www.nature.com/mp

Molecular Psychiatry

IMMEDIATE COMMUNICATION OPEN

Genome-wide tandem repeat expansions contribute to schizophrenia risk

Bahareh A. Mojarad[1], Worrawat Engchuan[1,2], Brett Trost[1,2], Ian Backstrom[1], Yue Yin[1], Bhooma Thiruvahindrapuram[1,2], Linda Pallotto[1], Aleksandra Mitina[1], Mahreen Khan[1,3], Giovanna Pellecchia[1,2], Bushra Haque[1,3], Keyi Guo[1], Tracy Heung[4,5], Gregory Costain[1,3,6], Stephen W. Scherer[1,2,3,7], Christian R. Marshall[8], Christopher E. Pearson[1,3], Anne S. Bassett[4,5,9,10] ✉ and Ryan K. C. Yuen[1,3,10] ✉

Tandem repeat expansions (TREs) can cause neurological diseases but their impact in schizophrenia is unclear. Here we analyzed genome sequences of adults with schizophrenia and found that they have a higher burden of TREs that are near exons and rare in the general population, compared with non-psychiatric controls. These TREs are disproportionately found at loci known to be associated with schizophrenia from genome-wide association studies, in individuals with clinically-relevant genetic variants at other schizophrenia loci, and in families where multiple individuals have schizophrenia. We showed that rare TREs in schizophrenia may impact synaptic functions by disrupting the splicing process of their associated genes in a loss-of-function manner. Our findings support the involvement of genome-wide rare TREs in the polygenic nature of schizophrenia.

Molecular Psychiatry (2022) 27:3692–3698; https://doi.org/10.1038/s41380-022-01575-x

INTRODUCTION

Schizophrenia is a major neuropsychiatric disorder with heritability estimated at ~79% [1]. Previous studies have revealed the role of copy number variants (CNVs) [2, 3] and small nucleotide variants in its pathogenesis [4]. However, those analyses were not designed to interrogate repetitive regions of the genome owing to challenges in detecting and interpreting such variation. In a separate study [5], we used a comprehensive analytic strategy to investigate tandem DNA repeats, which constitute ~6% of the human genome, and showed that genome-wide TREs are associated with the risk of autism spectrum disorder (ASD), a complex neurodevelopmental disorder with genetic risk that overlaps that of schizophrenia [6]. Our recent genome sequence analysis provided evidence supporting the involvement of TREs in schizophrenia through the identification of potentially damaging TREs in known disease-associated loci from a cohort of unrelated adults with schizophrenia [7].

Here, we used ExpansionHunter Denovo (EHdn) [8] and our established analytic approach to analyze TREs in the genomes of 257 unrelated adult cases with schizophrenia of European ancestry, 225 ancestry- and sequence-pipeline-matched individuals with no major neuropsychiatric disorders (non-psychiatric controls) [7], and in 2504 individuals from the 1000 Genomes Project [9] to estimate population frequency of TREs in cases and controls (*Methods*). Our study was driven by the historical observation of anticipation in schizophrenia [10], and the fact that most of the known tandem repeat disorders are caused by rare TREs [11, 12]. Therefore, we specifically assessed for large rare TREs in schizophrenia. Our approach interrogates the entire genome irrespective of prior knowledge of the presence or expected sequence of tandem repeats in any given region, and focuses on tandem repeats having motifs of 2–20 bp for which the total repeat tract length is greater than the sequencing read length (i.e., >150 bp) [8]. We define a tandem repeat to be expanded when its tract length is an outlier compared to lengths at that loci in other individuals [5] (*Methods*).

MATERIALS AND METHODS

Ethics statement

This study was approved by the Research Ethics Board at the Centre for Addiction and Mental Health (CAMH) (151/2002-02) and other local REBs. Written informed consent was obtained for all participants [7].

Samples sequencing, and genome alignment

We used genome sequencing data from Canadian individuals of European descent (257 with schizophrenia, and 225 with congenital heart disease (CHD) and no psychotic illness), as well as 2504 samples from the 1000 Genomes Project (1000G) [9]. The schizophrenia samples and non-psychiatric controls were assessed for quality, and prepared using TruSeq DNA library prep kits. These samples were sequenced on the Illumina

[1]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada. [2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada. [3]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [4]Clinical Genetics Research Program, Centre for Addiction and Mental Health, Toronto, ON, Canada. [5]The Dalglish Family 22q Clinic for Adults with 22q11.2 Deletion Syndrome, Toronto General Hospital, University Health Network, Toronto, ON, Canada. [6]Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada. [7]McLaughlin Centre, University of Toronto, Toronto, ON, Canada. [8]Genome Diagnostics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada. [9]Department of Psychiatry, University of Toronto, Toronto General Hospital Research Institute and Campbell Family Mental Health Research Institute, Toronto, ON, Canada. [10]These authors contributed equally: Anne S. Bassett, Ryan K.C. Yuen. ✉email: anne.bassett@utoronto.ca; ryan.yuen@sickkids.ca

Received: 30 December 2021 Revised: 7 April 2022 Accepted: 12 April 2022
Published online: 12 May 2022

HiSeq X platform (2 × 150 bp paired-end reads) at The Centre for Applied Genomics (TCAG, Toronto, Canada) and processed for alignment and genomic variant calling as previously described [7, 13]. The 1000 G samples were sequenced on the Illumina NovaSeq platform (2 × 150 bp paired-end reads). The 1000 G genome sequencing data are publicly available and we downloaded them via Amazon Web Services (s3://1000genomes/1000G_2504_high_coverage/data). All samples were aligned to the GRCh38/hg38 reference genome using BWA-mem [14]. The study protocol was approved by the Research Ethics Boards of The Hospital for Sick Children and CAMH. Informed consent was obtained from all participants at the recruitment locations.

Genome-wide identification of tandem repeats

Genome-wide detection of tandem repeats was performed as previously described [5]. Briefly, we used ExpansionHunter Denovo (EHdn; https://github.com/Illumina/ExpansionHunterDenovo) [8] to estimate the size and location of genomic tandem repeats. For a tandem repeat to be detected by EHdn, it must be larger than the sequence read length (for example, >150 bp). As a result, samples that did not meet this minimum size for a given region were left without size estimation by EHdn. EHdn estimates the size of a tandem repeat by counting the number of anchored in-repeat reads (IRRs), which are read pairs in which the first read (the IRR) contains repetitive sequence and the second read (the anchor) contains non-repetitive sequence that can be uniquely mapped to the reference genome, thus allowing the repeat's location to be determined. Although the EHdn sizes cannot be interpreted as exact numbers of base pairs, they are proportional to the number of base pairs comprising the repeat (see Fig. 1 of ref. [8]). To account for samples with different overall depths of coverage, the anchored IRR counts are normalized by the overall read depth of a given sample. We compared the tandem repeats identified by EHdn to tandem repeats in the human reference genome, derived from Tandem Repeats Finder (TRF) [15]. To support the accuracy of EHdn-predicted tandem repeat sizes, we genotyped the 8 rare exon-proximal repeat loci (9 motifs) identified in 13 individuals, along with 6 selected rare intronic repeat loci, using ExpansionHunter v.3.0.2 [16, 17], which estimates allele-specific tandem repeat sizes for each genomic coordinate and motif supplied by the user with high accuracy (precision = 0.91, recall = 0.99) [16, 17]. EHdn has been shown to be both sensitive (i.e., successfully rediscovered TREs in several disease-associated genes and was able to detect 77% of repeats >150 bp discovered by long-read sequencing [8], and specific (TREs that EHdn detects have been validated using orthogonal methods in several different studies [5, 7, 8, 18, 19]).

To determine more precise coordinates for input to ExpansionHunter, we identified coordinates from TRF that overlapped each locus. For each combination of TRF coordinates and EHdn motif, we used ExpansionHunter to estimate motif-specific (as detected by EHdn) tandem repeat sizes for the samples involved. We then calculated the Spearman correlation coefficient and *P* value between the EHdn-predicted tandem repeat sizes and the size estimated by ExpansionHunter (defined as either the size of the longest allele or the sum of the two allele sizes), aggregated over all of the EHdn-detected motifs for that locus (Supplementary Table S7). We also manually evaluated the presence of tandem repeat expansions (TREs) and the corresponding motif by inspecting reads from the BAM file for tandem repeats that were found to be expanded by EHdn for all.

Detection of rare tandem repeat expansions and sample quality assessment

We excluded tandem repeats with different size distribution between schizophrenia and CHD samples (two-sided Wilcoxon's signed-rank test $P < 0.05$) in order to avoid any potential technical biases in estimating tandem repeat size between the two cohorts. To detect rare TREs, we followed Trost et al. [5] by applying the non-parametric Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to identify outliers based on EHdn estimated tandem repeat size at each locus. We optimized two DBSCAN parameters as well as the population frequency cut-off for rare TRE identification (Supplementary Fig. S3; Supplementary Table S8). Based on a different set of rare TREs identified using different DBSCAN parameters and population frequency cut-off, burden test of exonic TREs and intergenic TREs were performed with the total number of rare TREs as a covariate. The DBSCAN parameters set (minclust = 11, eps = 2 × mode of EHdn sizes) and population frequency cut-off (frequency < 0.05) that provided the strongest signal in the exonic TREs burden test and weakest signal in intergenic TREs burden test was selected for the next step of the analysis. Sample quality assessment was done by inspecting the counts of total number of tandem repeat loci and TREs per sample. Anscombe transformation was done on the counts to put the count distribution closer to the gaussian distribution. Three control samples and five schizophrenia samples with the transformed counts exceeding 3 standard deviations from the mean of the transformed counts were tagged as outliers and excluded from the analysis. Rare TRE identification was then performed on the remaining set.

Burden analysis

To compare the prevalence of rare TREs in individuals with and without schizophrenia, we performed a logistic regression analysis by regressing the number of rare TREs on the affected status (unaffected = 0, affected = 1). For this analysis, we only included tandem repeats on autosomal chromosomes to avoid sex bias. Biological sex and the total number of rare TREs per individual were used as covariates. To test the burden of TREs in different functional elements (for example, exons and introns), we separated the genome (RefSeq, GRCh38) into different functional elements: upstream (1 kb upstream of the transcriptional start site, TSS), 5′ untranslated regions (5′ UTR), exon, core splice site, intron, 3′ UTR and downstream (1 kb downstream of transcription termination sites) [4]. If any rare TRE affected more than one functional element, we prioritized the effects based on their impact on the corresponding genes predicted by ANNOVAR (October 2019 release) [20]. One-sided Wald test was performed, assuming a higher burden of rare TREs in cases with schizophrenia. Empirical *p* values from 10,000 case-control label permutations were reported.

Experimental validation of tandem repeat expansions

Validation of the *SHANK1* TREs detected by EHdn was completed by PCR, gel electrophoresis, and Sanger sequencing. We designed primers flanking the repeat of interest with the following sequences: 5′-CCTATCTCCTATGAATGGACGAC-3′ and 5′-GATGCCGTTAAATGCGAGTTTC-3′. We performed PCR on the samples using HotStarTaq DNA polymerase (Qiagen), a primer annealing temperature of 63 °C, and an elongation time of 2 min. We then ran the PCR products on a 1.2% agarose gel to confirm the size of the repeats, and performed Sanger sequencing to confirm the sequence. Other DNA samples from this cohort without a predicted expansion in *SHANK1* were run under the same conditions as negative controls (Supplementary Fig. S2).

Validation of the *DAB1* TREs detected by EHdn was completed by PCR, gel electrophoresis, and Sanger sequencing. Primers flanking the repeat of interest were used with the following sequences: 5′-ATTTGCCCTTTGCTGATTGA-3′ and 5′-TGAAACTGAGGCTCAAAATGA-3′ [21]. We performed PCR on the samples using PrimeSTAR GXL DNA Polymerase (Takara), a primer annealing temperature of 61 °C, and an elongation time of 2 min. We then ran the PCR products on a 1.2% agarose gel to confirm the size of the repeats, and performed Sanger sequencing to confirm the correct region was amplified. Other DNA samples from this cohort without a predicted expansion in *DAB1* were run under the same conditions as negative controls.

Statistical comparison of clinical features

We hypothesized that individuals with schizophrenia compared with CHD and no psychotic illness (Fig. 1A, B), and differentially within schizophrenia those with clinical features, including family history of schizophrenia in first degree relatives, intellectual disability and syndromic features (Fig. 2A and 2C), would have a greater contribution from the genic and exonic (i.e., exon-proximal, within 300 bp of exon junctions) TREs identified. Therefore, we used the non-parametric one-sided Wilcoxon signed-rank test to compare the datasets unless stated otherwise.

Enrichment in common variant risk

For the 193 TRE-associated genes that we identified in this schizophrenia sample, we used MAGMA v.1.09b11 as described previously [22] to determine whether they were enriched in common variant risk loci for schizophrenia and other traits. Specifically, we compared our 193 TRE-associated gene set against summary statistics from genome-wide association studies (GWASs) for schizophrenia [23], ASD [24], attention deficit/hyperactivity disorder [25], educational attainment [26] and (as a negative control) height [27] (Fig. 2B, Supplementary Table S9). We also tested for an enrichment of 193 TRE-associated genes in the 655 genes within 270 refined genome-wide significant loci that involve fewer than 4 causal variants in the latest schizophrenia GWAS of 69,369 cases and 236,642 controls [14]. We applied one-sided Fisher's Exact test to compare

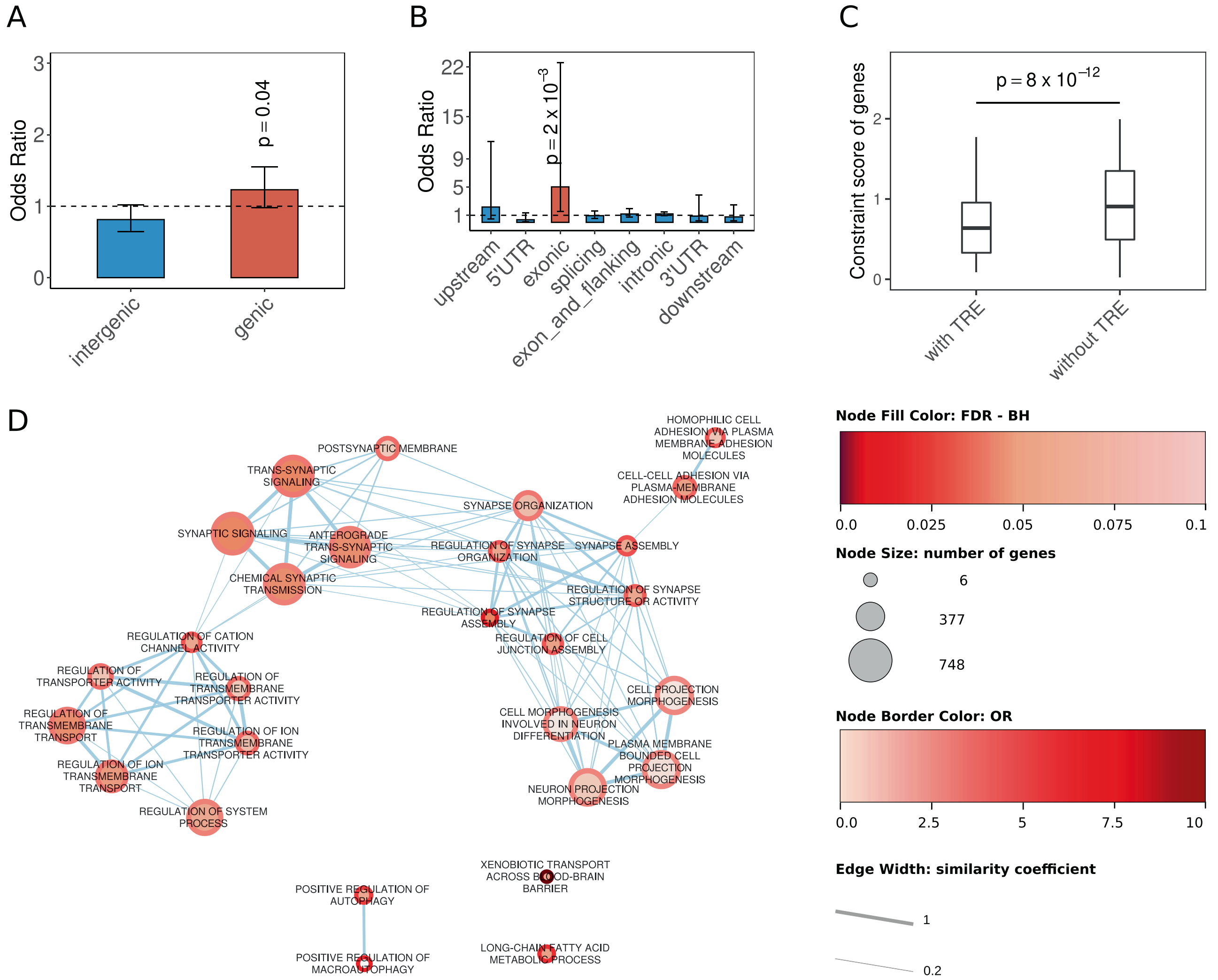

Fig. 1 Genetic features and functional impact of rare TREs. A Burden analysis of rare TREs located near (genic) or outside (intergenic) genes. **B** Burden analysis of rare TREs with different genomic features in individuals with schizophrenia relative to non-psychiatric controls. Red bar indicates significant enrichment of exon-proximal (labelled as exonic) rare TREs in individuals with schizophrenia. Horizontal dashed line represents odds ratio = 1. **C** Distributions of gnomAD observed/expected (o/e) upper bound LOEUF values are shown for 182 genes with rare TREs (182 of 193 TRE-associated genes have scores) in the 220 (of 257) individuals with schizophrenia, compared with 18,990 genes with no such TREs identified in the schizophrenia cohort (one-sided Wilcoxon rank-sum test). Minima and maxima indicate 3× the interquartile range-deviated o/e upper bounds from the median and the centre indicates the median of the o/e upper bound values. **D** A map of gene functions enriched in genes associated with rare TREs. Each node represents a gene-set with its size proportional to the number of genes involved. The significant enrichment of a particular function was assessed by one-sided Fisher's exact test. The false discovery rate and odds ratio represented by the color gradient and the width of edge is proportional to the similarity coefficient between gene functions. Synaptic mechanisms prevail.

the enrichment of GWAS signals in TRE-associated genes ($n = 193$) against other genes that are not associated with rare TRE ($n = 19,276$).

Functional enrichment analysis

For the functional enrichment test, we used a one-sided Fisher's exact test and gene ontology terms (from the R Bioconductor library org.Hs.eg.db v3.13.0), restricting the sets to those with a number of annotated genes between 5 and 1000. The p values were then corrected for multiple comparisons with the Benjamini Hochberg false discovery rate procedure. The results were loaded in Cytoscape v3.8.1 with the Enrichment map plugin v3.3.3 [28], filtering out sets with false discovery rate >0.1 and similarity coefficient (combined coefficient $k = 0.5$) lower than 0.2, to retain only the most significant terms.

RESULTS

After quality assessment and parameter optimization (*Methods*), we performed a burden analysis comparing rare TREs (<0.5% frequency in 1000 Genomes Project individuals) in individuals with schizophrenia and in non-psychiatric controls. We identified 583 rare TREs in 436 distinct regions in 220 individuals with schizophrenia (Supplementary Table S1); 199 of these distinct regions were genic (involving 193 genes, hereafter referred to as TRE-associated genes, including 6 genes with multiple repeat motifs/regions identified). In individuals with schizophrenia, rare TREs tend to be located within genes ($p = 0.04$, Fig. 1A), and more likely to be at exon junctions (odds ratio (OR) = 5.03, $p = 2 \times 10^{-3}$, Fig. 1B). Fine mapping of the eight exon-proximal TREs revealed their precise locations to be in intronic or untranslated regions with close proximity (<300 bp) to protein-coding exons. This included a CTG expansion in myotonic dystrophy-linked *DMPK* we reported previously [7] (Supplementary Information and Supplementary Table S2). The proportion of schizophrenia cases with at least one rare exon-proximal TRE is 5.16%, while the proportion of controls with at least one rare one exon-proximal TRE, after

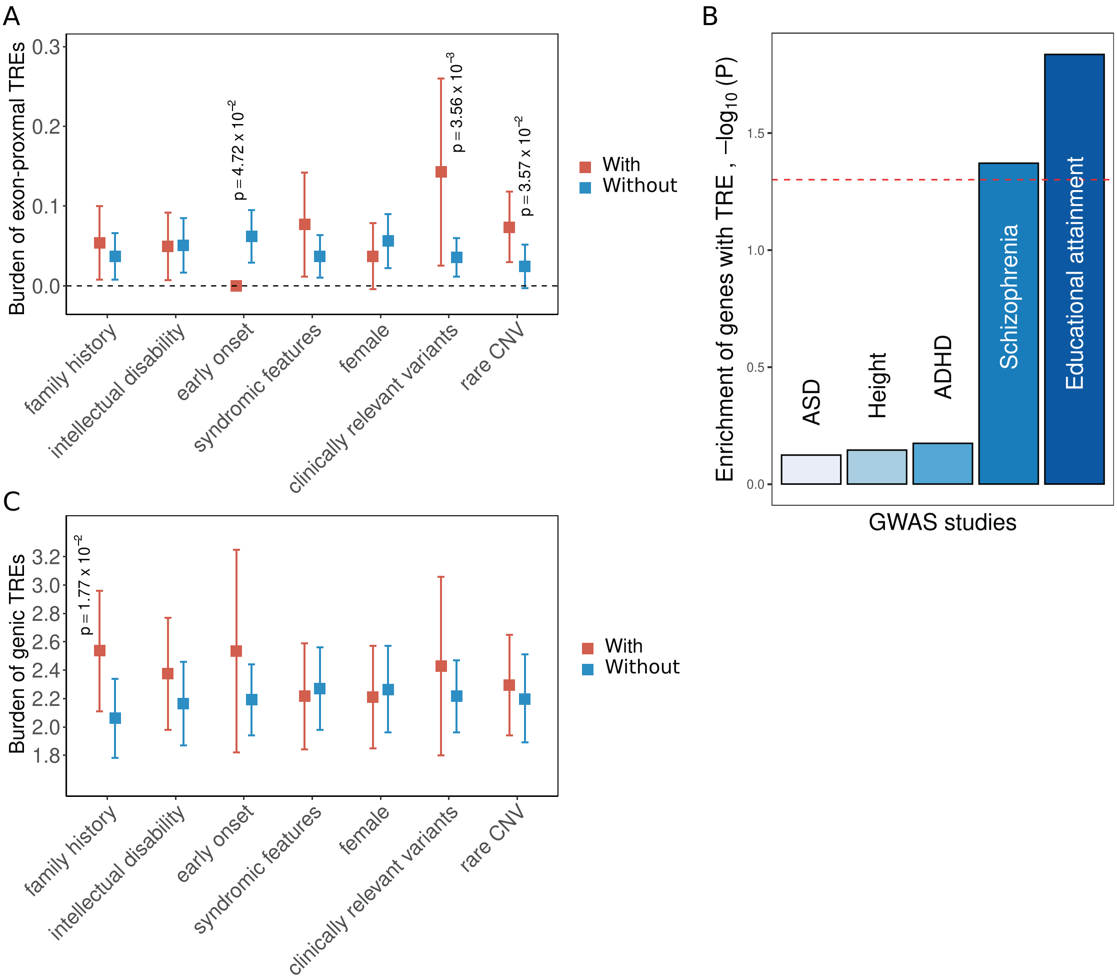

Fig. 2 Genetic and clinical features involved in rare genic TREs in schizophrenia. The burden of (**A**) rare exon-proximal and (**C**) rare genic TREs in schizophrenia was analyzed with respect to presence/absence of seven variables (*x*-axis). Orange and blue colored boxes indicate results for TRE-containing individuals, with and without each of the seven variables, respectively, with vertical bars representing 95% confidence intervals; *p* values are provided above variables noting significant differences. No individuals with a rare exon-proximal TRE had an early age at onset of schizophrenia. **B** MAGMA was used to assess the 193 genes with rare TREs detected by our pipeline against proximity (<10 kb) to common risk variants from GWAS studies for schizophrenia, autism spectrum disorder (ASD), attention deficit/hyperactivity disorder (ADHD), educational attainment, and (as a negative control) height (*Methods*). The dashed red line represents association with *p* value equal to 0.05.

correcting for bias in the intergenic region, is 1.20%. Thus, we estimate from this study that the rare exon-proximal TREs may collectively account for 3.96% of the risk in schizophrenia.

In 139 of the individuals with schizophrenia, there were 222 rare intronic TREs in 160 distinct regions of 155 genes (including 5 genes with multiple repeat motifs/loci), representing the largest subcategory of the genic region. Of these, 51 individuals had rare intronic TREs in one or more of the 38 genes that are associated with neurological abnormality, abnormal behavior or nervous system abnormality in Mammalian Phenotype Ontology, including genes previously associated with schizophrenia such as *DCLK1*, *ERBB4*, *GRIK4*, *GRIN2A*, *SHANK1*, and *VIPR2* (Supplementary Table S3). Gene-set analysis of rare exon-proximal and intronic TREs identified a significant enrichment of genes involved in postnatal brain expression, such as *GRIN2A* and *SHANK1* (OR = 1.83, $p = 8.6 \times 10^{-3}$, false discovery rate = 0.2) (Supplementary Table S4 and Supplementary Fig. 1).

The motifs involved in the TREs are diverse in terms of size and sequence (Supplementary Table S1). The GC content of the involved motifs of the rare TREs studied is significantly lower than in the unexpanded tandem repeats, or in the known pathogenic repeats, but is higher than that found in ASD (Supplementary Fig. S2). TRE-associated genes were significantly more constrained than other genes as measured by the GnomAD loss-of-function observed/expected upper bound fraction (LOEUF) [29] ($p = 3 \times 10^{-10}$) (Fig. 1C). Rare TREs are found more frequently in regions that are closer to the splice junction (Supplementary Fig. 3), and the TRE-associated genes are predominantly involved in synaptic functions and signaling pathways (Fig. 1D), which are commonly known to play a role in brain development and believed to be involved in the etiology of schizophrenia [2, 4, 30, 31]. Collectively, these findings suggest that the rare TREs in schizophrenia may impact synaptic functions by disrupting the splicing process of their associated genes in a loss-of-function manner.

Our previous analysis of this same community-based cohort of adults with schizophrenia included 33 individuals with clinically relevant CNVs and small nucleotide variants, constituting 12.8% of individuals studied [7, 32, 33]. Of the 13 individuals with the eight rare exon-proximal TREs, five had at least one other clinically relevant (non-TRE) rare variant (i.e., small nucleotide/CNVs) (Supplementary Table S5); [7, 32, 33] a significant association ($p = 3.56 \times 10^{-3}$) (Fig. 2A, and Supplementary Information). This supports schizophrenia as a complex disorder involving multiple genetic risk factors [30].

Next, we used MAGMA [22] to integrate summary statistics from GWASs of five traits (ASD, height, attention deficit hyperactivity disorder, schizophrenia, and educational attainment), and examined whether common genetic variation influencing these traits were located within 10 kb of the 193 TRE-associated genes we identified (*Methods*). We determined that signals for schizophrenia and for educational attainment (where the GWAS signal also correlates with schizophrenia), but not the other three GWAS signals tested (*Methods*), showed significant enrichment for our TRE-associated genes (Fig. 2B), further supporting their contribution to the polygenic risk of schizophrenia. Eleven of the 193 TRE-associated genes, including *DPYD* and *EFNA5*, were disproportionally found amongst the 655 genes potentially tagged by genome-wide significant common variant signals at 270 loci in the latest schizophrenia GWAS ($\mathrm{OR} = 2.65$, $p = 5 \times 10^{-3}$), but only two (*GRIN2A* and *MYT1L*) were in the 114 protein-coding genes prioritized using current variant mapping and expression methods [31].

In our schizophrenia cohort, we also found that individuals with a family history of schizophrenia were more likely to carry rare TREs in genic regions ($p = 1.77 \times 10^{-2}$) (Fig. 2C), suggesting that these individuals may have inherited the rare TREs. To investigate the involvement of rare genic TREs in families with a history of schizophrenia, we performed genome sequencing on 63 additional individuals (30 of whom were diagnosed with schizophrenia) from 14 independent families with an extended family history of schizophrenia [34]. In two probands of these 14 families, we detected rare TREs in three genic regions (*CALCOCO2* and *FXN* in one family, and *SHANK1* in the other family) that were also identified in our primary schizophrenia cohort (Supplementary Table S6). In these two families, all five genome-sequenced individuals with schizophrenia carried the detected rare genic TREs, even though the same expansions can be found in some of the unaffected family members (Supplementary Table S6). These include two individuals with the AAAAG expansion in *SHANK1* that targeted genotyping delineated to have originated from the paternal side of the family (Supplementary Fig. S4). These findings further substantiate the contribution of rare TREs to the heritable risk of schizophrenia.

Across both the original and familial cohorts, we identified and validated three known disease-associated tandem repeats. In the family that has rare TREs in *FXN*, individual III-1 has an expanded repeat size in *FXN* in the pathogenic range for Friedreich ataxia, and about three times larger on average than those detected in the individuals from the previous generation, with a diagnosis of this autosomal recessive disorder confirmed in clinical records (Supplementary Table S6). In the unrelated primary schizophrenia cohort, we previously reported a > 200 repeat-long CTG expansion in *DMPK*, within the known pathological range, and consistent with a history of myotonic dystrophy in the individual's family [7]. Also, for the three unrelated individuals with rare intronic TREs in *DAB1* (Supplementary Table S3), encoding a reelin adaptor protein, we confirmed the sizes of expanded ATTTT repeats (Supplementary Fig. S5), which are comparable to rare TREs reported by others at this locus for spinocerebellar ataxia type 37 [21]. Consistent with having no diagnosis, or clinical signs, of spinocerebellar ataxia, however, the three individuals in our study had no ATTTC repeat insertions in the expanded repeat tract in *DAB1* [21].

DISCUSSION

We demonstrate that rare TREs, in particular those that are intronic and close to exons, are an important class of variants contributing to the etiology of schizophrenia. The functional and constraint profiles of the implicated genes, the proximity of these genes to GWAS signals for schizophrenia, and the proximity of the rare repeats to coding sequence and to splice junctions, are consistent with the relevance of rare intronic and exon-proximal TREs to schizophrenia-related mechanisms. We estimate from this study that the rare exon-proximal TREs may collectively account for 3.96% of the risk in schizophrenia.

While epigenetic modifications are known as a gene-disrupting mechanism in some well-known TREs [12, 35], such as CGG repeat expansions in *FMR1*, we found that CG-containing motifs are uncommon in the tandem repeats expanded in schizophrenia (Supplementary Table S1), and in fact are significantly less common than the unexpanded tandem repeats or the known pathogenic repeats (Supplementary Fig. S2). This suggests that epigenetic modifications are unlikely to represent the main mechanism involved in schizophrenia. Instead, our results show that rare TREs in schizophrenia differentially impact synaptic functions (Fig. 1D), and that the mechanism is likely to involve disrupting the splicing process of their associated genes in a loss-of-function manner (Fig. 1C, Supplementary Fig. S3).

One example of the synaptic genes recurrently affected by rare TREs is *SHANK1*, a postnatal brain-expressed gene that encodes scaffold proteins that are required for the development and function of neuronal synapses [36]. Genetic variants, including rare CNVs encompassing *SHANK1*, have been observed in individuals with non-syndromic ASD [37, 38] (Supplementary Information). Further studies are warranted to delineate the mechanisms of TREs in regulating the expression of *SHANK1*, and for the other TRE-associated genes identified, during brain development.

Family history of schizophrenia was the only clinical factor assessed that was significantly associated with rare genic TREs (Fig. 2C). This may be consistent with inheritance of TREs and with the historical observation of clinical anticipation in familial schizophrenia [10]. Notably, a higher degree of polygenic risk for schizophrenia is also associated with positive family history [7]. As for most studies of genetic variants in schizophrenia, there was no association of rare TREs with age at onset. Consistent with findings for other rare variants associated with schizophrenia, and with incomplete penetrance, we found the same TREs to be present in some of the unaffected members in our family studies (Supplementary Table S6). For individuals with clinically relevant variants, or with rare CNVs who a priori were deliberately oversampled in this cohort [7], there was elevated burden of rare exon-proximal TREs (Fig. 2A), suggesting the possibility that rare TREs may act additively to increase the risk of schizophrenia.

Our approach may have increased the power to detect the rare TREs that collectively contribute to schizophrenia in a relatively small sample. We chose an unbiased genome-wide assessment of large rare TREs (>150 bp). Short de novo tandem-repeat expansions and contractions (e.g., of repeat size ≤150 bp) may impose additional schizophrenia risk [39]. These can be detected from sequence data by standard small variant calling algorithms as small insertion/deletions (i.e., indels), thus their contribution could well have been captured by previous exome or genome sequencing studies of schizophrenia [4, 40]. Complementary designs, including studies with a much larger sample size, are required to assess variants likely to have small effect sizes, such as common tandem repeats.

Due to the rarity of the TREs studied, determining the penetrance for individual expanded loci is impossible in this study. However, some of the rare TREs identified, such as the CTG repeat expansion in *DMPK*, have also been found in ASD [5]. This suggests a pleiotropic effect of TREs, which is consistent with many other schizophrenia-associated genetic variants [2, 40]. Further characterization of their effects and inheritance across large cohorts of multiple neuro-psychiatric/developmental disorders, and across generations within families, may help resolve the pleiotropic effects and penetrance for individual TREs. Future studies should also examine the potential impact of somatic TREs, which we have not assessed here due to the limitations of existing algorithms, and the use of blood, not brain, samples.

Involvement of genome-wide TREs in schizophrenia may help explain the clinical genetic anticipation that has long been recognized in schizophrenia and suspected to be related to tandem repeats, with the additional possibility that multiple risk variants accumulate over generations [10, 41]. The current study adds to support for both possibilities. The enrichment of common variant signals for schizophrenia GWAS at the TRE-associated genes identified further supports TRE contributions to the genetic architecture of schizophrenia. Our findings suggest rare TRE as a potential source of some of the missing heritability for schizophrenia, and highlight the necessity of further genome sequencing studies of TREs in other complex disorders for which missing heritability remains to be identified [42].

DATA AVAILABILITY

The 1000G genome-sequencing data are publicly available via Amazon Web Services (s3://1000genomes/1000G_2504_high_coverage/data).

CODE AVAILABILITY

Code used in this paper is available from the corresponding author upon reasonable request.

REFERENCES

1. Hilker R, Helenius D, Fagerlund B, Skytthe A, Christensen K, Werge TM, et al. Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. Biol Psychiatry. 2018;83:492–8.
2. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017;49:27–35.
3. Rees E, Walters JT, Georgieva L, Isles AR, Chambert KD, Richards AL, et al. Analysis of copy number variations at 15 schizophrenia-associated loci. Br J Psychiatry. 2014;204:108–14.
4. Howrigan DP, Rose SA, Samocha KE, Fromer M, Cerrato F, Chen WJ, et al. Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. Nat Neurosci. 2020;23:185–93.
5. Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature. 2020;586:80–6.
6. McCarthy SE, Gillis J, Kramer M, Lihm J, Yoon S, Berstein Y, et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. Mol Psychiatry. 2014;19:652–8.
7. Mojarad BA, Yin Y, Manshaei R, Backstrom I, Costain G, Heung T, et al. Genome sequencing broadens the range of contributing variants with clinical implications in schizophrenia. Transl Psychiatry. 2021;11:84.
8. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt J, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020;21:102.
9. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
10. Bassett AS, Honer WG. Evidence for anticipation in schizophrenia. Am J Hum Genet. 1994;54:864–70.
11. Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? Am J Hum Genet. 2021;108:764–85.
12. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. Genome Res. 2022;32:1–27.
13. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20:602–11.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
15. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
16. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35:4754–6.
17. Dolzhenko E, van Vugt J, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27:1895–903.
18. Qaiser F, Sadoway T, Yin Y, Zulfiqar Ali Q, Nguyen CM, Shum N, et al. Genome sequencing identifies rare tandem repeat expansions and copy number variants in Lennox-Gastaut syndrome. Brain Commun. 2021;3:fcab207.
19. Rafehi H, Szmulewicz DJ, Bennett MF, Sobreira NLM, Pope K, Smith KR, et al. Bioinformatics-Based Identification of Expanded Repeats: a Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS. Am J Hum Genet. 2019;105:151–65.
20. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
21. Seixas AI, Loureiro JR, Costa C, Ordonez-Ugalde A, Marcelino H, Oliveira CL, et al. A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. Am J Hum Genet. 2017;101:87–103.
22. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219.
23. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
24. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51:431–44.
25. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019;51:63–75.
26. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018;50:1112–21.
27. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173–86.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
29. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
30. Birnbaum R, Weinberger DR. Genetic insights into the neurodevelopmental origins of schizophrenia. Nat Rev Neurosci. 2017;18:727–40.
31. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022;604:502–8.
32. Costain G, Lionel AC, Merico D, Forsythe P, Russell K, Lowther C, et al. Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays. Hum Mol Genet. 2013;22:4485–501.
33. Lowther C, Merico D, Costain G, Waserman J, Boyd K, Noor A, et al. Impact of IQ on the diagnostic yield of chromosomal microarray in a community sample of adults with schizophrenia. Genome Med. 2017;9:105.
34. Brzustowicz LM, Hodgkinson KA, Chow EW, Honer WG, Bassett AS. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. Science. 2000;288:678–82.
35. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19:286–98.
36. Mossa A, Pagano J, Ponzoni L, Tozzi A, Vezzoli E, Sciaccaluga M, et al. Developmental impaired Akt signaling in the Shank1 and Shank3 double knock-out mice. Mol Psychiatry. 2021;26:1928–44.
37. May HJ, Jeong J, Revah-Politi A, Cohen JS, Chassevent A, Baptista J, et al. Truncating variants in the SHANK1 gene are associated with a spectrum of neurodevelopmental disorders. Genet Med. 2021;23:1912–21.
38. Sato D, Lionel AC, Leblond CS, Prasad A, Pinto D, Walker S, et al. SHANK1 Deletions in Males with Autism Spectrum Disorder. Am J Hum Genet. 2012;90:879–87.
39. Hannan AJ. Repeat DNA expands our understanding of autism spectrum disorder. Nature. 2021;589:200–2.
40. Rees E, Han J, Morgan J, Carrera N, Escott-Price V, Pocklington AJ, et al. De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia. Nat Neurosci. 2020;23:179–84.
41. Husted J, Scutt LE, Bassett AS. Paternal transmission and anticipation in schizophrenia. Am J Med Genet. 1998;81:156–62.
42. Hannan AJ. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet. 2010;26:59–65.

ACKNOWLEDGEMENTS

We are grateful to all patients and their families for their participation in this study. We thank The Centre for Applied Genomics (TCAG, a node of CGEn), which is supported by the Canada Foundation of Innovation, Genome Canada, the Hospital for Sick Children, and partners. RKCY is supported by The Hospital for Sick Children's Research Institute, SickKids Catalyst Scholar in Genetics, Brain Canada, The Azrieli

Foundation, the University of Toronto McLaughlin Centre and the Nancy E.T. Fahrner Award. This work was also supported by the Canadian Institutes of Health Research (CIHR) (MOP-89066 to ASB, MOP-111238 to ASB, PJT-175329 to RKCY, PJT-178161 to ASB and RKCY), and a Canada Research Chair in Schizophrenia Genetics and Genomic Disorders (Tier 1, 2009-2016 to ASB). ASB holds the Dalglish Chair in 22q11.2 Deletion Syndrome at the University Health Network and University of Toronto. BT was funded by the Canadian Institutes for Health Research Banting Postdoctoral Fellowship and the Brain Canada Canadian Open Neuroscience Platform Research Scholar Award. BAM was supported by the Restracomp Award from The Hospital for Sick Children. SWS holds the Northbridge Chair in Pediatric Research at The Hospital for Sick Children and the University of Toronto.

AUTHOR CONTRIBUTIONS

RKCY conceived and coordinated the study. WE and BT processed genome data for tandem repeat detection and analysis. IB, LP, BH, KG, AM, and MK performed laboratory experiments. WE, YY, and GP performed statistical analyses with additional contributions from BT, and BThiru. YY and TH performed genotype-phenotype correlation and analyses. BAM, WE, and GP performed other miscellaneous analyses. ASB managed, recruited, diagnosed and with GC examined the recruited participants. RKCY, BAM, and ASB wrote the paper with contributions from WE, BT, IB, SWS, CRM, GC, and CEP. All authors read, reviewed and approved the final paper.

COMPETING INTERESTS

SWS is on the Scientific Advisory Committee of Population Bio, serves as a Highly Cited Academic Advisor for King Abdulaziz University, and intellectual property from aspects of his research held at The Hospital for Sick Children are licensed to Athena Diagnostics and Population Bio.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41380-022-01575-x.

Correspondence and requests for materials should be addressed to Anne S. Bassett or Ryan K. C. Yuen.

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.