Long-Read Sequencing Identifies the First Retrotransposon Insertion and Resolves Structural Variants Causing Antithrombin Deficiency

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[8,9] Marja Puurunen[10] Anetta Undas[11] Francisco Vidal[12,13,14] Frances Lucy Raymond[3,15] Vicente Vicente[1] Willem H. Ouwehand[2,3] Javier Corral[1] Alba Sanchis-Juan[2,3] NIHR BioResource

[1] Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Universidad de Murcia, Murcia, Spain
[2] Department of Haematology, NHS Blood and Transplant Centre, University of Cambridge, Cambridge, United Kingdom
[3] NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom
[4] National Health Service Blood and Transplant (NHSBT), Cambridge Biomedical Campus, Cambridge, United Kingdom
[5] BHF Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, United Kingdom
[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
[9] C2VN, INRAE, INSERM, Aix-Marseille Université, Marseille, France
[10] The Framingham Heart Study, National Heart, Lung and Blood Institute, Framingham, Massachusetts, United States
[11] Department of Experimental Cardiac Surgery, Anesthesiology and Cardiology, Institute of Cardiology, Jagiellonian University Medical College and John Paul II Hospital, Kraków, Poland
[12] Banc de Sang i Teixits, Barcelona, Spain
[13] Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona (VHIR-UAB), Barcelona, Spain
[14] CIBER de Enfermedades Cardiovasculares, Madrid, Spain
[15] Department of Medical Genetics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom

Address for correspondence Alba Sanchis-Juan, University of Cambridge, Department of Haematology, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, United Kingdom (e-mail: as2635@cam.ac.uk).

Javier Corral, University of Murcia, Centro Regional de Hemodonación, Calle Ronda de Garay s/n, Murcia 30003, Spain (e-mail: javiercorraldelacalle@gmail.com).

Thromb Haemost 2022;122:1369–1378.

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, who 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 eight SVs detected by multiple ligation-dependent probe amplification, and identified

Keywords
- long-read sequencing
- antithrombin deficiency
- structural variants
- SVA retrotransposon

received
September 16, 2021
accepted after revision
January 10, 2022
published online
June 28, 2022

DOI https://doi.org/10.1055/s-0042-1749345.
ISSN 0340-6245.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

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 2.4 kb SINE-VNTR-Alu retroelement, which was characterized by de novo assembly and verified by specific polymerase chain reaction 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.

Introduction

Antithrombin deficiency is the most severe congenital thrombophilia first 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 (odds ratio: 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, while 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 considers the full spectrum of SVs and does not provide nucleotide-level resolution, which is important for confirming causality and reveal insights into SV formation.[7,9,10] These limitations may now be addressed by long-reads, which 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 (REs) in and around *SERPINC1* (where 35% of sequence are interspersed repeats),[15] which hinders SV 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 polymerase chain reaction [PCR]) were also selected ($N = 8$) (►**Table 1**). Detailed information of these procedures is shown in Supplementary Methods (►**Supplementary Material** [available in the online version]). Measurements of antithrombin levels and function were performed for all participants as previously described.[16,17]

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, which were then merged per sample.

Table 1 Cohort of individuals included in this study—demographic, antithrombin values, and genetic results

Participant	Antithrombin		Family history	Gender	MLPA *SERPINC1*	PGM	CGHa	LR-PCR and Illumina sequencing	WGS ONT	Algorithm	Genotype	Coordinates	Length (bp)
	Anti-FXa%	Ag (%)											
P1	30	30	Yes	M	Deletion exon 1	–	Negative	Deletion exon 1	Deletion exon 1	Nanosv; sniffles; svim	Het	1:173916704–173935703	18,999
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	3,737; 6,884
P3	44	41	Yes	F	Complete deletion	–	Deletion 2 genes	–	Deletion 2 genes	Nanosv; sniffles	Het	1:173879820–173925989	46,169
P4	45	38	No	M	Complete deletion	–	Deletion 20 genes	–	Deletion 20 gene	Nanosv; sniffles	Het	1:173847847–174816147	968,005
P5	36	50	Yes	F	Complete deletion	–	–	–	Deletion 5 genes	Nanosv	Het	1:173850996–173950174	99,178
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	11,404
P7	45	38	No	M	Deletion exons 1–5	–	Deletion exons 1–5 + 1 gene	–	Deletion 2 genes	Nanosv; sniffles	Het	1:173908334–174103015	194,389
P8	52	37	Yes	F	Deletion exons 2–5	–	–	–	Deletion exons 2–5	Nanosv; sniffles	Het	1:173908218–173915405	7,187
P9	56	61	Yes	F	Negative	Negative	–	Negative	Insertion SVA	Nanosv	Het	1:173905922	2,440
P10	50	46	Yes	F	Negative	Negative	–	Negative	Insertion SVA	Visual inspection	Het	1:173905922	2,440
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				

Abbreviations: Ag, antigen; bp, base pair; Het, heterozygous.
Note: *SERPINC1* gene-driven tests include MLPA, PGM sequencing (Ion Torrent) and long-range PCR (LR-PCR) amplification, and Miseq 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.

Data Processing and SV Identification

We used the Snakemake library to develop an *in-house* multimodal 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 ►**Fig. 1A**. Detailed information is provided in Supplementary Methods (►**Supplementary Material** [available in the online version]).

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 sub-elements.

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 4,499 bp and median genome coverage of 16× (►**Fig. 1B**). After a detailed quality-control analysis

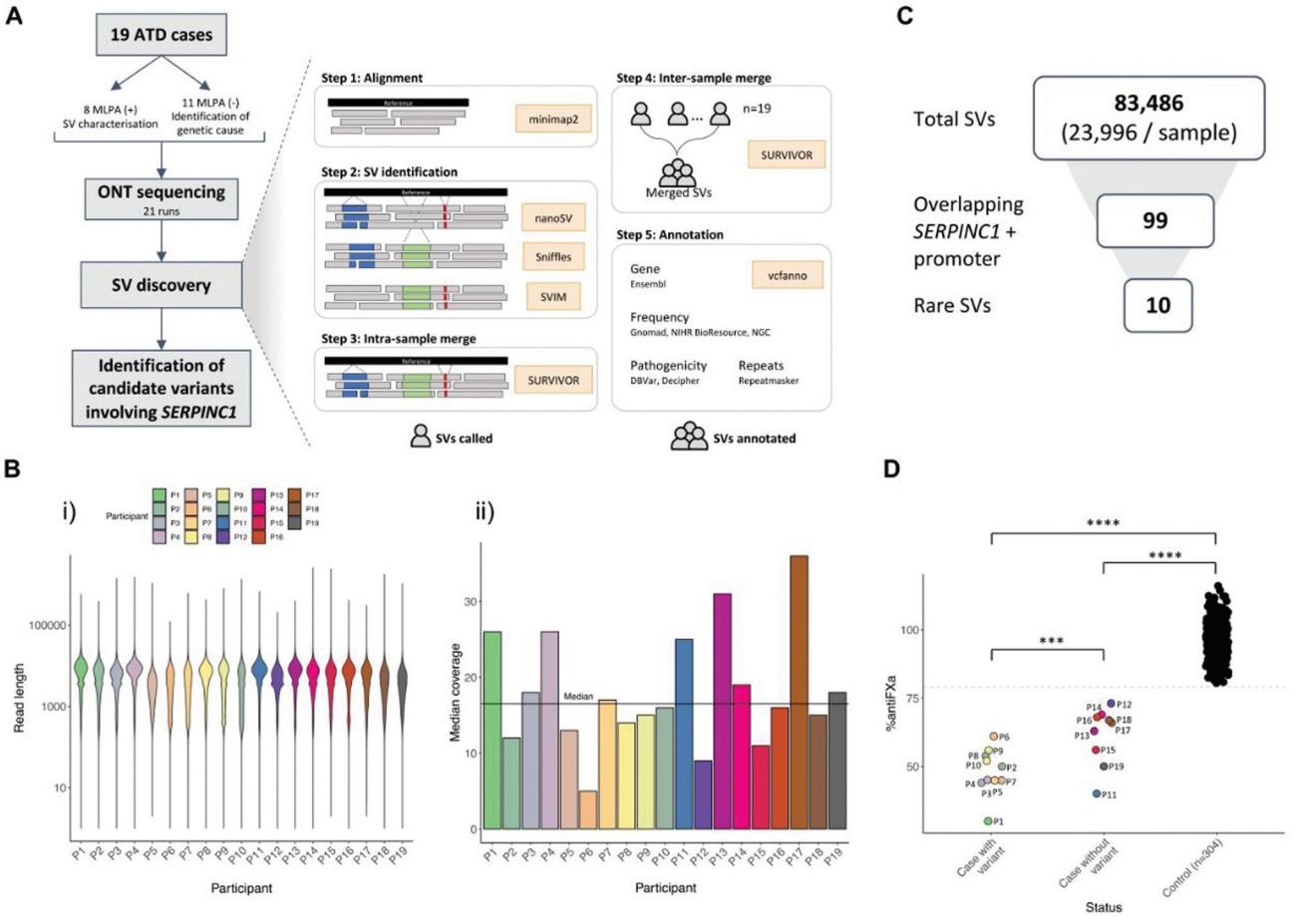

Fig. 1 Long-read sequencing workflow and results. (**A**) Overview of the general stages of the SVs discovery workflow. Algorithms used are depicted in yellow boxes. (**B**) Nanopore sequencing results. (i) Sequence length template distribution. Average read length was 4,499 bp (SD ± 4,268); the maximum read length observed was 2.5 Mb. (ii) Genome median coverage per participant. The average across all samples was 16× (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 $^{***}p < 0.001$, $^{****}p \leq 0.0001$. *p*-Values calculated by one-way ANOVA with Tukey's posthoc test for repeated measures. ATD, antithrombin deficiency; ONT, Oxford Nanopore Technologies; SV, structural variant.

(►**Fig. S1**, available in the online version), 83,486 SVs were identified, consistent with previous reports using LR-WGS (►**Fig. S2**, available in the online version).[11] Focusing on rare variants (allele count $\leq$ 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 (►**Fig. 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 eight individuals (P1–P8). SVs were identified independently of their size (from 7 to 968 kb, restricted to *SERPINC1* or involving neighboring genes) and their type (six deletions, one tandem duplication, and one complex SV) (►**Fig. 2** and ►**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 (►**Fig. 3**). Specific PCR amplification and Sanger sequencing validated this complex SV 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 (►**Fig. 4**). The tandem duplication of exons 1 to 5 was observed to be present in the affected son of P6, also with antithrombin deficiency.

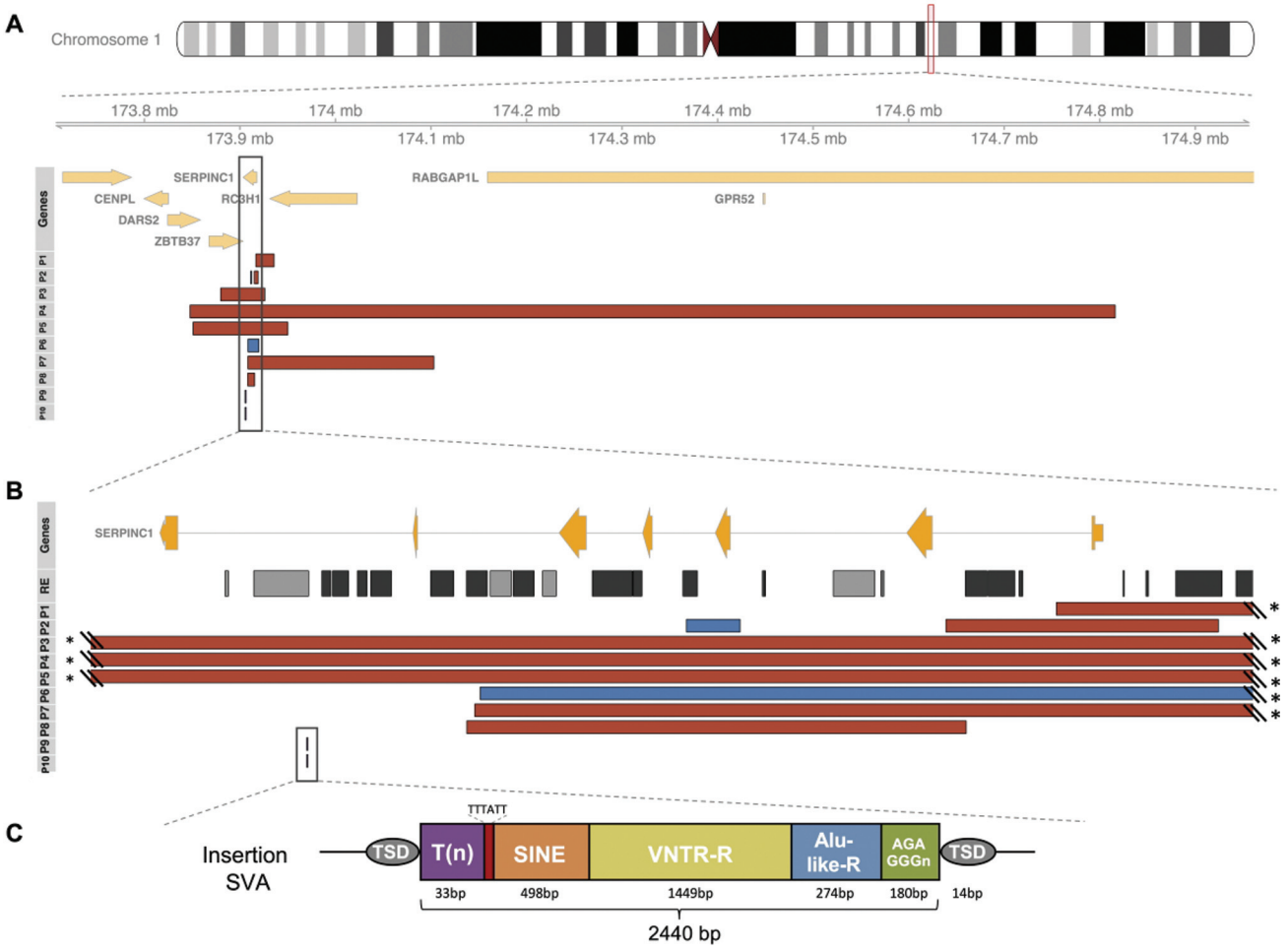

Fig. 2 Candidate SVs identified by long-read sequencing. (**A**) Schematic of chromosome 1 followed by protein coding genes falling in the zoomed region (1q25.1). SVs for each participant (P) are colored in *red* (deletions) and *blue* (duplications). The insertion identified in P9 and P10 is shown with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

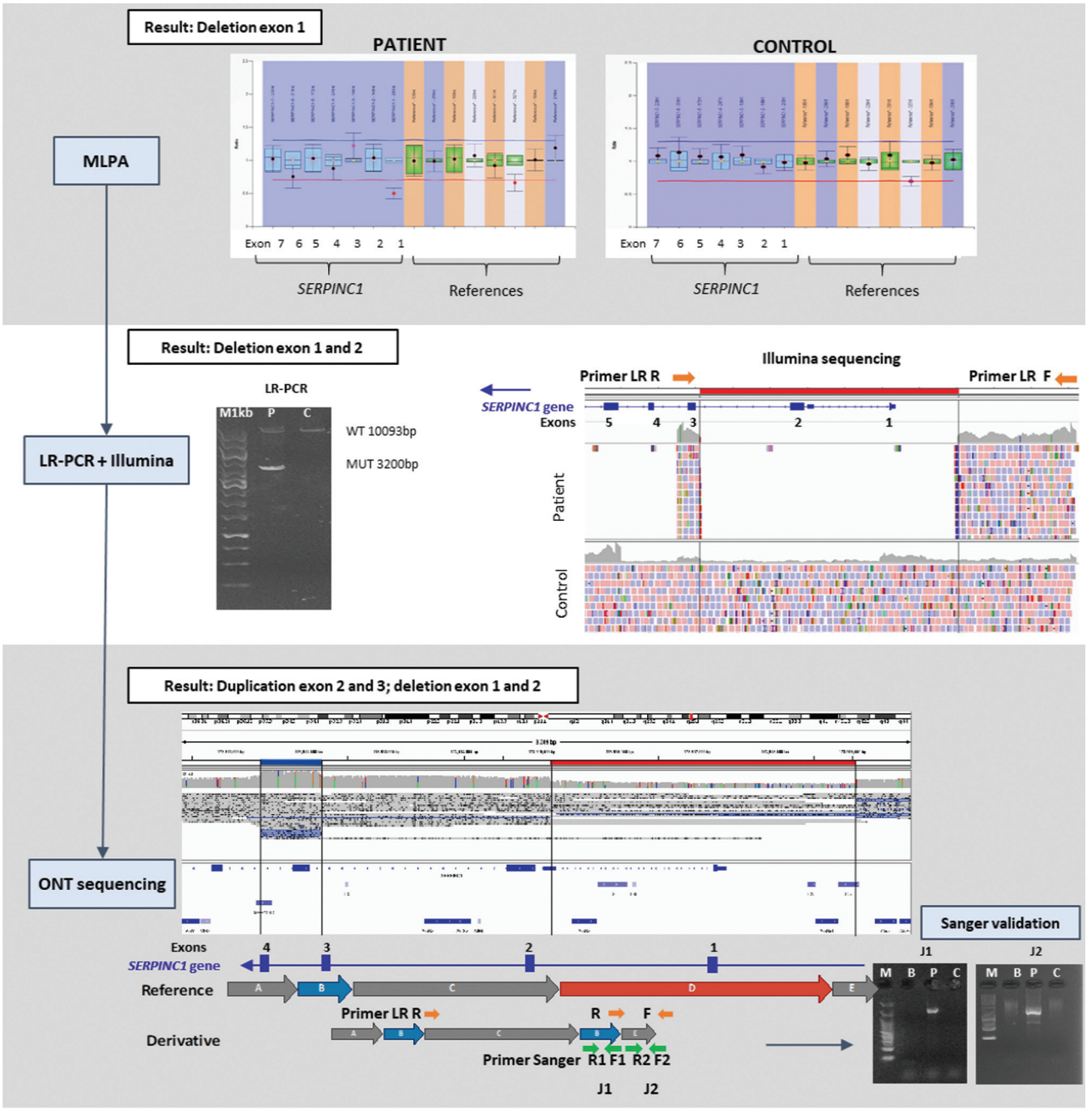

Fig. 3 Resolution of a complex SV. Schematic representation of genetic diagnostic methods used to characterize the SVs in participant P2. Results from MLPA, LR-PCR, and nanopore are shown in *white boxes*. Primers used for both LR-PCR and Sanger validation experiments are shown representing the genetic 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 newly formed junctions described in ►**Fig. S5**. J = new junction; M1k = 1 kb molecular weight marker; M = 100 bp molecular weight marker; P = patient; C = control; B = blank. LR-PCR, long-range polymerase chain reaction; MPLA, multiple ligation-dependent probe amplification; SV, structural variant.

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 2,440 bp in intron 6. Blast analysis of the inserted sequence revealed a new SINE-VNTR-Alu retroelement (►**Fig. 2** and ►**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 14 bp (►**Fig. 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 (►**Fig. S3A**, available in the online version) and it was observed to be closest to the SINE-VNTR-Alu E in the phylogenetic tree

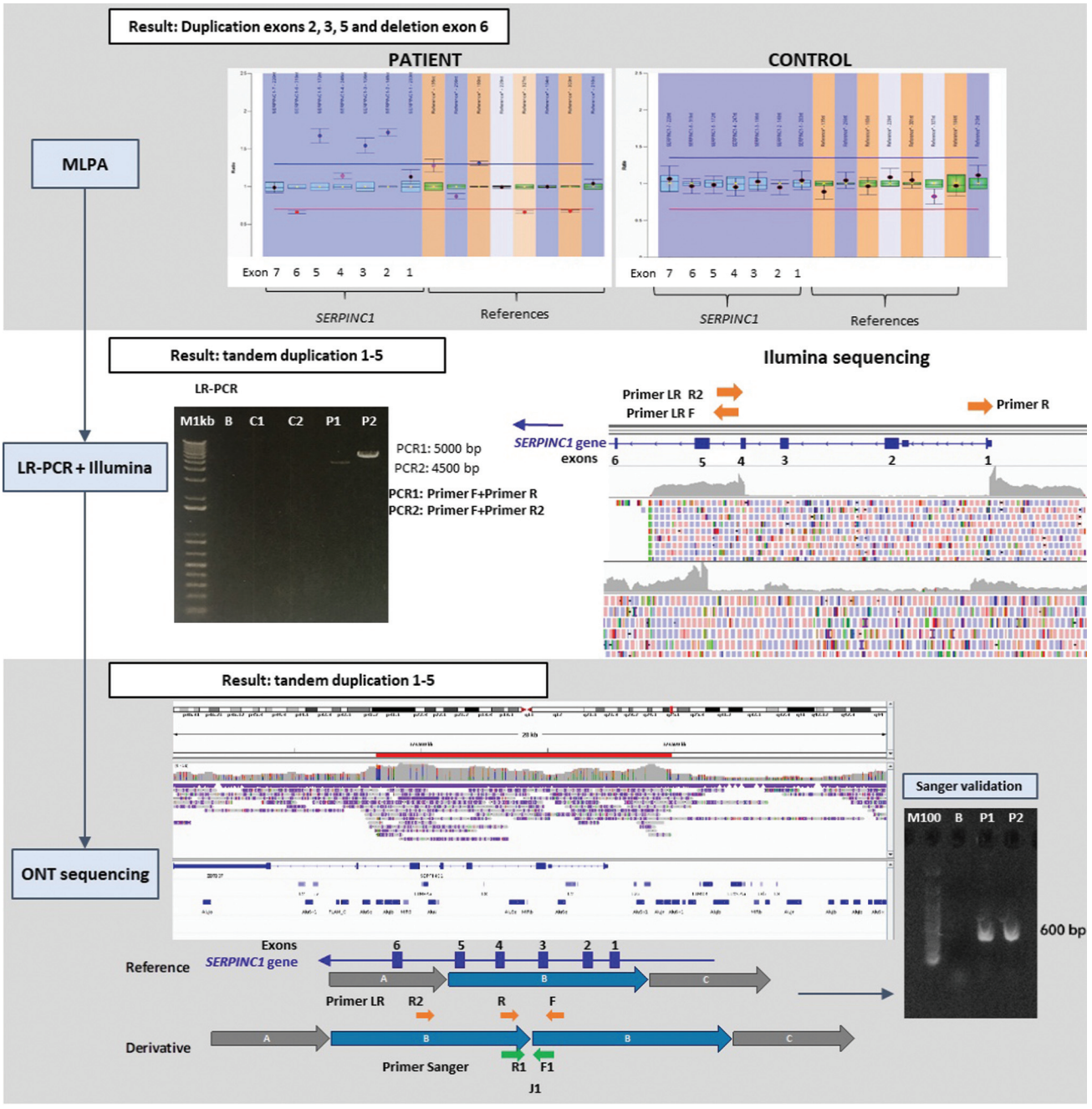

Fig. 4 Schematic representation of genetic diagnostic methods used to characterize the SVs in participant P6. Results from MLPA, LR-PCR, and nanopore are shown in *white boxes*. Primers used for both LR-PCR and Sanger validation experiments are shown representing the genetic location of each one with *orange* and *green arrows*, respectively. *SERPINC1* gene in the IGV screenshot is represented in *blue* and exons are indicated. J1 corresponds to the newly formed junctions described in ►**Fig. S5**. J = new junction; M = molecular weight marker 1 kb or 100 b; 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). LR-PCR, long-range polymerase chain reaction; MPLA, multiple ligation-dependent probe amplification; SV, structural variant.

(►**Fig. S3B**, available in the online version). Moreover, the VNTR sub-element harbored 1,449 bp, which was longer than the typical approximately 520 bp-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, whose design was facilitated by the nanopore data, was able to amplify the breakpoint (►**Fig. S4**, available in the online version). This method was used to confirm the insertion in P9 and 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 (►**Fig. S4**, available in the online version).

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 newly formed junctions, demonstrating a 100% accuracy in 7/10 (70%) SVs called. RE were detected in all the SVs, with Alu elements being the most frequent (16/24, 67%) (►**Table S1**, available in the online version). Additionally, breakpoint analysis identified microhomologies (7/11, 64%) and insertions, deletions, or duplications (7/11, 64%) (►**Fig. S5** and ►**Table S2**, available in the online version). Importantly, we observed a nonrandom 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 (►**Fig. 2B** and ►**Table S1** [available in the online version]).

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 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 2.4 kb 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 REs 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 nonrandom 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 three cases (P6, P7, P8), highlighting the important role that RE plays in the formation of nonrecurrent, but nonrandom, SVs. These results highlight that *SERPINC1* might be a hotspot for SVs given the high number of REs in this gene and show

how LR-WGS can be used to investigate and resolve events occurring in repetitive genes and regions.

In total, nine 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 has significantly higher anti-FXa activity than those with SVs (►**Fig. 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 counseling, 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.

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 fourth-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 disease-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.

Author Contributions

B.M.-B., W.H.O., J.C., and A.S.-J. designed the study. M.M.B., L.S., J.P., A.M., N.G., F.L.R., and V.V. helped with the study design. B.M.-B., M.M.B., J.P., and A.M. performed laboratory experiments and analyzed the experimental data. J.S. performed sample preparation and executed long-read sequencing. A.S.-J. developed the analysis workflow for long-read sequencing, applied this to data processing, and performed the computational and statistical analyses. B. M.-B. performed computational analyses and variant validation. J.J.L.C. and F.V. provided valuable insight into microarray and NGS data analysis. A.U., M.F., M.P., and P. M. recruited participants and collected the clinical data and samples. B.M.-B., W.H.O., J.C., and A.S.-J. wrote the manuscript. All authors read and approved the final version of the manuscript.

Data and Code Availability

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

Patient Consent

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

Ethical Approval

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 to the principles of the Declaration of Helsinki and their later amendments.

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 (Instituto de Salud Carlos III, FEDER & Next Generation and the 21642/PDC/21 project (Fundación Séneca).

Conflict of Interest

None declared.

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 Program. 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. L.S. received support from the British Heart Foundation. (RE/18/1/34212) The laboratory of WHO is supported by grants from the Addenbrooke's Charitable Trust, International Society on Thrombosis and Haemostasis, Medical Research Council, the National Institute for Health Research England, NHS Blood and Transplant and Thermo Fisher Scientific.

References

1 Egeberg O. Thrombophilia caused by inheritable deficiency of blood antithrombin. Scand J Clin Lab Invest 1965;17:92

2 Corral J, de la Morena-Barrio ME, Vicente V. The genetics of antithrombin. Thromb Res 2018;169:23–29
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
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(06):1193–1200
5 Bravo-Pérez C, Vicente V, Corral J. Management of antithrombin deficiency: an update for clinicians. Expert Rev Hematol 2019;12 (06):397–405
6 Stenson PD, Ball EV, 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(02):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(05):1015–1033
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(05):715–721
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(01):47–55
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(01):95
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. Nat Genet 2021;53(06):779–786
12 Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 2018;15(06):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(01):1326
14 French CE, Delon I, Dolling H, et al; NIHR BioResource—Rare Disease Next Generation Children Project. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. Intensive Care Med 2019;45(05):627–636
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):e1002384
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(05):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(08):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(07):1178–1187
19 Li H, Handsaker B, Wysoker A, et al; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–2079
20 Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020;17(02):155–158
21 Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–3100
22 Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med 2010;61(01): 437–455
23 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(06): 841–842
24 Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 2018;34(20):3600
25 Turro E, Astle WJ, Megy K, et al; NIHR BioResource for the 100,000 Genomes Project. Whole-genome sequencing of patients with rare diseases in a national health system. Nature 2020;583 (7814):96–102
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 nonrecurrent breakpoints. Genome Biol 2014;15(06):R80
27 Payer LM, Burns KH. Transposable elements in human genetic disease. Nat Rev Genet 2019;20(12):760–772
28 Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. Annu Rev Genet 2012;46:651–675
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(01):134–139
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(07):1051–1055
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(05):897. e21–909.e21
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
33 Hancks DC, Kazazian HH Jr. Roles for retrotransposon insertions in human disease. Mob DNA 2016;7(01):9
34 Kazazian HH Jr, Moran JV. Mobile DNA in health and disease. N Engl J Med 2017;377(04):361–370
35 Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 2016;17 (04):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(02):200–205
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(01):45–53