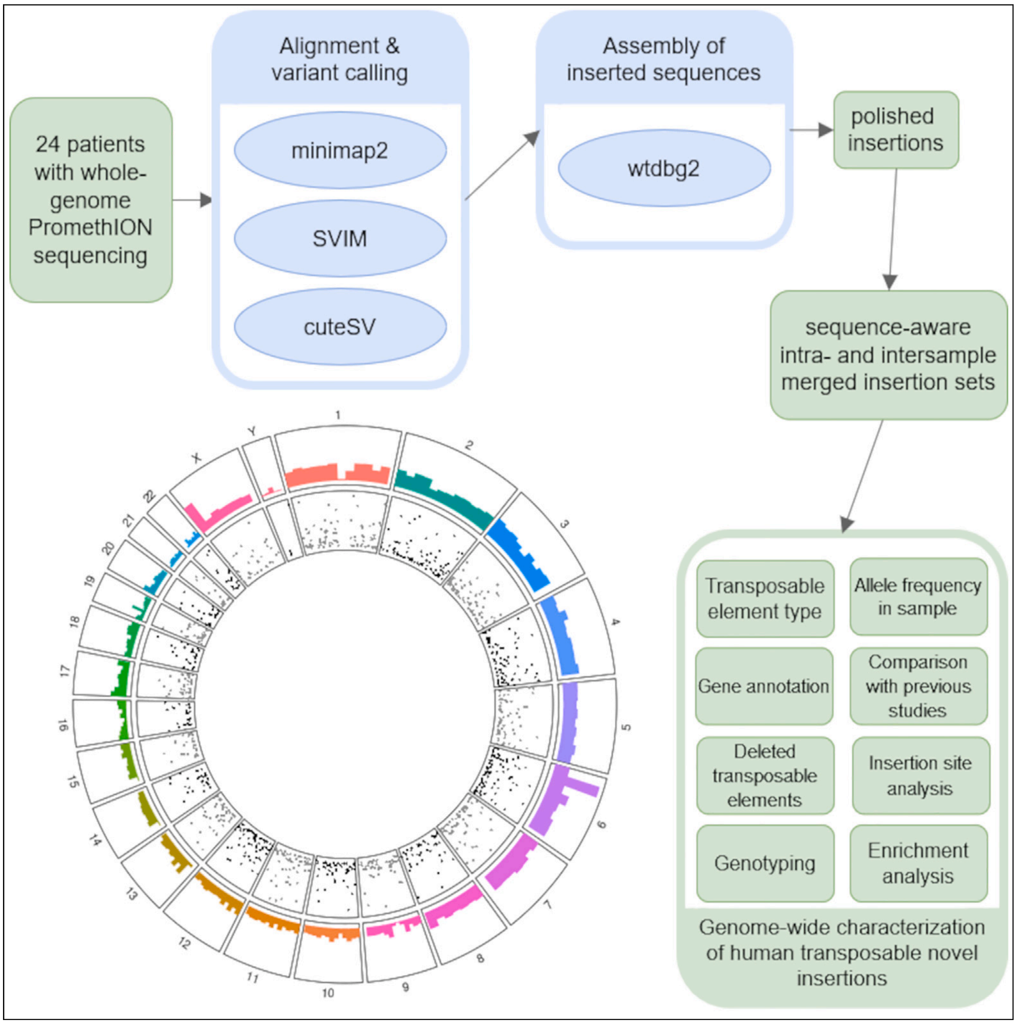
iScience

CellPress
OPEN ACCESS

Article

Detection and annotation of transposable element insertions and deletions on the human genome using nanopore sequencing


24 patients with whole-genome PromethION sequencing
Alignment & variant calling
minimap2
SVIM
cuteSV
Assembly of inserted sequences
wtdbg2
polished insertions
sequence-aware intra- and intersample merged insertion sets
Transposable element type
Allele frequency in sample
Gene annotation
Comparison with previous studies
Deleted transposable elements
Insertion site analysis
Genotyping
Enrichment analysis
Genome-wide characterization of human transposable novel insertions
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Javier Cuenca-Guardiola, Belén de la Morena-Barrio, Esther Navarro-Manzano, ..., Nicholas S. Gleadall, Javier Corral, Jesualdo Tomás Fernández-Breis

javier.corral@carm.es (J.C.)
jfernand@um.es (J.T.F.-B.)

Highlights

7344 TE insertions and 3056 deletions were detected in 24 patients with AT deficiency

2926 TE insertions were not present in previous datasets of TE insertions

6 TE insertions were found in exons, 3 929 in introns, and 147 in promoters

Neuron related functions and autism linked to the genes affected by TE insertions

Cuenca-Guardiola et al., iScience *26*, 108214
November 17, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.108214

iScience

CellPress
OPEN ACCESS

Article

Detection and annotation of transposable element insertions and deletions on the human genome using nanopore sequencing

Javier Cuenca-Guardiola,[1] Belén de la Morena-Barrio,[2] Esther Navarro-Manzano,[2] Jonathan Stevens,[3,4] Willem H. Ouwehand,[3,4,5,6] Nicholas S. Gleadall,[3,4] Javier Corral,[2,*] and Jesualdo Tomás Fernández-Breis[1,7,*]

SUMMARY

Repetitive sequences represent about 45% of the human genome. Some are transposable elements (TEs) with the ability to change their position in the genome, creating genetic variability both as insertions or deletions, with potential pathogenic consequences. We used long-read nanopore sequencing to identify TE variants in the genomes of 24 patients with antithrombin deficiency. We identified 7 344 TE insertions and 3 056 TE deletions, 2 926 were not previously described in publicly available databases. The insertions affected 3 955 genes, with 6 insertions located in exons, 3 929 in introns, and 147 in promoters. Potential functional impact was evaluated with gene annotation and enrichment analysis, which suggested a strong relationship with neuron-related functions and autism. We conclude that this study encourages the generation of a complete map of TEs in the human genome, which will be useful for identifying new TEs involved in genetic disorders.

INTRODUCTION

Mobile genetic elements are repetitive sequences abundant in genomes from all taxa, probably due to the fact that genes encoding transposases are the most prevalent genes in nature.[1] In the human genome, these repetitive regions amount to 45% of its content.[2] Some of these repetitive regions are transposable elements (TEs) and have (or had at some point) the ability to change positions in the genome. Traditionally, two classes of mobile elements have been described: class I and II. Class I elements require an RNA intermediary and retrotranscriptase activity, while class II elements or DNA transposons, move without retrotranscriptase activity. While some class I elements are active on the human genome, mainly Alu, SVAs and L1s, there is no evidence for class I ones being presently active on humans.[2] Importantly, due to their mobility, these elements are involved in the generation of structural variants affecting the genome.[3]

On the other hand, non-mobile repetitive elements include sequences such as satellites, centromeres, or RNA genes repeated in the genome such as tRNA. Repetitive elements can be identified in nucleotide sequences with ease using RepeatMasker.[4] RepeatMasker depends on a database of DNA repeats, such as the open access Dfam, which includes sequences for many species.[5,6] RepeatMasker classifies repetitive elements on three levels in the manner “level 1/level 2-level 3,”[6] in which level 2 and 3 are not always present. Additionally, it has a “name” that specifies the element itself. In this work, we refer to level 1 and 2 as TE families and subfamilies, in addition to the class I/II distinction. For example, a class I element could be named *AluY*, belong to the family *SINE* and to the subfamily *Alu*. An element considered *Tigger1* would be classified under subfamily *TcMar*, and like all class II repeats, it would be in the *DNA* family. Dfam itself offers a different system with more levels, constructed considering molecular mechanisms, cladistic classification and structure.[6]

Detection of TEs is of great relevance from both a genetic and clinical point of view as they are key elements in generating structural variations in humans. Thus, TEs are involved in the evolution of genomes[3]; retrotransposition of active class I elements can affect gene function with pathological consequences,[7] and inactive elements, such as class II ones, can promote DNA recombination, also with clinical consequences.[8] Unfortunately, the abundance and repetitive nature of these elements makes their identification and localization difficult, particularly for new ones when using short read sequencing technology. Although recent tools for short reads allow for good results on humans,[9,10]

[1]Departamento de Informática y Sistemas, Universidad de Murcia, CEIR Campus Mare Nostrum, IMIB-Pascual Parrilla, Facultad de Informática, Campus de Espinardo, Murcia 30100, Spain
[2]Servicio de Hematología, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Pascual Parrilla, CIBERER-III, Ronda de Garay S/N, Murcia 30003, Spain
[3]Department of Haematology, University of Cambridge, CB2 0PT, Cambridge Biomedical Campus, Cambridge, Cambridge, England, UK
[4]Blood and Transplant, National Health Service (NHS), CB2 0QQ, Cambridge Biomedical Campus, Cambridge, England, UK
[5]British Heart Foundation Cambridge Centre of Excellence, Division of Cardiovascular Medicine, Cambridge Heart and Lung Research Institute, Cambridge Biomedical Campus, Cambridge, England CB2 0AY, UK
[6]University College London Hospitals, NHS Foundation Trust, London, England, UK
[7]Lead contact
*Correspondence: javier.corral@carm.es (J.C.), jfernand@um.es (J.T.F.-B.)
https://doi.org/10.1016/j.isci.2023.108214

long-read based sequencing methods have been proved as the best approach for TE detection,[11,12] and have been used to resolve and characterize these variants.[13–16]

In this article, we report a study done in patients suffering from antithrombin deficiency, a monogenic, autosomal, dominant disease, mostly caused by mutations on *SERPINC1*,[17] the gene that encodes antithrombin. Although most cases with antithrombin deficiency are explained by single nucleotide variants (SNVs) or small insertions or deletions (INDELs), up to 5% of cases are due to a range of structural variants (SVs), heterogeneous both in type (duplications, deletions, or insertions) and size, that cover *SERPINC1*.[18] For this study, we used the data of whole genome sequencing done by nanopore technology in a cohort of 24 patients with antithrombin deficiency, 11 with different causal SVs fully characterized. Interestingly, two of these cases carried a 2.4 kb TE from the SVA family inserted in intron 6 of *SERPINC1*. The validation of this SVA insertion and its segregation with antithrombin deficiency was performed in the proband and relatives by specific PCR amplification.[18,19] Other four cases had causal SNPs, and the remaining 9 cases did not possess a genetic diagnosis for antithrombin deficiency. We considered this congenital disease with mendelian inheritance as a good starting point for genetic studies that can be generalized to genome wide application.

Previous genome wide studies of TEs have dealt with their detection, distribution, or even functional impact.[9,10,12,20] We intend to complement and expand their findings using nanopore sequencing and focusing on the biological implications on a new cohort of patients. In addition to the list of the variants that we have detected, this study also supplies characteristics of new TEs that may be useful for further studies, such as observed allele frequencies (AFs).

We expect that these results will be useful for generating a more complete map of TEs in the human genome, which might help to unravel their role in human genetic diseases.[7,21]

RESULTS

Variant calling approach

Minimap2[22] produced alignments as input for two variant callers: SVIM[23] and cuteSV.[24] The choice of tools was based on a benchmark we performed on public data and published works by others (see the STAR Methods). The intersection of variant callers was used to generate a high confidence dataset.[25]

For insertions, *de novo* assembly was used to reconstruct the inserted sequence. Then, the insertions were merged, genotyped and annotated, and their size was compared to consensus TE sequences. The annotation was used for general functional impact, e.g., to test whether an insertion is located on an exon, and for enrichment analysis. We also performed the calling, merging, genotyping, TE and gene annotations on deletions, and finally, we compared our TE dataset with two other studies.

Variant calling results

The combination of SVIM and CuteSV called a mean of 55 011 ± 9 679 SVs per patient using a lax criterion of at least 3 supporting reads with one variant caller reporting them. A stricter criterion, which required both callers to report the same SV, one (or both) of them with 3 or more supporting reads, left 26 483 ± 2 866 SVs per patient. Deletions and insertions made up most of the SVs called, with 45.9% and 51.5%, respectively. Duplications and inversions are a minority, with 2.1%, 0.5%, respectively. The use of the stringent filter slightly modified the distribution of SVs: deletions and insertions were the majority (49.9% and 48.7%, respectively), 1.3%% were duplications, and 0.13% inversions, and no translocations were identified. Table S1 shows SV counts and percentages per patient.

After variant calling, to improve the quality of the reported TE insertions, the surrounding region was assembled *de novo* using the supporting reads for each insertion. This created a consensus sequence more likely to avoid punctual errors from a single read, instead of resorting to variant callers' results, which are limited to reporting the inserted sequence from a single read, at the time of writing this article. With the reconstructed sequences, insertions were annotated for TE sequences using RepeatMasker (see STAR Methods).

To analyze insertions, and particularly TE insertions, we considered as "unique" insertions the merged result after the allele-aware merging from both callers for each patient. Then, the total occurrence of these identified unique insertions was examined across patients. This study identified 13 297 unique TE insertions, which amounted to a total of 72 858 occurrences among the 24 patients, with an average of any insertion appearing with a median of 3 occurrences/insertion. The average insertion count per patient was 3 035.8 ± 317.2.

The stringent criterion left 7 752 TE insertions reported by both callers, and a total of 49 773 calls across patients, with a median prevalence of 3, and each patient had an average of 2 073.9 ± 118.1insertions per patient.

As further characterization of the TE insertions, we compared the length reported by the variant callers against their respective consensus sequences. 55.6% had a length of at least 85% their respective TE sequence. The subfamily that presented most insertions of adequate length was Alu, with 80.7% (Table S2). This was not the case for the rest of the TE subfamilies, suggesting that the insertions were not complete elements.

Polymerase chain reaction validation

To evaluate the confidence of our analysis, we selected 17 TE insertions using the criteria detailed in the STAR Methods to validate by PCR and sequencing. We could confirm the presence of 14 out of 17 insertions (82.4%) (Table S3). These belonged to all the categories created for validation: insertions on reference TE sequences, incomplete TEs, and insertions inside reference TE sequences of the same type.

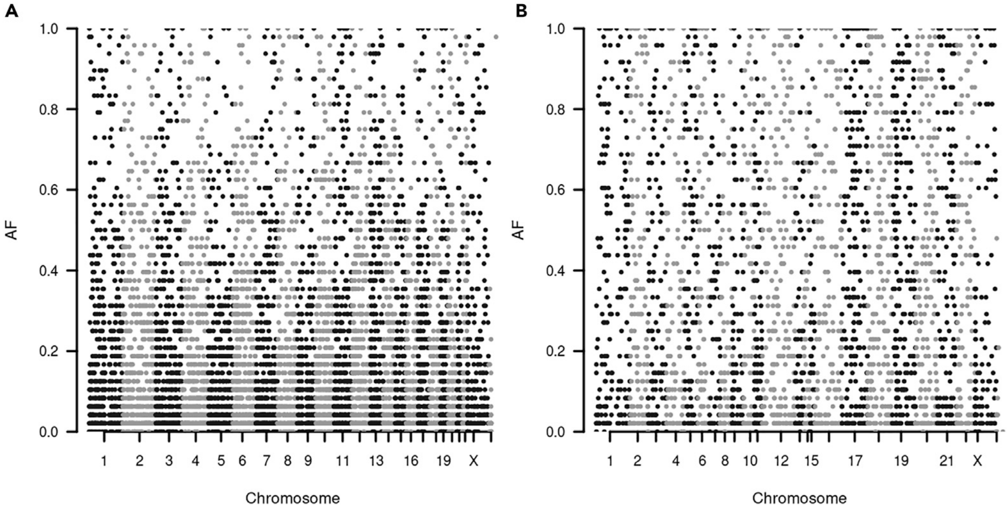


Figure 1. Manhattan plot showing allele frequencies of detected TE variants that passed the stringent filtering on each chromosome

(A) Insertions.

(B) Deletions.

Genotypes of insertions

The TE insertions were genotyped across patients to estimate AFs on the sample set using the Sniffles 2's genotyping method. For the following results, genotyping was considered a gold standard to determine false or true positives and negatives, that is, a variant call that passed either the lax or strict criterion is a true positive when it was assigned a genotype of heterozygous or homozygous and a false positive when it was genotyped as homozygous for the non-inserted allele (reference). The inverse applies for true and false negatives.

First, we observed that both the stringent and lax criteria produced erroneous calls, for example, calls that were genotyped as homozygous for the reference allele (0/0). We observed that 1 092 unique insertions that passed the strict filtering (14.1% of the unique insertions), had at least one occurrence genotyped as 0/0. This amounted to 1 816 total (3.6% of filter-passing occurrences). It is remarkable that 408 of these unique insertions were genotyped as 0/0 for all their occurrences, amounting to 628 insertion counts. Of these fully wrong calls, 330 were only found on a single patient each (80.9% of the 408). On the other hand, the strict criteria missed calls for 1 692 unique insertions (21.8%), a total of 6 589 calls, a sizable amount compared to the 49 773 calls that passed the stringent filter. It is also worth mentioning that 942 unique insertions had more false negative occurrences than positive ones (Table S4). The discrepancies between genotyping and the lax filtering are also summarized in Table S4. The lax criteria did not yield false negatives because it includes all the insertions analyzed.

Once genotyped, AFs for insertions could be computed. The results shown in this section only include unique insertions with at least one occurrence passing the strict filtering. 316 insertions (4.1%) were determined to have an AF of 0. This number is smaller than the amount of 408 reported by genotyping because it considers calls that did not pass the stringent filter, since genotyping was performed independently from the filtering. 5 955 insertions had an AF ≥ 0.062, supporting that they were found in more than one patient, and thus they might be considered as polymorphisms. However, it is interesting to point out that the most common AF was 0.0208, for 1 757 insertions, which means they were found as heterozygous and on a single patient, while 912 insertions were found with an AF ≥ 0.5. Additionally, 166 insertions were found on all cases with the strict method, but after genotyping, it was found that 201 were genotyped with at least one insertion allele on every patient. Of these, 41 were found as homozygous for the alternative allele in all cases, with an AF of 1. The AFs have been plotted on Figure 1A, and data from the 41 ubiquitous insertions are shown on Table S5.

In view of these results, the total number of new TE insertions identified in this study to be evaluated for the following sections is 7 344, as the 408 insertions that contained only false calls have been removed from the stringent dataset.

Class I transposable elements

With the analysis of new inserted sequences by RepeatMasker, we found 7 313 different LINEs, LTRs, retroposons, and SINEs, identifying up to 48 829 occurrences across patients when using strict criteria (Table 1). The most frequent TE insertion corresponded to SINEs from the Alu subclass (4 927 unique ones), followed by SVA retroposons (1 303 unique ones). The distribution of these TE insertions in each chromosome is shown in Figure 2A.

We annotated these TEs to identify genes affected by these insertions. This analysis revealed 3 915 unique TE insertions (25 708 total occurrences) located in introns of 2 938 different genes. Another 145 unique insertions (933 total occurrences) were located on the 1 kb upstream region (promoter) of 135 genes (Figure 2B). Chromosome 1 contained most of these TE insertions in genes, and chromosome 21 had the lowest number of TE insertions, as expected considering the size and number of genes of these chromosomes. The maximum number of

Table 1. Number and type of class I and II insertions and deletions identified in this study

		Insertions			Deletions		
Repeat family	Repeat subfamily	Unique	Total (lax)	Total (strict)	Unique	Total (lax)	Total (strict)
DNA	hAT	15	164	159	42	218	192
	TcMar	16	185	157	17	141	136
LINE	CR1	1	24	24	3	24	24
	L1	597	5359	4863	402	3974	3137
	L2	17	196	182	49	262	241
	RTE	–	–	–	1	3	3
LTR	ERV1	320	2787	2219	62	385	304
	ERVK	41	383	298	14	198	168
	ERVL	101	877	801	122	824	697
	Gypsy	–	–	–	4	25	23
Retroposon	SVA	1303	11274	5915	61	851	756
SINE	Alu	4927	35556	34466	2242	29511	29320
	MIR	6	67	61	35	236	223
	tRNA	–	–	–	2	27	24
Total	–	7344	56872	49145	3056	36679	35248

For each subfamily of TE insertions, the number of unique insertions and the total count across patients using both criteria are reported.

inserted TEs in a region was found in chr6:30 689 239-33 144 378, where 87 TEs insertions were recorded, overlapping 120 coding sequences, introns, or promoters (Figure 2B).

Interestingly, 6 class I TEs, all SINEs and LINEs, Alu and L1 sequences specifically, were inserted in the coding sequences of 6 genes. Table 2 shows the genes affected and the type and location of the inserted TE. Most of these insertions were unique, found in a single patient, but two were identified in more than one patient, supporting new polymorphisms. Indeed, the Alu sequence affecting *CHMP4A*, a gene encoding for a membrane protein in multivesicular bodies,[26] was present in 22 (23 according to genotyping) out of 24 patients, being identified in most of them (16) in a homozygous state. Interestingly, consulting available information of this region in gnomAD,[27] revealed that the coverage in that region is higher than in its surroundings (Figure S1). This may point to an error in the reference sequence, which should contain this Alu sequence, at least in our population. Additionally, the affected exon is present in only one of the coding transcripts[27] (Figure S1).

The other exonic TE insertions affected genes associated with different types of cancer (PRAMEF4,[28] BRCA2,[29] and PRSS23,[14] the latest also linked to retinopahty[30]; hematopoietic differentiation (C12orf29[31]); or sepsis-induced acute respiratory distress syndrome[32] and deafness,[30] for ANKRD36. The alignment for the Alu insertion on BRCA2 is shown on Figure S2.

It is important to point out that the SVA insertions affecting *SERPINC1* passed the stringent filter only for P3, while in P18 it was supported by a single read and did not match any of the quality criteria. In both cases, the SVA was experimentally validated and detected in relatives with antithrombin deficiency.[19]

Class II transposable elements

A total of 31 unique DNA transposons sequences insertions were annotated (316 considering repetitions across patients), none found in coding sequences; 2 (28 across samples) on the promoter region for 2 genes (*LRCH3* and *NT5DC3*); and 14 (161 counting repetition) were found on introns of 14 genes (Table S6). Again, 5 of these insertions were identified in many patients (>18 patients), mostly in homozygous state (the 82 bp TcMar sequence affecting an intron of *LRRC4C* was present in homozygous state in all 24 patients). Unlike the results observed with the insertion in *CHMP4A*, the mean coverage in gnomAD was lower in the regions affected by these class II elements, with small peaks reaching the values of its surroundings (Figure S3).

Enrichment analysis on genes with transposable element insertions

In the previous sections, we have shown that numerous TE insertions were located on genes. To further investigate the possible implications, a gene enrichment analysis was performed, to see if a particular process was overrepresented among the annotated genes.

Using GO molecular function,[33] we found that the most significant terms were linked to GTPase regulation and binding, with "GTPase activator activity" having the lowest p value (Figure 3A). Other significant terms were linked to channel activity (Figures 3B and S4). With GO biological process, the enriched categories were related to the regulation of GTPase mediated signals, NK regulation and neuron growth (Figure S5). The small GTPase mediated signal transduction was linked to both regulation of immune cells and, related to neurons, synapses and growth (Figure S6). The cellular component enrichment pointed to synapse-related locations, specifically the synaptic membrane (Figures S7 and S8).

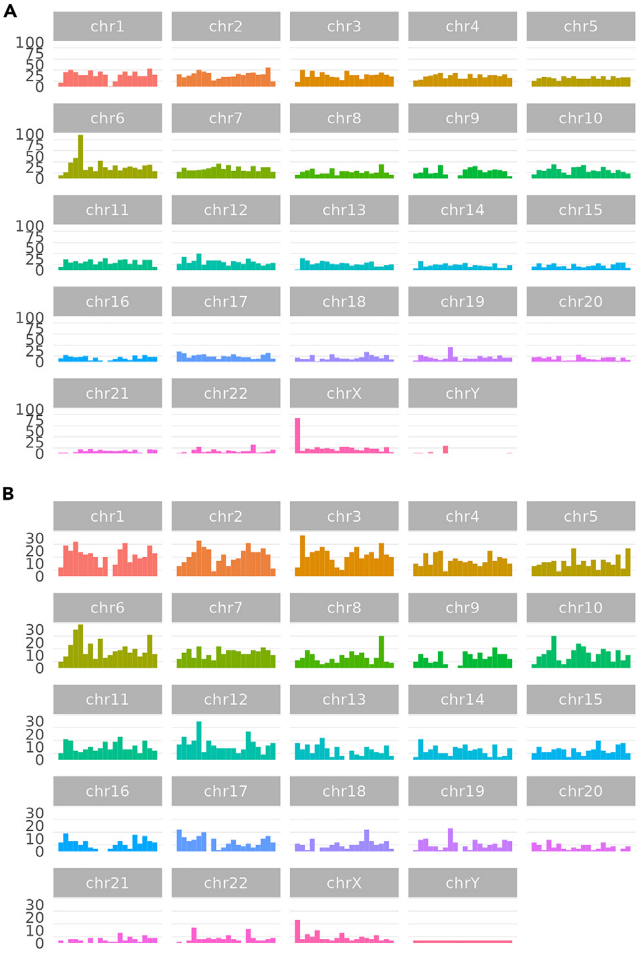


Figure 2. Number of insertions and their annotation across the genome

(A) Insertion counts for class I and II TE insertions in each chromosome.

(B) Annotation counts for class I and II TE insertions in chromosomes overlapping coding sequences, introns, and promoters.

Other ontologies were also used, and the results were coherent with GO's: KEGG[34] pointed to different synapses and neuron signaling annotations, morphine addiction, and cardiomyopathy (Figure S9). The results from disease ontologies were less numerous, but also interesting: DO, for diseases in general,[35] returned "autism spectrum disorder" with the lowest p value, which agrees with all the results related to neurons.

Insertions are more likely to appear on repetitive-rich genes

Insertions were not randomly distributed along chromosomes ($p < 2.2e-16$, chi-squared). Had the distribution been uniform, 136 insertions would have been found on chromosome Y, but only 23 were so. Conversely, chromosome 6, chromosome 17 and chromosome 19 had at least 30% more insertions than expected by a uniform distribution.

We identified a median of 1 TE insertion per gene. Out of the 32 641 genes present in the annotation used for this work, we found at least 1 insertion in 2 991 of them (9.2%). 2 190 genes were affected by a single insertion, while the gene with most insertions on its sequence was *ZCWPW2*, with 13 different unique insertions that amounted to 67 occurrences across patients.

Genes with insertions shared a common characteristic: a higher percentage of repetitive sequences characterized by RepeatMasker. Thus, genes with insertions had a median of 47.5% of repetitive sequences, while these sequences represented a median of 39.4% in genes without insertions ($p < 2.2e-16$, Mann-Whitney) (Figure 4A). The percentage of repetitive sequences among genes with multiple or single insertions was very similar, with median percentages of 47.6% and 47.4%, respectively.

It was observed that for all subfamilies of TE, at least 50% of their insertions were found on repeats annotated by RepeatMasker as TE sequences. Among subfamilies with more than 10 unique insertions in our dataset, SVA was the one with the highest percentage of insertions in repetitive sequences (95.6%). On the other hand, Alu was the subfamily with the lowest (53.4%). Information of validated TE insertions inside other TE sequences, including of their same type, is shown in Table S3.

Table 2. Insertions of class I TEs found on coding sequences

Position	Length	Type	Gene	Patients
chr1:12 881 890	311	SINE (Alu)	PRAMEF4	3
chr2:97 209 695	698	LINE (L1)	ANKRD36	1
chr11:86 939 037	305	SINE (Alu)	PRSS23	1
chr12:88 042 861	992	LINE (L1)	C12orf29	1
chr13:32 319 156	309	SINE (Alu)	BRCA2	1
chr14:24 212 603	163	SINE (Alu)	CHMP4A	22

The length (bp) and type of TE, the affected gene and the number of patients carrying the insertion are shown.

It is remarkable that, for most families, the percentage of insertions in the same type of element, e.g., an AluY insertion found inserted in an AluY sequence in the reference, was similar to the percentage of insertions found in sequences of the same subfamily or family, with the exceptions of Alu and L1, the latter presenting the highest difference (Table 3). The percentages found in repeats of the same family and the same subfamily were similar (Table 3). Similarly, the percentage of insertions found on any repeat was similar (difference lower than 15%) to the one for insertions in the same family, with few exceptions (ERVK, and Alu).

For the individual chromosomes, the insertion of TEs was uniform on chromosomes 3, 4, 5, 7, 8, 10, 11, and 20 (p values of 5.16, 5.38, 0.21, 13.67, 1.23, 2.86, 4.11, and 0.06, respectively, chi-squared). For the rest, the distribution was not uniform (for chromosomes 1, 2, 6, 9, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, and X, p values were 1e-05, 2e-02, 3e-59, 2e-07, 5e-04, 1e-07, 1e-04, 1e-06, 3e-02, 4e-08, 2e-02, 9e-08, 5e-02, 9e-08, 8e-72, 1e-32, chi-squared). The varying amount of insertions along chromosome length can be seen on Figure 2.

Insertions also found outside this study

Our results have been compared against two previous studies evaluating TEs on different cohorts. Out of the 7 344 TE insertions reported in our study, 3 424 were also found in at least one out of 9 subjects evaluated by the HGSVC.[20] That work presents the results of sequencing the whole genome by multiple techniques, including NGS, PacBio, and optical mapping; and identified a total of 28 415 insertions, of which 6 261 were annotated by RepeatMasker with TE sequences. Most of the insertions in common were classified as Alu (2 584) and SVA (399). We observed that the insertions present only in our dataset had low AF, 90% of them below 0.292 (Figure 4B). The number of insertions shared by these two studies are shown in Table S7. With a range of ±60 bp for coordinate matching, the median difference in coordinates for matching insertions was 6 bp, while for length the median difference was 13 bp.

The other dataset we could access contained 21 970 Alu insertions detected by NGS in more than 1 000 individuals[10] from the Indigen project.[36] We found 3 013 Alu insertions in common with our data (Table S7). That study was focused on Alu insertions, so most (99.9%)

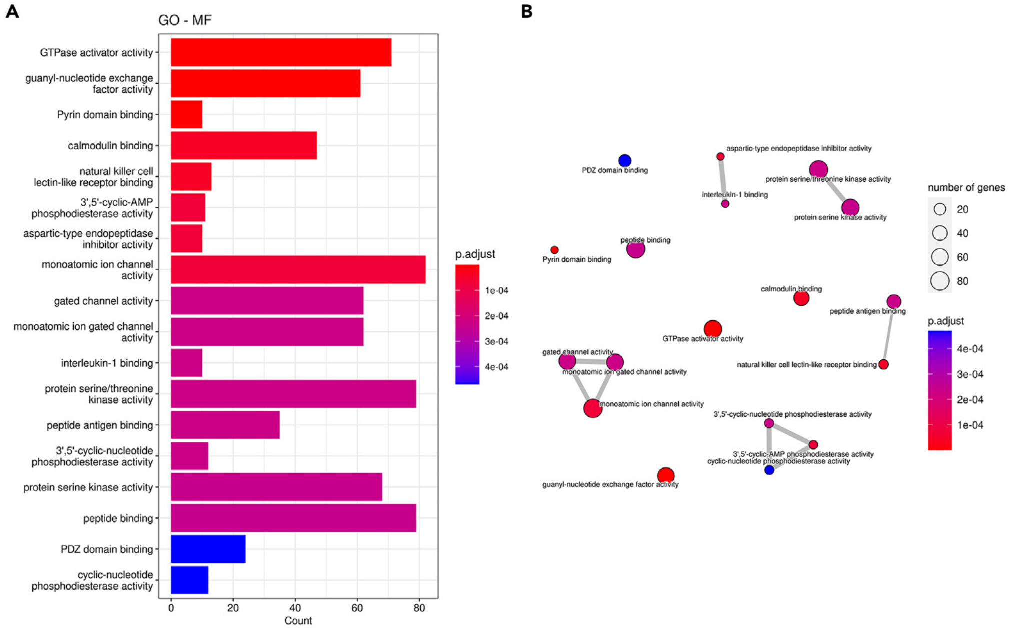


Figure 3. Annotation with GO's molecular function (MF)

(A and B) MF categories with lowest adjusted p values found for genes affected by TE (A) and their connections (B) via shared genes.

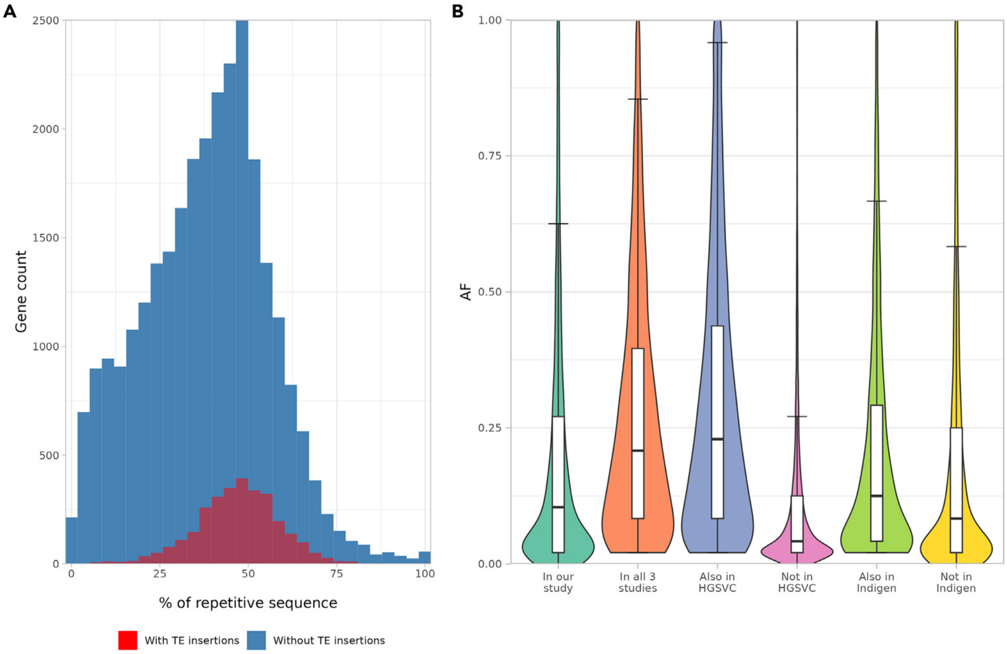


Figure 4. Distributions of TE insertions data

(A) Plot of the percentage of genes made up of repetitive elements for genes affected and not affected by TE insertions. Columns are superposed, not stacked.
(B) Boxplots for AF distribution for insertions grouped by their presence in other datasets. Boxes indicate limits for the Q1-Q3 range, and dashed lines extend to percentiles 5 (bottom) and 95 (top). The thick horizontal lines indicate medians, and the colored shapes, the distribution's shape.

common insertions were annotated by RepeatMasker as Alu in our data. The other 2 were L1 insertions according to our annotation. Using ±60 bp as the maximum for coordinate range, the mean difference in coordinates for matching insertions was 9, while for median length difference was 31 bp.

The insertions not found on the HGSVC dataset had lower AF than those which did (Figure 4B). The difference between insertions shared with the indigenous set and those not shared with the indigenous set was smaller (Figure 4B). We also checked the intersection of the three datasets, and found 2 020 TE insertions in common, with 1 411 additional ones shared only with HGSVC, and 985 only with Indigen, and 370 insertions in both Indigen and HGSVC but not in our data.

Deleted transposable element sequences in the reference genome

While the previous sections have focused on TE insertions, this one summarizes those TEs already described in the reference genome that might have been deleted as consequence of an imbalanced rearrangement or any other mechanism. 3 086 TE deletions were

Table 3. Repetitive sequences affected by TE insertions in each subfamily of TE

Repeat family	Repeat subfamily	Total	In repeats	In same family	In same subfamily	In same element
DNA	hAT	15	11 (73.3%)	11 (73.3%)	11 (73.3%)	9 (60.0%)
	TcMar	16	13 (81.2%)	11 (68.8%)	11 (68.8%)	9 (56.2%)
LINE	CR1	1	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	L1	599	404 (67.4%)	340 (56.8%)	336 (56.1%)	70 (11.7%)
	L2	16	11 (68.8%)	10 (62.5%)	9 (56.2%)	3 (18.8%)
LTR	ERV1 ERVK	313 41	294 (93.9%) 32 (78.0%)	288 (92.0%) 24 (58.5%)	288 (92.0%) 22 (53.7%)	220 (70.3%) 12 (29.3%)
	ERVL	98	81 (82.7%)	73 (74.5%)	73 (74.5%)	46 (46.9%)
Retroposon	SVA	1294	1236 (95.5%)	1172 (90.6%)	1172 (90.6%)	377 (29.1%)
SINE	Alu	4909	2616 (53.3%)	1041 (21.2%)	952 (19.4%)	145 (3.0%)
	MIR	6	3 (50.0%)	3 (50.0%)	3 (50.0%)	3 (50.0%)

For each subfamily, the number of unique insertions detected is shown, and then the absolute value and percentages of unique insertions found in repetitive sequences, or in elements of the same family, subfamily, or name.

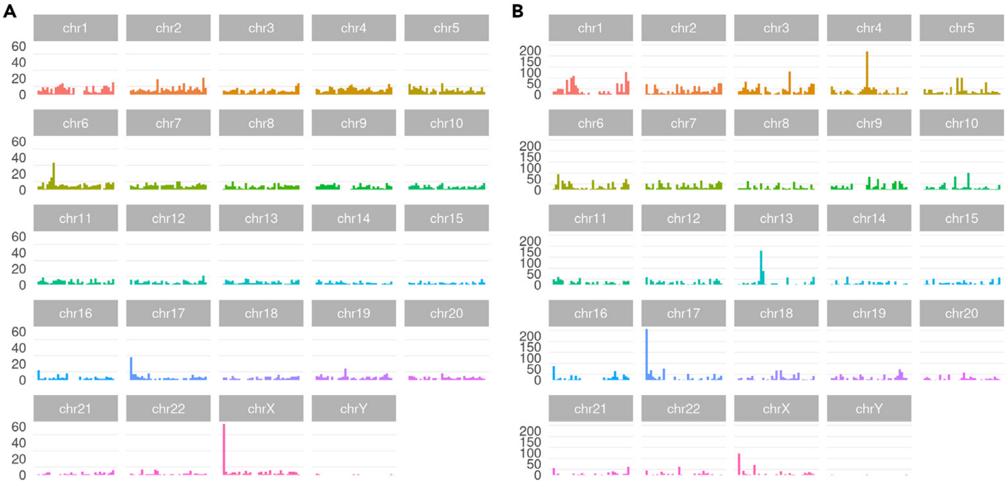


Figure 5. Number of deletions and their annotation across the genome

(A) Deletions of repetitive sequences in the reference human genome masked as TE sequences that were identified in our study.

(B) Gene annotations for TE deletions on coding sequences, introns, and promoters.

found, adding up to 35 304 occurrences among patients. These deletions had a bimodal prevalence (p = 0, Amejeiras-Alonso) per patient (modes of 1.60 and 23.1, densities of 0.0734 and 0.0528, respectively; antimode of 11.6, density of 0.0245). The deletion length was in the range of 217–328 bp for 50% of them (Q1-Q3, both included), and the maximum length found was 7 472, for a deleted L1 sequence.

These deletions were also genotyped, revealing that they were on average more frequent than insertions ($p < 2.2e-16$, Mann-Whitney). The median AF for deletions was 0.27, with 114 (all of them passed the stringent filtering) deletions appearing in a homozygous state in all patients (Figure 1B). On the other hand, 30 deletions were genotyped as 0/0 on all cases, and have not been included in the rest of this section.

We found that both class I and class II sequences could be deleted. The number and counts of each type and subtype of TEs deleted that our study identified are detailed in Table 1. The most frequent ones were classified under the Alu subfamily.

Interestingly, 4 deleted TEs were located in the coding regions of 4 genes (*ZMPSTE24*, *NPIPB12*, *FSTL4*, and *SPINK14*), 79 on <1 kb promoter of 74 genes, and 1 633 on introns of 1 535 genes (Figure 5).

DISCUSSION

Recently, an assembly of the human genome that includes the centromeres has been completed.[37] Whole genome sequencing by NGS methods have also made it possible to have available a complete map of the genomic variability involving SNVs in different populations.[38] However, there is still an important limitation: the identification, mapping, and population distribution of TEs across the genome, which has been difficult due to the abundance of these elements in the genome, their repetitive nature, and their mobility.[2,3] This information may be crucial to detect pathogenic insertions of TEs, which may be the cause of many disorders in patients with unknown molecular base.[39] While other tools such as[14–16] have been developed to detect TE insertions, our method has been developed with a focus on the human genome and is therefore tailored to human data. This specialization in the human genome has allowed us to develop a method that not only detects the insertions, but also describes and characterizes them in more detail and depth.

Using long-read WGS from nanopore sequencing we have explored the SVs occurring on 24 subjects with a monogenic mendelian disorder: antithrombin deficiency. The general proportions of SVs found agrees with what has been previously described, the most frequent ones being deletions and insertions, the latter being a significant difference with NGS.[11] We report that 25.1% of found SVs were annotated with TE by RepeatMasker on a meaningful (≥85%) fraction of their length. Moreover, our study also found a considerable proportion of TEs reported in the reference genome that were deleted in these patients.

Mobility of transposable element sequences

7 313 of the detected insertions were annotated as class I, and 31 as class II. This difference likely arises from the fact that some class I elements are active in the human genome, and class II are not.[2] It is remarkable that many TE insertions (54.0% and 45.2%, for class I and II respectively) were located on genes with a relatively high (≥15%, mean of 47.1%) content of repetitive regions, which probably played a role in their origin. Additionally, we observed that chromosome Y was underpopulated with insertions, while chromosomes 6, 17 and 19 had more than expected. The sex chromosomes are enriched in L1 elements,[40] and we did find 16% more insertions than expected from a uniform distribution, but these are only a fraction of our data and differences are to be expected, additionally, not all patients carry Y chromosomes, so sex chromosomes can be found in hemizygosity. On this, it is also worth mentioning that most elements

seem to target repeats of their own family, except for Alu. We observed that the number of insertions in repeats of the same subfamily and of the same element were similar, although L1, L2, and SVA insertions behaved differently in this regard, with differences of 44.7%, 41.2%, and 61.2%, respectively. These patterns may be explained by non-random insertion patterns for actual retrotransposition[40] or by sequence similarity for other molecular mechanisms.

However, we also report that a considerable fraction of insertions are not long enough to contain full elements, a situation that is more frequent on class II and non-mobile class I elements, while most Alu and several SVA do appear complete, which may point to true retrotransposition events. In relation to this, we find relevant that the group with more full-length insertions, Alu, is also the most active,[41] and the second one, SVAs, has a recently updated insertion rate's estimation of one in 63 births.[42] On the other hand, for the rest of the insertions, it is far more likely that they have occurred by different mechanisms. For example, the TcMar element *Hsmar2* is associated with human rare diseases via hotspots for recombination,[8] although it is true that truncation can occur during transposition.[43] In this manner, ectopic recombination may explain the insertions for non-active elements and incomplete sequences. Another explanation, which does not apply only to incomplete sequences, is that the ancestral human genome included these sequences, at some point in time a deletion occurred, and the reference genome now includes the deletion, with the inserted and deleted alleles coexisting as polymorphic.

Regardless of the mechanism of origin, it is important to state that we have validated the presence of insertions, both of active elements with complete elements, as well as class II elements, inactive class I ones, or shorter sequences.

Potential functional and pathological consequences of transposable elements insertions

Interestingly, 6 TEs affected coding regions with evident functional consequences. Remarkably, 4 TEs affected genes involved in cancer. These insertions might constitute germline defects predisposing these individuals and relatives carrying these TEs to develop cancer. This is new evidence supporting the role that TE insertions have in cancer.[44]

The enrichment analysis also linked the genes affected by insertions to autism. Recently, both intronic and exonic insertions have been linked to autism spectrum disorder.[39] This further encourages the screening of TEs for autism cases, and may have additional impact, since autism spectrum disorders' targets overlap with schizophrenia and intellectual disability.[45]

The number of TEs with potential pathogenic relevance may be higher as the low coverage generated by PromethION when sequencing the whole genome, and the strict conditions of identification used in this study make it possible to miss real TEs insertions. Indeed, among the 24 cases included in this study, we have demonstrated that the insertion of a TE in an intron of *SERPINC1* identified in two cases is pathogenic, causing antithrombin deficiency and increased risk of thrombosis,[18] which was detected in this study only in one case of two cases, due to low coverage on that region. A laxer screening, even sustained in only one read, may be useful for identifying potential pathogenic TEs insertions in candidate genes, such as *SERPINC1* in patients with antithrombin deficiency and unknown molecular bases after conventional molecular analysis. This screening might increase the number of diseases caused by TEs insertions, which probably are currently underestimated due to the difficulty to detect these genetic variants.

Polymorphic transposable elements. Estimated allele frequencies

After calling insertions with SVIM and CuteSV, it was interesting to get numeric values for their AFs in addition to patient occurrence. Since SVIM does not do this (its genotyping is not reliable per its documentation), and it was interesting to genotype insertions that had not been supported by CuteSV, we preferred to adapt the genotyping algorithm from Sniffles 2, which is based on genotype likelihood.[46,47] Prior to this, force-calling the insertions of interest and genotyping with Sniffles was attempted, but the most recent version failed to detect most of them, so no useful data was reported. However, the algorithm itself, when fed with the read support values by the other callers and coverage values on those regions did provide genotypes more in agreement with variant callers. Sniffles' failure to do this most likely stems from it not considering the supporting reads as evidence for the insertions.

As has been shown in detail in the Results section, the results for genotyping did not perfectly agree with either filtering criteria. This was expected, since the stringent one left out calls for some patients, mainly because one of the two callers did not report them, while the laxer one did add many spurious calls that should not be trusted. To solve this, a set of high-confidence TE insertions was created with those having one call passing the stringent filter in at least one patient. These insertions were then genotyped considering supporting reads in every patient, including the reports which did not pass. The basis for this is that, by using a stringent screening, the selected SVs are likely to be real, so it is reasonable to trust the calls in other patients.

It is interesting to mention that high-confidence insertions that were genotyped as 0/0 across all cases passed the filters because they occurred on regions with high depth coverage, rendering the support value of 3 trivial. Using a higher threshold would have still kept these, while leaving out many useful calls, so this path was not followed. Additionally, it is possible that the fraction of supporting reads is low because they are somatic mutations.

Regarding the values themselves of the AFs reported, two points are worth discussing. First, the study was done on a group of patients with a hereditary disease, and it could be argued that this may result in an inflated frequency when extrapolating the presented data. This is attenuated by the fact that most patients had a different genetic diagnosis explaining the antithrombin deficiency, 10 carrying SVs and 4 SNVs. The AF for the SVA insertion in *SERPINC1* is not reported for this reason. The second one is also related to insertions with a high frequency: the ones found in all patients. SVs, insertions included, are detected by comparing sequencing reads to the reference genome. There are two possibilities that may explain this phenomenon. First, as seen for the insertion affecting *CHMP4*, it is possible that the correct sequence is longer than the reference, explaining why mean coverage is higher in that region on gnomAD. However, the coverage data was different for the class II insertions that affected genes, so this may not be valid for all cases. A second explanation arises: while most patients from our study are Spanish, the reference genome was assembled from a more diverse group, which explains why all our cases carry a given sequence not found in the reference.

Transposable elements insertions may vary with population

The comparison of our data with previous studies that have evaluated TEs though the genome by using different methods[10,20] shows that 4 418 insertions were found in common with those studies, which is additional evidence for their existence. However, there are 2 926 TEs that were present only in our study. Different explanations to the new TEs identified in this study may be proposed: from the strength of nanopore sequencing to detect long SVs involving repetitive elements,[11] to genetic variability across populations, already described for SNPs[38] that can also involve TEs (not only insertions, but also deletions), and of course, *de novo* insertions/deletions of TEs only detected in this study. The method used to identify the TEs may also explain the differences observed in these studies, which could be able to uncover a different fraction of the insertions present in any individual's genome. Independently of this controversy, our study can help to provide a more complete map of potentially polymorphic TE insertions, which may help future studies involving potentially pathological TE insertions in patients with genetic diseases.

In conclusion, the long-read based nanopore sequencing of the whole genome in a substantial number of patients has allowed a dissection of TE insertions and deletions on a genome-wide scale, with coordinates and inserted sequences. By using available, benchmarked tools we identified and characterized many TEs, demonstrating that the number of new insertions is remarkably high, mainly involving class I TEs and affecting genes rich in repetitive elements. Our study also suggests that most of these TEs are polymorphic, but we also identified pathogenic TE insertions. This study provides new evidence that helps to understand the functional relevance of mobile elements in the human genome and will assist further studies investigating the role of TEs in genetic disorders.

Limitations of the study

Our study has two limitations: a) the sample size, which is modest when compared to that of studies using NGS short-read based methods,[9,10] but sizable for studies using long-read based sequencing methods[12,20]; and b) the low coverage of nanopore sequencing, which also limits the identification of TE insertions. Thus, the number of new TEs in the genome may be significantly higher than that indicated in this study, and the role of these elements in the genetic diversity of populations and to the development of genetic disorders may be higher than that described in this study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Long-read whole genome sequencing
  - ○ Calling of structural variants
  - ○ PCR validation
  - ○ Assembly of insertion sequences
  - ○ Genotyping insertions
  - ○ Annotation of insertions
  - ○ Size assessment of TE sequences
  - ○ TE insertions in other datasets
  - ○ Overlap with reference repetitive sequence
  - ○ Gene enrichment analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108214.

ACKNOWLEDGMENTS

We gratefully acknowledge the NIHR BioResource centers and staff for their contribution. We thank the National Institute for Health Research and NHS Blood and Transplant. 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.

Javier Cuenca-Guardiola is funded by the Ministerio de Universidades through grant FPU19/03662. This work has been possible thanks to the funding of the Fundación Séneca (grant number 21642/PDC/2) and of Instituto de Salud Carlos III (ISCIII) through the projects "PI21/00174" and "PMP21/00052," which were co-funded by the European Union and by the European Union - Next Generation EU, respectively. Belén de la Morena has a postdoctoral contract from CIBERER.

AUTHOR CONTRIBUTIONS

J.C.G, B.M.B, J.C., W.O, and J.T.F.B conceived the experiment(s). J.S., B.M.B and E.N.M. conducted the experiments; J.C.G, B.M.B, and N.S.G. analyzed the results; J.C.G, B.M.B, J.C. and J.T.F.B wrote the article; J.C.G, B.M.B, J.S., W.O, N.S.G., J.T.F.B. and J.C. reviewed the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 6, 2023
Revised: July 28, 2023
Accepted: October 11, 2023
Published: October 14, 2023

REFERENCES

1. Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res. *38*, 4207–4217. https://doi.org/10.1093/nar/gkq140.
2. Kazazian, H.H., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N. Engl. J. Med. *377*, 361–370. https://doi.org/10.1056/NEJMra1510092.
3. Ayarpadikannan, S., and Kim, H.-S. (2014). The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. Genomics Inform. *12*, 98–104. https://doi.org/10.5808/GI.2014.12.3.98.
4. Smit A.F.A., Hubley R., Green P. (2013). RepeatMasker Open-4.0. http://www.repeatmasker.org.
5. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. *44*, D81–D89. https://doi.org/10.1093/nar/gkv1272.
6. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA *12*, 2. https://doi.org/10.1186/s13100-020-00230-y.
7. Burns, K.H. (2020). Our Conflict with Transposable Elements and Its Implications for Human Disease. Annu. Rev. Pathol. *15*, 51–70. https://doi.org/10.1146/annurev-pathmechdis-012419-032633.
8. Gil, E., Bosch, A., Lampe, D., Lizcano, J.M., Perales, J.C., Danos, O., and Chillon, M. (2013). Functional Characterization of the Human Mariner Transposon Hsmar2. PLoS One *8*, e73227. https://doi.org/10.1371/journal.pone.0073227.
9. Niu, Y., Teng, X., Zhou, H., Shi, Y., Li, Y., Tang, Y., Zhang, P., Luo, H., Kang, Q., Xu, T., and He, S. (2022). Characterizing mobile element insertions in 5675 genomes. Nucleic Acids Res. *50*, 2493–2508. https://doi.org/10.1093/nar/gkac128.
10. Prakrithi, P., Singhal, K., Sharma, D., Jain, A., Bhoyar, R.C., Imran, M., Senthilvel, V., Divakar, M.K., Mishra, A., Scaria, V., et al. (2022). An Alu insertion map of the Indian population: identification and analysis in 1021 genomes of the IndiGen project. NAR Genom. Bioinform. *4*, lqac009. https://doi.org/10.1093/nargab/lqac009.
11. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nat. Rev. Genet. *21*, 597–614. https://doi.org/10.1038/s41576-020-0236-x.
12. Chu, C., Borges-Monroy, R., Viswanadham, V.V., Lee, S., Li, H., Lee, E.A., and Park, P.J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat. Commun. *12*, 3836. https://doi.org/10.1038/s41467-021-24041-8.
13. Ewing, A.D., Smits, N., Sanchez-Luque, F.J., Faivre, J., Brennan, P.M., Richardson, S.R., Cheetham, S.W., and Faulkner, G.J. (2020). Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Mol. Cell *80*, 915–928.e5. https://doi.org/10.1016/j.molcel.2020.10.024.
14. Han, S., Dias, G.B., Basting, P.J., Viswanatha, R., Perrimon, N., and Bergman, C.M. (2022). Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. Nucleic Acids Res. *50*, e124. https://doi.org/10.1093/nar/gkac794.
15. Mohamed, M., Sabot, F., Varoqui, M., Mugat, B., Audouin, K., Pélisson, A., Fiston-Lavier, A.-S., and Chambeyron, S. (2023). TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. Genome Biol. *24*, 63. https://doi.org/10.1186/s13059-023-02911-2.
16. Disdero, E., and Filée, J. (2017). LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. Mob. DNA *8*, 5. https://doi.org/10.1186/s13100-017-0088-x.
17. Corral, J., de la Morena-Barrio, M.E., and Vicente, V. (2018). The genetics of antithrombin. Thromb. Res. *169*, 23–29. https://doi.org/10.1016/j.thromres.2018.07.008.
18. de la Morena-Barrio, B., Orlando, C., Sanchis-Juan, A., García, J.L., Padilla, J., de la Morena-Barrio, M.E., Puruunen, M., Stouffs, K., Cifuentes, R., Borràs, N., et al. (2022). Molecular Dissection of Structural Variations Involved in Antithrombin Deficiency. J. Mol. Diagn. *24*, 462–475. https://doi.org/10.1016/j.jmoldx.2022.01.009.
19. de la Morena-Barrio, B., Stephens, J., de la Morena-Barrio, M.E., Stefanucci, L., Padilla, J., Miñano, A., Gleadall, N., García, J.L., López-Fernández, M.F., Morange, P.-E., et al. (2022). Long-Read Sequencing Identifies the First Retrotransposon Insertion and Resolves Structural Variants Causing Antithrombin Deficiency. Thromb. Haemost. *122*, 1369–1378. https://doi.org/10.1055/s-0042-1749345.
20. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784. https://doi.org/10.1038/s41467-018-08148-z.
21. Pfaff, A.L., Singleton, L.M., and Kõks, S. (2022). Mechanisms of disease-associated SINE-VNTR-Alus. Exp. Biol. Med. *247*, 756–764. https://doi.org/10.1177/15353702221082612.

22. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
23. Heller, D., and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. Bioinformatics *35*, 2907–2915. https://doi.org/10.1093/bioinformatics/btz041.
24. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189. https://doi.org/10.1186/s13059-020-02107-y.
25. De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., and Van Broeckhoven, C. (2019). PromethION WGS Data of NA19240 (Run 1) (European Nucleotide Archive).
26. Tsang, H.T.H., Connell, J.W., Brown, S.E., Thompson, A., Reid, E., and Sanderson, C.M. (2006). A systematic analysis of human CHMP protein interactions: additional MIT domain-containing proteins bind to multiple components of the human ESCRT III complex. Genomics *88*, 333–346. https://doi.org/10.1016/j.ygeno.2006.04.003.
27. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.
28. Li, B., Li, T., Pignon, J.-C., Wang, B., Wang, J., Shukla, S.A., Dou, R., Chen, Q., Hodi, F.S., Choueiri, T.K., et al. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. Nat. Genet. *48*, 725–732. https://doi.org/10.1038/ng.3581.
29. Le, H.P., Heyer, W.-D., and Liu, J. (2021). Guardians of the Genome: BRCA2 and Its Partners. Genes *12*, 1229. https://doi.org/10.3390/genes12081229.
30. Rappaport, N., Fishilevich, S., Nudel, R., Twik, M., Belinky, F., Plaschkes, I., Stein, T.I., Cohen, D., Oz-Levi, D., Safran, M., and Lancet, D. (2017). Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect. Biomed. Eng. *16*, 1–4.
31. Alliance of Genome Resources Consortium (2020). Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res. *48*, D650–D658. https://doi.org/10.1093/nar/gkz813.
32. Lin, Q., Liang, Q., Qin, C., and Li, Y. (2021). CircANKRD36 Knockdown Suppressed Cell Viability and Migration of LPS-Stimulated RAW264.7 Cells by Sponging MiR-330. Inflammation *44*, 2044–2053. https://doi.org/10.1007/s10753-021-01480-5.
33. Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., et al. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. *49*, D325–D334. https://doi.org/10.1093/nar/gkaa1113.
34. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. *49*, D545–D551. https://doi.org/10.1093/nar/gkaa970.
35. Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. *47*, D955–D962. https://doi.org/10.1093/nar/gky1032.
36. Jain, A., Bhoyar, R.C., Pandhare, K., Mishra, A., Sharma, D., Imran, M., Senthivel, V., Divakar, M.K., Rophina, M., Jolly, B., et al. (2021). IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. Nucleic Acids Res. *49*, D1225–D1232. https://doi.org/10.1093/nar/gkaa923.
37. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. https://doi.org/10.1126/science.abj6987.
38. Gudmundsson, S., Singer-Berk, M., Watts, N.A., Phu, W., Goodrich, J.K., Solomonson, M., Genome Aggregation Database Consortium, Rehm, H.L., MacArthur, D.G., and O'Donnell-Luria, A. (2022). Variant interpretation using population databases: Lessons from gnomAD. Human Mutation *43*, 1012–1030. https://doi.org/10.1002/humu.24309.
39. Borges-Monroy, R., Chu, C., Dias, C., Choi, J., Lee, S., Gao, Y., Shin, T., Park, P.J., Walsh, C.A., and Lee, E.A. (2021). Whole-genome analysis reveals the contribution of non-coding de novo transposon insertions to autism spectrum disorder. Mob. DNA *12*, 28. https://doi.org/10.1186/s13100-021-00256-w.
40. Graham, T., and Boissinot, S. (2006). The Genomic Distribution of L1 Elements: The Role of Insertion Bias and Natural Selection. J. Biomed. Biotechnol. *2006*, 75327. https://doi.org/10.1155/JBB/2006/75327.
41. Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703. https://doi.org/10.1038/nrg2640.
42. Feusier, J., Watkins, W.S., Thomas, J., Farrell, A., Witherspoon, D.J., Baird, L., Ha, H., Xing, J., and Jorde, L.B. (2019). Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res. *29*, 1567–1577. https://doi.org/10.1101/gr.247965.118.
43. Ardeljan, D., Taylor, M.S., Ting, D.T., and Burns, K.H. (2017). The human LINE-1 retrotransposon: an emerging biomarker of neoplasia. Clin. Chem. *63*, 816–822. https://doi.org/10.1373/clinchem.2016.257444.
44. Chenais, B. (2015). Transposable elements in cancer and other human diseases. Curr. Cancer Drug Targets *15*, 227–242. https://doi.org/10.2174/1568009615666150317122506.
45. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221. https://doi.org/10.1038/nature13908.
46. 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. https://doi.org/10.1038/s41592-018-0001-7.
47. Smolka, M., Paulin, L.F., Grochowski, C.M., Mahmoud, M., Behera, S., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S.W., Carvalho, C.M.B., et al. (2022). Comprehensive Structural Variant Detection: From Mosaic to Population-Level. Preprint at bioRxiv. https://doi.org/10.1101/2022.04.04.487055.
48. UCSC. RepeatMasker (2014). Out File (USCS).
49. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., and Fan, X. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.
50. Prakrithi, P., Singhal, K., Sharma, D., et al. (2022). Indigen Alu Final Geno10 All 22K VCF File.
51. D'Antonio, M., Pendino, V., Sinha, S., and Ciccarelli, F.D. (2012). Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. Nucleic Acids Res. *40*, D978–D983. https://doi.org/10.1093/nar/gkr952.
52. Ren, J., and Chaisson, M.J.P. (2021). lra: A long read aligner for sequences and contigs. PLoS Comput. Biol. *17*, e1009078. https://doi.org/10.1371/journal.pcbi.1009078.
53. Tham, C.Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M.J., Koh, B.T.H., Wang, W., Ng, C.H., Chng, W.J., Thiery, A., Tenen, D.G., and Benoukraf, T. (2020). NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol. *21*, 56. https://doi.org/10.1186/s13059-020-01968-7.
54. 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*, 14061. https://doi.org/10.1038/ncomms14061.
55. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
56. Pysam, developers. Pysam. (2021). (Pysam developers).
57. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3—new capabilities and interfaces. Nucleic Acids Res. *40*, e115. https://doi.org/10.1093/nar/gks596.
58. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158. https://doi.org/10.1038/s41592-019-0669-3.
59. D Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. JOSS *3*, 731. https://doi.org/10.21105/joss.00731.
60. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.
61. Cavalcante, R.G., and Sartor, M.A. (2017). annotatr: genomic regions in context. Bioinformatics *33*, 2381–2383. https://doi.org/10.1093/bioinformatics/btx183.

62. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

63. Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics *31*, 608–609. https://doi.org/10.1093/bioinformatics/btu684.

64. Genome Reference Consortium (2013). Genome Reference Consortium Human Build 38 (GRCh38) (Genome Reference Consortium).

65. De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., and Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. *29*, 1178–1187. https://doi.org/10.1101/gr.244939.118.

66. Bonfield, J.K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. GigaScience *10*, giab007. https://doi.org/10.1093/gigascience/giab007.

67. Dowle M., Srinivasan A. (2021). data.table: Extension of 'data.Frame'. https://r-datatable.com.

68. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

69. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

70. Ameijeiras-Alonso, J., Crujeiras, R.M., and Rodriguez-Casal, A. (2021). multimode: An R Package for Mode Assessment. J. Stat. Soft. *97*, 1–32. https://doi.org/10.18637/jss.v097.i09.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Insertion data	This study.	https://doi.org/10.5281/zenodo.8375338
Dfam	Storer et al.[6]	https://www.dfam.org
NA19240's PromethION data	De Coster et al.[25]	ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR258/ERR2585112/NA19240_run1.fastq.gz
GO	Gene Ontology Consortium[33]	https://geneontology.org/
KEGG	Kanehisa et al.[34]	https://www.genome.jp/kegg/
DO	Schriml[35]	https://disease-ontology.org/
RepeatMasker output for the human genome	USCS[48]	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.out.gz
HGVSC's SV set	Chaisson et al.[49]	nstd152
Indigen Alu dataset	Prakrithi et al.[50]	https://clingen.igib.res.in/indigen/download/Indigen_Alu_final_geno10_all_22K.vcf
NCG	D'Antonio et al.[51]	http://ncg.kcl.ac.uk/
Oligonucleotides		
DNA primers for PCR amplification	This study.	See Table S3 for sequence.
Software and algorithms		
TE analysis pipeline	This study.	https://doi.org/10.5281/zenodo.8375338
RepeatMasker	Smit et al.[4]	https://www.repeatmasker.org/
minimap2	Li[22]	https://github.com/lh3/minimap2
SVIM	Heller and Vingron[23]	https://github.com/eldariont/svim
cuteSV	Jiang et al.[24]	https://github.com/tjiangHIT/cuteSV
NGMLR	Sedlazeck et al.[46]	https://github.com/philres/ngmlr
Sniffles	Sedlazeck et al.[46]	https://github.com/fritzsedlazeck/Sniffles
Lra	Ren and Chaisson[52]	https://github.com/ChaissonLab/LRA
NanoVar	Tham et al.[53]	https://github.com/benoukraflab/NanoVar
SURVIVOR	Jeffares et al.[54]	https://github.com/fritzsedlazeck/SURVIVOR
bcftools, bgzip and tabix	Danecek et al.[55]	https://samtools.github.io/bcftools/bcftools.html
pysam	Pysam developers.[56]	https://github.com/pysam-developers/pysam
Primer3 (python bindings)	Untergasser et al.[57]	https://pypi.org/project/primer3-py/
wtdbg2	Ruan and Li[58]	https://github.com/ruanjue/wtdbg2
qqman	Turner[59]	https://cran.r-project.org/web/packages/qqman/index.html
mosdepth	Pedersen and Quinlan[60]	https://github.com/brentp/mosdepth
annotatr	Cavalcante and Sartor[61]	https://doi.org/doi:10.18129/B9.bioc.annotatr
clusterProfiler	Yu et al.[62]	https://doi.org/doi:10.18129/B9.bioc.clusterProfiler
DOSE	Yu et al.[63]	https://doi.org/doi:10.18129/B9.bioc.DOSE

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jesualdo Tomás Fernández Breis (jfernand@um.es).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The sequencing data that support the findings of this study are available from NIHR BioResource (study code DAA067, restrictions apply). The insertion data has been deposited at Zenodo along with the code and is publicly available as of the date of publication. DOI is listed in the key resources table. This paper also analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The study was done in 24 patients with thrombosis that had AT quantitative type I deficiency. 11 of them carried SVs affecting SERPINC1, 2 of which were SVA insertions. 4 other cases presented causal SNPs, and the remaining eight did not have a molecular diagnosis. All 24 cases were analyzed. The study protocol was approved by the Clinical Research Ethics Committee at Morales Meseguer University Hospital (code EST: 31/18) and conducted in accordance with the 1964 Declaration of Helsinki and their later amendments. All included subjects gave their written informed consent to enter the study.

The cohort is comprised of non-Finnish European patients, 7 of them of Polish ancestry, 3 of French, 1 of Belgium, the rest of them Spanish. Information on sex, age and ancestry is listed on Table S8. Ancestry was determined by country of origin.

All sequencing experiments yielded depth coverage above 10x, with a mean of 17.6 ± 5.1. The maximum genome-wide mean depth was 28.8, and the lowest, 10.3 (Table S1). An average of 84.5 ± 8.4% of the genome was covered by at least 10x depth. Individual values are shown on Table S1.

METHOD DETAILS

Long-read whole genome sequencing

Long-read whole genome sequencing (WGS) was done using the PromethION platform (Oxford Nanopore Technologies, Oxford Science Park, OX4 4DQ, UK). Samples were prepared using the 1D ligation library prep kit (SQK-LSK109), and genomic libraries were sequenced on R9 flow cells. Read sequences were extracted from basecalled FAST5 files by Guppy (versions3.0.4to3.2.8) to generate FASTQ files.

Calling of structural variants

After aligning with minimap2[22] against the human reference genome hg38,[64] structural variants were called with cuteSV[24] and SVIM.[23]

To select these tools, we benchmarked 3 aligners (minimap2, NGMLR,[46] and lra[52]) and 4 variant callers (cuteSV, SVIM, Sniffles,[46] and Nanovar[53]), choosing the one with best recall for insertions (Table S10). As data for this benchmark, nanopore reads from a single PromethION cell from a publicly available dataset of NA19240[25] were used. Although NanoVar scored better than SVIM, it does not have an option to return SV alleles' sequences, so it was discarded in favor of the latter. It is worth mentioning that cuteSV and SVIM have been proven to be well-performing tools for variant calling, particularly for some situations, are the best.[24]

Variant callers were configured to consider only reads with mapping quality ≥30. The SV calls generated were filtered with two criteria to produce two different datasets. First, a laxer filter required a minimum of 3 supporting reads. This threshold was chosen from the insertions that were found on *SERPINC1* and confirmed experimentally. Second, a more stringent one required that both cuteSV and SVIM detected the SV, at least one of them with 3 supporting reads. In summary, the union of call sets generates a laxer, more sensitive set, while the intersection results in more stringent, higher-confidence one.[65]

After this, SVs from both call sets were merged with SURVIVOR[54] to generate general metrics for all types of SVs. For VCF manipulation, bcftools,[55] bgzip,[55] tabix,[55] and the python library pysam[56,66] were used.

For specifically merging insertions we developed an algorithm that considers SV type and genomic coordinates, like SURVIVOR, but also takes the inserted sequence into account. For two insertions to be considered the same one, the inserted sequences must be no more dissimilar in length than 5%, and to have a Levenshtein ratio of 65% or above, calculated with the python packages fuzzy-wuzzy and python-Levenshtein. This ratio is equivalent to a global alignment score without penalties for gaps, and was selected to ensure that the sequences of merged insertions were similar, while considering the relatively low precision of nanopore reads. We used this method to combine the cuteSV and SVIM results for each patient, and after that, to generate an inter-patient merged set.

PCR validation

17 TE insertions were selected for validation. For selection, we included the six insertions present in exons, as well as the one in *SERPINC1*'s intron. Then, we divided class I insertions (minus the exonic) into three groups: ubiquitous insertions, with an AF of 1; common, with 0.75 < maf <1; and rare, with AF $\leq$ 0.4, and selected randomly two variants from each group. For class II variants, since they were fewer, instead of stratifying by AF, we randomly selected four.

Validation was performed by PCR, by designing primers that would only allow for amplification if the insertion was present. We used the sequences reported by nanopore, one primer aligning against the insertion and the other one against the upstream flanking region, which means the amplified sequences, while indicative of the insertion's presence, do not match its full sequence or length. Primers were designed with python bindings for Primer3.[57]

PCRs were done in potential carriers of the inserted TE from our cohort, and in healthy controls from the Spanish general population. An additional blank control was also used in all PCR reactions.

Assembly of insertion sequences

For each insertion, the identifiers for supporting reads were retrieved, and trimmed to the region surrounding the insertion. After this, they were assembled using wtdbg2[58] with the preset for nanopore data. The assembled sequence was aligned against one of the inserted sequences on the supporting reads (the one closest in length to what the variant caller reported), to get the coordinates of the insertion inside the assembly, which was then used for downstream analysis.

Genotyping insertions

To genotype the insertions, the method from Sniffles 2[47] was ported to R. A few considerations had to be made: the genotyping was made after the variant callers' execution, which mainly had the limitation of not having the exact count of reads covering the insertion point available. Instead, coverage around the insertion coordinates ±20 bp was measured. This margin was chosen because the alignment pointed to slightly different coordinates for the same insertion. Then, like Sniffles does, the probability for each genotype was calculated modeling the insertion as a binomial distribution, in which the sample size is the total number of reads at the insertion's location, the number of successes is the number of supporting reads, and the success chance is 0.5 for heterozygous SVs, 0.05 for homozygous reference allele, and 0.95 (1–0.05, the previous value) for homozygous non-reference allele. When either the supporting or total reads were higher than 250, they were normalized to add up to that number. Probabilities were calculated using the binomial distribution function (not the R base implementation, which does not allow for floating point values, and cannot process normalized data), except the binomial coefficient calculation was skipped, since it returns the same value for all genotypes of any given SV. The Manhattan plot of the AFs was generated by qqman.[59] Coverage for genotyping was obtained with mosdepth.[60]

Alternative AFs could then be calculated so that a heterozygous variant implies carrying a single allele of the insertion; a homozygous one, two; and a genotype of 0/0, no inserted alleles.

Annotation of insertions

The insertions were annotated for repetitive elements using RepeatMasker 4.1.2-p1[4] and Dfam 3.3.[5,6] For those insertions which were annotated with more than one repetitive element, the annotation with highest score (which RepeatMasker obtains with a Smith-Waterman alignment) was selected. Finally, only annotations that covered at least 85% of the inserted sequence were considered.

For gene annotation, annotatr[61] was used. The annotations for hg38 genes were used for the data. For repetitive sequences on the reference genome, RepeatMasker output was downloaded from UCSC.[48] To calculate the percentage of repetitive elements on genes, the overlapping regions were then condensed as one, e.g., an annotation spanning chr1:15 100-15 150 and another on chr1:15 125-15 200 would have been merged into one with coordinates chr1:15 100-15 200, to avoid counting bases with multiple annotations more than one time. To get gene coordinates, for a given gene in the annotation, the lowest and highest coordinates among promoters, exons, and introns were selected as start and end, respectively. Then the two sets of coordinates were merged with data.table.[67]

Size assessment of TE sequences

The length of TE insertions was compared to that of the consensus sequences, requiring at least 85% of the consensus' length to be considered full. The reference length used for Alu was 300 bp; for SVAs, the mean of SVA A to F retrieved from Dfam,[6] for L1, 6 kb,[43] and for the rest, they were taken from Dfam.

TE insertions in other datasets

The HGSVC[49] and Indigen[50] datasets were downloaded. Chromosome names were changed from the format "1" to "chr1". Additionally, for HGSVC variants, the file was splitted by case, into a total of 9 files, which were then merged using the custom code without considering sequence. After processing, the files were merged without considering insertion sequence.

Overlap with reference repetitive sequence

RepeatMasker results from UCSC were also used to determine which deletions and insertions overlapped repetitive sequences. For insertions, it was considered whether they were found on the same element, family, etc. Since repetitive sequences may overlap, any insertion could match coordinates with multiple reference sequences, in that case, all were considered, so if at least one matched, it was considered as the same type. For deletions, they had to be aligned with a TE sequence, with a difference of coordinates up to 15%.

Gene enrichment analysis

For enrichment, gene symbols from genes affected by at least one insertion were chosen. Analysis was done using clusterProfiler[62] and DOSE,[63] with ontologies GO,[33,68] KEGG,[34] DO,[35] and NCG.[51] The p value threshold for all enrichment was 0.01, except for NCG, for which 0.05 was used.

QUANTIFICATION AND STATISTICAL ANALYSIS

For distribution comparison, R[69] implementation of Wilcoxon two-sample test (also knwon as MannWhitney test) was used, and for multimodal analysis, R's multimode[70] with the default method (Amejeiras-Alonso). To check for significant differences in distribution across chromosomes, the chi-squared test was used. To check the distribution of insertions inside the same chromosome, each chromosome was divided into evenly sized bins and tested in the same manner. To account for these 24 comparisons, Bonferroni correction was used. For each statistical result in text, the p value and test used is indicated.