Schizophrenia risk from locus-specific human endogenous retroviruses

Rodrigo R.R. Duarte[a-c], Matthew L. Bendall[c,d], Miguel de Mulder[c], Christopher E. Ormsby[e], Greta A. Beckerle[c], Sashika Selvackadunco[f], Claire Troakes[f], Gustavo Reyes-Terán[e], Keith A. Crandall[d], Deepak P. Srivastava[†b,g], Douglas F. Nixon[†c] and Timothy R. Powell*[†a,c]

† Shared senior authors.

[a] Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.

[b] Department of Basic & Clinical Neuroscience, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.

[c] Division of Infectious Diseases, Weill Cornell Medicine, Cornell University, New York, NY, USA.

[d] Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA.

[e] Center for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico.

[f] MRC London Neurodegenerative Diseases Brain Bank, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom.

[g] MRC Centre for Neurodevelopmental Disorders, King's College London, London, United Kingdom.

* Correspondence to: Dr. Timothy R. Powell, Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London SE5 8AF, United Kingdom. E-mail: timothy.1.powell@kcl.ac.uk. Tel: +44 (0)20 7848 5361.

Abstract

Schizophrenia genome-wide association studies highlight the substantial contribution of risk attributed to the non-coding genome where human endogenous retroviruses (HERVs) are encoded. These ancient viral elements have previously been overlooked in genetic and transcriptomic studies due to their poor annotation and repetitive nature. Using a new, comprehensive HERV annotation, we found that the fraction of the genome where HERVs are located (the 'retrogenome') is enriched for schizophrenia risk variants, and that there are 148 disparate HERVs involved in susceptibility. Analysis of RNA-sequencing data from the dorsolateral prefrontal cortex of 259 schizophrenia cases and 279 controls from the CommonMind Consortium showed that HERVs are actively expressed in the brain (n = 3,979), regulated in *cis* by common genetic variants (n = 1,759), and differentially expressed in patients (n = 81). Convergent analyses implicate LTR25_6q21 and ERVLE_8q24.3h as HERVs of etiological relevance to schizophrenia, which are co-regulated with genes involved in neuronal and mitochondrial function, respectively. Our findings provide a strong rationale for exploring the retrogenome and the expression of these locus-specific HERVs as novel risk factors for schizophrenia and potential diagnostic biomarkers and treatment targets.

Introduction

Human endogenous retroviruses (HERVs) are remnants of genetic material acquired through our evolutionary past which originated from the infection of germline cells with ancient retroviruses. These viruses multiplied through a copy-and-paste mechanism and were eventually endogenized (i.e., vertically transmitted), and now constitute approximately 8% of the genome[1,2]. These repetitive sequences were generally assumed to be transcriptionally inactive in the modern genome, having a purely regulatory function due to the retainment of the viral promoters (long-terminal repeats, LTRs). However, certain HERVs were co-opted to serve novel specialized roles, including in the regulation of embryonic development[3,4] and neural progenitor cells[5,6]. They have also been implicated in neuropsychiatric conditions such as amyotrophic lateral sclerosis[7-9], major depressive disorder, bipolar disorder, and schizophrenia[10-12]. Despite their abundance in the genome and relevance to disease and fundamental aspects of human biology, the location and function of most HERVs remain elusive to-date.

Recent large genome-wide association studies (GWAS) comparing schizophrenia cases and non-affected individuals enabled the identification of polymorphisms mediating risk for this disorder[13-15]. These studies highlight the substantial contribution of risk attributed to the non-coding genome, where HERVs are encoded, which are often overlooked in both genomic and transcriptomic studies[16]. Until recently, there was no comprehensively annotated map of HERVs in the genome, and there were no computationally efficient tools to analyze the expression of these repetitive sequences with single-locus resolution. Consequently, previous studies were unable to test whether locus-specific HERVs were genetically associated with traits of interest, or to assess their expression in conditions of interest, or to distinguish expressed from dormant HERVs from within the same family, which contributed to the generation of inconclusive findings. For example, Karlsson and colleagues[17] reported decreased ERV9 expression in the brain of schizophrenia patients, whereas Diem and collaborators[18] found the opposite. While the heterogeneity of schizophrenia is also likely to be a contributing factor for these contradictory findings, there are approximately 120 copies of ERV9 in the genome[19,20] that likely also confounded these reports.

Recently, there have been significant advances in the annotation of HERVs in the genome[20-24], and in the development of tools that are able to map repetitive sequences in RNA-sequencing data to their most likely source of origin in the genome[20,25,26]. These improvements allow for the identification and quantification of specific HERV copies associated with traits of interest. These advances, combined with modern population genetic methods, enabled us to identify the contribution of the 'retrogenome' (the full genomic complement of HERVs) and expression of locus-specific HERVs to schizophrenia etiology and neurobiology.

Results

The contribution of polymorphisms in the retrogenome to schizophrenia risk

A recently developed annotation of putatively functional HERV elements describes 14,968 elements dispersed throughout the genome, which were defined based on the presence of LTRs and remnants of genetic elements which code for viral proteins like *env, gag* or *pol*[20]. We assessed the contribution of common genetic variants within HERVs to schizophrenia susceptibility using GARFIELD[27]. This pipeline quantifies the co-localization of GWAS results with variants in annotation categories (i.e. variants in the retrogenome), assessing significance using linear models that control for minor allele frequency and linkage disequilibrium. For comparison, enrichment was also calculated for heritable neuropsychiatric traits and non-neuropsychiatric phenotypes: height, body mass index, coronary artery disease, Crohn's Disease, type 2 diabetes, neuroticism, eczema, major depressive disorder, bipolar disorder, Alzheimer's disease, attention deficit hyperactivity disorder, amyotrophic lateral sclerosis, and autism spectrum disorder (**Figure 1**; full results on **Supplemental Table 1**). Polymorphisms within HERVs were significantly enriched for genome-wide significant variants ($P < 5.00 \times 10^{-8}$) associated with schizophrenia ($P_{enrichment} = 1.88 \times 10^{-5}$, $P_{corrected}$ [for 14 traits and 2 GWAS thresholds tested] $= 5.27 \times 10^{-4}$, $\beta = 0.90$, 95% CI [0.49, 1.31]). Analysis of variants associated with these traits under a more relaxed P cut-off ($P < 5.00 \times 10^{-5}$) also showed an enrichment for schizophrenia only ($P_{enrichment} = 4.56 \times 10^{-7}$, $P_{corrected} = 1.28 \times 10^{-5}$, $\beta = 0.57$, 95% CI [0.35, 0.80]). These findings suggest a role for the retrogenome in the etiology of this neurodevelopmental disorder.

<<< **Figure 1** >>>

The contribution of locus-specific HERVs and HERV families to schizophrenia susceptibility

We co-localized polymorphisms within individual HERVs and those associated with schizophrenia by GWAS using MAGMA[28], which calculates gene-level statistics and weighted p-values based on summary statistics whilst adjusting for gene size, single nucleotide polymorphism (SNP) density and linkage disequilibrium. This gene-level enrichment analysis revealed that 148 HERVs from multiple families were significantly enriched for risk variants implicated in schizophrenia, after correcting for

the number of HERVs tested ($P_{corrected}$ < 0.05 / 12,389 HERVs in chromosomes 1-22, excluding those at the major histocompatibility locus; **Supplemental Table 2**). The quantile-quantile plot highlights the contribution of many locus-specific HERVs associated with risk, compared to an expected normal distribution (**Figure 2A**), and the Manhattan plot shows their diverse genomic location (**Figure 2B**).

To explore the contribution of HERV families towards schizophrenia risk, we investigated whether any of the 60 families defined in the HERV annotation were overrepresented in the list of schizophrenia-associated HERVs using a gene-set enrichment analysis in MAGMA. Each HERV was assigned to a family based on the RepBase model that most closely matched the internal region sequences[20,29]. We observed a nominal association between schizophrenia and the HERVL40 family (P = 0.02, β = 0.14 ± 0.02, SE = 0.07), but this did not survive multiple testing correction ($P_{corrected}$ [for 60 families] > 0.05; **Supplemental Figure 1**). These findings suggest that risk for schizophrenia attributed to the retrogenome may occur via locus-specific sequences, as opposed to entire HERV families.

<<< **Figure 2** >>>

HERV expression in the dorsolateral prefrontal cortex

To advance our understanding of HERV expression in the adult brain, we assessed global HERV expression (the retrotranscriptome) in the dorsolateral prefrontal cortex (DLPFC) of 593 post-mortem individuals (N = 593), including 279 unaffected controls, 259 schizophrenia patients, 47 bipolar disorder patients and 8 cases broadly diagnosed with an affective disorder. This was achieved by applying the Telescope pipeline[20] to the RNA-seq data from the CommonMind Consortium (CMC) dataset. Analysis of these samples revealed that 3,979 HERVs were consistently expressed in the DLPFC according to DESeq2 independent filtering criteria (mean normalized counts = 40.79 [37.8, 43.78]; **Figure 3A**). We performed RT-qPCR to confirm the expression of five arbitrarily selected HERVs in an independent post-mortem cohort of control individuals from the London Neurodegenerative Diseases Brain Bank (N = 10; **Supplemental Material; Supplemental Figure 2**).

<<< **Figure 3** >>>

Cis-acting HERV eQTLs in the DLPFC

To understand how HERVs are regulated in the brain, we performed an expression quantitative trait loci (eQTL) analysis using all samples (N = 593). Such analysis aims to identify SNPs that explain variation in the expression of HERVs residing in close proximity, to inform about the basic processes responsible for HERV regulation, and to complement our genetic enrichment analyses. The genomic coordinates from the expressed HERVs were remapped to hg19 positions to match genotype information, and only HERVs from chromosomes 1-22 were analyzed (total = 5,349 HERVs, which includes lowly expressed HERVs, according to DESeq2's internal filtering criteria). This analysis revealed that 1,759 HERVs were regulated in *cis* by 1,622 SNPs located within a 1 Mb window upstream or downstream the start site of the HERV annotation, under the false discovery rate of 5% ($q < 0.05$). The majority of eQTLs were located within a 10 kb window upstream or downstream of the annotation start site from the HERV they regulate (**Figure 3B**). Of the 148 HERVs that were enriched for schizophrenia variants, we observed that 19 were significantly regulated by eQTLs in the DLPFC (**Supplemental Tables 2 and 3**, highlighted in green; **Figure 3C**).

Case-control differences in HERV expression

We observed 81 HERVs as differentially expressed between cases (N = 259) and unaffected individuals (N = 279) under the false discovery rate of 5%, independent of a genetic association with schizophrenia ($q < 0.05$ [corrected for the 3,979 expressed HERVs]; **Supplemental Figure 4, Supplemental Table 4**). Our analysis showed that 36 HERVs were downregulated and 45 upregulated in cases, and that expression differences were subtle, with log2 fold-changes of 0.18 ± 0.10 on average.

Convergent analyses implicate ERVLE_8q24.3h and LTR25_6q21 as robust risk factors for schizophrenia

To make a more informative assessment of HERV expression differences associated with disease status, we first explored whether HERVs identified as genetic risk factors for this disorder were expressed in the DLPFC. Of the 148 HERVs identified via gene-level enrichment analysis, only 65 were consistently expressed across samples according to DESeq2 internal filtering criteria, with mean normalized counts of 57.5, CI 95% [12.59, 102.3]. Of these, we observed that two were significantly upregulated (ERV316A3_12q24.11, log2 fold-change = 0.16, standard error = 0.06; LTR25_6q21, log2 fold-change = 0.09, standard error = 0.04), and one downregulated (ERVLE_8q24.3h; log2 fold-change = -0.09, standard error = 0.04) in schizophrenia cases relative to unaffected controls, under the false discovery rate of 5% ($q < 0.05$ [corrected for 65 HERVs]; **Figures 3C and D; Supplemental Table 5**).

Of the three HERVs enriched for schizophrenia variants and differentially expressed in cases, we observed that two were modulated by eQTLs (**Figures 3C and D**, highlighted in blue), which we initially hypothesized could explain the case-control differences observed. These included **ERVLE_8q24.3h** and **LTR25_6q21**, which were regulated by the top eQTLs, **rs4875048** and **rs174399,** respectively**.** Importantly, **rs4875048** is associated with schizophrenia and is in linkage disequilibrium with the top association signal at this locus, rs10552126; **rs174399** is in linkage disequilibrium with a risk variant at the locus, rs11153302 (**Table 1**)[14]. Strikingly, no gene within a 1 Mb window upstream or downstream of either of the two top eQTLs was associated with schizophrenia according to a recent analysis by PsychENCODE (http://resource.psychencode.org/)[30].

We observed complex regulatory mechanisms governing expression of these two HERVs. **ERVLE_8q24.3h** was significantly downregulated in cases, but, contrary to what was expected, the risk (A-) allele of **rs4875048** was associated with increased expression of this HERV ($\beta = -0.35$, $P_{\beta\ dist.} = 0.002$, $q = 0.007$, **Figure 4A**). Similarly, **LTR25_6q21** was upregulated in patients, but the risk (G-) allele of **rs174399** was associated with reduced expression of this HERV ($\beta = 0.29$, $P_{\beta\ dist.} = 0.004$, $q = 0.02$; **Figure 4B**). We investigated the effect of these eQTLs in cases and control individuals separately, which revealed that they influenced HERV expression exclusively in

unaffected individuals, suggesting there is a compensatory mechanism in patients counteracting the effects of the eQTLs (**Figures 4A and B; Supplemental Table 6**).

To understand about the regulation of these HERVs during neurodevelopment, we tested their expression in an *in vitro* model of cortical development, which consisted of neural stem cells from the CTX0E16 cell line and cells differentiated for 28 days[31-33] **(Supplemental Material, Supplemental Figure 5**). We observed that both HERVs were differentially regulated during differentiation, suggesting that risk to schizophrenia pertaining these HERVs starts during neurodevelopment. Interestingly, we also observed that both HERVs are located in the antisense strand of *SLC16A10* and *IQANK1* introns, respectively. Moreover, LTR25_6q21 is exclusively present in humans, whereas ERVLE_8q24.3h is shared with primates (**Supplemental Figure 6**), according to the UCSC Genome Browser[34].

<<< **Figure 4** >>>

ERVLE_8q24.3h and LTR25_6q21 are co-regulated with genes implicated in mitochondrial and synaptic function in the adult brain, respectively

To explore the potential function of ERVLE_8q24.3h and LTR25_6q21, we determined the genes that are co-expressed with these HERVs by applying Weighted Correlation Network Analysis (WGCNA)[35] to the RNA-seq data from schizophrenia patients and controls combined. This systems biology approach is based on the hypothesis that genes within the same co-regulated network share the same function[36]. We observed 19 modules of co-expression in these data (**Supplemental Tables 7 and 8, Supplemental Figure 7)**. LTR25_6q21 was assigned to the green module (module membership statistic (MM) = 0.41, P = 3.55 x 10^{-23}; gene significance (GS) statistic in relation to case-control status: 0.12, GS P = 0.005), whereas ERVLE_8q24.3h was assigned to the turquoise module (MM = 0.46, P = 4.81 x 10^{-29}; GS = -0.15, GS P = 5.14 x 10^{-4}). Gene Ontology (GO) analysis of the green module indicates that this gene set is associated with neuronal function, with significant terms including “synapse organization”, “presynapse” and “vesicle-mediated transport in synapse”

(q < 0.05, **Figure 5A**, upper panel, **Supplemental Table 9**). The turquoise module, in turn, was significantly associated with mitochondrial function, with terms including the "respiratory chain" and "mitochondrial matrix", and "NADH dehydrogenase complex assembly" (q < 0.05, **Figure 5A**, lower panel, **Supplemental Table 9**).

A correlation between the first principal component capturing the variability within each module (module eigengene) and case-control status ('Profile') was performed in WGCNA, which revealed that the green module is positively associated with case-control status (r = 0.13, P = 0.003), whereas the turquoise module was negatively associated (r = -0.15, P = 4 x 10^{-4}; **Figure 5B**). The co-expression modules were detected assuming a signed network, which means that a positive module-trait correlation entails higher expression of genes in the module in association with disease status, and vice-versa, corroborating the case-control differences observed for each HERV in the previous analysis (**Figures 4A and B**, left panels).

To understand the function of other HERVs associated with schizophrenia from the gene-level enrichment analysis performed using MAGMA (**Supplemental Table 2**), we identified the co-expression modules associated with each enriched HERV that was expressed in the brain (**Supplemental Table 10**). We observed that several of these HERVs were assigned to modules implicated in synaptic organization, including MER41_1p33 and MER41_5p12 alongside LTR25_6q21 in the green module, as well as synaptic function, including HERVL_14q24.2d, HERVL40_8q21.3b, HERVW_6q21c, HML3_5p12a and MER41_2q31.2 in the red module. Interestingly, we observed that these two modules were clustered together in an unsupervised hierarchical clustering analysis performed in WGCNA, corroborating a related or shared function (**Supplemental Figure 8**).

<<< Figure 5 >>>

LTR25_6q21 belongs to a module associated with increased neuronal counts, whereas ERVLE_8q24.3h does not robustly correlate with major neural cell types

We investigated the major cell types associated with each modules eigengene (the first principal component of the module) to infer the cell types associated with ERVLE_8q24 and LTR25_6q21. We estimated the proportion of neural cell types in the RNA-sequencing data by applying the R package BRETIGEA[37] to the normalized gene counts of the RNA-sequencing data. BRETIGEA estimates the cell type proportions that constitute the sequenced material based on a database of single-cell RNA-sequencing data, ultimately generating coefficients that represent the proportion of astrocytes, microglia, endothelial cells, oligodendrocytes, oligodendrocyte progenitor cells and neurons in the sequenced samples[37]. To infer the cell types associated with each module, we performed correlations between each module and the cell-type proportion coefficients (**Figure 5B**). The module assigned to LTR25_6q21 (green) was significantly correlated with the expression of neuronal markers ($r = 0.24$, $P = 3 \times 10^{-8}$), and negatively associated with the expression of all other cell types ($q < 0.05$, corrected for the six cell types tested and the two modules of interest), consistent with a role for this module in the regulation of neuronal function, as indicated by the GO analysis. The module assigned to ERVLE_8q24.3h (turquoise) was negatively correlated with the expression of oligodendrocyte progenitor cell markers ($r = -0.21$, $P = 6 \times 10^{-7}$, $q < 0.05$), but showed no correlation with the other cell types, which may suggest a specific role for this HERV in non-dividing cells. Ultimately, these data indicate that LTR25_6q21 and ERVLE_8q24.3h are part of co-regulated networks implicated in neuronal and mitochondrial function, respectively, suggesting their involvement in these cell functions.

Discussion

HERVs are ancient viral genetic elements scattered throughout the genome, with previously hypothesized influences on neurodevelopment and risk for schizophrenia. We took a comprehensive approach to reconsider the role of HERVs at the omics level, leveraging on the recent advances in the genomic annotation of HERVs, single-locus resolution quantification, systems biology methods, and population genetic tools.

We investigated the combined contribution of HERVs (the retrogenome) to risk for schizophrenia and other complex polygenic traits, by assessing the overlap between SNPs implicated in these traits and those in genomic locations encompassing HERVs. We were surprised to find that schizophrenia was the only tested trait significantly enriched for common variants within the retrogenome, especially since there is evidence linking HERVs to conditions like amyotrophic lateral sclerosis[7-9], Crohn's disease[38], major depressive disorder and bipolar disorder[10,11]. The fact that the retrogenome is significantly enriched for polymorphisms implicated in schizophrenia suggests that HERVs comprise an important set of risk factors for this disorder, within the 'non-coding' genome.

We investigated which HERV families and locus-specific HERVs could be particularly important in moderating risk for schizophrenia at the genetic level by co-localizing risk variants with individual HERV loci from across 60 families. We identified 148 specific HERVs enriched for schizophrenia-associated SNPs. A subsequent gene-set analysis did not find a particular HERV family enriched for schizophrenia variants, further suggesting that disparate HERVs scattered throughout the genome moderate susceptibility rather than whole families. These findings have several implications for the interpretation of previous studies in the literature, and may explain how our results differ from previous data implicating other HERVs or HERV families as risk factors, which may not have been identified here. Most HERV expression research to-date has been performed using microarrays, antibodies or RT-qPCR probes, which do not provide sufficient specificity to assess single HERVs[7-11]. Therefore, previous studies may have captured the expression of multiple HERVs concomitantly due to their repetitive sequences. Based on our findings, it is unlikely that individual HERVs (even from within the same family) contribute equally to risk. The HERV annotation developed by Bendall

and colleagues[20], as well as the Telescope pipeline, and the application of modern population genetic methods, enabled us to revisit the role HERVs play in relation to schizophrenia risk on a genome-wide scale, and allow for more robust inferences regarding their etiological relevance.

Next, to better understand HERV expression regulation in the brain, we analyzed RNA-sequencing data from the CommonMind Consortium, which confirmed widespread HERV expression in the adult brain and cis-regulatory mechanisms. We found 3,979 HERVs actively expressed in a key brain area linked to schizophrenia pathophysiology, the DLPFC[39]. We further identified 1,759 HERVs that were regulated by 1,622 short-range (cis-) eQTLs in the DLPFC. The identification of eQTLs that impact HERV expression informs us about the basic processes responsible for HERV regulation, and can be useful for the interpretation of GWAS findings in the context of the retrogenome[40,41]. In addition, these results suggest that SNPs within HERVs are not simply affecting the expression of neighboring protein-coding genes via their LTRs, rather it demonstrates that common genetic variation impacts locally on HERV expression.

We used a complementary set of analyses to identify the most robust HERVs implicated in schizophrenia. Two HERVs, **ERVLE_8q24.3h** and **LTR25_6q21**, identified from the gene-level enrichment analysis, were found to be regulated by schizophrenia-associated eQTLs, and were found to be differentially expressed in the DLPFC of patients and in an *in vitro* model of cortical neurodevelopment. These findings suggest a potential risk mechanism for schizophrenia that starts during neurodevelopment and persists through to adulthood, as observed for schizophrenia risk genes such as *NT5C2, AS3MT,* and *BORCS7*[33,42]. Importantly, we observed a complex regulation of both HERVs, whereby the lead eQTLs only exerted their regulatory effects in unaffected individuals. This suggests that compensatory mechanisms (e.g. epigenetic alterations) may be acting to correct for the effects of the risk variants on the expression of these HERVs in the DLPFC of schizophrenia patients.

We also describe here, for the first time, the co-regulation of several HERVs with known genes in the adult brain, and the GO terms associated with each co-expression module. Our findings suggest

that LTR25_6q21 is implicated in neuronal function, whereas ERVLE_8q24.3h is involved in mitochondrial function. WGCNA has been successfully used to predict the biological function of unknown genes or non-coding RNAs in different organisms[43,44], and to identify clinically relevant cell types when in combination with cell-type deconvolution analysis[45], and thus represents a powerful approach to functionally characterize the HERVs expressed in the brain. For a long time HERVs were assumed to be mere regulatory DNA sequences, but the discovery of their expression and co-regulation with several other genes implicated in multiple biological processes in the brain, ranging from neuronal, glial and mitochondrial regulation to splicing and cell motility (**Supplemental Table 9**), is a landmark for HERV research, and adds an extra layer of complexity to our understanding of human neurobiology.

There are limitations to this study which should be acknowledged. Schizophrenia is a highly polygenic, heterogeneous disorder, and as such large sample sizes are required for appropriate comparisons. The CommonMind Consortium provides the largest and best characterized cohort of schizophrenia cases and unaffected individuals with RNA-sequencing data to-date, but it might be underpowered for case-control comparisons considering the heterogeneity of schizophrenia. Nevertheless, we complemented case-control comparisons with genomic and eQTL analyses to provide additional insights. Another limitation to our study is that it investigated RNA-sequencing data from bulk DLPFC tissue only, which is composed of a heterogeneous mixture of several types of neurons and glial cells, and it is possible that HERVs expressed in particular cell types, or in other brain regions, are more relevant to risk[46]. To address this, we performed cell type deconvolution using BRETIGEA to determine cell-type specific effects, but ultimately the analysis of data from other brain areas, developmental time points and single-cell datasets has the potential to reveal important insights about the etiology of schizophrenia in relation to HERV expression. In addition, our post-mortem and *in vitro* work suggest HERV expression is important in the DLPFC and during its development, but we still do not know the function of these HERVs. To infer function we performed WGCNA which provides insight into which processes risk HERVs moderate, but future functional studies are required to definitively characterize how HERVs, particularly ERVLE_8q24.3h and LTR25_6q21, influence the transcriptome, neural stem cell proliferation or neuronal differentiation in

359 the context of schizophrenia risk, as is currently being investigated in relation to protein-coding risk
360 genes[32,33,47,48].
361
362 The development of a retrogenome annotation, and advances in modern population genetic methods
363 and transcriptomic tools, now allows us to investigate HERVs at the omics level, in the context of
364 risk for many biological traits. Our work studying the role of HERVs in the brain, and their relationship
365 to schizophrenia ignites a new, provocative line of thought implicating HERVs as biological risk
366 factors for schizophrenia and confirms that these previously assumed ‘dormant’ sequences in the
367 brain may not be dormant after all.

Online Methods

We used a combination of gene expression, genetic and *in vitro* analyses to identify the most robust HERVs implicated in schizophrenia risk (**Figure 6**). Further details are provided in the **Supplemental Material.**

<<< Figure 6 >>>

Genetic enrichment analyses

We estimated the contribution of genetic polymorphisms within the retrogenome towards risk of developing multiple traits using GARFIELD[27]. We downloaded summary statistics from well-powered genome-wide association studies, including of schizophrenia (N = 105,318 individuals)[14], height (N = 693,529)[49], body mass index (N = 681,275)[49], coronary artery disease (N = 547,261)[50], Crohn's Disease (N = 59,957)[51], type 2 diabetes (N = 659,316)[52], neuroticism (N = 2,370,390)[53], eczema (N = 103,066)[54], major depressive disorder (N = 480,359)[55], bipolar disorder (N = 51,710)[56], Alzheimer's disease (N = 74,046)[57], attention deficit hyperactivity disorder (N = 53,293)[58], amyotrophic lateral sclerosis (N = 36,052)[59], and autism spectrum disorder (N = 46,350)[60], which analyzed European cohorts only. GARFIELD performs greedy pruning of SNPs in GWAS summary statistics (those in linkage disequilibrium, with $R^2 > 0.1$), and quantifies enrichments using odds ratios, assessing their significance by employing generalized linear model testing, controlling for minor allele frequency, and number of linkage disequilibrium proxies ($R^2 > 0.8$). Linkage disequilibrium and allele frequency information were calculated based on the UK10K study. Enrichments were calculated based on summary statistics from each trait using two $P_{association}$ thresholds: $P < 5 \times 10^{-8}$, to test the enrichment of HERVs within genome-wide significant variants; and a more relaxed threshold, $P < 5 \times 10^{-5}$, to allow signal capture in less powered GWAS. The enrichment significance was corrected for the number of tests performed [14 traits and two P-value thresholds tested]. For consistency with the GWAS data, the HERV annotation used in the expression analysis (hg38) was remapped to hg19 coordinates using liftOver[34].

To identify locus-specific HERVs and potential HERV families associated with schizophrenia, we used MAGMA 1.07b[28]. Briefly, MAGMA calculates gene-level enrichment by generating a gene-wide

397 statistic from summary statistics, adjusting for gene size, variant density, and linkage disequilibrium
398 using the 1000 Genomes Phase 3 European reference panel. SNPs from the summary statistics
399 were assigned to HERVs using an annotation window of 10 kb upstream and downstream of each
400 HERV (as suggested by the authors)[28]. A Bonferroni correction was applied to identify significantly
401 enriched HERVs ($P_{cut\text{-}off} < 4.03 \times 10^{-6}$ [0.05 / 12,393 HERVs in chromosomes 1-22, excluding the
402 major histocompatibility locus]). Q-Q and Manhattan plots were generated using qqman 0.1.4[61].
403 Gene-set enrichment analysis was additionally performed using MAGMA, to test whether HERVs
404 associated with schizophrenia in the previous step were enriched for any of the 60 HERV families
405 (excluding HERVs located in sex chromosomes or at the MHC locus – chromosome 6: 26 – 34 Mb).
406
407 **The CommonMind Consortium dataset**
408 To identify HERV expression differences in schizophrenia patients, or HERV expression quantitative
409 trait loci (eQTL) in the dorsolateral prefrontal cortex, we analyzed RNA-sequencing data from the
410 CommonMind Consortium (release 1.0, N = 593 individuals, https://doi.org/10.7303/syn2759792)[39].
411 This dataset consisted of dorsolateral prefrontal cortex (DLPFC) samples from 279 unaffected
412 individuals, 259 schizophrenia cases, 47 bipolar disorder patients and 8 cases broadly diagnosed
413 with an affective disorder. Access to this dataset, which includes expression, genotype and clinical
414 data, was granted under a Material Transfer Agreement with the NIMH Repository and Genomics
415 Resources. Briefly, autopsy samples from the Mount Sinai NIH Brain Bank and Tissue Repository,
416 the University of Pennsylvania Brain Bank of Psychiatric illnesses and Alzheimer's Disease Core
417 Center, and The University of Pittsburgh Brain Tissue Donation Program, were sent to the Icahn
418 School of Medicine at Mount Sinai for nucleic acid isolation and sequencing. Individuals were
419 diagnosed according to the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, as
420 determined in consensus conferences after review of medical records and interviews of family
421 members and care providers. Total RNA was extracted from autopsy tissue using the RNeasy kit
422 (QIAGEN, Hilden, Germany). Ribosomal RNA was depleted using the Ribo-Zero Magnetic Gold kit
423 (Illumina, San Diego, California, United States), libraries were constructed using the TruSeq RNA
424 Sample Preparation Kit v2 (Illumina), and samples were sequenced on an Illumina HiSeq 2500. For
425 whole-genome genotyping, DNA was extracted using the DNeasy Blood and Tissue Kit (QIAGEN)

426 according to the manufacturer's protocol, and samples were genotyped using Illumina Infinium
427 HumanOmniExpressExome 8 1.1b chips. Further details on quality control and sample processing
428 are described in Fromer and colleagues[39].
429
430 **RNA-sequencing data processing and HERV expression quantification**
431 Bam files containing mapped and unmapped RNA-sequencing reads aligned to the human reference
432 genome (hg19) using TopHat 2.0.9 and Bowtie 2.1.0, were downloaded to the King's College London
433 High Performance Computer Cluster Rosalind, using the synapse client (1.7.5). Bam files were
434 merged, and fastq files were extracted using samtools 1.5[62] and the flag '-F 0x100'. Trimmomatic
435 0.38[63] was used to prune Illumina adaptors, low quality bases (leading/trailing sequences with phred
436 score < 3, or those with average score < 15 every four bases), or reads below 36 bases in length.
437 Trimmed reads were mapped to the human genome hg38 using bowtie2[64] and the parameters --
438 very-sensitive-local, -k 100, and --score-min L,0,1.6. Subsequently, Telescope 1.0.2 was used to
439 quantify expression of 14,968 HERVs (annotation version hg38), which we defined as the
440 retrogenome[20]. We analyzed HERVs with counts > 10 across 4 samples at least, to avoid inflation
441 driven by lowly expressed elements. Case-control expression differences (N = 538 individuals in
442 total) were calculated in R[65] using Wald tests in DESeq2[66], where data was normalized (median of
443 ratios) and controlled for the main confounders of gene expression estimated by Fromer and
444 colleagues[39], which included institution of sample origin, RNA integrity number, gender, post-mortem
445 interval, age (determined in five bins: #1 = 13-29 years, #2 = 30-49 years, #3 = 50-69 years, #4 =
446 70-89 years, #5 = 90+ years), as well as the first five population covariates estimated using
447 multidimensional scaling in PLINK 1.9[67], and the first ten hidden HERV expression confounders
448 estimated using sva[68], considering schizophrenia and unaffected individuals only.
449
450 **Whole-genome genotype data processing**
451 Markers with zero alternate alleles, genotyping call rate < 0.98, Hardy-Weinberg $P < 5 \times 10^{-5}$, or
452 individuals with genotyping call rate < 0.90, were removed from the analysis, as described by Fromer
453 and colleagues[39]. PLINK files were generated containing genotype information for 958,178 variants
454 for the 593 subjects. Marker alleles were phased to the forward strand, and ambiguously stranded

markers were removed. Additional genotype information was imputed from the 1000 Genomes Phase 1 reference panel using minimac3 and Eagle v2.3 phasing with the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html). Genotype information from the 22 autosomes was concatenated using bcftools 1.9 (https://samtools.github.io/bcftools/bcftools.html), non-single nucleotide polymorphisms (SNPs) were excluded, as well as sites with an imputation R^2 < 0.8, minor allele frequency < 0.05, or Hardy-Weinberg $P < 5 \times 10^{-5}$.

eQTL analysis

Normalized HERV counts per sample were obtained using DESeq2 (N = 593), and HERV expression was tested for the effect of genotype at all variants located within a 1 Mb window upstream or downstream from the annotation start site of each HERV using QTLtools[69], according to the authors' manual. We covaried for the effect of case-control status, institution of sample origin, RNA integrity number, gender, post-mortem interval, age (five bins, as described previously), the first five population covariates, and ten hidden expression confounders estimated using sva[68] (**Supplemental Figure 3**). eQTL P-values were corrected through estimation of a beta distribution using a minimum of 1,000 permutations and maximum of 10,000, and were further corrected for the number of HERVs tested using the false discovery rate method (q < 0.05).

Weighted Correlation Network Analysis (WGCNA)

WGCNA is a systems biology approach that enables the identification of co-expressed genes in transcriptomic data, which we used here to identify the genes co-expressed with schizophrenia HERVs in order to infer their biological function[35]. We used this tool to construct a signed network consisting of HERVs and genes, which was created based on an adjacency matrix that informs about the co-expression similarity observed between all pairs of genes and HERVs in the expression data (i.e. genes and genes, genes and HERVs, HERVs and HERVs). Normalized HERV and gene counts were variance-stabilized in DESeq2, and were further adjusted for all confounders previously described, using the *removebatcheffect* function in limma[70]. To achieve this, we combined gene and HERV counts obtained from the brain of 538 individuals (279 unaffected individuals and 259 schizophrenia cases), and filtered out genes and HERVs that were lowly expressed, i.e. those with

484 < 10 counts in < 80% of samples), as these can drive spurious correlations[35]. The normalized counts
485 were variance stabilized transformed using DESeq2 and adjusted for institution of sample origin,
486 gender, case-control status, age bins, post-mortem interval, the first five population dimensions
487 (estimated in plink), and RIN, using limma[70]. WGCNA identifies modules by applying hierarchical
488 clustering to the adjacency matrix, further filtering spurious relationships through the application of a
489 topological overlap approach. We used an R^2 cut-off of 0.8, which corresponds to a $\beta = 12$, to
490 construct the network. Each module was assigned a color, and genes or HERVs not belonging to
491 any module were assigned to the gray module. The relationship between modules and specific cell
492 types was tested based on the correlation between the module eigengenes (ME), defined as the first
493 principal component of the module, and cell count estimates, as described below. We applied the
494 false discovery rate method to correct for the module-cell type associations ($q < 0.05$). Plots were
495 generated by WGCNA.
496
497 **Cell type estimates and module correlations**
498 We performed a BRain cEll Type specIfic Gene Expression Analysis (BRETIGEA)[37] to estimate the
499 abundance of major neural cell types in the 538 samples analyzed by WGCNA. Briefly, BRETIGEA
500 uses expression data from single cell RNA-sequencing data sets to identify the proportion of
501 astrocytes, microglia, endothelial cells, oligodendrocytes, oligodendrocyte progenitor cells and
502 neurons, in bulk brain gene expression data. More specifically, this tool uses a panel of 50 well-
503 established cell type-specific markers to generate coefficients that represent the proportion of each
504 cell type per sample, which were tested for association with each module.
505
506 **Gene Ontology (GO) analyses**
507 We performed GO analyses using the WEB-based GEne SeT AnaLysis Toolkit (Webgestalt)[71] to
508 identify the function of the genes co-regulated with the schizophrenia HERVs in the brain, and thus
509 infer the potential function of these HERVs. All genes inputted to WGCNA were used as background
510 (reference) gene set. We used the false discovery rate method to correct for the GO enrichment
511 analyses within Webgestalt ($q < 0.05$) and report up to 10 significant GO terms per module. Volcano
512 plots were generated in Webgestalt.

513

514

515 **Statistical analysis and data visualization**

516 The co-localization of GWAS-supported variants with the retrogenome was calculated using linear
517 regressions in GARFIELD[27], and the gene-level and gene-set enrichment analyses were calculated
518 in MAGMA[28]. Findings were corrected for multiple testing using the Bonferroni method. The case-
519 control HERV expression differences (N = 538 individuals) and effects of eQTLs on HERV
520 expression (N = 593 individuals) were calculated, respectively, using Wald tests in DESeq2, and
521 stepwise linear regressions in QTLtools, respectively. These were corrected using the false
522 discovery rate ($q < 0.05$), a more permissive multiple testing correction method, to increase our
523 detection power. The effect of genotype on specific HERVs within cases and control groups,
524 separately or combined, was calculated using linear regressions in IBM Statistics SPSS 25 (IBM
525 Corp., Armonk, NY, United States). Other analyses were performed in R[65]. Graphs were generated
526 in R or Graph Pad Prism 7 (GraphPad Software, San Diego, CA, United States).

527

Acknowledgements

The work was supported in part by US National Institutes of Health grants: CA206488 (DFN), AI076059 (DFN) and UL1TR001876 (KAC). This work was also supported by a grant from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil, Science without Borders award no. BEX1279/13-0) and an NIHR Maudsley Biomedical Research Centre Career Development Award to RRRD. TRP is funded by a Medical Research Council (MRC) Skills Development Fellowship (MR/N014863/1). This study represents independent research part funded by the NIHR-Wellcome Trust King's Clinical Research Facility and the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. Data for this publication were obtained from NIMH Repository & Genomics Resource, a centralized national biorepository for genetic studies of psychiatric disorders. Data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881, AG02219, AG05138, MH06692, R01MH110921, R01MH109677, R01MH109897, U01MH103392, and contract HHSN271201300031C through IRP NIMH. Brain tissue for the study was obtained from the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories, and the NIMH Human Brain Collection Core. CMC Leadership: Panos Roussos, Joseph Buxbaum, Andrew Chess, Schahram Akbarian, Vahram Haroutunian (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Enrico Domenici (University of Trento), Mette A. Peters, Solveig Sieberts (Sage Bionetworks), Thomas Lehner, Stefano Marenco, Barbara K. Lipska (NIMH). Tissue samples used in RT-qPCR experiments were supplied by The London Neurodegenerative Diseases Brain Bank, which receives funding from the Medical Research Council and as part of the Brains for Dementia Research programme, jointly funded by Alzheimer's Research UK and Alzheimer's Society. We thank Professor Cathryn Lewis for her comments on the manuscript and Daniel Bean (King's College

London) for his assistance with coding. We also thank Richard “Brad” Jones, Mario Ostrowski, and Nathaniel Bachtel for their comments on this manuscript.

Author contributions

Study design: TRP, RRRD. Performed analyses and experiments: RRRD, TRP. Contributed reagents, biological material, revised the manuscript: MLB, MM, CEO, GAB, SS, CT, GRT, KAC, DPS, DFN. Wrote the paper: RRRD, TRP.

Conflict of Interest

The authors declare no conflict of interest.

Data availability

Telescope and the HERV annotation are available at http://github.com/mlbendall/telescope. Access to the CommonMind Consortium dataset can be requested to the NIMH Repository & Genomics Resource via https://www.nimhgenetics.org/resources/commonmind.

References

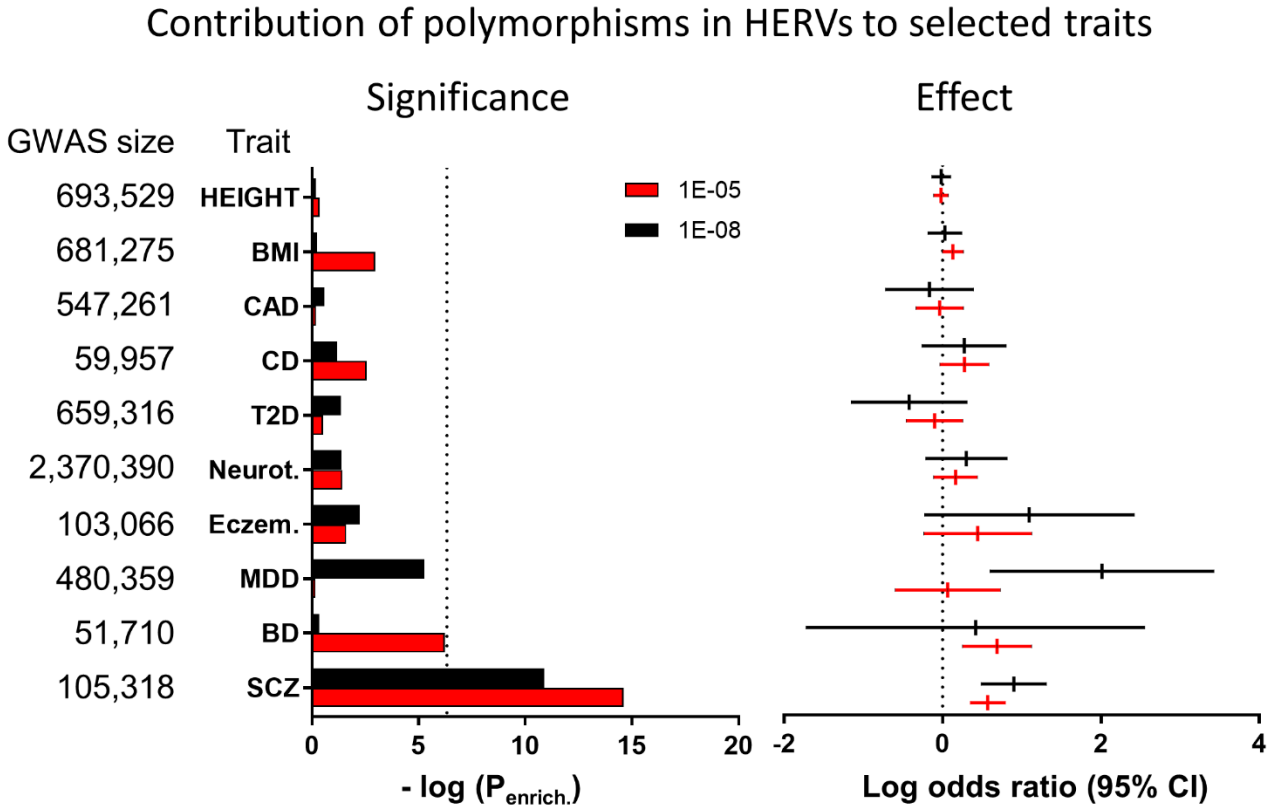
1. Gifford, R. & Tristem, M. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* **26**, 291-315 (2003).
2. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
3. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785-9 (2000).
4. Grow, E.J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221-225 (2015).
5. Brattas, P.L. *et al.* TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. *Cell Rep* **18**, 1-11 (2017).
6. Fasching, L. *et al.* TRIM28 Represses Transcription of Endogenous Retroviruses in Neural Progenitor Cells. *Cell reports* **10**, 20-28 (2015).
7. Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* **7**, 307ra153 (2015).
8. Douville, R., Liu, J., Rothstein, J. & Nath, A. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. *Annals of neurology* **69**, 141-151 (2011).
9. Garson, J.A. *et al.* Quantitative analysis of human endogenous retrovirus-K transcripts in postmortem premotor cortex fails to confirm elevated expression of HERV-K RNA in amyotrophic lateral sclerosis. *Acta Neuropathol Commun* **7**, 45 (2019).
10. Perron, H. *et al.* Molecular characteristics of Human Endogenous Retrovirus type-W in schizophrenia and bipolar disorder. *Transl Psychiatry* **2**, e201 (2012).
11. Weis, S. *et al.* Reduced expression of human endogenous retrovirus (HERV)-W GAG protein in the cingulate gyrus and hippocampus in schizophrenia, bipolar disorder, and depression. *J Neural Transm (Vienna)* **114**, 645-55 (2007).
12. Yolken, R.H., Karlsson, H., Yee, F., Johnston-Wilson, N.L. & Torrey, E.F. Endogenous retroviruses and schizophrenia. *Brain Res Brain Res Rev* **31**, 193-9 (2000).
13. Roussos, P. *et al.* A role for noncoding variation in schizophrenia. *Cell reports* **9**, 1417-1429 (2014).
14. Pardiñas, A.F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics* **50**, 381-389 (2018).
15. Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421 (2014).
16. Goke, J. & Ng, H.H. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO Rep* **17**, 1131-44 (2016).
17. Karlsson, H. *et al.* Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc Natl Acad Sci U S A* **98**, 4634-9 (2001).
18. Diem, O., Schäffner, M., Seifarth, W. & Leib-Mösch, C. Influence of Antipsychotic Drugs on Human Endogenous Retrovirus (HERV) Transcription in Brain Cells. *PLoS ONE* **7**, e30054 (2012).
19. Svensson, A.C. *et al.* Chromosomal distribution, localization and expression of the human endogenous retrovirus ERV9. *Cytogenet Cell Genet* **92**, 89-96 (2001).
20. Bendall, M.L. *et al.* Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLOS Computational Biology* **In press**, 398172 (2018).
21. Subramanian, R.P., Wildschutte, J.H., Russo, C. & Coffin, J.M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**, 90-90 (2011).

22. Grandi, N., Cadeddu, M., Blomberg, J. & Tramontano, E. Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. *Retrovirology* **13**, 67-67 (2016).
23. Grandi, N. *et al.* Identification of a novel HERV-K(HML10): comprehensive characterization and comparative analysis in non-human primates provide insights about HML10 proviruses structure and diffusion. *Mobile DNA* **8**, 15 (2017).
24. Vargiu, L. *et al.* Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **13**, 7 (2016).
25. Tokuyama, M. *et al.* ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proceedings of the National Academy of Sciences* **115**, 12565-12572 (2018).
26. Jeong, H.-H., Yalamanchili, H.K., Guo, C., Shulman, J.M. & Liu, Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. in *Biocomputing 2018* 168-179.
27. Iotchkova, V. *et al.* GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature Genetics* **51**, 343-353 (2019).
28. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).
29. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
30. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**(2018).
31. Anderson, G.W. *et al.* Characterisation of neurons derived from a cortical human neural stem cell line CTX0E16. *Stem Cell Res Ther* **6**, 149 (2015).
32. Deans, P.J.M. *et al.* Psychosis Risk Candidate ZNF804A Localizes to Synapses and Regulates Neurite Formation and Dendritic Spine Structure. *Biol Psychiatry* **82**, 49-61 (2017).
33. Duarte, R.R.R. *et al.* The Psychiatric Risk Gene NT5C2 Regulates Adenosine Monophosphate-Activated Protein Kinase Signaling and Protein Translation in Human Neural Progenitor Cells. *Biol Psychiatry* **86**, 120-130 (2019).
34. Hinrichs, A.S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-8 (2006).
35. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
36. van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J.P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* **19**, 575-592 (2018).
37. McKenzie, A.T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Scientific Reports* **8**, 8868 (2018).
38. Klag, T. *et al.* Human Endogenous Retroviruses: Residues of Ancient Times Are Differentially Expressed in Crohn's Disease. *Inflammatory Intestinal Diseases* **3**, 125-137 (2018).
39. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience* **19**, 1442-1453 (2016).
40. O'Brien, H.E. *et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome biology* **19**, 194-194 (2018).
41. Bray, N.J. & O'Donovan, M.C. The genetics of neuropsychiatric disorders. *Brain and Neuroscience Advances* **2**, 2398212818799271 (2018).
42. Duarte, R.R.R. *et al.* Genome-wide significant schizophrenia risk variation on chromosome 10q24 is associated with altered cis-regulation of BORCS7, AS3MT, and NT5C2 in the human brain. *Am J Med Genet B Neuropsychiatr Genet* **171**, 806-14 (2016).

43. Liu, W. *et al.* Construction and Analysis of Gene Co-Expression Networks in Escherichia coli. *Cells* **7**, 19 (2018).
44. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J.P. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* **19**, 575-592 (2017).
45. Inkeles, M.S. *et al.* Cell-type deconvolution with immune pathways identifies gene networks of host defense and immunopathology in leprosy. *JCI Insight* **1**(2016).
46. Tansey, K.E. & Hill, M.J. Enrichment of schizophrenia heritability in both neuronal and glia cell regulatory elements. *Translational Psychiatry* **8**, 7-7 (2018).
47. Hill, M.J. *et al.* Knockdown of the schizophrenia susceptibility gene TCF4 alters gene expression and proliferation of progenitor cells from the developing human neocortex. *Journal of psychiatry & neuroscience : JPN* **42**, 181-188 (2017).
48. Hill, M.J., Jeffries, A.R., Dobson, R.J., Price, J. & Bray, N.J. Knockdown of the psychosis susceptibility gene ZNF804A alters expression of genes involved in cell adhesion. *Hum Mol Genet* **21**, 1018-24 (2012).
49. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
50. Harst, P.v.d. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research* **122**, 433-443 (2018).
51. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
52. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications* **9**, 2941 (2018).
53. Baselmans, B.M.L. *et al.* Multivariate genome-wide analyses of the well-being spectrum. *Nature Genetics* **51**, 445-451 (2019).
54. The Early Genetics Lifecourse Epidemiology Eczema Consortium *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics* **47**, 1449 (2015).
55. Wray, N.R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics* **50**, 668-681 (2018).
56. Stahl, E.A. *et al.* Genome-wide association study identifies 30 Loci Associated with Bipolar Disorder. *bioRxiv*, 173062 (2018).
57. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).
58. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics* **51**, 63-75 (2019).
59. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* **48**, 1043-8 (2016).
60. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics* **51**, 431-444 (2019).
61. Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*, 005165 (2014).
62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
63. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
64. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).
65. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2018).

66. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
67. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
68. Jeffrey T. Leek *et al.* sva: Surrogate Variable Analysis. *R package* **Version 3.30.1, https://bioconductor.org/packages/release/bioc/html/sva.html**(2019).
69. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nature Communications* **8**, 15452 (2017).
70. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47-e47 (2015).
71. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research* **45**, W130-W137 (2017).

741 **Figures and Table**



742

743 **Figure 1.** Variants within the retrogenome are enriched with schizophrenia-associated

744 polymorphisms. No association with any other trait was observed (see **Supplemental Table 1** for

745 full results). Calculated using GARFIELD[27].

746

747

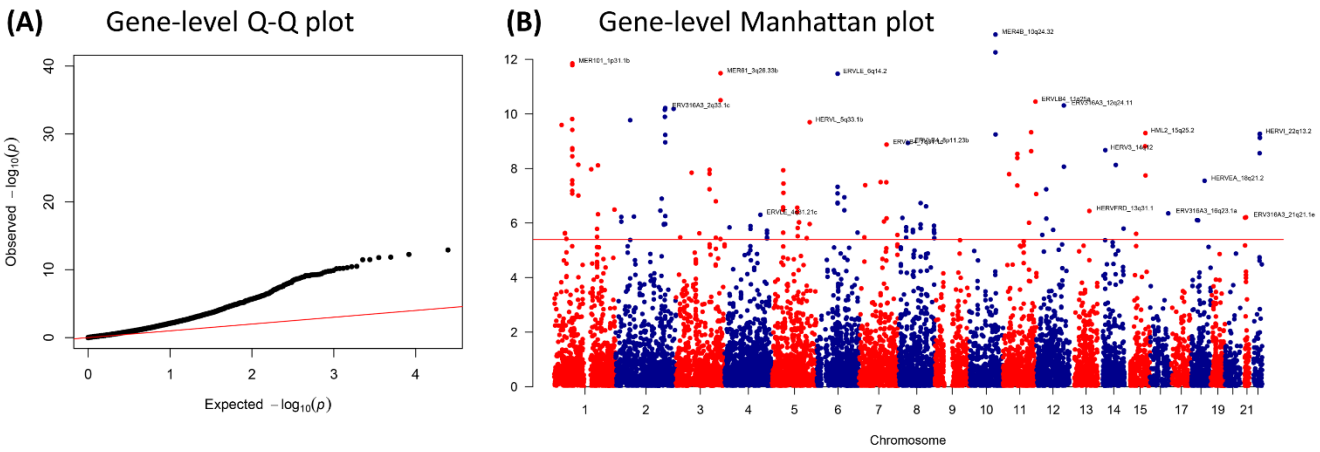


Figure 2. Gene-level enrichment analysis of the schizophrenia GWAS summary statistics using the HERV annotation developed by Bendall and colleagues[20], calculated using MAGMA[28]. Chromosomes 1-22 only, extended MHC region excluded (chromosome 6, from 26-34 Mb). **(A)** Quantile-quantile plot showing the contribution of several HERVs to schizophrenia genetics compared to an expected normal distribution (red line). **(B)** Manhattan plot showing the location of the HERVs associated with schizophrenia. Plots created using qqman[61]. All enriched HERVs are shown on **Supplemental Table 2**.

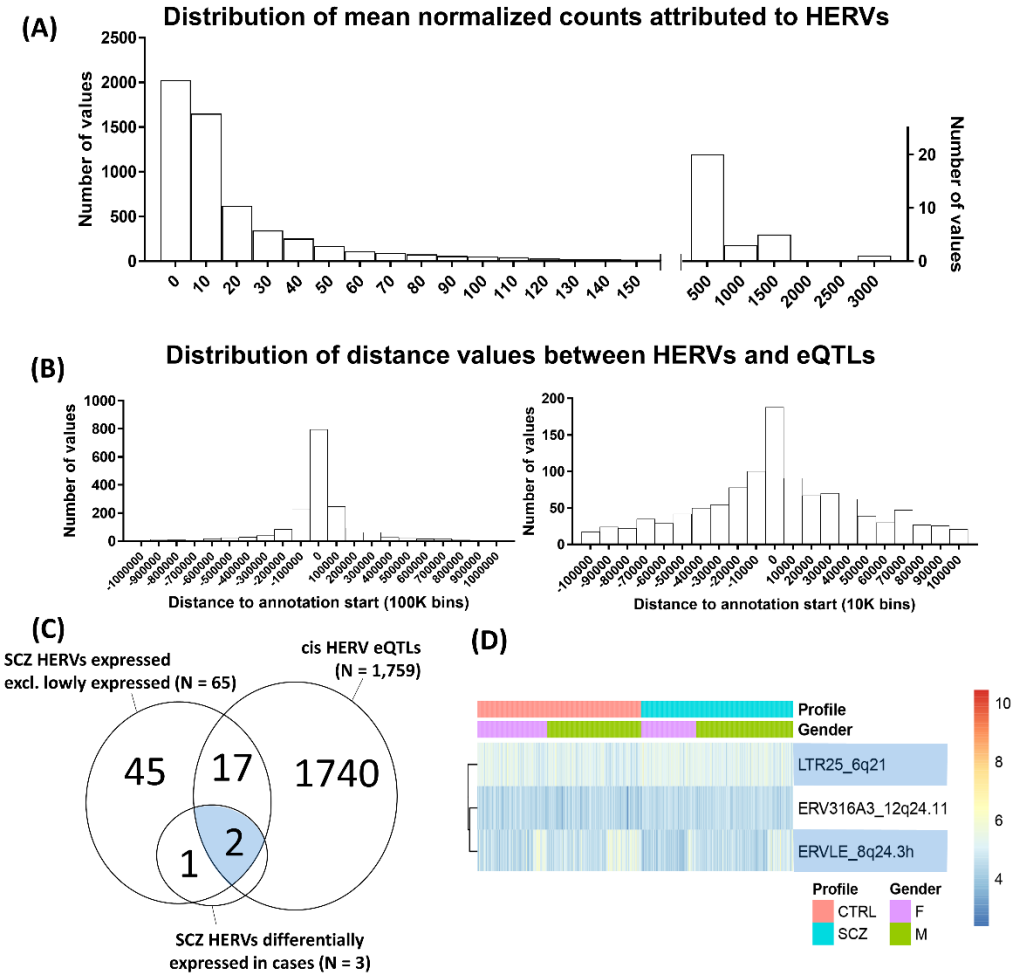
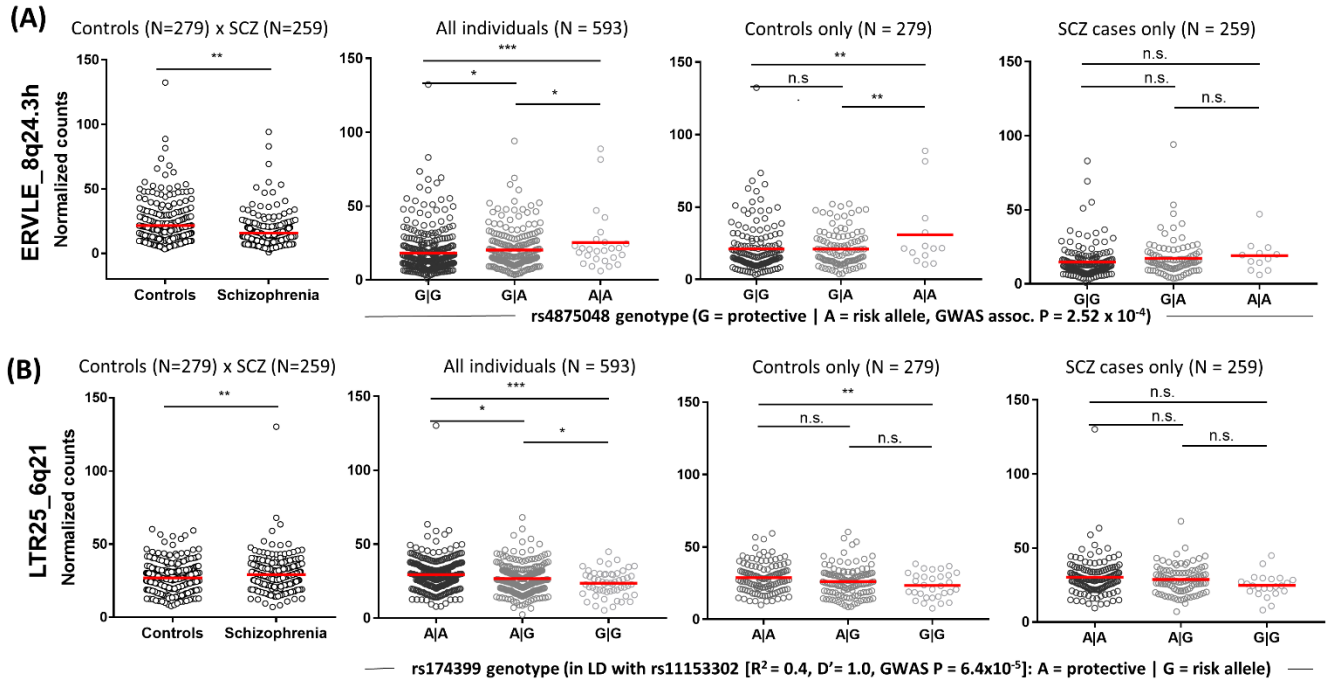


Figure 3. HERV expression in the dorsolateral prefrontal cortex, based on an analysis of the CommonMind Consortium dataset with Telescope. **(A)** Frequency distribution of the mean normalized counts per HERV across samples showing that the majority of HERVs are lowly expressed, according to an analysis of 593 post-mortem brains. Normalized counts do not include adjustments utilized in the analyses. **(B)** Distribution of values representing the distance between the eQTL for a HERV and the start site for that HERV. Left panel shows all data in bins of 100,000, and the right panel only data nearer the start site of the HERVs, in bins of 10,000. A large proportion of eQTLs was located within a 10kb window upstream or downstream the start coordinates of the HERV they regulate. **(C)** Overlap between the 65 schizophrenia-associated HERVs expressed in the brain, the 1,759 HERVs modulated by eQTLs, and the three HERVs enriched for schizophrenia variants and additionally differentially expressed between schizophrenia cases (N = 259) and unaffected individuals (N = 279). **(D)** Heatmap of the six HERVs enriched for schizophrenia variants and further differentially expressed in patients, separated by gender. In blue, HERVs that are modulated by eQTLs, as demonstrated in **Figure 4**.



773

774 **Figure 4.** HERVs enriched for schizophrenia variants and differentially expressed between cases
775 and controls, and according to genotype. eQTLs modulated HERV expression exclusively in control
776 individuals. Graphs show normalized counts associated with case-control status (left; Wald tests,
777 **P < 0.01) or per genotype considering all individuals, control individuals or schizophrenia patients
778 only (last three graphs, respectively; ANOVAs followed by pairwise comparisons corrected using the
779 Bonferroni method (*P < 0.05, **P < 0.01, ***P < 0.001; n.s.: not significant; **Supplemental Table
780 6**). Data shown are for **(A)** ERVLE_8q24.3h and its top eQTL in the DLPFC, rs4875048, and **(B)**
781 LTR25_6q21 and its top eQTL, rs174399. Values are uncorrected for the factors and covariates
782 included in the eQTL analysis.

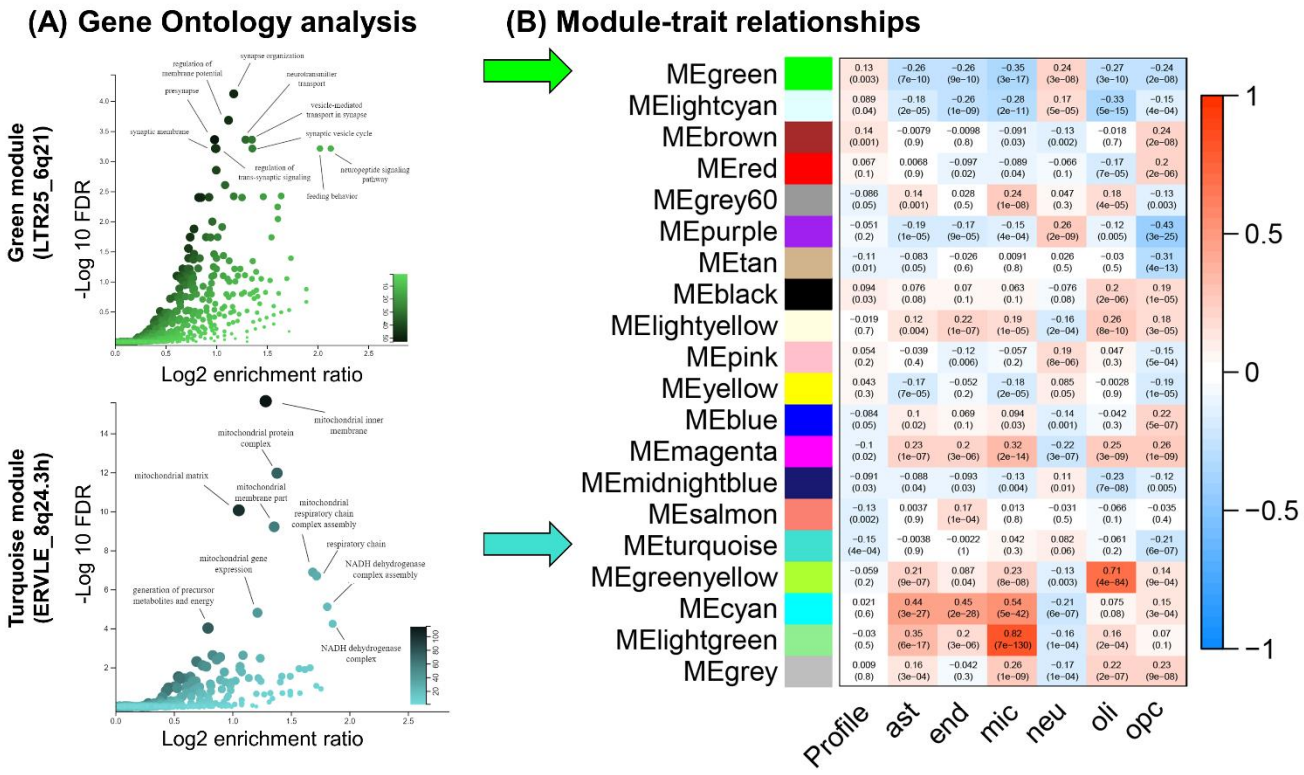


Figure 5. Gene ontology enrichment analysis of the co-expression modules associated with the green (LTR25_6q21) and turquoise modules (ERVLE_8q24.3h), and correlations between modules eigengenes, case-control status (Profile) and coefficients associated with cell counts for major neural cell types. **(A)** LTR25_6q21 belongs to the green module, which is significantly enriched for GO terms associated with neuronal function, as shown in the Volcano plot. ERVLE_8q24.3h, in turn, belongs to the turquoise module, which is significantly enriched for GO terms associated with mitochondrial function. **(B)** The green module is positively associated with neuronal counts, whereas the turquoise module does not correlate strongly with a specific cell type, apart from a negative correlation with oligodendrocyte progenitor cell counts. Each cell of the heatmap contains the Pearson's r coefficient followed by significance of the correlation (P). Ast: astrocytes, end: endothelial cells, mic: microglia, neu: neurons, oli: oligodendrocytes, opc: oligodendrocyte progenitor cells.

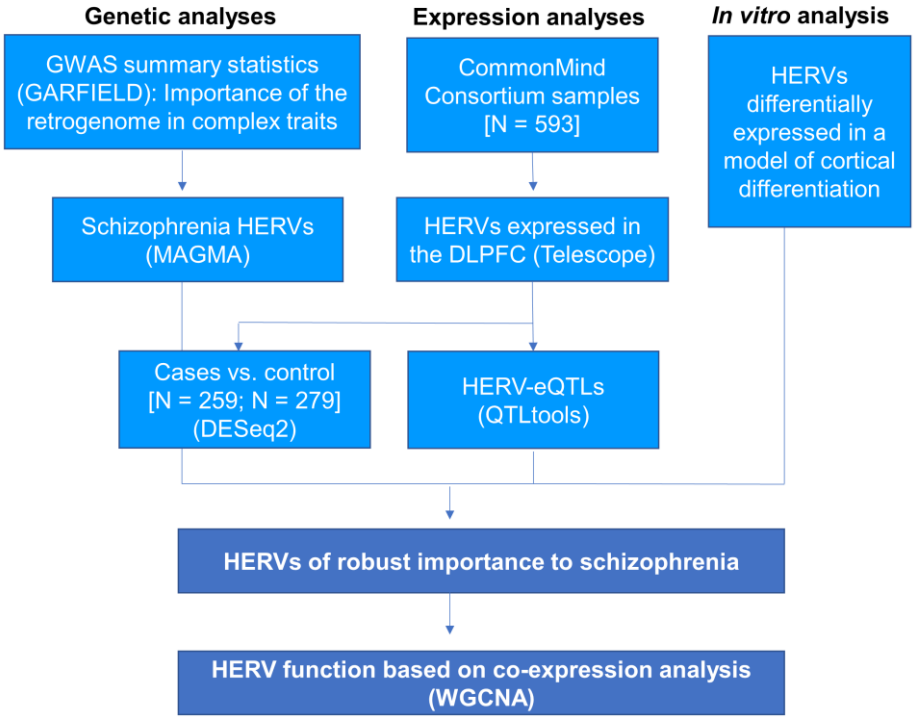


Figure 6. Analysis strategy. We performed a series of analyses using post-mortem brain RNA-sequencing data, genetic enrichment analyses and *in vitro* cortical differentiation data to identify HERVs of robust importance in schizophrenia. Bioinformatic tools used are indicated in parentheses. DLPFC: dorsolateral prefrontal cortex.

Table 1. Association of the eQTLs (and variants in linkage disequilibrium) with schizophrenia.

HERV	SNP	Chromosome	Position	Risk allele	Other allele	Frq cases	OR	SE	P	R^{2} ¥	D' ¥
ERVLE_8q24.3h	rs4875048	8	144826671	A	G	0.1869	1.0553	0.011855	5.6E-06	-	-
	rs10552126	8	144844056	CAT	C	0.2018	1.05919	0.0116	7.1E-07	0.8954	1
LTR25_6q21	rs174399	6	111919619	G*	A	0.2565	1.0025	0.010488	0.81	-	-
	rs11153302	6	111918869	A	G	0.506	1.0397	0.00973	6.4E-05	0.38	1

* This SNP is not associated with schizophrenia, but is in linkage disequilibrium with another variant that is associated with this disorder.

¥ Linkage disequilibrium statistics in relation to the top association signal at the locus which is in LD with the HERV-eQTL.