCCGGGCAGAGGCTGCAATCTCGGCTCTCCGGGAGGCCAAGGCAGGCGGCTGGGAGGTGGCT 2000
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-2001       GCGGAGCCGAGATCACGCCACTGCACTCCAGCCTGGGCACCATTGAGCACTGAGTGAACGGACTCCATCTGCAATCCGGCACCCCCGGGGAGGCCGAGGC 2100
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-2101       TGGCGGATCACTCGCGGCCAGGAGCTGGAGACCAGCCCGGCCAACACAGCGAAACCCCATCTCCACCAAAAAAACGAAAACCAGTCAGGCGTGGGCGGCG 2200
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-2201       CCTGCAATCGCAGGCACTCGGCAGGCTGAGGCAGGAGAATCAGGCAGGAGGTGCAGTGAGCGAGATGGCAGCAGTACCGTCCAGTTTTGGGCTCGGCATG 2300
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-2301       AGAGGGAGAGGGAGACGGGAGAGGGAGAGGGAGACGGAGAGGGAGAGGGAGACGGGAGAGGGAGGGGGGAGCGGGGAGAGGGAGAGGGGAGAGGGGGGCG 2400
1-173905926   ---------------------------------------------------------------------------------------------------- 173905926
P9-2401       GGGAGGGGGAGAGGGAGACGGGAGGCAGAGGGAGCGGGGGGAGGGTGAGGGGGGCGGGAGGGAGGAGGGAGACGGGAGAGGGAGACGGGGCCACCTTGTG 2500

1-173905926   ------CAAGCTATCACATTTCCTCGCTCCCGTTAAAATTCCACTTTCCTCTGA     173905970
                    ||||||||||||||| |||||  |||||||||||| ||||||||||||
P9-2501       TTCTTTCAAGCTATCACTTTTCCCTGCTCCCGTTAAA-TTCCACTTTCCTCTGA     2600
```

SUPLEMENTAL BIBLIOGRAPHY

1. de la Morena-Barrio, M.E., Martínez-Martínez, I., de Cos, C., Wypasek, E., Roldán, V., Undas, A., van Scherpenzeel, M., Lefeber, D.J., Toderici, M., Sevivas, T., et al. (2016). Hypoglycosylation is a common finding in antithrombin deficiency in the absence of a SERPINC1 gene defect. J. Thromb. Haemost. *14*, 1549–1560.
2. De la Morena-Barrio, B., Borràs, N., Rodríguez-Alén, A., Morena-Barrio, M.E., García-Hernández, J.L., Padilla, J., Bravo-Pérez, C., Miñano, A., Rollón, N., Corral, J., et al. (2019). Identification of the first large intronic deletion responsible of type I antithrombin deficiency not detected by routine molecular diagnostic methods. Br. J. Haematol. *186*, e82–e86.
3. Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.
4. Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods *15*, 461–468.
5. Cretu Stancu, M., Van Roosmalen, M.J., Renkens, I., Nieboer, M.M., Middelkamp, S., De Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat. Commun. *8*, 1–13.
6. Heller, D., and Vingron, M. (2019). SVIM: Structural variant identification using mapped long reads. Bioinformatics *35*, 2907–2915.
7. Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. *8*, 1–11.
8. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
9. Sanchis-Juan, A., Stephens, J., French, C.E., Gleadall, N., Mégy, K., Penkett, C., Shamardina, O., Stirrups, K., Delon, I., Dewhurst, E., et al. (2018). Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. *10*, 95.
10. Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., and Mesirov, J.P. (2017). Variant review with the integrative genomics viewer. Cancer Res. *77*, e31–e34.
11. Katoh, K., Misawa, K., Kuma, K.I., and Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.
12. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.
13. Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. *47*, W256–W259.

Participant	Bkp in SV	Segments	Window	Repeat coordinates	Repeat	Family	Start in query	End in query	Repeat length
P1	Deletion start	A3'-B5'	1:173916554-173916854	-					
	Deletion end	B3'-C5'	1:173935553-173935853	1:173935699-173935738	A-rich	Low_complexity	147	186	40
P2	Duplication start	A3'-B5'	1:173911229-173911529	1:173911329-173911536	MER103C	DNA/hAT-Charlie	101	300	201
	Duplication end	B3'-C5'	1:173912001-173912301	-					
	Deletion start	C3'-D5'	1:173914965-173915265	-					
	Deletion end	D3'-E5'	1:173918884-173919184	1:173918822-173919087	L2a	LINE/L2	1	204	204
P3	Deletion start	A3'-B5'	1:173879670-173879970	1:173879670-173879957	AluSx1	SINE/Alu	1	288	288
	Deletion end	B3'-C5'	1:173925839-173926139	1:173926103-173926323	AluJb	SINE/Alu	265	300	37
P4	Deletion start	A3'-B5'	1:173847697-173847997	1:173847804-173847844	HAL1	LINE/L1	108	148	41
				1:173847845-173848152	AluY	SINE/Alu	149	300	153
	Deletion end	B3'-C5'	1:174815997-174816297	1:174816148-174816445	AluSp	SINE/Alu	152	300	150
P5	Deletion start	A3'-B5'	1:173850846-173851146	1:173850956-173851264	AluSx1	SINE/Alu	111	300	191
				1:173850606-173850904	AluSx	SINE/Alu	1	59	59
	Deletion end	B3'-C5'	1:173950024-173950324	1:173949885-173950176	AluSx	SINE/Alu	1	153	153
				1:173950177-173950297	L1ME3C	LINE/L1	154	121	121
				1:173950298-173950593	AluJr	SINE/Alu	275	300	27
P6	Duplication start	A3'-B5'	1:173908262-173908562	1:173908214-173908512	AluSx1	SINE/Alu	1	251	251
				1:173908555-173908859	AluJb	SINE/Alu	294	300	8
	Duplication end	B3'-C5'	1:173919640-173919940	1:173919609-173919890	AluSz	SINE/Alu	1	251	251
				1:173919924-173920236	AluSx1	SINE/Alu	285	300	17
P7	Deletion start	A3'-B5'	1:173908184-173908484	1:173908214-173908512	AluSx1	SINE/Alu	31	300	271
	Deletion end	B3'-C5'	1:174102865-174103165	1:174102793-174102889	Tigger1	DNA/TcMar-Tigger	1	25	25
				1:174102890-174103191	AluSp	SINE/Alu	26	300	276
P8	Deletion start	A3'-B5'	1:173908068-173908368	1:173908214-173908512	AluSx1	SINE/Alu	147	300	155
	Deletion end	B3'-C5'	1:173915255-173915555	1:173915394-173915705	AluSz	SINE/Alu	140	300	162
P9	Insertion site		1:173905771-173906071	1:173905736-173905941	L2	LINE/L2	1	171	171
P10	Insertion site		1:173905771-173906071	1:173905736-173905941	L2	LINE/L2	1	171	171

Table S1. Repetitive elements at the SV breakpoints. Segments column match to those in Figure S5, and refers to the relative location in the segments where the SV breakpoint is, where 5' and 3' correspond to the upstream and downstream parts respectively. Window was done for +- 150 bp from the breakpoint. Start and End in query are the relative positions to the query sequence where the repeat starts and ends respectively. Bkp=Breakpoint.

Participant	New junction	Microhomology at breakpoint	Microhomology at breakpoint length	Deletion at breakpoint	Deletion at breakpoint length	Insertion at breakpoint	Insertion at breakpoint length	Duplication at breakpoint	Duplication at breakpoint length	Expected reference sequence	Observed sequence (Sanger)
P1	J1	-	-	-	-	GGAGCAAGGAAC	12	-	-	CCAAGACAGTTGTGCCACCACCCCGTTTCTAGCCTCCTGTGACAACTCA*AAGGAAAAGGAAAAAAGCAAAAAACAAAAAACAAAAAACATTTTGGGGGTGTTAG	CCAAGACAGTTGTGCCACCACCCCGTTTCTAGCCTCCTGTGACAACTCA*{GGAGCAAGGAAC}AAGGAAAAAAAGCAAAAAACAAAAAACAATTTTGGGGGTGTTAG
P2	J1	GTTGC	5	GTTGC	5	-	-	-	-	CCCGGGACAGGTTCAGTCCTAGACTTCTTGCCAGGGGACAGTTCAGTTGC*GTTGCAGAGTTTCTCAGAACCTTCATAGGCCCATGTGCTTCATGAATCTCC	CCCGGGACAGGTTCAGTCCTAGACTTCTTGCCAGGGGACAGTTCAGTTGC*-----AGAGTTTCTCAGAACCTTCATAGGCCCATGTGCTTCATGAATCTCC
	J2	AGC	3	AGC	3	-	-	-	-	AATAGCCACCTCAGTTGTTAACTCCTTCAAGCACAGTACCTGGGACAGC*AGCTGTCCAAAGAGAACTCAGAAATACAAAGACCTTGAGGTTTCAGGAACA	AATAGCCACCTCAGTTGTTAACTCCTTCAAGCACAGTACCTGGGACAGC*---TGTCCAAAGAGAAATCAGAAATACAAAGACCTTGAGGTTTCAGGAACA
P3	J1	TGGN	3	-	-	GCC	3	-	-	GTGAAACCCCATCTCTACTAAATATAAAACATTAGCCAGGCATGGTGGT*TGGACAAGGCTATTCTTTGGAGGCTGAGCAAGAGAGAGTTGTGTGGGGTCA	GTGAAACCCCATCTCTACTAAATATAAAACATTAGCCAGGCATGGTGGT{GCC}*TGGACAAGGCTATTCTTTGGAGGCTGAGCAAGAGAGAGTTGTGTGGGGTCA
P4	J1	TTTTTTTTTTTTTTTTTTTTTTGAGA	26	TTT	3	-	-	-	-	CCTCACCTAGTTACTCTGTACCTTCTGCACAGCTGTCTTTCTGAACCTTT*TTTTTTTTTTTTTTTTTTTTTTGAGATAGAGTTTTGCTCTTGTTGCCCAG	CCTCACCTAGTTACTCTGTACCTTCTGCACAGCTGTCTTTCTGAACC---*TTTTTTTTTTTTTTTTTTTTTTGAGATAGAGTTTTGCTCTTGTTGCCCAG
P5	J1	-	-	-	-	-	-	-	-	CTTAAGGGTTGGCCGGGCATGGTGGCTAACGCCTGTAATCCCAGCACTTT*AAAGTACACATCTTAAAAGCAATGCTAAGTGAAAACATGAAAATGCAAGCT	CTTAAGGGTTGGCCGGGCATGGTGGCTAACGCCTGTAATCCCAGCACTTT*AAGTACACATCTTAAAAGCAATGCTAAGTGAAAACATGAAAATGCAAGCT
P6	J1	GGAGGCTGAGGCAGGAGAATCACTTGAA	28	-	-	-	-	-	-	ACCTGTAATCTCAGCTACTCAGGAGGCTGAGGCAGGAGAATCACTTGAA*TCTGGGGGGTGGAGGTTGCAGTGAGCCAAGATTGCACCACTGCAC	GCCTGTAATCTCAGCTACTCAGGAGGCTGAGGCAGGAGAATCACTTGAA*TCTGGGGGGTGGAGGTTGCAGTGAGCCAAGATTGCACCACTGCAC
P7	J1	CAAAAATTAGCC	12	-	-	-	-	-	-	TTCAAGACCAGCCTGGCCAACATAGTGAAACCCCGTTTCTACTAAAAAT*ACAAAAATTAGCCAGGCATGGTGGTGAGCGCCTGTAATCCCAGCTACTCCA	TTCAAGACCAGCCTGGCCAACATAGTGAAACCCCGTTTCTACTAAAAAT*{A}CAAAAATTAGCCAGGCATGGTGGTGAGCGCCTGTAATCCCAGCTACTCCA
P8	J1	GTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGC	36	-	-	-	-	-	-	GGCAACATTTTTAGTTTTCAAAACTAGCTTAAAGACAGCCCAATAGCACA*GTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCTGAGGCATGCAGA	GGCAACATTTTTAGTTTTCAAACTAGCTTAAAGACAGCCCAATAGCACA*GTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCTGAGGCATGCAGA
P9	J1	-	-	-	-	-	-	ACCTTGTGTTCTTT	14	AAATTATATTAATAGCATCCTTTTCTCAGATTATAACCTTGTGTTCTTT*CAAGCTATCACATTTCCTCGCTCCCGTTAAAATTCCAC	AAATTATATTAATAGCATCCTTTTCTCAGATTATAACCTTGTGTTCTTT*{INS_SVA}**ACCTTGTGTTCTTT**CAAGCTATCACTTTTCCCTGCTCCCGTTAAA-TTCCACTTTCCTCTG
P10	J1	-	-	-	-	-	-	ACCTTGTGTTCTTT	14	AAATTATATTAATAGCATCCTTTTCTCAGATTATAACCTTGTGTTCTTT*CAAGCTATCACATTTCCTCGCTCCCGTTAAAATTCCAC	AAATTATATTAATAGCATCCTTTTCTCAGATTATAACCTTGTGTTCTTT*{INS_SVA}**ACCTTGTGTTCTTT**CAAGCTATCACTTTTCCCTGCTCCCGTTAAA-TTCCACTTTCCTCTG

Table S2. Structural variants breakpoint details. New junction numbers correspond to Figure S5. Expected reference sequence refers to the putative derivative sequence as shown in Figure S5. Microhomology sequences (from 3bp to 36bp) at breakpoints are highlighted in red, inserted sequences are in blue and duplicated fragments are in bold. Breakpoint is marked with *.

What is known about this topic.

Antithrombin deficiency is mainly caused by SNV, small indels, and structural variants in *SERPINC1*, usually identified by sequencing and MLPA. Up to 25% of cases had an unknown molecular base. Nanopore sequencing is an emerging 4th generation sequencing method that obtains long reads, which are ideal for identification and characterization of gross gene defects.

What does this paper add.

Long-read whole-genome nanopore sequencing resolved all types and sizes of structural variants causing antithrombin deficiency, and identified the first causal complex structural variant. This method also found a new causing mechanism: the insertion of a new SVA retrotransposon in 2 out of 11 unknown cases. This result enlarges the catalogue of genetic disorders caused by retrotransposon insertions.