

Hybridization capture long-read sequencing and de novo assembly of homologous haplotypes: a comprehensive hemophilia test

Boyan Liu,^{1,*} Ruixia Xu,^{2,*} Siqian Ma,^{1,*} Mengnan Gu,² Lingyin Kong,^{2,3} Lu Zhou,⁴ Haoning Liu,¹ Shujin Chen,¹ Yuyan Yang,² Ziqiang Yu,^{5,6} Bo Liang,^{2,7} and Miao Jiang^{1,5,6}

¹Hematology Department, The Fourth Affiliated Hospital of Soochow University, Suzhou, China; ²Basecare Medical Device Co, Ltd, Suzhou, China; ³School of Food and Biology Engineering, Jiangsu University, Zhenjiang, China; ⁴Hematology Department, Affiliated Hospital of Nantong University, Nantong, China; ⁵National Clinical Medical Research Center of Blood Diseases, Jiangsu Institute of Hematology, The First Affiliated Hospital of Soochow University, Suzhou, China; ⁶Key Laboratory of Thrombosis and Hemostasis of National Health Commission of People's Republic of China, Suzhou, China; and ⁷State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Key Points

- A comprehensive genetic testing program was developed using PacBio LRS for hemophilia diagnosis.
- Complex structure variants were successfully detected, improving diagnostic accuracy over traditional methods.

Hemophilia is an X-linked bleeding disorder caused by defects in the *F8* or *F9* genes. Given the wide variety of *F8* variants, conventional genetic testing typically requires a combination of multiple methods, and detecting rearrangements in the intron 22 homologous regions (int22hs) remains a challenging task. In this study, we developed a comprehensive hemophilia testing program using the PacBio long-read sequencing (LRS) platform. Experimentally, we established a standard operating procedure for hybridization capture LRS (hc-LRS), which generates reads longer than 5 kilobases. Analytically, we used a suite of bioinformatics tools to identify variants associated with hemophilia, including the detection of int22h-related rearrangements through de novo assembly of homologous haplotypes. Our approach successfully identified pathogenic variants in patients with, and carrier of, hemophilia, encompassing both single-nucleotide variants and structural variations, with full concordance to validated methods. Moreover, the program identified complex int22h rearrangements in several samples, which were previously difficult to detect using traditional techniques. Compared with conventional methods, hc-LRS is more cost-effective, convenient, and capable of detecting various variants in a single test. This approach provides a powerful tool for the genetic diagnosis of hemophilia, particularly in patients with unknown genetic backgrounds or complex variants. In conclusion, our comprehensive testing program represents a significant advancement in the genetic diagnosis of hemophilia.

Introduction

Hemophilia comprises a group of common X-linked recessive disorders. Deficiencies in coagulation factor VIII (FVIII) and FIX are classified as hemophilia A and hemophilia B, respectively. Together with von Willebrand disease, they account for 95% to 97% of all inherited coagulation factor deficiencies.¹ Patients with hemophilia typically exhibit a lifelong tendency toward hemorrhaging, which can lead to severe or even life-threatening bleeding during trauma, surgical procedures, or childbirth.²

Submitted 12 February 2025; accepted 10 June 2025; prepublished online on *Blood Advances* First Edition 15 July 2025. <https://doi.org/10.1182/bloodadvances.2025016200>.

*B. Liu, R.X., and S.M. contributed equally to this study.

Data are available on request from the corresponding authors, Miao Jiang (jiangmiao@suda.edu.cn) and Bo Liang (boliang880@alumni.sjtu.edu.cn).

The full-text version of this article contains a data supplement.

© 2025 American Society of Hematology. Published by Elsevier Inc. Licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), permitting only noncommercial, nonderivative use with attribution. All other rights reserved.

Numerous genetic variants associated with hemophilia have been identified. The *F8* gene, responsible for encoding coagulation FVIII, spans 186 kilobases (kb) and has a wide variety of variant types. According to the European Association for Haemophilia and Allied Disorders database, >3000 variants in *F8* have been documented to cause hemophilia A.^{3,4} Single-nucleotide variants (SNVs) and indels have been identified across all 26 exons, splice regions, and untranslated regions of *F8*.³⁻⁵ In addition, rearrangements between intron 22 homologous region 1 (int22h-1), located within intron 22 of *F8*, and its homologous sequences, int22h-2 and int22h-3, result in intron 22 inversions (Inv22; Inv22 type I and Inv22 type II), accounting for ~40% of severe hemophilia A cases.^{4,6} Other int22h-related variants include deletions, duplications, and complex rearrangements.⁷⁻¹² Additional types of *F8* variants include Inv1 and large structural variants (SVs).⁴ In hemophilia B, most pathogenic variants are SNVs in *F9*, although SVs have also been observed.¹³

The accurate genetic diagnosis of hemophilia remains a significant challenge. Conventional genetic testing for hemophilia A involves a series of steps, starting with long-range polymerase chain reaction (LR-PCR) or inverse-PCR (IS-PCR) to detect int22h-related rearrangements and Inv1,¹⁴⁻¹⁶ followed by next-generation sequencing (NGS) to identify SNVs and indels,¹⁷ and concluding with additional methodologies to assess for large deletions and duplications such as multiplexed ligation-dependent probe amplification, quantitative real-time PCR or array-based comparative genomic hybridization.¹⁸ However, this series of testing procedures is costly, labor-intensive, and inefficient, resulting in low rates of genetic testing for hemophilia.

In recent years, gene therapy has emerged as a promising approach for treating hemophilia.¹⁹⁻²¹ The advent of preimplantation genetic testing (PGT) has led to a paradigm shift in the management of inherited diseases, shifting strategies from treatment toward prevention.²²⁻²⁴ However, the successful implementation of these techniques depends on an accurate diagnosis of genetic defects.

Long-read sequencing (LRS), exemplified by the PacBio platform, has the potential to identify variants with high precision. The technology behind this platform is single-molecule real-time (SMRT) sequencing,²⁵ which is capable of detecting large SVs. High fidelity (HiFi) reads from circular sequencing enable the detection of SNVs with an accuracy of 99.9%,²⁶ comparable with that of NGS.

LRS has typically been associated with high costs. To reduce costs and increase throughput, pre-enrichment of target regions is essential. Although targeted sequencing is well-established in short-read sequencing platforms, the targeted enrichment of long DNA fragments still presents significant challenges. Current methods include hybridization capture, PCR enrichment, and CRISPR associated protein (Cas) mediated enrichment. Hybridization capture with nucleic acid probes is a simple and efficient method, but it is limited by the length of the final library (typically <5 kb).²⁷⁻²⁹ Given that the int22h regions are 9.5 kb in length, conventional bioinformatics tools may produce erroneous results when processing such short fragments of highly homologous DNA. The adaptation of PCR enrichment to LRS has been facilitated by the commercial availability of long-range DNA polymerases.³⁰ However, multiplex PCR must be compatible with each

reaction condition. An increase in primer types or product length may introduce additional bias, making reaction conditions more challenging.³¹ It is generally preferable to divide the same sample into multiple reaction systems.³² Given that PCR primers are fixed to hot spot mutation regions, such as exons, this approach may overlook opportunities to discover new mutation sites.^{33,34} In addition, the CRISPR-Cas9 system, a prominent tool in genome editing, can be used to obtain targeted DNA fragments that retain epigenetic information.³⁵⁻³⁷ However, it currently faces limitations in efficiency and multiplexing capabilities.

To address the challenges of genetic testing for hemophilia and the limitations of long-read targeted sequencing, we have developed a comprehensive testing program using the PacBio LRS platform. This program combines hybridization capture LRS (hc-LRS) to generate sequencing reads of >5 kb, and a set of bioinformatics tools for analyzing various variant types, including the identification of int22h-related rearrangements through the de novo assembly of homologous haplotypes (DAHh). The program has been validated through the analysis of 18 samples from 14 families with hemophilia. Furthermore, we have investigated cases involving complex variants, and novel variants were identified.

Methods

Samples

The testing samples consisted of 18 peripheral blood samples collected from patients or their relatives, representing 14 hemophilic family lines (12 with hemophilia A and 2 with hemophilia B). The clinical diagnosis for all patients was confirmed through standardized coagulation activity tests for FVIII or FIX. These samples were archived at The Fourth Affiliated Hospital of Soochow University between 2021 and 2024.

All procedures adhered to the ethical principles outlined in the 1964 Declaration of Helsinki and its subsequent amendments, with approval granted by the institutional review board of The Fourth Affiliated Hospital of Soochow University. Informed consent was obtained from each participant or their legal guardian before inclusion in the study.

Probes for targeted enrichment

The target regions include the entire *F8*, *F9*, and *VWF* genes, 3 homologous regions (int22h-2, int22h-3, and int1h-1), and *SRY* on the Y chromosome, which is used to identify the sex of the sample. A total of 979 probes, targeting 383 kb of loci of interest and 28 kb of flanking regions, were designed and manufactured by iGeneTech, China. A comprehensive list of all probes is provided in supplemental Methods.

hc-LRS

We established a standard operating procedure for hc-LRS. A detailed description can be found in the supplemental Methods. Briefly, the first step is the extraction of genomic DNA from the blood. Next, Tn5 transposase randomly cleaves double-stranded DNA while inserting primer-binding sequences at both ends, and the resulting DNA fragments are then ready for PCR amplification. The subsequent step involves precapture amplification using long-range DNA polymerases, followed by size selection using beads to remove DNA fragments of <5 kb from the PCR products.

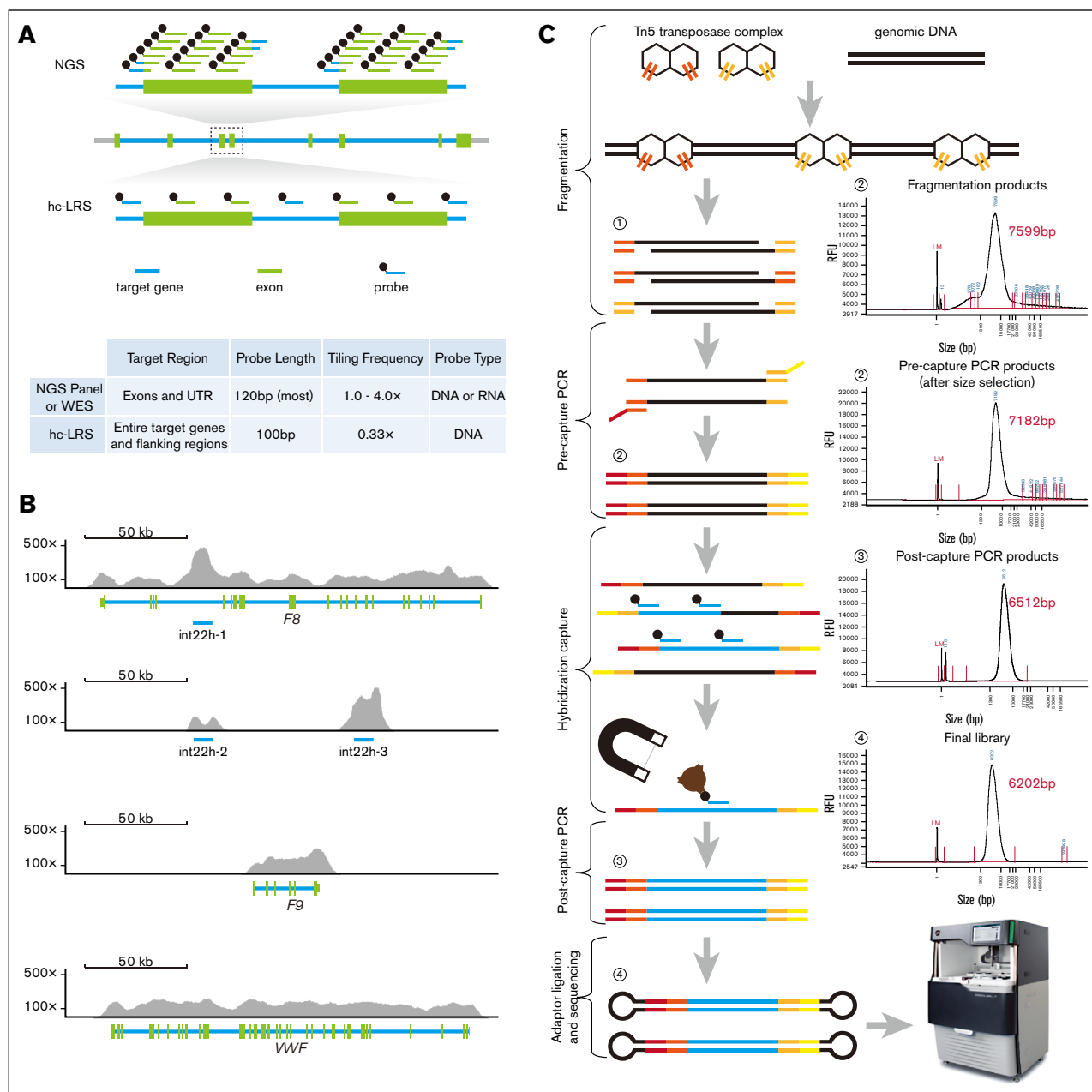


Figure 1. Probe design and experimental procedure for hc-LRS. (A) Comparison of the probes used for hc-LRS in this study with those used for NGS. (B) Sequencing depth and coverage of the target regions. (C) The experimental process from genomic DNA to the final sequencing library involves the following steps: fragmentation, precapture PCR, hybridization capture, postcapture PCR, and adapter ligation. The figure illustrates the results of capillary electrophoresis for the fragmentation products, precapture PCR products, postcapture PCR products, and the final library, showing a trend of decreasing fragment length throughout the experiment. The red and orange bands indicate the fragments introduced by Tn5 transposase, the dark red/red and yellow/orange bands represent the primers for precapture amplification, the blue bands indicate the targeted regions, and the black bands represent the hairpin adapter. UTR, untranslated regions; WES, whole-exome sequencing; RFU, relative fluorescence unit.

Biotinylated oligonucleotide probes are then hybridized to the target DNA and captured by streptavidin magnetic beads. This is followed by a rinse to remove uncaptured DNA. The postcapture amplification system is then constructed to amplify the target DNA fragments. The amplification product is purified and ligated to hairpin adapters, creating the final library, which is ready for sequencing on the PacBio Sequel II sequencing system.

Data analysis process

This study did not develop any unique code or algorithms but used a combination of published bioinformatics tools adapted for analyzing various types of genetic variants. The data analysis process consisted of several steps, including tag removal, sample splitting, alignment with reference sequences, analysis of SNVs/indels and SVs, de novo assembly, and visualization. The PacBio Sequel II

sequencing system generates HiFi reads with a base accuracy of Q30, eliminating the need for additional quality control.

Initially, sequence tags were excised from the raw sequencing data using the Demultiplex barcodes module of SMRT Link (version 13.1.0.221970). The sequence data were then divided into sample-specific files in Fastq format. Next, the split sequence data files were aligned to the reference genome (hg38) using Minimap2 (version 2.17-r974-dirty) and indexed.³⁸ SNVs and indels were identified using DeepVariant (version 1.6.0),³⁹ whereas structural variations and copy number variations were analyzed with Pbsv (version 2.8.0).²⁶ Paraphase (version 3.1.0) was used to detect int22h-related recombination through DAHH.⁴⁰ Finally, Hifiasm (version 0.19.3-r572) was used to identify additional structural variations or genomic recombination events, if necessary.⁴¹ The resulting assemblies were visualized using the Integrative Genomics Viewer.

Methods for validation

We performed Sanger sequencing on samples from patients in whom SNVs were detected, and their relatives. For int22h-related rearrangements, LR-PCR was performed following the methodology described by Bagnall et al.¹⁵ For deletions and other complex SVs, we used optical genome mapping (OGM; A12 and B2, as described in our previous study)⁴² or PacBio whole-genome sequencing (A9).

Results

Establishment of hc-LRS method

The comprehensive testing program combines hc-LRS with a suite of bioinformatics tools, aiming to detect various variant types in a single test.

First, we designed probes specifically suited for capturing long DNA fragments, which differ from those used in NGS platforms (Figure 1A). The biotinylated DNA probes are 100 base pairs (bp) in length and have a 0.33× tiling (200-bp gaps between probes). The regions of interest include the entire *F8*, *F9*, and *VWF* genes and 3 homologous regions (int22h-2, int22h-3, and int1h-1). Following the standard operating procedure of this study, it was determined that this probe set achieves complete coverage of all target regions (Figure 1B).

Next, we established the standard operating procedure for hc-LRS (Figure 1C). The overall design of the experimental workflow was inspired by the general approach used in targeted sequencing on NGS platforms, following the steps of fragmentation, pre-amplification, hybridization capture, and postamplification. However, because LRS requires much longer DNA molecules than NGS, the conditions for fragmentation and amplification needed to be adjusted and are more stringent. Because longer sequencing reads lead to more accurate genome assemblies and a higher possibility

Table 1. Sample information and the results of the comprehensive testing program

Number	Sex/age, y	Phenotype	Average length of reads	On-target rate, %	Depth of sequencing	Type of variant	Nucleotide site	Validation
A1	M/34	Severe HA	5631	26.0	128	Inv22 type I		LR-PCR
A2	M/57	Severe HA	5872	26.1	148	Inv22 type I		LR-PCR
A3	M/32	Severe HA	5707	26.0	147	Nonsense	NM_000132.4(<i>F8</i>): c.1804C>T (p.Arg602Ter)	Sanger sequencing
A4	M/27	Severe HA	5610	27.2	153	Inv22 type I		LR-PCR
A4-2	F/51	A4' mother	5583	28.1	152	Inv22 type I carrier		LR-PCR
A5	M/15	Severe HA	5735	26.7	112	Nonsense	NM_000132.4(<i>F8</i>): c.6496C>T (p.Arg2166Ter)	Sanger sequencing
A5-2	F/40	A5' mother	5576	27.6	151	No identical A5 variant found		Sanger sequencing
A6	M/58	Severe HA	5409	26.2	140	Inv22 type I		LR-PCR
A7	M/31	Severe HA	5457	25.9	127	Inv22 type I		LR-PCR
A8	M/34	Severe HA	5595	25.8	136	Nonsense	NM_000132.4(<i>F8</i>): c.6714G>A (p.Trp2238Ter)	Sanger sequencing
A9	M/7	Severe HA	5597	25.2	135	int22h-related complex rearrangement		LR-PCR PacBio WGS
A10	M/47	Severe HA	5795	24.8	126	Inv22 type I		LR-PCR
A10-2	F/23	A10' daughter	5346	28.4	151	Inv22 type I carrier		LR-PCR
A11	M/50	Severe HA	5633	26.8	140	Missense	NM_000132.4(<i>F8</i>): c.5530C>T (p.Pro1844Ser)	Sanger sequencing
A12	F/32	Mild HA	5643	28.0	169	int22h-related complex rearrangement carrier		LR-PCR OGM
B1	M/59	Severe HB	5719	28.3	170	Missense	NM_000133.4(<i>F9</i>): c.205T>G (p.Cys69Gly)	Sanger sequencing
B2	M/13	Severe HB	5690	22.8	107	<i>F9</i> complete deletion		OGM
B2-2	F/36	B2' mother	5864	24.9	166	<i>F9</i> complete deletion carrier		

F, female; HA, hemophilia A; HB, hemophilia B; M, male; WGS, whole-genome sequencing.

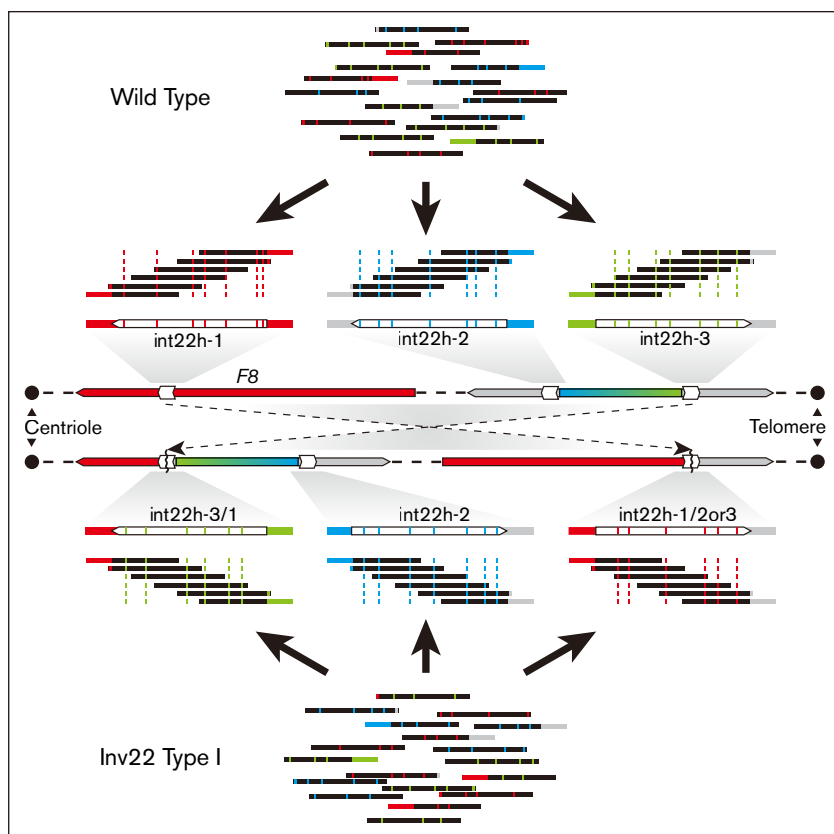


Figure 2. Diagram of *int22h*-related rearrangements analysis through the DAHH. Wild type and Inv22 type I as examples. *int22h-1* is located within intron 22 of *F8*. *int22h-2* and *int22h-3* are located within the 2 arms of a stem-loop structure. Inv22 type I represents the inversion of the fragment between *int22h-1* and *int22h-3*. Red bands represent *F8*, blue bands represent the upstream sequence of *int22h-2*, green bands represent the upstream sequence of *int22h-3*, and gray bands represent the downstream sequence of *int22h-2* and *int22h-3*.

of identifying structural variations, it was essential to preserve fragments as long as possible. Therefore, we optimized the conditions for each experimental step, with particular emphasis on the following key stages: DNA fragmentation (supplemental Figure 1), selection of long-range DNA polymerases (supplemental Figure 2), and DNA size selection (supplemental Figure 3). Following the standard operating procedure, we obtained sequencing reads with an average length of ~5.5 kb and a capture efficiency of ~25%, whereas the average sequencing depth exceeded 100x in all target regions.

The samples used for testing comprised 18 specimens derived from patients or relatives within 14 families with hemophilia. Table 1 summarizes the results for all samples. Full concordance was observed between the pathogenic SNVs and SVs identified by the comprehensive testing program and validated methods.

DAHH and analysis of *int22h*-related rearrangements

Identifying *int22h*-related rearrangements is of particular significance in genetic testing for hemophilia. A single sequencing read cannot fully cover the 9.5-kb-long *int22h* regions, whereas de novo assembly offers a feasible solution for distinguishing homologous regions.

Paraphrase, a Python tool developed to differentiate the homologous genes *SMN1* and *SMN2*, which are associated with spinal muscular atrophy,⁴⁰ was used in this study to identify *int22h* regions (Figure 2). Paraphrase first extracted all HiFi reads

containing the *int22h-1*, *int22h-2*, and *int22h-3* regions, then aligned them with the *int22h-2* region of the reference genome. The same single-nucleotide polymorphisms were identified and ligated to assemble the HiFi reads into haplotypes. These haplotypes were then assigned to *int22h-1*, *int22h-2*, or *int22h-3* based on differences in sequences upstream and downstream of the homologous region. Male samples typically assemble 3 sets of haplotypes, whereas female samples assemble 6 (Figure 3A). *Int22h*-related rearrangements were indicated by the presence of fusion fragments, in which the upstream and downstream sequences originate from different *int22h* regions. In this study, the fusion fragment formed by the anterior part of *int22h-3* and the posterior part of *int22h-1* is designated *int22h-3/1*, whereas the fusion fragment formed by the anterior part of *int22h-2* and the posterior part of *int22h-1* is designated *int22h-2/1*. *Int22h-2* and *int22h-3* have the same downstream sequence.⁴³ The fused fragment formed by the anterior part of *int22h-1* and the posterior part of either *int22h-2* or *int22h-3* is designated *int22h-1/2or3*.

Through DAHH, we identified 10 samples with *int22h*-related rearrangements, some of which are shown in Figure 3A. The fusion fragments *int22h-1/2or3* and *int22h-3/1* were detected in 8 of these samples, in which normal *int22h-2* was also identified in male patients (A1, A2, A4, A6, A7, and A10), whereas female carriers contained normal *int22h-1*, *int22h-2*, and *int22h-3*, indicating a normal X allele (A5-2, A10-2). This variant type is considered to be Inv22 type I, the most common type of SV leading to severe hemophilia A.⁴ In addition, A9 and A12 are considered complex variants, which will be described in the following part of this article.

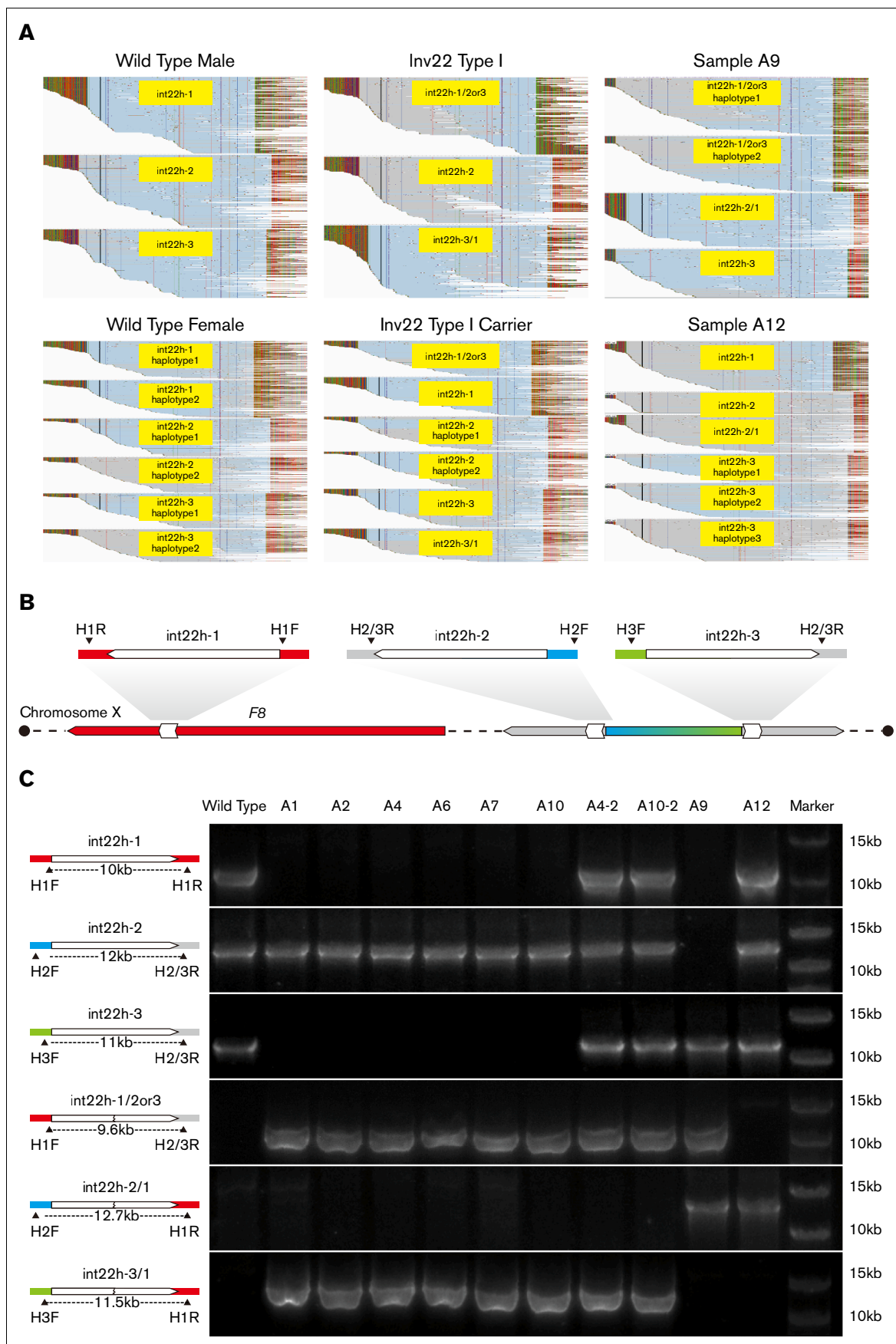


Figure 3.

We conducted LR-PCR on int22h-1, int22h-2, and int22h-3, as well as on 3 types of fusion fragments, using a combination of 5 primers (H1F, H2F, H3F, H1R, and H2/3R), in accordance with the methodology proposed by Bagnall et al (Figure 3B).¹⁵ The results for the samples identified as int22h-related rearrangements were identical (Figure 3C). However, PCR is only capable of characterizing normal and fusion fragments and cannot provide copy number information for each type.

In addition, we evaluated the potential of DAHH for application to other homologous loci by applying it to distinguish the *VWF* gene from its pseudogene *VWFP1*. As a result, we successfully resolved 4 sets of haplotypes corresponding to *VWF* and *VWFP1*, respectively (supplemental Figure 4).

Detection of SNVs

Different pathogenic SNVs were identified in 5 samples, including 3 nonsense mutations located in exons 12, 23, and 24 of *F8*; and 2 missense mutations, 1 in exon 16 of *F8* and the other in exon 2 of *F9* (Table 1). One of the variants in *F8* (NM_000132.4:c.6714G>A) is a novel variant resulting in p.Trp2238Ter. It has been documented that an adjacent SNV (NM_000132.4:c.6713G>A) causes translation to terminate prematurely at the same position, leading to hemophilia A. All other SNVs are documented pathogenic variants in the database (<https://dbs.eahad.org/FVIII>; <https://dbs.eahad.org/FIX>). Additionally, patient A5, a 15-year-old adolescent boy with hemophilia A, was found to carry a nonsense mutation in *F8* (NM_000132.4:c.6496C>T). However, this SNV was not identified in his mother (A5-2), suggesting that it is a de novo variant in A5. For all relevant aforementioned samples, Sanger sequencing was performed, confirming complete consistency with the results of the targeted sequencing.

Detection of other SVs

A large deletion of the *F9* gene was identified in a male patient with hemophilia B and his mother (B2 and B2-2; Figure 4A). No reads corresponding to the *F9* gene were identified in the sequencing results of patient B2, suggesting a complete deletion of the *F9* gene in the probe-covered region. Although relevant reads of *F9* gene were found in B2-2, the de novo assembly of the *F9* gene only yielded a single haplotype, whereas 2 haplotypes should have been assembled in females with normal genotype. This suggests that B2-2 is a carrier of the *F9* deletion. Additionally, in our previous study,⁴² OGM was performed on samples from patient B2, revealing a 451-kb deletion in the Xq27.1 region that includes the *F9* gene (supplemental Figure 5).

In addition, heterozygous insertions of 140 bp and 144 bp were detected in the *VWF* gene in individuals A2 and A9, respectively (Figure 4B). Both insertions are located within the CATA repeat region of intron 47 and are not predicted to exert any functional effects on *VWF*.

Further exploration of complex variants

One set of int22h-3, 1 set of int22h-2/1, and 2 sets of int22h-1/2or3 were identified in a male patient with hemophilia A (patient A9), indicating complex int22h-related rearrangements, including copy number variants (Figure 3A). To further investigate this complex variant, we conducted PacBio whole-genome sequencing on samples from patient A9, which generated HiFi reads with an average length of 12 kb and a depth of 15×. Although the data did not yield a contiguous assembly of the entire region, several major fragments were successfully assembled. No fusion reads outside the homologous regions were identified, suggesting that all rearrangements occurred within the homologous regions. Based on the sequencing results, we constructed a model for the evolution of this variant. It is possible that an inversion between int22h-2 and int22h-3, another inversion between int22h-1 and int22h-2 (distal int22h), and a duplication from int22h-2 to int22h-1/2or3 occurred sequentially (Figure 5A). This inference is consistent with the results of DAHH and LR-PCR.

The DAHH results of female patient A12 showed 1 set of int22h-2/1, 1 set of int22h-1, 1 set of int22h-2, and 3 different sets of int22h-3 (Figure 3A). An additional set of fusion fragments was identified, chrX:155,471,846(+)-chrX:154,872,422(+), located outside int22h. The OGM results for the same sample in our previous study indicated a SV in Xq28,⁴² in which the fragment between int22h-2 and int22h-3 was inverted and repeatedly inserted into int22h-1 (supplemental Figure 5). Because of resolution limitations, OGM cannot resolve the nucleotide positions of the breakpoints. Synthesizing all the information, it can be concluded that 1 of the breakpoints of the insertion is at hg38 chrX:155,471,846; whereas the other is located within int22h-2. In addition, a minor duplication occurs from hg38 chrX:154,872,422 to int22h-1, with the insertion positioned at the junction of this duplication (Figure 5B). This conclusion is generally consistent with the results of DAHH and LR-PCR, but there is a minor issue. Because of the duplication, there should be 8 sets of int22h, whereas DAHH only assembled 6 sets. This discrepancy, in which the number of assembled haplotypes is fewer than the predicted number, was also observed for patient A10-2. This may be because the nucleotide sequences of int22h on the 2 alleles are too similar to be distinguished.

Discussion

Hemophilia is a hereditary bleeding disorder with a long history of research, yet the genetic variations of the disease continue to be elucidated. Among the known causative genes, *F8* is particularly challenging because of its large genomic size, complex internal structure, and unique location at the telomeric end of the X chromosome's long arm. These features contribute to the wide variety of variant types observed in hemophilia A and complicate genetic testing. NGS technologies have limitations in detecting SVs, especially those involving highly homologous regions.

Figure 3. Analysis of int22h-related rearrangements. (A) Results of DAHH. Three sets of haplotypes were assembled in wild-type males, and 5 sets of haplotypes were assembled in wild-type females. The fusion fragments int22h-1/2or3 and int22h-3/1 were identified in samples with Inv22 type I (use A1 as an example for the male Inv22 type I, and A4-2 as an example for the female Inv22 type I carrier.) A9 and A12 are considered int22h-related complex variants. (B) LR-PCR for int22h-related rearrangements. The locations of the 5 primers (H1F, H2F, H3F, H1R, and H2/3R) in the reference genome. (C) Results of LR-PCR. Five primers were used in different combinations to amplify int22h-1, int22h-2, int22h-3, and the fusion fragments int22h-1/2or3, int22h-2/1, and int22h-3/1. Samples A1, A2, A4, A6, A7, and A10 were identified as Inv22 type I; samples A4-2 and A10-2 were identified as Inv22 type I carriers; and samples A9 and A12 are int22h-related complex variants.

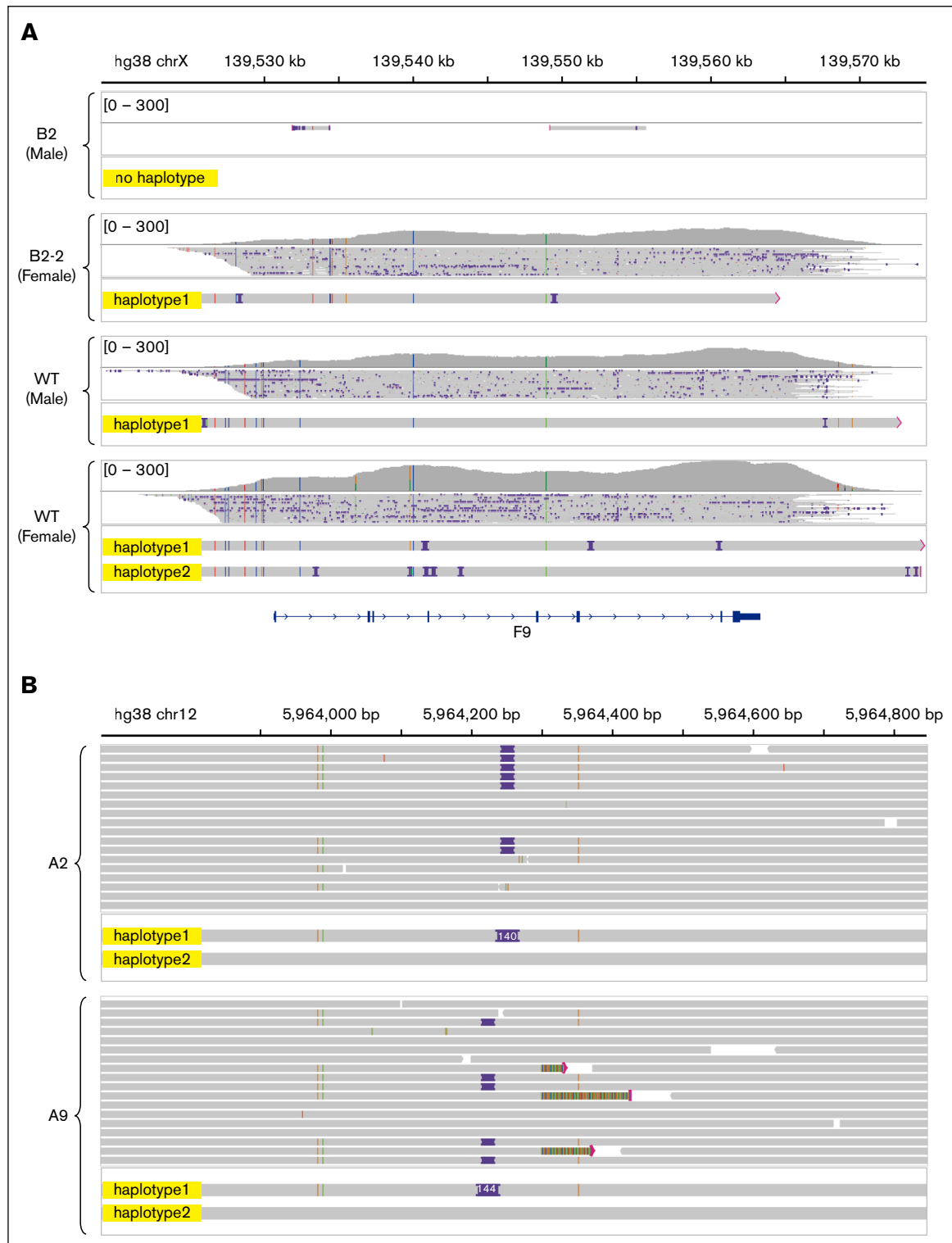


Figure 4. Detection of other SVs. (A) Sequencing results and de novo assembly of the *F9* gene in B2 and B2-2. No haplotypes were assembled in B2, and only 1 set of haplotypes was assembled in B2-2, whereas 1 set of haplotypes is present in WT males, 2 sets of haplotypes are present in WT females. (B) Sequencing results and de novo assembly showed the insertions in *VWF* intron 47 in A2 and A9. These variants do not affect the function of *VWF*. Chr12, chromosome 12; ChrX, X chromosome; WT, wild-type.

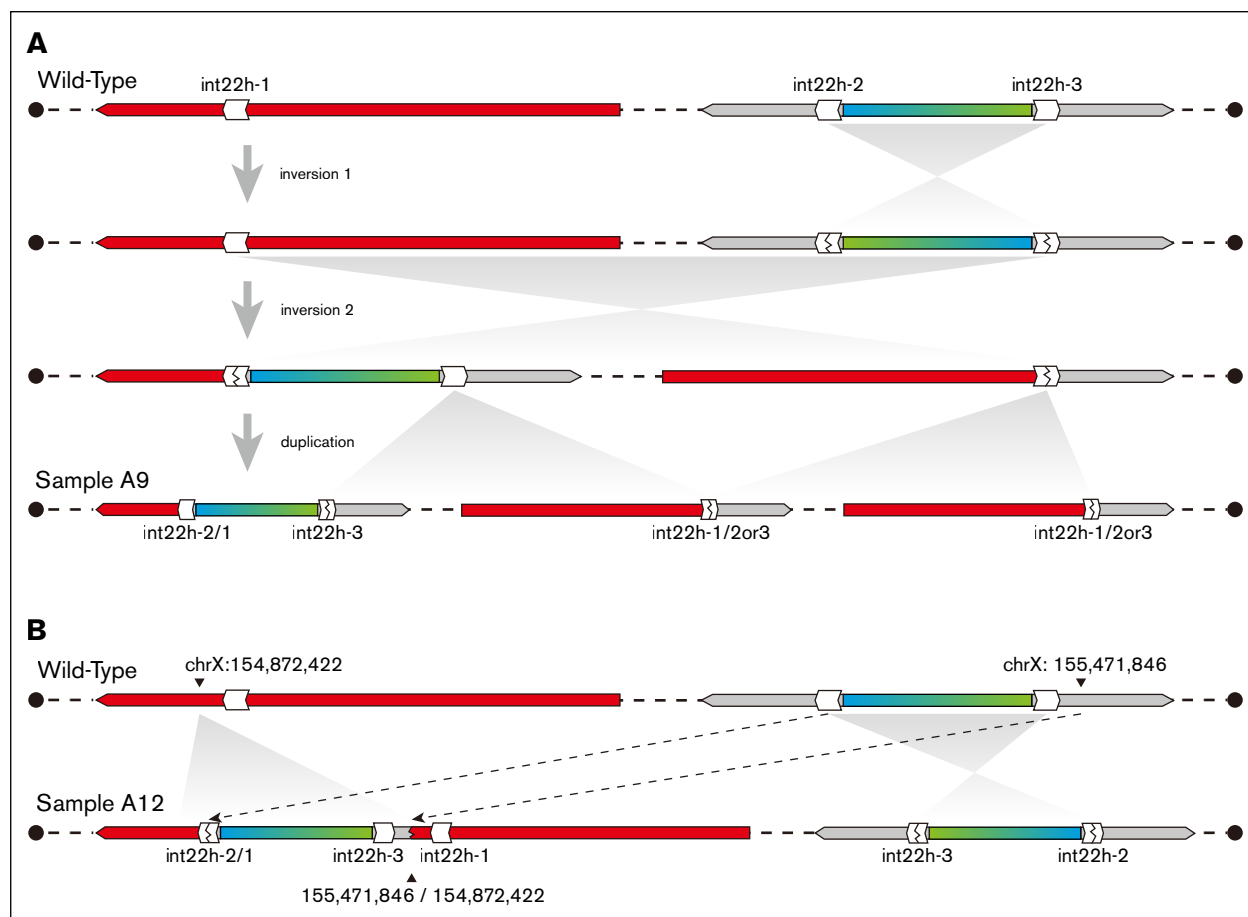


Figure 5. Complex variants in samples A9 and A12. (A) A model for the evolution of the complex variant in A9. An inversion occurred between int22h-2 and int22h-3, another inversion occurred between int22h-1 and int22h-2 (distal int22h), and a duplication of int22h-2 and int22h-1/2or3 repeats occurred sequentially. (B) The structure of the complex variant in A12. The fragment between int22h-2 and int22h-3 was inverted, and a repeat fragment was inserted into F8. One breakpoint of the insertion fragment is located at hg38 chrX:155,471,846; whereas the other is situated within int22h-2. In addition, a minor duplication of the sequence from hg38 chrX:154,872,422 to int22h-1 occurred, with the insertion site located at the junction of this duplication.

Conventional genetic testing for hemophilia A typically follows a stepwise approach, initially detecting int22h-related rearrangements and Inv1 using LR-PCR or IS-PCR,¹⁴⁻¹⁶ followed by NGS to identify SNVs and indels.¹⁷ When PCR and sequencing are inconclusive, or when exon deletions are suspected, multiplexed

ligation-dependent probe amplification is used for further analysis.¹⁸ In this study, we developed a comprehensive testing program for hemophilia based on the PacBio LRS platform. The program was validated with 18 clinical samples from 14 hemophilia pedigrees, with results entirely concordant with those

Table 2. Methods of genetic testing for hemophilia

	SNVs/ indels	Inv 22	int22h-related complex variants	Inv 1	Large deletions/ duplications	Distinguish VWD	Cost per sample, \$	Comment
NGS	✓				Possible	✓	100	
LR-PCR/IS-PCR		✓	✓	✓			30	Complex variants may be misdiagnosed
MLPA					✓		100	
OGM		✓	✓	✓	✓		800	Demanding high-quality DNA; resolution limited
PacBio WGS	✓	✓	✓	✓	✓	✓	1350	Low sequencing depth
Comprehensive testing program	✓	✓	✓	✓	✓	✓	100	Identification of fusion fragments and CNVs in homologous regions

✓, capable; CNV, copy number variation; MLPA, multiplexed ligation-dependent probe amplification; VWD, von Willebrand disease.

obtained from validated methods. Leveraging the high single-base accuracy and long-read capability of the PacBio system, our approach enables simultaneous detection of various variant types in a single assay. Experimentally, we established a standard operating procedure for hc-LRS, consistently generating high-quality data, with read lengths exceeding 5 kb. Analytically, the DAHH pipeline enabled detection of int22h-related rearrangements and copy number analysis within homologous regions. Compared with conventional approaches, our method offers notable advantages in cost-effectiveness, workflow simplicity, and the capacity to identify complex variants. A comparative analysis with several alternative methods is summarized in Table 2.

LRS enables the exploration of genomic regions previously inaccessible with conventional sequencing techniques. To fully leverage its potential, targeted sequencing is essential. Multiplex LR-PCR has recently been used for targeted LRS in hemophilia A.^{33,34} However, this method could miss variants outside primer cover sites, whereas addition of primer types may also increase amplification bias, especially when amplifying long DNA fragments. In contrast, hybridization-based probe capture is particularly effective for identifying complex rearrangements and novel variants. Benefit from optimized technical parameters (detailed in the supplemental Methods), our standard operating procedure completes the experimental process within 2 days and provides sequencing results within 5 days, while maintaining cost-efficiency. For the panel used in this study, 10 000 HiFi reads are sufficient to achieve an average target region coverage of $\geq 100\times$. Assuming a capture efficiency of 25% and 2 to 4 million HiFi reads from a single PacBio SMRT Cell 8M chip, each sample requires only 1% to 2% of the chip capacity and can be mixed proportionally with any other barcoded HiFi library, such as a barcoded whole-genome sequencing sample. In this study, the cost of hc-LRS is approximately \$100 per sample by sharing the flowcell with other samples, whereas whole-genome sequencing on the PacBio Sequel II platform costs up to \$1350. The newer PacBio Revio platform, with the 25M chip, is expected to further enhance cost-efficiency.

The hc-LRS and DAHH approach offers an opportunity to identify and characterize complex SVs. For patient A12, hc-LRS identified fusion reads and nucleotide-level breakpoints beyond the homologous regions. For patients A9 and A12, DAHH revealed unexplained copy number abnormalities of int22h, indicating the presence of complex int22h-related rearrangements. Integrative analysis combining long-read whole-genome sequencing and OGM revealed the complete structure of these variants. Interestingly, both complex variants involved fragments of int22h-2/1. Sequencing of the human X chromosome has revealed that the orientations of int22h-1 and int22h-2 are identical, whereas int22h-3 is oriented oppositely to the former 2.⁴⁴ As a result, inversions between int22h-1 and int22h-3 (Inv22 type I) are more likely to occur than inversions between int22h-1 and int22h-2 (Inv22 type II). Theoretically, the formation of Inv22 type II may require an inversion between int22h-2 and int22h-3 (as observed in patient A12), followed by another inversion, which ultimately forms the fusion fragments int22h-2/1 and int22h-1/2. The number of cases diagnosed with Inv22 type II was only one-fifth that of Inv22 type I, in severe hemophilia A.⁴ Int22h-related duplications or deletions mainly occur between int22h-1 and int22h-2, forming int22h-2/1, with only a few cases reported.⁷⁻¹² The identification of int22h-related rearrangements relies on LR-PCR or IS-PCR, but the

results could be misleading. For instance, if only PCR results are considered, A9 could be misclassified as Inv22 type II, and A12 could be misdiagnosed as a deletion between int22h-1 and int22h-2. Similar misclassifications have also been reported in other studies.⁸ These findings underscore that fusion fragments should not be automatically interpreted as simple inversions or deletions, as they may reflect more intricate genomic rearrangements.

Beyond *F8*, the hc-LRS and DAHH method also shows promise for analyzing other loci involving homologous regions. In this study, DAHH effectively distinguished the *VWF* gene from its pseudogene *VWFP1*, which shares 97% sequence identity with *VWF* exons 23 to 34 (supplemental Figure 4).^{45,46}

Moreover, the comprehensive testing program has potential applications in PGT. Currently, pedigree linkage analysis is the mainstay for PGT of int22h-related rearrangements.^{47,48} An emerging alternative involves LRS to construct parental haplotypes, followed by single-nucleotide polymorphism-based linkage analysis for embryo selection, eliminating the need for a proband. Such applications have already been reported.^{23,24} To ensure robust haplotype assembly, a minimum of 30 $\times$ coverage is recommended.^{26,49} Whereas whole-genome sequencing requires multiple sequencing chips to meet this requirement, hc-LRS can achieve it efficiently. In terms of haplotype assembly at partial loci, targeted and whole-genome LRS yield comparable results. Further studies are needed to assess the suitability of this approach for PGT in hemophilia.

This study has several limitations. Although hc-LRS and DAHH can indicate the presence of complex variants, OGM or long-read whole-genome sequencing remain necessary to resolve the complete structures. Expanding the target region to include sequences between *F8*, int22h-2, and int22h-3 may enable complete assembly of int22h-related rearrangements. In addition, the sample size was limited, and variants such as Inv1 or other non-int22h-related pathogenic SVs were not collected. Although the probe set includes int1h-2, a 1041-bp homologous region of *F8* intron 1 implicated in Inv1,⁵⁰ and theoretically a HiFi read can span and resolve this region, such rearrangements were not identified in our samples. Furthermore, although 2 heterozygous insertions in *VWF* were detected, both were nonpathogenic, leaving the method's utility in identifying heterozygous pathogenic SVs and breakpoints unverified. The comprehensiveness of this approach requires validation in larger cohorts with more diverse variant types.

Acknowledgments

This work was supported by funds from the Priority Academic Program Development of Jiangsu Higher Education Institutions (20KJA320001 [M.J.]), the Suzhou Science and Technology Project (SKY2022012 [M.J.], SZS2023014 [M.J.]), Jiangsu Provincial Research Hospital Project (YJXY202204 [L.Z.]), and the National Natural Science Foundation of China (82300151 [L.Z.]).

Authorship

Contribution: L.K., B. Liang, and M.J. designed the experiments; B. Liu, R.X., and Y.Y. carried out the molecular genetic studies; B. Liu and M.G. analyzed the data; B. Liu, H.L., S.C., and M.J. wrote the manuscript; and S.M., L.Z., and Z.Y. assisted in collecting clinical samples.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

ORCID profiles: B.L., 0009-0004-1957-5426; S.M., 0009-0006-6258-2706; M.J., 0000-0002-4929-4885.

Correspondence: Miao Jiang, The Fourth Affiliated Hospital of Soochow University, 9 Chongwen St, Suzhou 215021, China; email: jiangmiao@suda.edu.cn; and Bo Liang, Basecare Medical Device Co, Ltd, 77 Jingu St, Suzhou 215028, China; email: boliang880@alumni.sjtu.edu.cn.

References

1. Peyvandi F, Jayandharan G, Chandy M, et al. Genetic diagnosis of haemophilia and other inherited bleeding disorders. *Haemophilia*. 2006;12 (suppl 3):82-89.
2. Berntorp E, Fischer K, Hart DP, et al. Haemophilia. *Nat Rev Dis Primers*. 2021;7(1):45.
3. Johnsen JM, Fletcher SN, Huston H, et al. Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the My Life, Our Future initiative. *Blood Adv*. 2017;1(13):824-834.
4. Johnsen JM, Fletcher SN, Dove A, et al. Results of genetic analysis of 11 341 participants enrolled in the My Life, Our Future hemophilia genotyping initiative in the United States. *J Thromb Haemost*. 2022;20(9):2022-2034.
5. McVey JH, Rallapalli PM, Kemball-Cook G, et al. The European Association for Haemophilia and Allied Disorders (EAHAD) coagulation factor variant databases: important resources for haemostasis clinicians and researchers. *Haemophilia*. 2020;26(2):306-313.
6. Antonarakis SE, Rossiter JP, Young M, et al. Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood*. 1995;86(6):2206-2212.
7. Pegoraro E, Whitaker J, Mowery-Rushton P, Surti U, Lanasa M, Hoffman EP. Familial skewed X inactivation: a molecular trait associated with high spontaneous-abortion rate maps to Xq28. *Am J Hum Genet*. 1997;61(1):160-170.
8. El-Hattab AW, Fang P, Jin W, et al. Int22h-1/int22h-2-mediated Xq28 rearrangements: intellectual disability associated with duplications and in utero male lethality with deletions. *J Med Genet*. 2011;48(12):840-850.
9. Lannoy N, Grisart B, Eeckhoudt S, et al. Intron 22 homologous regions are implicated in exons 1-22 duplications of the F8 gene. *Eur J Hum Genet*. 2013;21(9):970-976.
10. Li S, He J, Chu L, et al. F8 gene inversion and duplication cause no obvious hemophilia A phenotype. *Front Genet*. 2023;14:1098795.
11. Fahiminiya S, Oikonomopoulos S, Rivard G, et al. Deciphering a novel complex inversion affecting F8 in a family with severe haemophilia A by optical genome mapping. *Haemophilia*. 2023;29(3):921-924.
12. Yuan S, Hu L, Zhong J, et al. Genetic analysis and reproductive interventions for two rare families affected by severe haemophilia A. *Haemophilia*. 2025;31(1):148-155.
13. Xu Z, Spencer HJ, Harris VA, Perkins SJ. An updated interactive database for 1692 genetic variants in coagulation factor IX provides detailed insights into hemophilia B. *J Thromb Haemost*. 2023;21(5):1164-1176.
14. Liu Q, Nozari G, Sommer SS. Single-tube polymerase chain reaction for rapid diagnosis of the inversion hotspot of mutation in hemophilia A. *Blood*. 1998;92(4):1458-1459.
15. Bagnall RD, Giannelli F, Green PM. Int22h-related inversions causing hemophilia A: a novel insight into their origin and a new more discriminant PCR test for their detection. *J Thromb Haemost*. 2006;4(3):591-598.
16. Rossetti LC, Radic CP, Larripa IB, De Brasi CD. Developing a new generation of tests for genotyping hemophilia-causative rearrangements involving int22h and int1h hotspots in the factor VIII gene. *J Thromb Haemost*. 2008;6(5):830-836.
17. Chen J, Li Q, Lin S, et al. The spectrum of FVIII gene variants detected by next generation sequencing in 236 Chinese non-inversion hemophilia A pedigrees. *Thromb Res*. 2021;202:8-13.
18. Rost S, Löffler S, Pavlova A, Müller CR, Oldenburg J. Detection of large duplications within the factor VIII gene by MLPA. *J Thromb Haemost*. 2008;6 (11):1996-1999.
19. George LA, Sullivan SK, Giermasz A, et al. Hemophilia B gene therapy with a high-specific-activity factor IX variant. *N Engl J Med*. 2017;377(23): 2215-2227.
20. Xue F, Li H, Wu X, et al. Safety and activity of an engineered, liver-tropic adeno-associated virus vector expressing a hyperactive Padua factor IX administered with prophylactic glucocorticoids in patients with haemophilia B: a single-centre, single-arm, phase 1, pilot trial. *Lancet Haematol*. 2022; 9(7):e504-e513.
21. Ozelo MC, Mahlangu J, Pasi KJ, et al. Valoctocogene roxaparvovec gene therapy for hemophilia A. *N Engl J Med*. 2022;386(11):1013-1025.
22. Laurie AD, Hill AM, Harraway JR, et al. Preimplantation genetic diagnosis for hemophilia A using indirect linkage analysis and direct genotyping approaches. *J Thromb Haemost*. 2010;8(4):783-789.
23. Madjunkova S, Sundaravadanam Y, Antes R, et al. Detection of structural rearrangements in embryos. *N Engl J Med*. 2020;382(25):2472-2474.
24. M M YC, Yu Q, Ma M, et al. Variant haplotyping by long-read sequencing: a new approach to preimplantation genetic testing workups. *Fertil Steril*. 2021;116(3):774-783.

25. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133-138.
26. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37(10):1155-1162.
27. Karamitros T, Magiorkinis G. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic Acids Res*. 2015;43(22):e152.
28. Lefoulon E, Vaisman N, Frydman HM, et al. Large enriched fragment targeted sequencing (LEFT-SEQ) applied to capture of Wolbachia genomes. *Sci Rep*. 2019;9(1):5939.
29. Bethune K, Mariac C, Couderc M, et al. Long-fragment targeted capture for long-read sequencing of plastomes. *Appl Plant Sci*. 2019;7(5):e1243.
30. Hook PW, Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. *Nat Rev Genet*. 2023;24(9):627-641.
31. Togi S, Ura H, Niida Y. Optimization and validation of multimodular, long-range PCR-based next-generation sequencing assays for comprehensive detection of mutation in tuberous sclerosis complex. *J Mol Diagn*. 2021;23(4):424-446.
32. Walczak M, Skrzypczak-Zielinska M, Plucinska M, et al. Long-range PCR libraries and next-generation sequencing for pharmacogenetic studies of patients treated with anti-TNF drugs. *Pharmacogenomics J*. 2019;19(4):358-367.
33. Liu Y, Li D, Yu D, et al. Comprehensive analysis of hemophilia A (CAHEA): towards full characterization of the F8 gene variants by long-read sequencing. *Thromb Haemost*. 2023;123(12):1151-1164.
34. Ling X, Pan L, Li L, et al. Detection of hemophilia A genetic variants using third-generation long-read sequencing. *Clin Chim Acta*. 2024;562:119884.
35. Tsai YC, Greenberg D, Powell J, et al. Amplification-free, CRISPR-Cas9 targeted enrichment and SMRT sequencing of repeat-expansion disease causative genomic regions. *bioRxiv*. Preprint posted online 16 October 2017. <https://doi.org/10.1101/203919>
36. Hafford-Tear NJ, Tsai YC, Sadan AN, et al. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat. *Genet Med*. 2019;21(9):2092-2102.
37. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38(4):433-438.
38. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*. 2021;37(23):4572-4574.
39. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983-987.
40. Chen X, Harting J, Farrow E, et al. Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am J Hum Genet*. 2023;110(2):240-250.
41. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18(2):170-175.
42. Liu B, Zhou L, Cao L, et al. Optical genome mapping identified deletions, inversions, and insertions in hemophilia. *Blood Adv*. 2025;9(2):360-364.
43. Bagnall RD, Giannelli F, Green PM. Polymorphism and hemophilia A causing inversions in distal Xq28: a complex picture. *J Thromb Haemost*. 2005;3(11):2598-2599.
44. Ross MT, Grafham DV, Coffey AJ, et al. The DNA sequence of the human X chromosome. *Nature*. 2005;434(7031):325-337.
45. Eikenboom JC, Vink T, Briët E, Sixma JJ, Reitsma PH. Multiple substitutions in the von Willebrand factor gene that mimic the pseudogene sequence. *Proc Natl Acad Sci U S A*. 1994;91(6):2221-2224.
46. Mancuso DJ, Tuley EA, Westfield LA, et al. Human von Willebrand factor gene and pseudogene: structural analysis and differentiation by polymerase chain reaction. *Biochemistry*. 1991;30(1):253-269.
47. Bui TMP, Tran VK, Nguyen TTH, et al. Preimplantation genetic testing (PGT) for hemophilia A: experience from one center. *Taiwan J Obstet Gynecol*. 2022;61(6):1009-1014.
48. Nguyen MT, Nguyen TT, Nguyen DB, et al. Robust preimplantation genetic testing of the common F8 Inv22 pathogenic variant of severe hemophilia A using a highly polymorphic multi-marker panel encompassing the paracentric inversion. *Thromb J*. 2023;21(1):108.
49. Peng C, Chen H, Ren J, et al. A long-read sequencing and SNP haplotype-based novel preimplantation genetic testing method for female ADPKD patient with de novo PKD1 mutation. *BMC Genomics*. 2023;24(1):521.
50. Fahiminiya S, Rivard G, Scott P, et al. A full molecular picture of F8 intron 1 inversion created with optical genome mapping. *Haemophilia*. 2021;27(5):e638-e640.