

ARTICLE OPEN



Evolutionarily recent retrotransposons contribute to schizophrenia

Giorgia Modenini^{1,5}, Paolo Abondio^{1,2,5}, Guia Guffanti³, Alessio Boattini¹✉ and Fabio Macciardi⁴✉

© The Author(s) 2023

Transposable elements (TEs) are mobile genetic elements that constitute half of the human genome. Recent studies suggest that polymorphic non-reference TEs (nrTEs) may contribute to cognitive diseases, such as schizophrenia, through a cis-regulatory effect. The aim of this work is to identify sets of nrTEs putatively linked to an increased risk of developing schizophrenia. To do so, we inspected the nrTE content of genomes from the dorsolateral prefrontal cortex of schizophrenic and control individuals and identified 38 nrTEs that possibly contribute to the emergence of this psychiatric disorder, two of them further confirmed with haplotype-based methods. We then performed in silico functional inferences and found that 9 of the 38 nrTEs act as expression/alternative splicing quantitative trait loci (eQTLs/sQTLs) in the brain, suggesting a possible role in shaping the human cognitive genome structure. To our knowledge, this is the first attempt at identifying polymorphic nrTEs that can contribute to the functionality of the brain. Finally, we suggest that a neurodevelopmental genetic mechanism, which involves evolutionarily young nrTEs, can be key to understanding the ethio-pathogenesis of this complex disorder.

Translational Psychiatry (2023)13:181; <https://doi.org/10.1038/s41398-023-02472-9>

INTRODUCTION

Transposable elements (TEs) are DNA sequences that have the ability to move around in the genome. TEs constitute 53–60% of the human DNA [1, 2] and are essential elements in driving genome evolution [3]. Among non-LTR retrotransposons (Alu, LINE, and SVA), only LINE-1 (L1) can actively transpose, while Alu and SVA rely on L1's machinery to mobilize themselves [4]. While the vast majority of TEs are no longer transpositionally active, they can still play a functional role as exapted enhancers or transcriptional start sites [5–8], by inserting transcription factor binding sites (TFBS) [9, 10] or by acting as novel RNA genes such as long non-coding RNAs (lnc-RNAs) [11]. Therefore, TEs participate in regulating the expression of nearby genes, at transcriptional and post-transcriptional levels, providing a crucial role as both cis- and trans-regulatory RNA sequences [12]. Baillie and colleagues [13] also found that protein-coding loci are disproportionately affected by TEs, with over-representation of L1s in introns and Alus in exons. Overall, TEs seem to predominantly affect neurogenesis and synaptic function, with studies suggesting a putative regulatory role of TEs in the neural genome [14–18]. Nonetheless, only initial research work has been systematically performed on this issue and the TE-controlled regulatory architecture of the human genome still needs to be better explored and investigated. According to recent studies, TE's insertion polymorphisms “can be mapped as cis-expression quantitative trait loci with substantial effects on gene expression, especially at loci involved in immune response and cognitive function” [19, 20]. A polymorphic TE insertion can be exapted as a functional element depending on its site of insertion within the

genome, or it can disrupt an already existing enhancer [21]. Additionally, polymorphic TEs have been shown to be closely associated with complex phenotypes in GWAS investigations suggesting that polymorphic non-reference TEs (nrTEs) may contribute to disease phenotypes through cis-regulatory effects [22–24]. Interestingly, nrTEs are relatively young compared to fixed TEs, therefore they are likely to have had a role in the most recent phases of the evolution of our species, which particularly involved the brain and superior cognitive abilities.

In recent years, mounting data provided evidence that epigenetic mechanisms and TEs are playing a key role in schizophrenia and other neurological disorders [17, 25–27]. For example, Bundo and colleagues [15] described an increased number of somatic L1 retrotransposition in the dorso-lateral prefrontal cortex (DLPFC, Brodmann's area 46) of people affected by schizophrenia, observing that the total number of brain-specific L1 insertions tended to be higher in schizophrenia patients, an observation confirmed by Doyle et al. [18]. In our own previous work [17] and in a recent review [28], L1 insertion sites were also reported to be preferentially localized to synapse- and schizophrenia-related genes. Guffanti et al. [25] developed a method to quantify the tissue-specific expression of TEs (as well as other ncRNAs), and found more than 650,000 expressed TEs in the DLPFC of post-mortem human brains: about 114,000 TEs are differentially expressed between schizophrenia cases and healthy controls and mostly represented by primate- or human-specific elements.

A recent study suggests another potential key role for TEs in rewiring the local functional architecture of human accelerated

¹BiGeA Department, University of Bologna, Bologna, Italy. ²Department of Cultural Heritage, University of Bologna, Ravenna, Italy. ³Department of Psychiatry, McLean Hospital-Harvard Medical School, Belmont, MA, USA. ⁴Department of Medical Education (Neuroscience), CUSM, Colton, CA, USA. ⁵These authors contributed equally: Giorgia Modenini, Paolo Abondio. ✉email: alessio.boattini2@unibo.it; fmacciar@hs.uci.edu

Received: 13 January 2023 Revised: 2 May 2023 Accepted: 12 May 2023

Published online: 27 May 2023

regions (HARs) in Schizophrenia and bipolar disorder [29]. Indeed, HARs have been implicated in neurodevelopmental and neuropsychiatric disorders [30–32], and most HARs are known to act as developmental enhancers that are involved in controlling and regulating human cognition [33–37].

In this study, our goal was to identify polymorphic TEs that can potentially contribute to schizophrenia. We choose to look at the non-reference TE content of DLPFC genomes of schizophrenic individuals (SCZ), to investigate the brain tissue-specific presence of nrTEs and possibly disentangle their somatic or germ-line origin. To accomplish this task, we will (1) compare SCZ with control (CTRL) genomes; (2) check for the presence of nrTEs and the population-specific/geographic distribution of the identified variants in the 1000 Genome data; (3) perform haplotype-based association tests; (4) explore the possible functional roles of nrTEs as cis-regulatory elements of protein-coding genes and as putative modifiers of known HARs in silico.

SUBJECTS AND METHODS

DNA from the DLPFC of 10 schizophrenic patients and 10 psychiatrically healthy controls has been obtained from the UCI Brain Bank, following a UCI/IRB-approved protocol. After DNA extraction from brain tissue samples and QC controls, we outsourced the whole genome sequencing to Illumina (<https://www.illumina.com/services/sequencing-services.html>), which then returned the assembled genomes and the fastq raw reads. Using the fastq files provided, we then realigned the raw reads to the human reference genome *hs37d5* (<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/>) with BWA-mem [38]. After sorting and merging with Samtools [39], we applied the GATK best practices to generate VCF files that include single nucleotide polymorphisms (SNPs) as well as insertions/deletions (Indels) (<https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices>).

We searched for non-reference TEs (nrTEs: Alu, LINE1, and SVA) with the mobile element locator tool (MELT) v.2.1.5 [40], using MELT-Split with default parameters on our 20 high-coverage genomes (Supplementary Table 1). To analyze the possible presence of nrTEs—which will suggest their germline rather than somatic origin—and the geographic variability of the putative schizophrenia-related nrTEs, we additionally selected 125 samples from the “1000 Genomes Project” phase 3 [41]. These 125 samples were analyzed with MELT jointly with SCZ and CTRL samples. Only “PASS” sites were included in a single final VCF file and only nrTEs mapping in genic or regulatory regions (introns, exons, promoters, terminators, and UnTranslated Regions, UTRs) on autosomal chromosomes were considered for further analyses. Fisher tests of independence were performed to identify which nrTEs show significantly different frequencies in SCZ and CTRL. Tests were performed with one and two degrees of freedom, respectively, for allelic and genotype frequencies. nrTEs that yielded nominally significant tests ($p\text{-val} < 0.05$) at least for allele and/or genotype frequencies were considered as putatively associated with schizophrenia.

To assess the genetic relationships among the individuals included in our dataset, as well as their ancestry, we implemented a principal component analysis (PCA) and an ADMIXTURE analysis [42], both on the whole variant dataset (single nucleotide polymorphisms, SNPs, and nrTEs) and on the nrTE-based only. Quality control (QC) was performed with the PLINK software [43].

We also performed a haplotype reconstruction procedure with SHAPEIT v.1.9 [44] on the whole variant dataset to contextualize the polymorphic inserted nrTEs into their local genetic environment and evaluate the frequency of the corresponding haplotypes within the DLPFC cohort. We then performed a haplotype association test using Beagle v.3.3.2 [45] on the nrTEs with significantly different allelic and/or genotype frequencies between SCZ and CTRL.

As highlighted in the “Introduction” section, TEs can act as cis-regulatory elements by, for example, modifying the expression of nearby genes and/or inducing alternative splicing. Therefore, we checked if non-reference TEs may act as eQTLs and/or sQTLs, by comparing our significant results with those from Cao et al. [46], based on the GTEx dataset [47].

Moreover, we verified whether the statistically significant non-reference TEs (i.e., nrTEs with significantly different allele/genotype frequencies between cases and controls) are located close to genes previously studied in the context of schizophrenia.

We also compared our 38 nrTEs with the lists of known HARs as originally discovered by Pollard et al. [48, 49], Prabhakar et al. [50], Bird et al. [51], Capra et al. [52], and Gittelman et al. [53], as well as against the HAR genes proposed by Wei et al. [54] to check if some of the nrTEs we identified are located in those regions.

RESULTS

non-reference retrotransposon insertions

We identified 7952 nrTEs in genic/regulatory regions: 6542 Alu (82.3%), 1065 LINE-1 (13.4%), and 345 SVA (4.3%), as shown in Table 1, using MELT-Split on our 145 samples (10 SCZ, 10 CTRL and 125 normal individuals from 1000 Genomes Project, 1KGP).

We checked the chromosomal distribution of the 7952 nrTEs and found no significant difference (Fisher $p\text{-val} > 0.63$) between the expected and observed content of the different families of TEs (SVA, LINE-1, and Alu) (Supplementary Figs. 1–3).

Population structure of the dataset

To contextualize our 20 DLPFC samples in the worldwide genomic landscape, we performed PCA and ADMIXTURE analyses on nrTEs genotypes and found that the best estimate for K in the latter is 3, with CV error = 0.37350, including 125 1KGP samples from five populations. Our results (Figs. 1 and 2) show that 7338 nrTEs (93%) that we identified in our 20 DLPFC samples are also present in the 125 samples from 1000K Genomes, suggesting their germ-line rather than somatic origin, and are also useful predictors of the genomic structure of the different human populations, as confirmed by the intermediate position of Indians (ITU) between Europeans (CEU) and Chinese (CHB), as well as by the clear differentiation between Eurasian and African samples. The remaining 7.3% of the nrTEs we detected in the DLPFC DNAs of our sample ($n = 534$: 35 SVAs (10% of total non-reference SVAs and 1.75/subject), 56 LINE1s (5% and 2.8/subject) and 473 Alus (7% and 23.6/subject)) are unique and not shared across other samples nor are they listed in known reference databases, like euL1db [55] and gnomAD [56]. They may be regarded as somatic retrotranspositions or they may still be germline nrTEs with low frequencies since we cannot distinguish between the two possible origins. Even in case they are somatic rather than germline events, they still represent a minority of our observed nrTEs.

The ADMIXTURE plot of nrTEs shows that CTRL and SCZ share the same ancestral components of CEU, (violet). Han Chinese (orange) have their own ancestral component, as well as the two African populations. ITUs show a mixture of European and Asian components. These results, as a whole, are coherent with those obtained using SNPs [41, 57].

Accordingly, nrTEs show systematic differences in allele frequencies across populations: 3131 of 7952 non-reference TEs, 2711 Alu (41.4%), 332 LINE-1 (31.1%) and 88 SVA (25.5%), have a significant geographic stratification (Fisher $p\text{-val} < 0.05$) (Supplementary Table 2), with 2263 (28%) presenting with an allele frequency $> 5\%$. Among these, 1501 nrTEs are found only in African populations, 833 are exclusive of non-African populations

Table 1. Location of the identified non-reference TEs as defined by the MELT output.

Location	Alu	LINE1	SVA	Total
Intronic	2635	376	151	3162
Promoter	1949	343	97	2389
Terminator	1902	342	96	2340
3_UTR	44	2	0	46
5_UTR	7	1	0	8
Exon	5	1	1	7
Total	6542	1065	345	7952

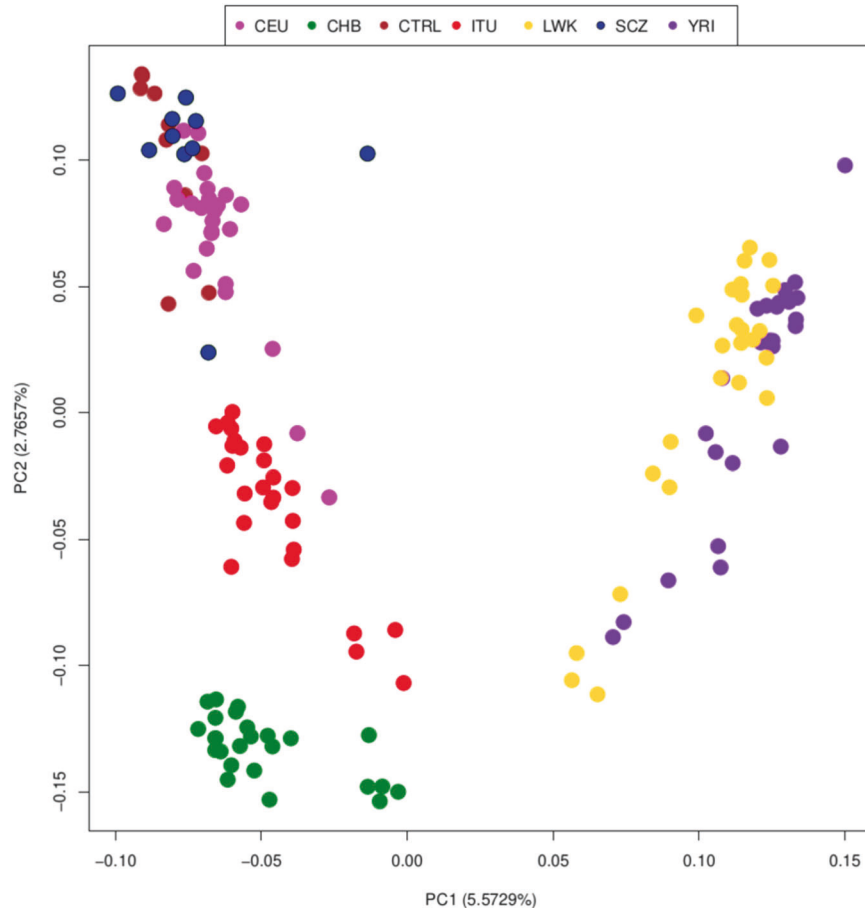


Fig. 1 Principal component analysis (PCA) of the DLPFC and 1KGP samples based only on non-reference TEs. Pink: Europeans (CEU); green: Han Chinese in Beijing (CHB); brown: controls from the DLPFC (CTRL); red: Indian Telugus (ITU); yellow: Luhya in Kenya (LWK); blue: schizophrenic individuals from the DLPFC (SCZ); violet: Yoruba in Nigeria (YRI).



Fig. 2 Admixture plot of the 20 DLPFC + 125 1KGP samples. ADMIXTURE plot based only on nrTEs. $K=3$ is shown (CV error = 0.37350).

(Europeans, Indian Telugus, and Chinese) and 955 are common to all five groups (see Supplementary Table 2 and Supplementary Fig. 6). Both methods (PCA and ADMIXTURE) highlight that DLPFC samples overlap with CEU, with the partial exception of a single SCZ sample, which presents signs of admixture with a Sub-Saharan African source (as represented by YRI and LWK).

These findings suggest that SCZ and CTRL are genetically homogeneous; consequently, variants associated with the disease condition do not depend on underlying population structure.

Comparison between SCZ and CTRL subjects

We then compared the distribution of allele and genotype insertion frequencies of nrTEs (herein, 'counts') between SCZ cases and normal CTRLs. We detected 38 non-reference TEs with significantly different allele/genotype counts between cases and controls: three LINE1s, three SVAs, and 32 Alus, that yielded significant Fisher tests for allele and/or genotype counts [58, 59] at the nominal p -value ≤ 0.05 (Table 2). Given the limited size of our sample, no correction for multiple testing was performed. All significant nrTEs belong to evolutionarily recent elements (L1Ta and AluY), with two exceptions: the L1 on chr12:126802943

(undetermined subfamily) and the Alu on chr7:141748320 (which belongs to the subfamily Sz, older than Y).

Of these 38 nrTEs, 11 show a significant difference in allele counts only, 14 in genotype counts only, and 13 in both allele and genotype counts. Interestingly, most of these TEs also show evidence of differential segregation (Fisher test, p -val < 0.05) in human populations (27 for allele counts, 24 for genotype counts, 23 for both) and are more common in European and Asian populations.

The most significant allele-wise results (p -val < 0.01) among the 38 significant nrTEs findings include three Alus and one SVA, whose insertion can be found on: chr4:17150918 and chr4:23511024 (both AluY and only observed in SCZ), chr11:40727097 (AluYa5, only observed in CTRL) and chr20:5268423 (SVA, more frequent in SCZ). The three Alus show a statistically significant geographical distribution (Fig. 3 and Supplementary Table 3), presenting variable insertion frequencies across populations, while the SVA on chr20:5268423 has similar allele frequencies in all the considered populations.

Haplotype-based association analysis

To better elucidate the potential association of the identified nrTEs with schizophrenia, we then performed a haplotype-based analysis using Beagle on the previously detected 38 variants (Table 2) and obtained two significant results. The first one was for a 188 bp haplotype that includes an AluYb on chr5:100497396 in the promoter of the ST8SIA4 gene. This haplotype is characterized by 4 polymorphisms: T+TG (where "+" points out to the presence of the nrTE and the other letters represent single

Table 2. List of 38 significant nrTEs that are different between SCZ and CTRL: 3 SVAs, 3 L1s, and 32 Alus.

Type	Family	Chr.	Bp	Gene	Location	SCZ(-)	SCZ(+)	CTRL(-)	CTRL(+)	Fisher	SCZ(---)	SCZ(-+)	SCZ(++)	CTRL(-+)	CTRL(++)	Fisher	
SVA-INS	SVA	9	33,130,559	BAGALT1	Intronic	12	8	16	4	0.301	2	8	0	7	2	1	0.023
SVA-INS	SVA	11	73,527,418	MRLP48	Intronic	20	0	15	5	0.047	10	0	0	5	5	0	0.033
SVA-INS	SVA	20	5,268,423	PROKR2	Terminator	7	13	16	4	0.010	1	5	4	6	4	0	0.017
LINE1-INS	L1Ta1d	3	38,626,082	SCNSA	Intronic	20	0	14	6	0.020	10	0	0	4	6	0	0.011
LINE1-INS	L1Ta	9	32,463,887	DXS8	Intronic	4	16	11	9	0.048	1	2	7	3	5	2	0.166
LINE1-INS	L1Ambig	12	126,802,943	LINC02347	Promoter	17	3	13	7	0.273	8	1	1	3	7	0	0.020
ALU-INS	AluYa3	1	70,091,072	LRRC7	Promoter	15	5	11	9	0.320	6	3	1	1	9	0	0.020
ALU-INS	AluYa4	1	163,314,443	NUF2	Intronic	12	8	6	14	0.111	5	2	3	0	6	4	0.038
ALU-INS	AluYb6	2	9,888,801	TAF1B	Promoter	9	11	2	18	0.031	1	7	2	0	2	8	0.023
ALU-INS	AluYb7	2	36,476,695	CRIM1	Terminator	15	5	20	0	0.047	6	3	1	10	0	0	0.087
ALU-INS	AluYa4	2	114,106,446	PAX8-AS1	Terminator	17	3	9	11	0.019	7	3	0	2	5	3	0.044
ALU-INS	AluYa5	3	169,951,024	PRKCI	Intronic	12	8	18	2	0.065	2	8	0	8	2	0	0.023
ALU-INS	AluYa4	4	17,150,918	QDPR	Terminator	11	9	20	0	0.001	4	3	3	10	0	0	0.011
ALU-INS	AluYb7	4	23,511,024	MIRS48AUJ	Promoter	13	7	20	0	0.008	3	7	0	10	0	0	0.003
ALU-INS	AluYa4	4	154,901,048	SRRP2	Promoter	15	5	11	9	0.320	5	5	0	5	1	4	0.047
ALU-INS	AluYb	4	183,647,531	TENM3	Intronic	10	10	15	5	0.191	0	10	0	5	5	0	0.033
ALU-INS	AluYb8	5	55,689,499	ANKRD55	Promoter	14	6	8	12	0.111	4	6	0	0	8	2	0.043
ALU-INS	AluYb8	5	84,516,075	EDIL3	Promoter	20	0	15	5	0.047	10	0	0	5	5	0	0.033
ALU-INS	AluYa	5	86,372,695	MIR4280	Terminator	7	13	7	13	1.000	3	1	6	0	7	3	0.013
ALU-INS	AluYb	5	100,497,396	ST8SIA4	Promoter	13	7	5	15	0.025	5	3	2	1	3	6	0.137
ALU-INS	AluYg5b3	5	159,122,155	ADRA1B	Promoter	10	10	18	2	0.014	1	8	1	8	2	0	0.005
ALU-INS	AluYc1	6	75,338,236	COL12A1	Terminator	17	3	10	10	0.041	7	3	0	2	6	2	0.103
ALU-INS	AluYa5	6	166,279,907	LINC00473	Terminator	11	9	18	2	0.031	3	5	2	8	2	0	0.095
ALU-INS	AluYa5	7	141,013,590	TMEM178B	Intronic	20	0	15	5	0.047	10	0	0	6	3	1	0.087
ALU-INS	AluSz	7	141,748,320	MGAM	Intronic	10	10	13	7	0.523	0	10	0	4	5	1	0.033
ALU-INS	AluYb6	8	56,033,091	XKR4	Intronic	14	6	20	0	0.020	5	4	1	10	0	0	0.033
ALU-INS	AluYb8	9	75,542,985	ALDH1A1	Intronic	15	5	7	13	0.025	6	3	1	1	5	4	0.093
ALU-INS	AluYb3a1	9	91,099,740	SPIN1	Terminator	15	5	20	0	0.047	5	5	0	0	0	0	0.033
ALU-INS	AluYb6	11	19,382,774	NAV2	Intronic	13	7	19	1	0.044	4	5	1	9	1	0	0.057
ALU-INS	AluYa5	11	40,727,097	LRRC4C	Intronic	20	0	12	8	0.003	10	0	0	2	8	0	0.001
ALU-INS	AluYg6	11	76,990,585	GDPD4	Intronic	13	7	19	1	0.044	4	5	1	9	1	0	0.057
ALU-INS	AluYa3	12	31,120,751	TSPAN11	Intronic	13	7	19	1	0.044	4	5	1	9	1	0	0.057
ALU-INS	AluYa5	12	81,315,235	LIN7A	Intronic	16	4	16	4	1.000	6	4	0	8	0	2	0.043
ALU-INS	AluYa3	13	108,669,030	FAM155A	Promoter	9	11	17	3	0.019	1	7	2	8	1	1	0.003
ALU-INS	AluYc1	15	39,691,605	C15orf54	Terminator	5	15	9	11	0.320	2	1	7	1	7	2	0.022
ALU-INS	AluYe	16	80,010,958	MAF	Promoter	15	5	20	0	0.047	6	3	1	10	0	0	0.087
ALU-INS	AluYb	17	60,376,780	TBC1D3P2	Promoter	18	2	13	7	0.127	9	0	1	3	7	0	0.003
ALU-INS	AluYe	21	28,221,356	ADAMTS1	Promoter	17	3	11	9	0.082	8	1	1	2	7	1	0.009

Presence or absence of the nrTE is defined by “+” and “-”, respectively. Genotypes are displayed as follows: “++”= homozygous, “+-”= heterozygous, “--”= absence.

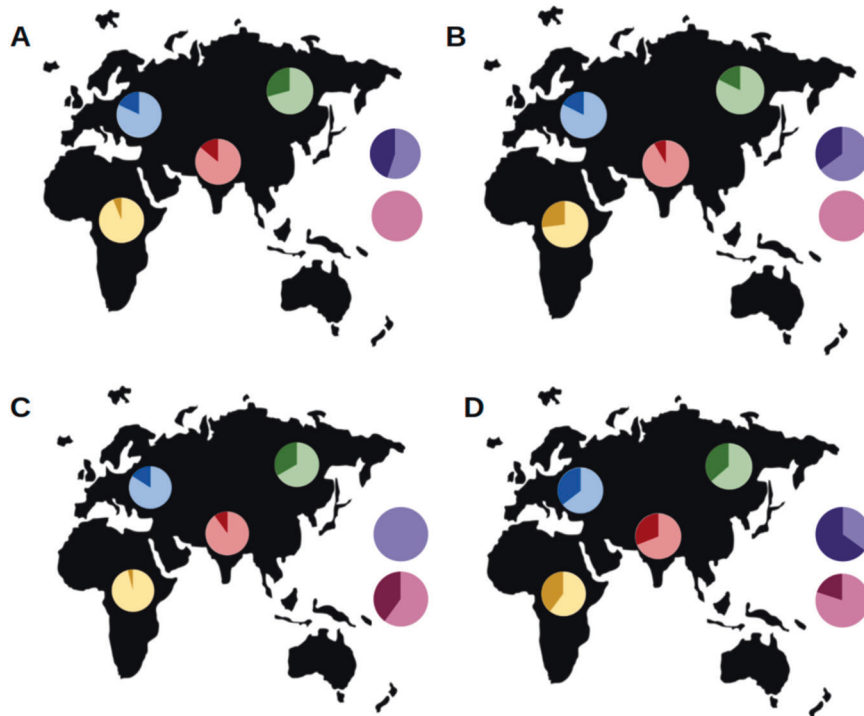


Fig. 3 Geographic distribution of the most significant (allele-wise) nrTEs. Allele frequencies of Alus on chr4:17150918 (A), chr4:23511024 (B) and chr11:40727097 (C), compared to SVA on chromosome chr20:5268423 (D). Darker colors indicate the presence of the TE (+), while lighter colors indicate the absence (-). Following populations are displayed: Europeans (blue), Indian Telugus (red), Chinese (green) and Africans (yellow), represented by Luhya in Kenya and Yoruba in Nigeria. Allele frequencies for schizophrenic individuals and healthy controls are shown in violet and pink, respectively.

nucleotide variants). It is present with 15 copies in CTRL and 2 in SCZ, suggesting a strong association ($p_{\text{val}} = 6.86 \times 10^{-5}$) between the presence of the haplotype and the absence of the disease.

The second significant result was for a 1 172 bp haplotype that includes the locus of an AluYb7 on chr4:23511024 in the promoter of MIR548AJ2. Interestingly, this haplotype is characterized by the absence of the insertion, with polymorphisms GC-TTI (where “I” stands for InDel and “-” indicates the absence of the nrTE) and was found with 19 copies in CTRL and 6 copies in SCZ ($p_{\text{val}} = 3.93 \times 10^{-5}$): the Alu is completely absent in CTRL samples (Table 2) and present in 7 SCZ, only in a heterozygous condition.

In-silico functional inferences for non-reference TEs

At least seven genes putatively mapped by our significant non-reference TEs have been already associated with schizophrenia: LRRC4C [60], LRRC7 [61–63], ST8SIA4 [64–66], MGAM [67], ADAMTS1 [68], MIR548AJ2 [69] and SCN5A, which is also linked to the Brugada syndrome [70–72].

We then compared our set of 38 significant nrTEs with the eQTLs and sQTLs TEs lists produced by Cao et al. [46] using the GTEx dataset [47] (Supplementary Table 3). Indeed, 27 TEs (3 LINE-1 and 24 Alu) (68.4%) were detected as potential eQTLs acting in different tissues, 7 of which are expressed in the brain. As for sQTLs, 13 (34.21%) of our TEs (2 LINE-1 and 11 Alu) were detected as potentially contributing to alternative splicing in different tissues, 2 of them supposedly acting in the brain (chr11:76990585 and chr2:36476695). All sQTLs TEs were also eQTLs, with the single exception of an Alu (chr5:159122155, in the promoter of ADRA1B) which works only as sQTL.

Last, we compared our set of 38 nrTEs with the lists of HARs produced by Pollard et al. [48, 49], Prabhakar et al. [50], Bird et al. [51], Capra et al. [52] and Gittelman et al. [53], as well as against the HAR genes proposed by Wei et al. [54] and found that 12 nrTEs are located within as many HAR-genes (ADAMTS1, ANKRD55,

CRIM1, EDIL3, LRRC4C, LRRC7, MAF, NAV2, QDPR, TENM3, TSPAN11, XKR4), which collectively show enrichment for “regulation of neuron projection development”.

DISCUSSION

Far from being “junk”, it has been shown that transposable elements, such as non-Long Terminal Repeats retrotransposons, can contribute to human genomic diversity in various ways. Mounting data suggest both a positive and a detrimental role of retrotransposons in shaping human cognitive traits [17] and in the development of brain and central nervous system (CNS) structures [73, 74]. Several authors also suggest that retrotransposons have an important role in neurological and psychiatric disorders, such as schizophrenia [15, 16, 18, 25, 26].

However, the full impact of TEs on the human genome is still unclear, both for technological/methodological limitations as well as our current lack of knowledge of their precise effects and interactions with other genetic/epigenetic elements. Since a relationship between cognitive disorders and reference TEs has been the subject of several recent studies [15, 25, 26] with this work, we aimed to provide the first investigation of non-reference TEs as risk factors that can potentially contribute to increasing the risk of developing schizophrenia.

In the first step, to evaluate whether nrTEs distribution can contribute to population substructure (=nrTEs frequency differences due to specific populations' origin/evolutionary trajectory) and lead to specific population patterns, we inspected the genetic distribution of traditional DNA variants (SNPs, Indels, CNVs) and of nrTEs in our 20 DLPFC subjects together with 125 worldwide individuals that we collected from the 1KGP. Our PCA and Admixture results show that nrTEs are present with mostly population-specific frequencies within our worldwide dataset, similarly to the well-known patterns previously detected in SNP-

Table 3. nrTEs with allele *p*-values < 0.01.

Type	Chr.	Bp	SCZ(-)	SCZ(+)	CTRL(-)	CTRL(+)	Fisher	ITU(-)	ITU(+)	CEU(-)	CEU(+)	CHB(-)	CHB(+)	YRI(-)	YRI(+)	LWK(-)	LWK(+)
SVA-INS	20	5268423	7	13	16	4	0.0095	29	13	31	17	28	16	37	11	22	18
ALU-INS	4	17150918	11	9	20	0	0.0012	43	7	41	9	34	14	45	5	47	1
ALU-INS	4	23511024	13	7	20	0	0.0083	33	3	33	7	28	6	20	14	34	6
ALU-INS	11	40727097	20	0	12	8	0.0033	45	5	42	8	32	16	49	1	47	3

The presence of the nrTE is defined by “+”, while the absence is defined with “-”.

based studies and we ultimately confirm that nrTEs too contribute to the higher genomic diversity of African compared to non-African populations [41, 57, 75].

Our results (Figs. 1 and 2) also show that our SCZ and CTRL individuals fall within the European genomic variability (represented by CEU in both PCA and Admixture analyses) and share a predominant European ancestral component, except for a single individual showing signs of admixture with a Sub-Saharan African source. The non-reference, polymorphic TE we have identified in our DLPFC samples are mostly present in more than one subject, therefore they are probably due to germline retrotranspositions. Only a limited number of events (7.3%) may be considered either somatic or low-frequency germline retrotranspositions because they are not shared across samples and are present in only one sample, thus these single events should be better considered private insertions (or singletons). The total number of singletons and their proportion per subject are concordant with the estimates provided by Watkins et al. [76] for the Caucasian/European population.

Therefore, our results confirm both that polymorphic nrTEs can be used as reliable markers for reconstructing the genomic structure (and potentially the history) of samples/populations that we analyzed, as previously suggested by others [40, 76, 77] and that our SCZ and CTRL sample presents a clear nrTEs genetic homogeneity, which allows to exclude spurious associations due to hidden population substructure.

Admittedly, our results are based on a relatively low number of subjects, and they must be considered only as a preliminary discovery investigation. However, we performed a careful evaluation of population structure (PCA and Admixture analyses) and observed that our SCZ and CTRL samples are genomically homogeneous, in addition overlapping with reference samples of European ancestry (as represented by CEU from 1000 Genomes Project). Therefore, spurious associations due to population substructure can be excluded. Then, in order to strengthen the potential association between genomic markers (nrTEs) and the investigated trait (schizophrenia), in addition to performing standard allele frequency-based association methods, we also applied haplotype-based analyses, which, as expected, yielded [78] highly significant results with a magnitude $pval \leq 10^{-5}$.

As outlined above, we first identified 38 nrTEs whose frequencies are significantly different between Schizophrenics and Controls. Even considering the low sample size, we observe that allele frequency differences for these TE are similar or even higher than those observed between the most different ‘control’ populations for the same insertions, making it highly improbable that the observed differences emerged by chance. These 38 nrTEs constitute our set of putative candidates associated with schizophrenia. Focusing only on the most significant results ($pval < 0.01$), they refer to three Alus and one SVA, which respectively fall on chr4:17150918, chr4:23511024, chr11:40727097 and chr20:5268423. The first two Alus are found only in SCZ, the third only in CTRL and the SVA on chromosome 20 is more frequent in SCZ subjects. The Alu on chr11:40727097 is completely absent in SCZ (but present in 27.4% of CTRLs) (Fig. 3C) and is located in the second intron of the Leucine Rich Repeat Containing 4C (LRRC4C) gene, which is highly expressed in the frontal cortex and has been associated with a positive response to antipsychotic therapy with lurasidone in SCZ patients [60]. In our case, the absence of this Alu insertion is preferentially associated with the schizophrenic condition (see Tables 2 and 3).

As highlighted in the “Introduction” section, TE can act in *cis*, for example, by altering the expression of a gene or by having an impact on its alternative splicing. Therefore, we also looked at the potential role of our 38 significantly different nrTEs by comparing them with the lists produced by Cao and colleagues [46] based on the GTEx dataset [47]. We found that 27 nrTEs act as eQTLs in different tissues, with 7 showing a putative eQTL effect in the brain. For instance, the AluYb3a1 on chr9:91099740 acts as eQTL

in the frontal cortex and is located in the terminator of SPIN1 (Spindlin 1). Moreover, 13 nrTEs act as sQTLs, two of them in the brain: the AluYg6 chr11:76990585 and the AluYb7 on chr2:36476695. These two Alus are located in the third intron of GDPD4 (Glycerophosphodiester Phosphodiesterase Domain-Containing Protein 4) and in the terminator of the CRIM1 gene, respectively. Interestingly, the AluYb7 on chr2:36476695 acts as sQTL in the frontal cortex, and CRIM1 encodes for the cysteine-rich neuron motor 1 protein, which is developmentally regulated and involved in CNS development and organogenesis [79], other than being part of the HAR-genes that are functionally relevant in brain networks implicated in cognition [54]. Actually, recent research suggests that TE s could change the local functional architecture of HARs in schizophrenia and bipolar disorder [29] and our present findings add a further layer of support to this hypothesis, showing that 12 of the 38 significant nrTEs fall within the ORF of genes that are enriched for a neurodevelopmental process, the “regulation of neuron projection development” (ADAMTS1, ANKRD55, CRIM1, EDIL3, LRRC4C, LRRC7, MAF, NAV2, QDPR, TENM3, TSPAN11, XKR4). It is further interesting that at least three of these genes (ADAMTS1, LRRC4C, and LRRC7) have already been associated with schizophrenia.

After inferring the haplotypic context surrounding the 38 nrTEs of interest, we performed a haplotype-based association test with Beagle, which returned two significant results: one is the AluYb on chr5:100497396, located in the promoter of the gene ST8 Alpha-N-Acetyl-Neuraminidase Alpha-2,8-Sialyltransferase 4 (ST8SIA4). In particular, the haplotype-based analysis revealed that the haplotype (T + TG) with the presence of the nrTE is found with 15 copies in CTRL and 2 in SCZ; indeed, there was a strong association ($p\text{val} = 6.86 \times 10^{-5}$) between the presence of the nrTE and the absence of the considered trait (schizophrenia). Furthermore, this insertion was shown to act as eQTL [46], therefore suggesting a potential functional mechanism. Accordingly, we could hypothesize that the haplotype with the nrTE has a protective role against the disease. Further in vitro or in vivo experiments could elucidate this potential relationship.

The other haplotype (GC-TTI) is characterized by the absence of the AluYb7 on chr4:23511024, one of the most significant variants also from the allele frequency point of view. This nrTE is completely absent in CTRL samples and present in 7 SCZ patients, only in heterozygous conditions. Therefore, our hypothesis is that the presence of the element is putatively related to an increased risk of developing schizophrenia. Moreover, the Alu is located in the promoter of MIR548AJ2, one of the 108 genome-wide significant loci for schizophrenia reported by Ripke and colleagues [69].

In conclusion, our analysis provides the first overview of nrTEs as DNA variants that are possibly related to an increased risk of developing schizophrenia. We have identified 38 nrTEs of interest, two of them being further confirmed by highly significant haplotype-based analyses. We then highlighted that several of these elements can have a remarkable impact on the expression, alternative splicing, and functionality of nearby genes by cross-checking our results with those available in recently published studies. We also defined two haplotypes in which the presence of the nrTE is either protective against the disease or associated with the schizophrenic condition. Our results, as well as those from other papers dealing with nrTEs, are based on presence/absence of TE s, i.e. considering them as biallelic markers; future research based on long-read sequencing will also need to include TE s sequence variability. We expect that a similar framework applied to a larger cohort of subjects could confirm and possibly extend our results, and experimental validation of the identified nrTEs will elucidate their effective impact on the cognitive genome.

Having identified both reference [25] and non-reference TE s associated with an increased risk to develop schizophrenia suggests that a neurodevelopmental genetic mechanism is at

play in the etiopathogenesis of this complex disorder. Under this hypothesis, and given that TE s controlling for the functional architecture of the neural genome are mostly evolutionarily recent (either human-only or primate-specific), schizophrenia can emerge as a trade-off between our ongoing cognitive evolution and possible molecular flaws of our not yet completed evolutionary process.

REFERENCES

- de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*. 2011;7:e1002384.
- Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science*. 2022;376:eabk3112.
- Guio L, González J. New insights on the evolution of genome content: population dynamics of transposable elements in flies and humans. In: Anisimova M (ed) *Evolutionary genomics*. New York, NY: Springer New York; 2019. pp. 505–30.
- Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. *Annu Rev Genom Hum Genet*. 2011;12:187–215.
- Rangwala SH, Zhang L, Kazazian HH. Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol*. 2009;10:R100.
- Deininger P. Alu elements: know the SINEs. *Genome Biol*. 2011;12:236.
- Su M, Han D, Boyd-Kirkup J, Yu X, Han, DJ J-. Evolution of Alu elements toward enhancers. *Cell Rep*. 2014;7:376–85.
- Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. *Mob DNA*. 2016;7:24.
- Emera D, Wagner GP. Transformation of a transposon into a derived prolactin promoter with function during human pregnancy. *Proc Natl Acad Sci USA*. 2012;109:11246–51.
- Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep*. 2015;10:551–61.
- Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*. 2015;11:1110–22.
- Ali A, Han K, Liang P. Role of transposable elements in gene regulation in the human genome. *Life*. 2021;11:118.
- Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*. 2011;479:534–7.
- Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. *Nature*. 2009;460:1127–31.
- Bundo M, Toyoshima M, Okada Y, Akamatsu W, Ueda J, Nemoto-Miyauchi T, et al. Increased L1 retrotransposition in the neuronal genome in schizophrenia. *Neuron*. 2014;81:306–13.
- Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci*. 2014;15:497–506.
- Guffanti G, Gaudi S, Klengel T, Fallon JH, Mangalam H, Madduri R, et al. LINE1 insertions as a genomic risk factor for schizophrenia: preliminary evidence from an affected family. *Am J Med Genet Part B*. 2016;171:534–45.
- Doyle GA, Crist RC, Karatas ET, Hammond MJ, Ewing AD, Ferraro TN, et al. Analysis of LINE-1 elements in DNA from postmortem brains of individuals with schizophrenia. *Neuropsychopharmacol Publ Am Coll Neuropsychopharmacol*. 2017;42:2602–11.
- Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol*. 2022;23:481–97.
- Goubert C, Zevallos NA, Feschotte C. Contribution of unfixed transposable element insertions to human regulatory variation. *Philos Trans R Soc B Biol Sci*. 2020;375:20190331.
- Wang L, Norris ET, Jordan IK. Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. *Front Microbiol*. 2017;8:1418.
- Payer LM, Steranka JP, Kryatova MS, Grillo G, Lupien M, Rocha PP, et al. Alu insertion variants alter gene transcript levels. *Genome Res*. 2021;31:2236–48.
- Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D et al. Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proc Natl Acad Sci USA*. 2017;114. <https://doi.org/10.1073/pnas.1704117114>.
- Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. *Nature*. 2016;530:177–83.

25. Guffanti G, Bartlett A, Klengel T, Klengel C, Hunter R, Glinsky G, et al. Novel bioinformatics approach identifies transcriptional profiles of lineage-specific transposable elements at distinct loci in the human dorsolateral prefrontal cortex. *Mol Biol Evol.* 2018;35:2435–53.
26. Misiak B, Ricceri L, Szaśadek MM. Transposable elements and their epigenetic regulation in mental disorders: current evidence in the field. *Front Genet.* 2019;10:580.
27. Macciardi F, Giulia Bacalini M, Miramontes R, Boattini A, Taccioli C, Modenini G, et al. A retrotransposon storm marks clinical phenocopy to late-onset Alzheimer's disease. *GeroScience.* 2022;44:1525–50.
28. Terry DM, Devine SE. Aberrantly high levels of somatic LINE-1 expression and retrotransposition in human neurological disorders. *Front Genet.* 2019;10:1244.
29. Erady C, Amin K, Onilgbo TOAE, Tomasik J, Jukes-Jones R, Umrانيا Y, et al. Novel open reading frames in human accelerated regions and transposable elements reveal new leads to understand schizophrenia and bipolar disorder. *Mol Psychiatry.* 2022;27:1455–68.
30. Cheung JP, Tubbs JD, Sham PC. Extended gene set analysis of human neuropsychiatric traits shows enrichment in brain-expressed human accelerated regions across development. *Schizophr Res.* 2022;246:148–55.
31. Doan RN, Bae B-I, Cubelos B, Chang C, Hossain AA, Al-Saad S, et al. Mutations in human accelerated regions disrupt cognition and social behavior. *Cell.* 2016;167:341–54.e12.
32. Hubisz MJ, Pollard KS. Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr Opin Genet Dev.* 2014;29:15–21.
33. Boyd JL, Skove SL, Rouanet JP, Pilaz L-J, Bepler T, Gordân R, et al. Human–chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr Biol.* 2015;25:772–9.
34. Franchini LF, Pollard KS. Human evolution: the non-coding revolution. *BMC Biol.* 2017;15:89.
35. Girskis KM, Stergachis AB, DeGennaro EM, Doan RN, Qian X, Johnson MB, et al. Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron.* 2021;109:3239–51.e7.
36. Levchenko A, Kanapin A, Samsonova A, Gainetdinov RR. Human accelerated regions and other human-specific sequence variations in the context of evolution and their relevance for brain development. *Genome Biol Evol.* 2018;10:166–88.
37. Ziffra RS, Kim CN, Ross JM, Wilfert A, Turner TN, Haeussler M, et al. Single-cell epigenomics reveals mechanisms of human cortical development. *Nature.* 2021;598:205–13.
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinforma Oxf Engl.* 2009;25:1754–60.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinforma Oxf Engl.* 2009;25:2078–9.
40. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* 2017;27:1916–29.
41. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
42. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19:1655–64.
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
44. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 2013;10:5–6.
45. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81:1084–97.
46. Cao X, Zhang Y, Payer LM, Lords H, Steranka JP, Burns KH, et al. Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues. *Genome Biol.* 2020;21:185.
47. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45:580–5.
48. Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature.* 2006;443:167–72.
49. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, et al. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* 2006;2:e168.
50. Prabhakar S, Noonan JP, Pääbo S, Rubin EM. Accelerated evolution of conserved noncoding sequences in humans. *Science.* 2006;314:786.
51. Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, et al. Fast-evolving noncoding sequences in the human genome. *Genome Biol.* 2007;8:R118.
52. Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci.* 2013;368:20130025.
53. Gittelman RM, Hun E, Ay F, Madeoy J, Pennacchio L, Noble WS, et al. Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* 2015;25:1245–55.
54. Wei Y, de Lange SC, Scholtens LH, Watanabe K, Ardesch DJ, Jansen PR, et al. Genetic mapping and evolutionary analysis of human-expanded cognitive networks. *Nat Commun.* 2019;10:4839.
55. Mir G, Meer S, Cottrell D, McMillan D, House A, Kanter JW. Adapted behavioural activation for the treatment of depression in Muslims. *J Affect Disord.* 2015;180:190–9.
56. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature.* 2020;581:444–51.
57. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature.* 2016;538:201–6.
58. Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief Bioinform.* 2002;3:146–53.
59. Crespi B, Summers K, Dorus S. Adaptive evolution of genes underlying schizophrenia. *Proc Biol Sci.* 2007;274:2801–10.
60. Li J, Yoshikawa A, Brennan MD, Ramsey TL, Meltzer HY. Genetic predictors of antipsychotic response to lurasidone identified in a genome wide association study and by schizophrenia risk genes. *Schizophr Res.* 2018;192:194–204.
61. Hathy E, Szabó E, Vincze K, Haltrich I, Kiss E, Varga N, et al. Generation of multiple iPSC clones from a male schizophrenia patient carrying de novo mutations in genes KHSRP, LRRC7, and KIR2DL1, and his parents. *Stem Cell Res.* 2021;51:102140.
62. Hathy E, Szabó E, Varga N, Erdei Z, Tordai C, Czehlár B, et al. Investigation of de novo mutations in a schizophrenia case-parent trio by induced pluripotent stem cell-based in vitro disease modeling: convergence of schizophrenia- and autism-related cellular phenotypes. *Stem Cell Res Ther.* 2020;11:504.
63. Carlisle HJ, Luong TN, Medina-Marino A, Schenker L, Khorosheva E, Indersmitten T, et al. Deletion of Densin-180 results in abnormal behaviors associated with mental illness and reduces mGluR5 and DISC1 in the postsynaptic density fraction. *J Neurosci.* 2011;31:16194–207.
64. Curto Y, Alcaide J, Röckle I, Hildebrandt H, Nacher J. Effects of the genetic depletion of polysialyltransferases on the structure and connectivity of interneurons in the adult prefrontal cortex. *Front Neuroanat.* 2019;13:6.
65. Volk DW, Edelson JR, Lewis DA. Altered expression of developmental regulators of parvalbumin and somatostatin neurons in the prefrontal cortex in schizophrenia. *Schizophr Res.* 2016;177:3–9.
66. Kröcher T, Röckle I, Diederichs U, Weinhold B, Burkhardt H, Yanagawa Y, et al. A crucial role for polysialic acid in developmental interneuron migration and the establishment of interneuron densities in the mouse prefrontal cortex. *Dev Camb Engl.* 2014;141:3022–32.
67. Wu X, Huai C, Shen L, Li M, Yang C, Zhang J, et al. Genome-wide study of copy number variation implicates multiple novel loci for schizophrenia risk in Han Chinese family trios. *iScience.* 2021;24:102894.
68. Pantazopoulos H, Katsel P, Haroutunian V, Chelini G, Klengel T, Berretta S. Molecular signature of extracellular matrix pathology in schizophrenia. *Eur J Neurosci.* 2021;53:3960–67.
69. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511:421–7.
70. Spellmann I, Reinhard MA, Veverka D, Zill P, Obermeier M, Dehning S, et al. QTC prolongation in short-term treatment of schizophrenia patients: effects of different antipsychotics and genetic factors. *Eur Arch Psychiatry Clin Neurosci.* 2018;268:383–90.
71. Roden DM. The Brugada ECG and schizophrenia. *Circ Arrhythm Electrophysiol.* 2014;7:365–7.
72. Roberts E. GABAergic malfunction in the limbic system resulting from an aboriginal genetic defect in voltage-gated Na⁺-channel SCN5A is proposed to give rise to susceptibility to schizophrenia. *Adv Pharm San Diego CA.* 2006;54:119–45.
73. Suarez NA, Macia A, Muotri AR. LINE-1 retrotransposons in healthy and diseased human brain. *Dev Neurobiol.* 2018;78:434–55.
74. Ferrari R, Grandi N, Tramontano E, Dieci G. Retrotransposons as drivers of mammalian brain evolution. *Life Basel Switz.* 2021;11:376.
75. Byrka-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell.* 2022;185:3426–40.e19.
76. Watkins WS, Feusier JE, Thomas J, Goubert C, Mallick S, Jorde LB. The Simons Genome Diversity Project: a global analysis of mobile element diversity. *Genome Biol Evol.* 2020;12:779–94.

77. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. *Mob DNA*. 2015;6:21.
78. Sehgal D, Mondal S, Crespo-Herrera L, Velu G, Juliana P, Huerta-Espino J, et al. Haplotype-based, genome-wide association study reveals stable genomic regions for grain yield in CIMMYT spring bread wheat. *Front Genet*. 2020;11:589490.
79. Kolle G, Georgas K, Holmes GP, Little MH, Yamada T. CRIM1, a novel gene encoding a cysteine-rich repeat protein, is developmentally regulated and implicated in vertebrate CNS development and organogenesis. *Mech Dev*. 2000;90:181–93.

ACKNOWLEDGEMENTS

The authors would like to thank Melanie Oakes and Valentina Ciobanu for their valuable technical support, Dr. Maria Giulia Bacalini and Dr. Cristian Taccioli for their useful suggestions about the in-silico analyses. This study was funded by the NIH grant NIMH R21-MH115327-01.

AUTHOR CONTRIBUTIONS

FM and AB conceived the study. GM, PA, and AB performed the analyses. All authors contributed to writing the manuscript and approved the final version.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41398-023-02472-9>.

Correspondence and requests for materials should be addressed to Alessio Boattini or Fabio Macciardi.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023