RESEARCH ARTICLE

AMERICAN JOURNAL OF medical genetics PART B
Neuropsychiatric Genetics

LINE1 Insertions as a Genomic Risk Factor for Schizophrenia: Preliminary Evidence From an Affected Family

Guia Guffanti,[1] Simona Gaudi,[2] Torsten Klengel,[1] James H. Fallon,[3] Harry Mangalam,[3] Ravi Madduri,[4,5] Alex Rodriguez,[4,5] Paula DeCrescenzo,[6] Emily Glovienka,[6] Janet Sobell,[7] Claudia Klengel,[1] Michele Pato,[7] Kerry J. Ressler,[1] Carlos Pato,[7] and Fabio Macciardi[3,8,9]*

[1]Department of Psychiatry, McLean Hospital, Harvard Medical School, Belmont, Massachusetts

[2]Department of Infectious, Parasitic and Immune-Mediated Diseases, Italian National Institute of Health, Rome, Italy

[3]Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, California

[4]Division of Mathematics and Computer Science, Argonne National Laboratory, Lemont, Illinois

[5]Computation Institute, University of Chicago, Chicago, Illinois

[6]Department of Psychiatry, Columbia University Medical Center and New York State Psychiatric Institute, New York, New York

[7]SUNY Downstate, College of Medicine, Brooklyn, New York

[8]Center for Autism Research and Treatment (CART), University of California, Irvine, California

[9]Center for Epigenetics and Metabolism, University of California, Irvine, California

Manuscript Received: 16 April 2015; Manuscript Accepted: 11 February 2016

Recent studies show that human-specific LINE1s (L1HS) play a key role in the development of the central nervous system (CNS) and its disorders, and that their transpositions within the human genome are more common than previously thought. Many polymorphic L1HS, that is, present or absent across individuals, are not annotated in the current release of the genome and are customarily termed "non-reference L1s." We developed an analytical workflow to identify L1 polymorphic insertions with next-generation sequencing (NGS) using data from a family in which SZ segregates. Our workflow exploits two independent algorithms to detect non-reference L1 insertions, performs local *de novo* alignment of the regions harboring predicted L1 insertions and resolves the L1 subfamily designation from the *de novo* assembled sequence. We found 110 non-reference L1 polymorphic loci exhibiting Mendelian inheritance, the vast majority of which are already reported in dbRIP and/or euL1db, thus, confirming their status as non-reference L1 polymorphic insertions. Four previously undetected L1 polymorphic loci were confirmed by PCR amplification and direct sequencing of the insert. A large fraction of our non-reference L1s is located within the open reading frame of protein-coding genes that belong to pathways already implicated in the pathogenesis of schizophrenia. The finding of these polymorphic variants among SZ offsprings is intriguing and suggestive of putative pathogenic role. Our data show the utility of NGS to uncover L1 polymorphic insertions, a neglected type of genetic variants with the potential to influence the risk to develop schizophrenia like SNVs and CNVs. © 2016 Wiley Periodicals, Inc.

Key words: mobile elements; LINE1; retrotransposition; schizophrenia; next-generation sequencing

How to Cite this Article:
Guffanti G, Gaudi S, Klengel T, Fallon JH, Mangalam H, Madduri R, Rodriguez A, DeCrescenzo P, Glovienka E, Sobell J, Klengel C, Pato M, Ressler KJ, Pato C, Macciardi F. 2016. LINE1 Insertions as a Genomic Risk Factor for Schizophrenia: Preliminary Evidence From an Affected Family.
Am J Med Genet Part B 171B:534–545.

INTRODUCTION

Recent studies of retrotransposition in the human genome have revealed that at least for LINE1 (L1) elements, these occurrences

Conflict of interest: None.
*Correspondence to:
Fabio Macciardi, Department of Psychiatry and Human Behavior, School of Medicine, University of California, Irvine, 5251 California Ave, Suite 240, Irvine, CA 92617.
E-mail: fmacciar@uci.edu
Article first published online in Wiley Online Library (wileyonlinelibrary.com): 16 March 2016
DOI 10.1002/ajmg.b.32437

are more common than previously thought [Beck et al., 2010; Ewing and Kazazian, 2010, 2011; Huang et al., 2010]. Moreover, many evolutionary recent L1 insertions are polymorphic, being either present or absent across individuals, and are customarily termed non-reference L1s, given that they are not annotated in the current releases of the human genome [Ewing and Kazazian, 2010]. Non-reference L1s are present in fewer individuals than reference L1s, and are not significantly enriched in any particular chromosomal region [Kazazian, 2011], but they can influence the human transcriptome and can represent genetic variants contributing to important human phenotypes [Kines and Belancio, 2012; Cowley and Oakey, 2013; Kuhn et al., 2014]. In addition, *de novo* L1HS insertions have been also observed in single individuals, with retrotransposition rates similar to *de novo* mutations [Xing et al., 2009; Ewing and Kazazian, 2011].

Recently, different computational algorithms have been developed to identify non-reference L1 polymorphisms from NGS data [Keane et al., 2013; Wu et al., 2014]. The core feature of these algorithms is to overcome the daunting issue of identifying a new insertion in a genomic region where accurate mapping of the sequenced reads is made difficult by the highly repetitive nature of retrotransposon sequences themselves. Different methods rely on two main algorithms defined as read-pair (RP) and split-read (SR). The RP method consists in identifying uniquely mapped reads (also called "anchor" reads) spanning into an annotated reference L1 with a fragment length not consistent with the given library fragment length distribution. The SR method consists in identifying anchor reads that have unmapped mates that can be aligned to the sequence of annotated reference L1s elsewhere in the genome.

As a consequence of such technological advancements, L1s, and specifically L1HS, have been shown to play a key role in both CNS development and CNS disorders [Muotri et al., 2005; Singer et al., 2010; Baillie et al., 2011; Evrony et al., 2012; Reilly et al., 2013; Bundo et al., 2014; Guffanti et al., 2014; Insel, 2014].

In schizophrenia, Bundo et al. [2014] described an increased number of somatic L1 retrotranspositions in the dorsal-lateral prefrontal cortex (DLPFC, Brodmann's area 46) of patients, comparing L1 insertions in brain versus liver tissues for both cases and controls. The authors observed that the total number of brain-specific L1 insertions tended to be higher in schizophrenia patients, but their finding was not statistically significant, perhaps, due to their limited sample size ($n = 3$) and the high inter-individual variation. However, a further Gene Ontology analysis of the genes affected by somatic L1 retrotranspositions in the schizophrenic patients was enriched for terms related to brain functions, and these data ultimately suggested that instability of the neural genome in early development phases may account for an increased risk of disease.

We are interested in expanding the analysis of polymorphic non-reference L1 insertions in schizophrenia studying their transmission in affected families under the hypothesis that they also account for a still unknown proportion of the overall genomic risk in developing the disease. The high frequency of inter-individual structural variations due to polymorphic non-reference L1 germ cell insertions observed in the 1,000 Genomes Project [Kidd et al., 2010; Lam et al., 2010; Stewart et al., 2011] has been shown to rework the genome architecture not only through the insertion itself, but also by modifying the target site directly or with downstream post-insertional modifications [Gilbert et al., 2002; Han et al., 2005, 2008; Lee et al., 2012; Grandi et al., 2015]. The potential large effect of non-reference L1 insertions on genomic structure and function not only introduces both germline and somatic cells mosaicisms [Grandi and An, 2013], but also suggests that these insertions can contribute to the pathogenesis of schizophrenia, in addition to the proposed high amount of somatic transpositions that occur in neuronal cells [Baillie et al., 2011; Reilly et al., 2013]. To establish reliable methods to assess, detect, and computationally validate non-reference and *de novo* L1 insertions from next-generation sequencing data, we analyzed a family dataset chosen out of the schizophrenia cohorts available from the Genomic Psychiatric Cohort consortium [Pato et al., 2013].

MATERIALS AND METHODS

Whole Genome Sequencing

Whole genome sequences (WGS) from six members of a family in which schizophrenia segregates were obtained from the Genomic Psychiatric Cohort (GPC) consortium [Pato et al., 2013]. The GPC family analyzed in this study includes an unaffected father (HC), an affected mother (SZ), and four offsprings, of which one is unaffected (HC) and three are affected (SZ—two males and one female). All DNA samples were derived from whole blood. Deep (30–40×) WGS was performed at the Broad Institute using the Illumina HiSeq Platform (450–500 bp-insert library, 100 bp reads). Genomes were aligned to hg19 with BWA [Li and Durbin, 2009], and all subsequent analyses were performed with hg19 as the reference. Sequencing details are described in Supplementary Methods.

Detection of L1 Insertions

We developed a workflow to identify previously undetected L1 insertion sites—also called retrotransposon insertion polymorphisms (RIPs)—in the six members of our GPC family, without a priori knowing whether they would be non-reference L1 polymorphisms or *de novo* events. RIPs calls for each individual were made by running two publicly available programs: Retroseq [Keane et al., 2013] and Tangram [Wu et al., 2014]. Both programs rely on a combination of read-pair (RP) and split-read (SR) algorithms with some differences (see Supplementary Methods and Supplementary Fig. S1 for more details). The workflow can be implemented in Galaxy and Globus Genomics [Madduri et al., 2014].

Quality Control

To ensure replicability of RIPs calls and reduce false positives, we set up a set of conservative quality controls. RetroSeq assigns a quality score ranging from 1 to 8, called filter ($FL = 1–8$), to each RIP based on the following criteria: (1) depth too high in region, (2) not enough reads in cluster, (3) not enough total flanking reads, (4) not enough inconsistently mapped reads, (5) neither side passes ratio test, (6) one side passes ratio test, (7) distance too large at breakpoint, (8) passed all filters. Following recommendations of the authors of this program (https://github.com/tk2/RetroSeq/

wiki/1000-Genome-CEU-Trio-Analysis), we removed RetroSeq-called RIPs from the dataset if quality score (FL) was <6, if quality score (FL) = 6 and genotype quality (GQ, as specific in the VCF format) < 28, and if quality score (FL) = 7 or 8 and genotype quality <20. RIPs surviving the FL and GQ filtering process are classified as "QC1." In addition, RIP breakpoint positions mapping to a reference transposable elements (TE) annotated in the human reference sequence (GRCh37), including L1, Alu, or LTR (genomic coordinates catalogued in Repeatmasker [Smit et al., 1996] and retrieved were also discarded to limit false positive calls. RIPs surviving the reference TE overlap filtering process are classified as "QC2."

For each identified RIP, Tangram provides the number of supportive read-pair/split-read fragments from 5′ and 3′ end. Following recommendation of the authors of the program (personal communication), we initially considered only RIPs with at least one supporting fragment at both the 5′ end (SR5) and the 3′ (SR3). Tangram allows searching for the same RIP for as many subjects submitted to the analysis at the same time. As we deal with a set of six related subjects, we allowed for RIPs with at least one supporting fragment at either the 5′ end or at the 3′ end when at least one of the subjects fulfilled QC of both SR5 and SR > 0. The degree of confidence of Tangram calls depends on the number of supporting reads at the 5′ and 3′ end, with a higher number of supporting reads providing higher confidence. As for RetroSeq RIP calls, insertion breakpoint positions mapping to a reference TE annotated in the human reference sequence (GRCh37), including L1, Alu, or LTR (genomic coordinates catalogued in RepeatMasker and retrieved through UCSC/Galaxy) were also discarded to limit false positive calls. A RIP surviving the reference TE overlaps filtering process is classified as "QC2," as for Retroseq's criteria. For the purpose of inheritance estimates, we allowed Tangram RIPs to be defined by either 5′ and/or 3′ end supporting reads (i.e. QC1), which means that RIPs that have zero supporting reads on at least one end are still included in the analysis, after and as a consequence of the *de novo* local assembly.

For any potential *de novo* insertion in the children, we performed a manual search of the aligned reads at the L1 boundary to look for any evidence of a non-reference L1 element in the parent: this being the case, even if not fulfilling our a priori QC criteria, would be likely sufficient to establish the insertion as inherited as opposed to *de novo*.

De Novo Local Assembly

To establish a more stringent data set of high-confidence L1 calls, we performed computational validation on the data set of L1s identified by the two programs. Only non-reference L1 insertions computationally validated by local *de novo* assembly, notwithstanding by which algorithm (RetroSeq and/or Tangram) they have been identified, were considered for downstream analysis. The computational validation consists in performing *de novo* local assembly in the regions of a putative L1 insertion using the reads spanning the locus of the new insertion. If RetroSeq/Tangram are able to identify the insertion site of an L1, we should be able to identify the fragments of the *de novo* sequence of the newly inserted element that lead RetroSeq/Tangram to detect the insertion breakpoint. *De novo* alignment is required as the newly inserted L1 by definition should not be annotated in the reference genome; therefore, the insertion should have been missed by regular alignment procedures, even if truly existing. For the local *de novo* alignment, we used Velvet [Zerbino and Birney, 2008], invoked through a script included in the package SVmerge [Wong et al., 2010]. SVmerge consists of a suite of scripts that allows performing a *de novo* alignment in genomic regions a priori defined by the user, in our case the regions harboring the putative L1 insertion site. The suite of scripts included in the package SVmerge were originally created to identify structural variants (SV) and then to perform *de novo* alignment to refine the breakpoints in the regions of detected SVs. We extended the region of interest to include an interval of ±6 Kb around the insertion site.

Mapped reads and any unmapped mate-pairs were extracted from the BAM file within 6 Kb of a predicted insertion breakpoint and formatted to FASTA format with interleaved read pairs and assembled by Velvet (v1.2.08 including parameters hashlen = 29, ins_len = 220, exp_cov = 35 and cov_cutoff = 2). All contigs generated for each 12 Kb region harboring the predicted insertion site were aligned to the corresponding region in the reference genome using Exonerate (European Bioinformatics Institute v.2.2.0, http://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate), a pairwise sequence comparison program [Slater and Birney, 2005]. We used the Exonerate parameters recommended by SVmerge pipeline (with the following settings: model [affine:local], bestn [50], gapexend [−3], dnahspdropoff [10], hspfilter [200], gappedextension [FALSE], joinrangeext [300], score [15]).

Classification of L1s at the Insertion Sites Identified by RetroSeq/Tangram

The scaffolds generated by Velvet in the regions harboring the predicted insertion breakpoint were subsequently analyzed using RepeatMasker [Smit et al., 1996] to identify any transposable element that is not annotated in the hg19 release of the human reference genome.

Each Velvet contig/scaffold was aligned against the library of known TE deposited in the database RepBase Update [Jurka, 2000] of the Genetic Information Research Institute (GIRI) to assign family type (e.g., LINE/L1) and name (e.g., L1HS) to each TE mapping to the target region. As we predict that the new insertion is the result of the retrotransposition (i.e., "copy-and-paste") of an active L1, *de novo* alignment should identify a sequence which is not present in the reference genome at the insertion site predicted by Tangram/RetroSeq, but that has sequence similarity to an already annotated reference L1. By comparing the distribution of TEs mapping to the target region of our *de novo* assembled sequence against the reference genome, we can identify the sequence of the new insertion annotated by RepeatMasker.

We identified four different scenarios. First, Velvet assembles contig sequences into one scaffold that includes the sequence of the newly observed L1 at the predicted insertion breakpoint (in the middle of the 12 Kb region submitted for *de novo* local assembly). Second, Velvet assembles contig sequences into two scaffolds covering the 12 Kb of the target region and interrupted at the

predicted insertion breakpoint. If computational validation is successful, we find that one (or both) contig(s) ends with a sequence annotated as a L1 (e.g., L1HS). This may indicate that *de novo* local assembly process was not able to retrieve enough sequence information at the predicted insertion site to generate a unique scaffold, and gets interrupted in the correspondence of the repetitive sequence. Third, Velvet contigs are not assembled into one scaffold, but RepeatMasker still identifies annotated L1s within contig sequences. Although these contigs cannot fit into a unique scaffold, a L1 can nonetheless be positioned at the predicted insertion breakpoint. Fourth, Velvet generates multiple contigs that are not assembled together into scaffolds, possibly due to a region of highly repetitive sequences. Although RepeatMasker identifies L1s within these contigs, Exonerate aligns them to sequences that do not match the coordinates of the predicted insertion breakpoint. Alternatively, the predicted insertion site may match the coordinates of a sequence embedded in a region rich in simple repeats, which are highly similar to short sequences (~100 bp) of L1s. In both cases, as there is not sufficient information to confirm or exclude the presence of a "new" L1 insertion, we decided to define the results of this fourth scenario as "inconclusive" given the relatively poor resolution of *de novo* local assembly.

Comparison of L1 Calls by RetroSeq and Tangram

As the two programs use slightly different algorithms to identify RIPs, the exact position of the insertion site might be different in one program compared to the other. If the insertion site identified by RetroSeq is within 300 bp from the insertion site identified by Tangram, we assume they refer to the same RIP. We used Galaxy function "Fetch closest non-overlapping feature" (Tool: Operate on Genomic Intervals) to identify RetroSeq RIP within 300 bp from Tangram RIP.

Comparison of L1 Calls With Non-Reference L1 Polymorphisms Reported in dbRIP and euL1db

We compared the final set of computationally validated L1 calls by Retroseq and Tangram with RIPs reported in dbRIP (http://dbrip.brocku.ca/) [Wang et al., 2006] and euL1db (http://eul1db.unice.fr/db/searchmenu.jsp) [Mir et al., 2014]. For euL1db, we selected the catalog of "MRIPs" (meta-retrotransposon insertion polymorphisms), a list of meta-loci derived by merging overlapping or close RIPs. We compared coordinates of observed RIP calls in our study with dbRIP and MRIP loci within a window of 300 bp from the insertion site coordinates.

Identification of Inherited and *De Novo* L1

We evaluated the pattern of inheritance of each L1 identified by either RetroSeq or Tangram and computationally validated by *de novo* local assembly.

Tangram identifies insertions for all subjects submitted to the analysis at the same time and assigns genotype of homozygote no insertion (i.e., 0/0) whenever the supporting reads are equal to zero at both the 5′ and 3′. We observed cases in which one parent displayed >1 supporting reads at both the 5′ and 3′ end of the L1 insertion breakpoint, but the offspring displayed >1 supporting reads only at the 5′ or the 3′ or vice versa. For the purpose of heritability analysis, Tangram calls that were supported either at the 5′ or the 3′ by >2 reads for each single individual (e.g., $5' = 2$ and $3' = 0$ or $5' = 0$ and $3' = 2$) were included in the analysis upon computational validation.

RetroSeq identifies L1 insertions separately for each subject without taking into consideration the results of the L1 insertion detection of his/her family relatives. Therefore, in order to investigate the pattern of inheritance, we needed to identify the L1 insertions predicted at the same locus for more than one subject of the family. We assumed that two L1 insertions in two related individuals refer to the same L1 locus if the insertion site identified in one subject is within 100 bp from the insertion site identified in another subject, hence, considering L1 insertions identified within 100 bp from one another as belonging to the same locus. By assigning a unique ID to each locus, we could define a minimal set of discrete family-based L1 putative loci among the six members of our family.

We classified L1s into inherited and *de novo* calls, where the latter was defined as a L1 not present in the genomes of either parents but present in the genome of one of the four offsprings. We considered a L1 "private" when it was present in the genome of either parent but not present in the genome of any sibling. This type of L1 was also defined "untransmitted." Then we examined whether the inherited L1 were transmitted by one or both parents to the unaffected, the affected or both unaffected and affected offspring.

Experimental Validation of Non-Reference L1 Polymorphisms

Non-reference L1 polymorphisms not previously detected and not reported in any publicly available database were validated by PCR and Sanger sequencing. Briefly, PCR amplification was performed in 50 μl total reaction volume using 100 ng genomic DNA, 10 μl 5X Long Amp buffer (NEB), PCR primers at a final concentration of 0.2 μM (Life Technologies, Carlsbad, CA), dNTPs at a final concentration of 200 μM (NEB) and 5 U of Long Amp Taq polymerase (NEB) on a 96-well PTC-200 Thermal Cycler (MJ Research, Waltham, MA). Primers were designed to span the L1 site using the publically available Primer3 program. Primer sequences and specific PCR conditions are reported in Supplementary Materials (Supplementary Table SI). PCR products were visualized on a 1% ethidiumbromide agarose gel (Supplementary Fig. S2). PCR products for both alleles were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany). Finally, capillary sequencing was performed under standard conditions at Beckman Coulter Genomics on an ABI 3730XL (Applied Biosystems, Inc., Foster City, CA, http://www.beckmangenomics.com/). Ab1 sequence files were analyzed using 4peaks (http://nucleobytes.com/4peaks/). The aligned sequence products, which included some flanking regions along with the non-reference L1 insertion site, were screened in BLAST to determine the precise insertion coordinates comparing the heterozygous L1 amplicon with the reference allele. L1 subfamilies identification was determined along with percent divergence from the subfamily consensus sequences using Repeatmasker.

L1 Functional Annotation

For each L1 calls, we performed an initial gene annotation analysis. We used Galaxy "Join" function of the "Operate with genomic intervals" tool to map computationally validated L1s to a list of genes derived from GENCODE v19 [Harrow et al., 2012], including protein coding, pseudogenes, and RNA genes obtained through BioMart (http://www.ensembl.org/biomart). Then we performed a crude functional annotation of L1 insertions using Panther [Thomas et al., 2003] and the Reactome-database of protein functional interactions (http://www.reactome.org/). For each category of transmitted L1 insertions, we identified the list of represented GO terms.

RESULTS

Detection of L1 Polymorphisms by Whole-Genome Sequencing in a SZ Six Members Family

We set out to characterize the patterns of germline and somatic L1 insertions in the whole-genome sequences (WGS) of a six members family in which schizophrenia segregates. Detection of L1 insertions was performed using two different algorithms implemented in the programs RetroSeq and Tangram, respectively. The differences between the algorithms are fully described in the Supplementary Methods. After quality control procedures, we performed a comprehensive computational validation by target *de novo* assembly on the previously identified candidate insertions. Central to our approach were the refinement of the position of the predicted insertion and the identification of newly inserted sequences harboring LINE1s. New insertions and newly inserted sequences refer to genomic features that are not annotated in hg19 reference genome.

Across all the six family members, we identified a total of 234 non-reference L1 insertions using Tangram and 276 using RetroSeq (Table I). Of these, 45 are shared by both programs (19% for Tangram and 16% for RetroSeq). Computational validation confirmed these new insertions for 83 out of 234 Tangram calls (35.5%) and 224 out of 276 RetroSeq calls (81.1%) (for details by individual genome and program see Fig. 1). Eighty-four out of 234 Tangram calls and 35 out of 276 RetroSeq calls were not validated, that is, no sequences supported the predicted new insertion. For 24 Tangram calls and 17 RetroSeq calls, computational validation yielded inconclusive results (see description of criteria in Methods and Supplementary Materials) to define as "inconclusive" the results of computational validation). For 43 Tangram calls, the *de novo* assembly analysis failed, yielding no aligned scaffold for the analysis. *De novo* assembly fails whenever the number of reads exceeds 10,000 reads, which is the maximum number of reads for insertions to consider an insertion call not an artifact [Wong et al., 2010].

Classification of L1 Subfamily

RepeatMasker was used to determine the subfamily L1 classification within the *de novo* aligned sequences generated by Velvet. Full-length human L1s are 6 Kb in length and contain an internal promoter located in the 5′-untranslated region (5′UTR), and additional putative TSSs close to ORF1, two non-overlapping open reading frames (ORF1 and ORF2), a short 3′UTR, and a poly(A) tail. Although ORF1 encodes a non-specific RNA-binding protein, ORF2 encodes a protein with endonuclease (L1 EN) and reverse transcriptase (L1 RT) activities, required for L1 retrotransposition. All *de novo* sequences aligned at the insertion target site could be identified in the RepBase database and classified as 5′UTR, 3′UTR or ORF2 of L1HS, L1P1, L1PA2, and L1PA3 subfamilies in all cases with three exceptions, that is, one L1PA4 and two L1PA6 elements (for details by individual genome and program see Fig. 2). The vast majority (86%) of L1 inserted sequences were aligned to L1HS or L1P1 elements, including 26.7% to 5′UTR, 33.3% to 3′UTR, and 26% to ORF2.

Rate of L1 Insertions

The computationally validated calls (83 for Tangram and 234 for RetroSeq) point to a final dataset that includes 110 family-based

TABLE I. Summary of Non-Reference RIPs Detected by Tangram and RetroSeq

	Tangram			RetroSeq		
	Raw calls	QC1	QC2	Raw calls	QC1	QC2
ID	SR5 > 0 or SR3 > 0	SR5 > 0 and SR3 > 0	SR5 > 0 and SR3 > 0 and no overlap with ANY L1, Alu, LTR	No filters	FL > 6 and GQ > 28 FL > 7 or 8 and GQ > 20	No overlap with ANY L1, Alu, LTR
9414-01	30,016	65	39	88	83	49
9414-02	26,752	63	40	82	77	49
9414-05	28,733	65	41	89	85	47
9414-06	28,624	68	47	87	83	52
9414-07	20,469	48	32	76	71	42
9414-08	22,770	65	35	66	59	37
Total	157,364	374	234	488	458	276

For each member of the family, count of non-reference RIPs classified as raw calls, pre-quality control (QC); QC1, after first stage of quality control screening based on supporting reads at 5′ (SR5) and 3′ (SR3) end for Tangram and filter score (FL) and genotype quality (GQ) parameters for RetroSeq; and QC2, after second stage of quality control screening, which consists in removing non-reference MEI whose coordinates overlap with a reference (i.e., annotated in the hg19 release of the genome) L1, Alu, or LTR.

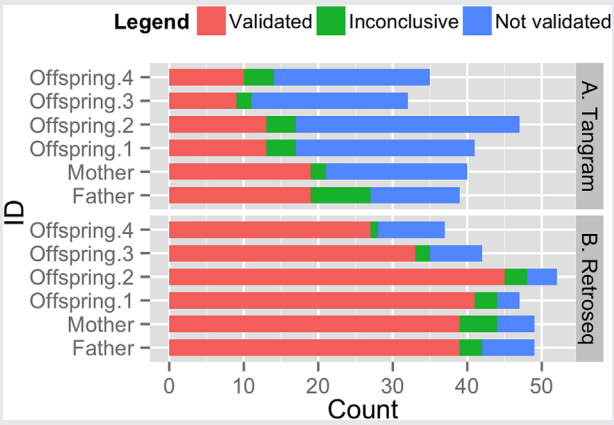

FIG. 1. Computational validation of non-reference RIPs detected by Tangram (A) and RetroSeq (B). Fraction of non-reference RIPs validated via local *de novo* assembly using Velvet and RepeatMasker for L1 sub-family designation (blue), faction of non-reference MEI not yielding inconclusive results (red), and fraction of non-reference RIPs not yielding evidence of *de novo* sequence at the predicted insertion site (green).

loci, 12 loci identified by Tangram, 78 identified by RetroSeq, and 20 identified by both algorithms. Overall, the number of RIPs per subject ranges between 39 and 46 in our sample. This estimate represents a lower bound of published estimates [Ewing and Kazazian, 2010, 2011] and may be, in part, due to our conservative L1 identification workflow.

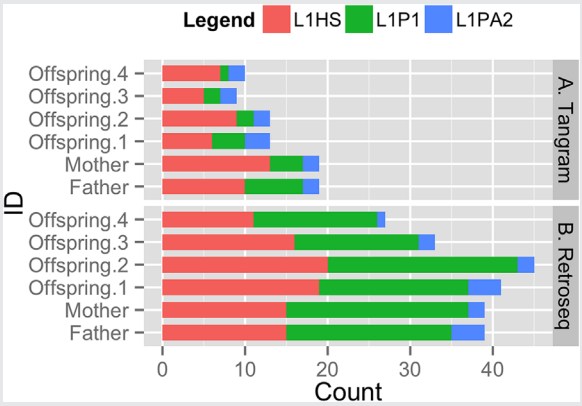

FIG. 2. L1 sub-family designation of local *de novo* assembly sequence at the predicted insertion site of non-reference RIPs detected by Tangram (A) and RetroSeq (B). Fraction of non-reference RIPs classified as LIHS (blue), L1P1 (red), and other L1 such as L1PA2, L1P, and L1PA6 (green).

Comparison With dbRIP and euL1db Catalog of Polymorphic L1 Insertions

We found that 40 over 110 of our non-reference L1 polymorphic insertions were already reported in the dbRIP catalog, which contains a total of 598 loci. For 17 of these 40 loci, coordinates were overlapping across datasets; for the remaining 23, it was possible to retrieve a dbRIP locus at <300 bp. The comparison with the MRIP catalog of euL1db (which contains 8,991 loci) yielded even more consistent results, with 105 over 110 of our non-reference L1 polymorphic insertions already reported in the MRIP catalog. Of these, only for 13 over 105 of our loci it was not possible to find a complete overlap with the coordinates of loci in the MRIP catalog, but a MRIP mapped <300 bp away. Overall, we found that only 5 over 110 loci were not reported either in dbRIP or euL1db.

Experimental Validation of Non-Reference L1 Polymorphisms

PCR analysis was performed for four over five suspected "novel" loci for each of the six members of the family to identify the presence/absence of the L1 polymorphism at the insertion site computationally predicted by our workflow. For the fifth locus, no PCR analysis was possible due to high complexity of the genomic region. The presence/absence of the insertion is determined by the size of the PCR product as expected based on the primer design (see Methods for more details). Two of the four novel non-reference L1 polymorphisms were computationally predicted in the genome of two affected siblings (mei044 and mei061) and two in the father's genome (mei070 and mei076). The PCR analysis confirmed the presence of the insertion site at the two predicted loci in the affected offspring's genomes and further identified insertion sites at the same locus in the genomes of the mother and the father, respectively. In both cases, the insertion site in the genome of the parents was originally identified by our analytical workflow, but was filtered out based on quality control procedures, at the QC1 in one case and after computational validation in the second case. The PCR analysis also confirmed the L1 polymorphisms predicted in the father's genome and further identified an insertion site at the same locus in the genome of one of the offsprings, in the unaffected and one of the affected, respectively. In both cases, the insertion site identified in the offspring by PCR analysis was not previously predicted based on computational assessment (see Supplementary Fig. S2). Overall, PCR analysis revealed that our analytical workflow is 100% sensitive, as all predicted insertion sites were confirmed by experimental validation, and 77% specific, as 14 out of 18 L1 polymorphic loci were correctly identified as not having the insertion site.

We report Sanger sequencing results for the insertion site of all four non-reference L1 polymorphisms, all of which map in supposedly gene desert regions. For each non-reference L1 polymorphism, the consensus sequence at the insertion site along with ~350 bp on average genomic flanking sequence are available in FASTA format (Supplementary Materials) and have been submitted to RepBase. Repeatmasker aligns the 5′ and 3′ end of the consensus sequence of mei061 to the 5′ and 3′ end of the consensus

sequence of L1HS with percent divergence of 3.7% and 30.9%, respectively. Similarly, the consensus sequence of mei070 was aligned to the 5′ and 3′ end of L1HS with percent divergence of 1% and 30.03%, respectively. For mei076, both 5′ and 3′ end of the consensus sequence aligned to 3′ end of the consensus sequence of L1HS with percent divergence of 33.27% and 2.96%, respectively. However, it should be noted that the two fragments of mei076 consensus sequence align to different and consecutive portions of the 3′ end of the L1HS consensus sequence.

For mei044, two PCR products with different size were obtained from the PCR analysis. The consensus sequence analysis of these two PCR products reveals a possible duplication in the consensus sequence obtained from the longer PCR product. For the purpose of validation, we focused on the analysis of the consensus sequence targeting the insertion site without the duplication and we found that the 5′ end consensus sequence of mei044 did not align to any L1 consensus sequence, while the 3′ end consensus sequence aligned to 3′ end of the consensus sequence of L1HS with percent divergence of 0.7%. To fully evaluate the extent of the duplication, we will need to use a different technique that allows sequencing longer fragments. Overall, the sequence analysis of the novel non-reference L1 polymorphisms confirmed the insertion sites predicted by our workflow further supporting the L1 signature of the fragments not yet reported in the reference genome.

Mendelian Inheritance of L1 Insertions

Heritability of non-reference L1 insertion events within the six members family was investigated at L1 insertion loci identified by Tangram and RetroSeq that were (1) confirmed by computational validation, (2) not confirmed by computational validation but overlapping with an MRIP from euL1db, (3) not confirmed by computational validation but present in the offspring with full QC'ed insertions and in the parents with insertions not fully fulfilling the QC2 criteria.

First, we found that 30 (27%) non-reference L1 polymorphisms are present in offspring only and did not show a pattern of Mendelian inheritance in our dataset. Rates of retrotransposition events exhibiting Mendelian discordance were substantially different between Tangram and RetroSeq. This category includes events originally identified by RetroSeq only and computationally validated by *de novo* assembly. The same type of event was also originally identified by Tangram, but none of them was confirmed by computational validation. Mendelian discordance could reflect the rate of *de novo* events at these loci or simply errors in the detection of new insertions either in the parents (lack of it) or in the offspring (false positive). Among the 30 putative *de novo* L1 polymorphisms, we found that L1 insertions were identified in at least 1 of the parents with parameters below QC thresholds for 4 over 8 shared by affected and unaffected siblings; for 3 over 5 shared by affected offspring only; for 2 of the 11 *de novo* events carried uniquely by any of the affected offspring; and for 3 over 6 carried uniquely by the unaffected sibling (see Supplementary Table SII). Comparison of the remaining loci with non-reference L1 polymorphisms reported in euL1db revealed that our supposed *de novo* L1 insertions were most likely false positives since the same loci were reported in the database. L1 polymorphisms supported by parental insertion were classified as transmitted insertions and inheritance patterns evaluated consistently. For those L1 insertions not supported by parental insertion but by comparison with euL1db polymorphisms, we ruled out the possibility they were *de novo* events, but we could not evaluate inheritance patterns. The inheritance patterns for this non-reference L1 polymorphisms was labeled "undefined." Second, of the 110 family-based non-reference L1 polymorphic loci, 61 (55.4%) exhibited a clear mendelian inheritance. Of these, 41 are transmitted to either affected or unaffected offspring, 4 are transmitted from both parents to affected offspring only, and 2 to the unaffected sibling only. Nine are transmitted from the affected mother and five from the unaffected father to affected offspring only. Third, we found 19 (17%) non-reference L1 insertion events carried only by the parents; these events were classified as "untransmitted." As uniquely present in the parents, for which is not possible to reconstruct the pattern of inheritance, we defined these non-reference insertions as "private." Of the 19 L1 insertions that were not transmitted from the parents to any offspring, the unaffected father carried 13, the affected mother carried 5, and 1 was shared by both parents.

Genes and Pathways Affected by Retrotransposition Events

We found that 52 of the 110 (47.2%) non-reference L1 insertions loci map to protein coding or RNA genes. In particular, a gene annotation analysis implicates 34 protein-coding genes (31%), 10 lincRNA (9%), 5 regions harboring protein-coding and RNA genes (4.5%), 2 processed transcripts (1.8%), and 1 pseudogene (0.9%; Fig. 3).

Although it was not possible to formally test the association of single genes or pathways enrichment with schizophrenia, we consulted the Reactome database to retrieve information on the gene ontology terms for molecular function, biological processes, and protein–protein functional interactions of the L1-disrupted genes.

We found that L1 polymorphic insertions in SZ subjects have an impact on genes previously implicated in schizophrenia, such as *GABRB1* [Fatemi et al., 2013; Mueller et al., 2014], *FHIT* [Sullivan et al., 2008], and *RYR3* [Matsuo et al., 2009; Stephens et al., 2012]. Moreover, in reconstructing, the functional interactions that links together the genes affected by a L1 insertion in schizophrenia offspring, our findings suggest that both the neuronal signal transduction cascade and the synaptic calcium signaling pathways are affected by non-reference L1 insertions. Details of the results related to these gene pathways are described in Supplementary Table SII and Supplementary Results.

DISCUSSION

Our data demonstrate that our workflow to identify non-reference L1 insertions from NGS data is able to detect and confirm L1 insertions as polymorphic loci previously detected in other studies in the human genomes of six members of a family with schizophrenia. The identification of 110 non-reference

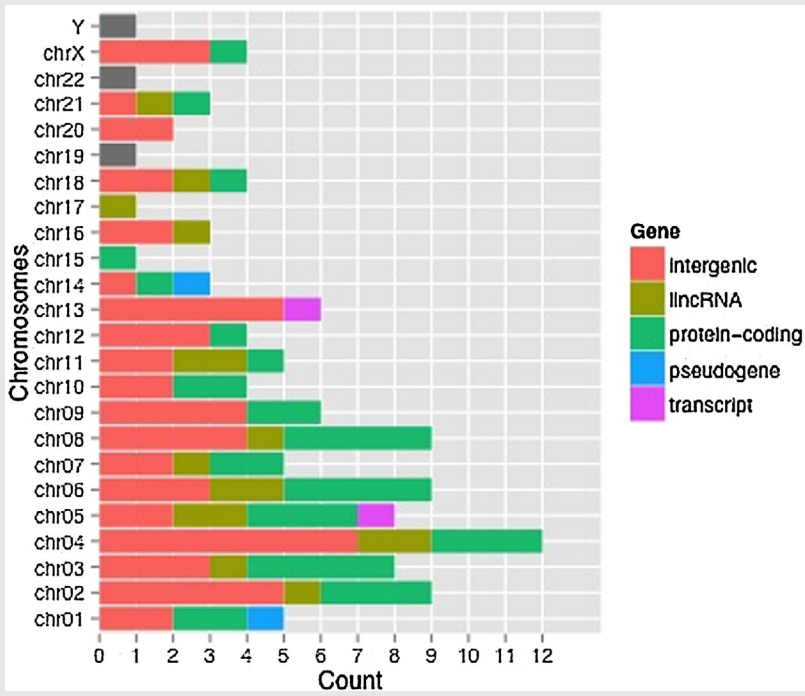

FIG. 3. Fraction of protein coding, RNA genes, pseudogenes, and processed transcripts affected by non-reference RIPs.

polymorphic L1 insertion loci demonstrates that retrotransposition events extend beyond the map of currently annotated transposable elements in the reference human genome (hg19), as previously shown also in several studies focusing on healthy subjects [Ewing and Kazazian, 2010, 2011; Iskow et al., 2010; Stewart et al., 2011] and different types of cancer [Tubio et al., 2014]. Our workflow is designed to allow computational identification and validation of L1 insertions from NGS data using publicly available algorithms that take advantage of different signals provided by NGS mapping. Previous studies used pipelines developed to identify RIPs from sequencing mapping signals derived from libraries constructed to specifically and uniquely target TE insertions. The extension of these pipelines to the continuously increasing number of NGS datasets available for different type of diseases would be cumbersome and difficult to achieve. Our findings, along with previous reports, underscore the relevance of identifying L1 RIPs (and for other transposable elements as a future extension) as much as any other type of structural variants. Algorithms capable of using NGS data to make such an investigation have the potential to allow the identification of RIPs in parallel to SNVs and CNVs in NGS datasets. Our approach is complemented with the application of *de novo* assembly as an independent method to computationally validate RIPs detected by RetroSeq and Tangram. Both software detected RIPs with high accuracy, although in our case, RetroSeq seems to perform better than Tangram, a trend already noted when dealing with heterozygous calls in deeper sequence coverage [Wu et al., 2014]. Local *de novo* alignment is traditionally used to refine the breakpoints of common structural variants such as deletions, duplications, and insertions. Here, the application of this method offers the advantage of assembling non-reference L1 sequences without the constraint of aligning reads to the reference genome. *De novo* assembled sequences putatively harboring L1 insertions were analyzed with RepeatMasker to determine the L1 sub family designation. Further, the inheritance patterns we observed in most of the identified L1 insertions provide support for the significant involvement of rare variants derived from retrotransposition in the susceptibility to schizophrenia. Under this scenario, L1 insertions might play a role in shaping the landscape of diversity and mutability underlying the genetic architecture of the disease. More than one third of the identified non-reference L1s is located within known protein-coding genes potentially affecting their function [Reilly et al., 2013; Erwin et al., 2014; Guffanti et al., 2014]. This estimate increases to one-half if we consider L1 insertions affecting RNA genes and pseudogenes with potential regulatory functions.

The high-coverage, full family-based, data we have used allow for the most precise estimates of transmission of RIPs, allowing also detection of *de novo* L1 insertions, that is, L1 insertions present in the offspring but not in the parents. *De novo* rates of retrotranspositions were previously analyzed in two trios (each including three individual genomes per trio for a total of six sequenced genomes) from the 1,000 Genomes Project [Stewart et al., 2011] and provided no evidence of *de novo* insertions. In line

with previous findings, we also ruled out the possibility that there were *de novo* insertion in our dataset. Even those events that do not exhibit Mendelian inheritance and are present solely in the offspring could not be confirmed as *de novo* as they were identified at the same meta-locus in dbRIP and/or euL1db.

Local *de novo* assembly, followed by RepeatMasker analysis, allows the classification of more than 86% of the sequences aligned at the target insertion site as regulatory (5′ and 3′ untranslated regions) or coding regions of L1HS and L1P1. The enrichment of these L1s elements supports recent L1 insertions in terms of evolutionary time [Smit et al., 1995; Khan et al., 2006]. More than 26% of the *de novo* assembled sequences align to 5′ untranslated regions specific to L1HS and L1P1. Although it is not possible to definitively reconstruct the full sequence of an inserted L1 with traditional NGS libraries, we argue that local *de novo* assembly allows decreasing the chances that the predicted insertion site derives from a mere mistake in the original alignment of the reads to the reference genome. Even if the inserted L1 *de novo* assembled sequences range between 75 and 100 bp, the identification of specific L1 signatures support the presence of a non-reference L1 insertions. Although Beck and colleagues proposed that hot L1 responsible for active retrotransposition can be more abundant than previously thought [Beck et al., 2010], the integration steps of retrotransposed L1 are still incompletely known (*see*, *e.g.*, [Zingler et al., 2005]), nonetheless showing that the majority of transposed L1s are truncated and no longer transpositionally active [Suzuki et al., 2009]. It is critically important to note, however, that even truncated L1s can modify the architecture of the transcriptome, frequently in a tissue- or disease-specific way and both in a sense or anti-sense direction [Matlik et al., 2006; Kines and Belancio, 2012; Guffanti et al., 2014].

A large fraction of non-reference L1 polymorphisms is located within the ORF of protein-coding genes (31%) and RNA genes, pseudogenes, and processed transcripts (16%). These estimates are consistent with previous reports on detection of L1 polymorphisms [Iskow et al., 2010]. The abundance of functional sequences disrupted by L1 insertions further supports that the contribution of TE polymorphic retrotransposition to shaping the genome is likely to be appreciable [Burns and Boeke, 2012]. The impact of non-reference L1 polymorphisms on local regulation of gene expression and function have already been described [Feschotte, 2008] and include different mechanisms of epigenetics relevance, for example, the creation of Transcriptional Start Sites, the modification of the methylation pattern in the chromosomal region where the insertion takes place or the creation of novel transcription factor binding sites [Reilly et al., 2013; Erwin et al., 2014; Guffanti et al., 2014].

What is the significance of L1 polymorphic insertions within the open reading frame of protein-coding genes in a family in which SZ segregates? Our hypothesis is that germline non-reference L1 polymorphic insertions account for a still unknown proportion of the overall genomic risk in developing schizophrenia. Disruption of a gene by insertion of an L1 may modify the organization of functional biological networks on the same ways as other polymorphic structural variants, which is a well-known phenomenon in schizophrenia [Rees et al., 2014]. The genes that we found presenting an L1 polymorphic insertion have been previously implicated in schizophrenia either by association studies or animal models. A simple functional gene annotation revealed common molecular function and biological processes implicated by genes affected by L1 insertions regardless of diagnosis status, mostly including integral protein of cell membrane with cell–cell communication and adhesion properties, as well as catalytic enzymes involved in signal transduction. The finding that genes affected by L1 polymorphic insertions are implicated in molecular functions like signal transduction, GTPase/Ras signaling, kinase activity regulation, cell–cell adhesion, comprehensively pointing to higher order gene-network processes such as regulation of synapse formation during early neural development and adult synaptogenesis is intriguing and suggestive of a pathogenic role. Interestingly, the same processes have been previously implicated by other structural variants findings, like CNVs, in schizophrenia [Walsh et al., 2008] and in the context of other neuropsychiatric diseases with neurodevelopmental component like autism [Pinto et al., 2010; Michaelson et al., 2012] and Alzheimer's disease [Guffanti et al., 2013]. It is plausible to contemplate a model in which the proportion of L1 polymorphic insertions is higher in the affected subjects than in the non-affected subjects, thus contributing to the genetic risk of schizophrenia.

Once we set up an efficient workflow for L1 polymorphic detection, the next step will be to perform a formal quantitative assessment of non-reference L1s in a larger sample to test this hypothesis. Here, our data demonstrate that these genetic variants exist and can be easily detected using NGS data and publicly available algorithms.

This identification of previously detected L1 polymorphic loci may simply be the ultimate result of a stochastic process where non-reference L1 insertion polymorphisms distribute randomly across the genome and act like harmless "passenger" mutations, still potentially maintaining a neutral evolutionary role, or it may suggest that the non-coding regions of CNS-related genes are the preferential sites of positive soft selective sweeps in the human genome [Taft et al., 2007], a frequently invoked evo-devo mechanism. To further investigate this hypothesis, we need to extend investigations of L1 insertions to population-scale samples rather than simple familial cases.

ACKNOWLEDGMENT

The authors would like to thank Dr. Thomas Keane for assistance with the Retroseq set of analyses.

REFERENCES

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479(7374):534–537.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. Cell 141(7):1159–1170.

Bundo M, Toyoshima M, Okada Y, Akamatsu W, Ueda J, Nemoto-Miyauchi T, Sunaga F, Toritsuka M, Ikawa D, Kakita A, Kato M, Kasai K, Kishimoto T, Nawa H, Okano H, Yoshikawa T, Kato T, Iwamoto K. 2014. Increased l1 retrotransposition in the neuronal genome in schizophrenia. Neuron 81(2):306–313.

Burns KH, Boeke JD. 2012. Human transposon tectonics. Cell 149(4): 740–752.

Cowley M, Oakey RJ. 2013. Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet 9(1):e1003234.

Erwin JA, Marchetto MC, Gage FH. 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci 15(8):497–506.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A, Park PJ, Walsh CA. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 151(3):483–496.

Ewing AD, Kazazian HH Jr. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 20(9):1262–1270.

Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res 21(6):985–990.

Fatemi SH, Folsom TD, Rooney RJ, Thuras PD. 2013. Expression of GABAA alpha2-, beta1- and epsilon-receptors are altered significantly in the lateral cerebellum of subjects with schizophrenia, major depression and bipolar disorder. Transl Psychiatry 3:e303.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9(5):397–405.

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110(3):315–325.

Grandi FC, An W. 2013. Non-LTR retrotransposons and microsatellites: Partners in genomic variation. Mob Genet Elemements 3(4):e25674.

Grandi FC, Rosser JM, Newkirk SJ, Yin J, Jiang X, Xing Z, Whitmore L, Bashir S, Ivics Z, Izsvak Z, Ye P, Yu YE, An W. 2015. Retrotransposition creates sloping shores: A graded influence of hypomethylated CpG islands on flanking CpG sites. Genome Res 25(8):1135–1146.

Guffanti G, Gaudi S, Fallon JH, Sobell J, Potkin SG, Pato C, Macciardi F. 2014. Transposable elements and psychiatric disorders. Am J Med Genet Part B Neuropsychiatr Genet 165B(3):201–216.

Guffanti G, Torri F, Rasmussen J, Clark AP, Lakatos A, Turner JA, Fallon JH, Saykin AJ, Weiner M, Vawter MP, Knowles JA, Potkin SG, Macciardi F. 2013. Increased CNV-region deletions in mild cognitive impairment (MCI) and Alzheimer's disease (AD) subjects in the ADNI sample. Genomics 102(2):112–122.

Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer MA. 2008. L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci USA 105(49):19366–19371.

Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, Liang P, Batzer MA. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res 33(13): 4040–4052.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res 22(9):1760–1774.

Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka JP, Valle D, Civin CI, Wang T, Wheelan SJ, Ji H, Boeke JD, Burns KH. 2010. Mobile interspersed repeats are major structural variants in the human genome. Cell 141(7):1171–1182.

Insel TR. 2014. Brain somatic mutations: The dark matter of psychiatric genetics? Mol Psychiatry 19(2):156–158.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, van Meir EG, Vertino PM, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141(7):1253–1261.

Jurka J. 2000. Repbase update: A database and an electronic journal of repetitive elements. Trends Genet 16(9):418–420.

Kazazian HH Jr. 2011. Mobile DNA transposition in somatic cells. BMC Biol 9:62.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: Transposable element discovery from next-generation sequencing data. Bioinformatics 29(3):389–390.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16(1):78–87.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 143(5):837–847.

Kines KJ, Belancio VP. 2012. Expressing genes do not forget their LINEs: Transposable elements and gene expression. Front Biosci (Landmark Ed) 17:1329–1344.

Kuhn A, Ong YM, Cheng CY, Wong TY, Quake SR, Burkholder WF. 2014. Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. Proc Natl Acad Sci USA 111(22):8131–8136.

Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol 28(1):47–55.

Lee J, Ha J, Son SY, Han K. 2012. Human genomic deletions generated by SVA-associated events. Comp Funct Genomics 2012:807270.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Madduri RK, Dave P, Sulakhe D, Lacinski L, Liu B, Foster IT. 2014. Experiences in building a next-generation sequencing analysis service using galaxy, globus online and Amazon web service. Concurr Comput 26(13):2266–2279.

Matlik K, Redik K, Speek M. 2006. L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol 2006(1): 71753.

Matsuo N, Tanda K, Nakanishi K, Yamasaki N, Toyama K, Takao K, Takeshima H, Miyakawa T. 2009. Comprehensive behavioral phenotyping of ryanodine receptor type 3 (RyR3) knockout mice: Decreased social contact duration in two social interaction tests. Front Behav Neurosci 3:3.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J. 2012. Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. Cell 151(7):1431–1442.

Mir AA, Philippe C, Cristofari G. 2014. EuL1db: The European database of L1HS retrotransposon insertions in humans. Nucleic Acids Res 43 (Database issue):D43–D47.

Mueller TM, Haroutunian V, Meador-Woodruff JH. 2014. N-Glycosylation of GABAA receptor subunits is altered in Schizophrenia. Neuropsychopharmacology 39(3):528–537.

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435(7044):903–910.

Pato MT, Sobell JL, Medeiros H, Abbott C, Sklar BM, Buckley PF, Bromet EJ, Escamilla MA, Fanous AH, Lehrer DS, Macciardi F, Malaspina D, McCarroll SA, Marder SR, Moran J, Morley CP, Nicolini H, Perkins DO, Purcell SM, Rapaport MH, Sklar P, Smoller JW, Knowles JA, Pato CN. 2013. The genomic psychiatry cohort: partners in discovery. Am J Med Genet Part B Neuropsychiatr Genet 162B(4):306–312.

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BH, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL, Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, Piven J, Ponting CP, Posey DJ, Poustka A, Poustka F, Prasad A, Ragoussis J, Renshaw K, Rickaby J, Roberts W, Roeder K, Roge B, Rutter ML, Bierut LJ, Rice JP, Salt J, Sansom K, Sato D, Segurado R, Sequeira AF, Senman L, Shah N, Sheffield VC, Soorya L, Sousa I, Stein O, Sykes N, Stoppioni V, Strawbridge C, Tancredi R, Tansey K, Thiruvahindrapduram B, Thompson AP, Thomson S, Tryfon A, Tsiantis J, Van Engeland H, Vincent JB, Volkmar F, Wallace S, Wang K, Wang Z, Wassink TH, Webber C, Weksberg R, Wing K, Wittemeyer K, Wood S, Wu J, Yaspan BL, Zurawiecki D, Zwaigenbaum L, Buxbaum JD, Cantor RM, Cook EH, Coon H, Cuccaro ML, Devlin B, Ennis S, Gallagher L, Geschwind DH, Gill M, Haines JL, Hallmayer J, Miller J, Monaco AP, Nurnberger JI Jr., Paterson AD, Pericak-Vance MA, Schellenberg GD, Szatmari P, Vicente AM, Vieland VJ, Wijsman EM, Scherer SW, Sutcliffe JS, Betancur C. 2010. Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466(7304):368–372.

Rees E, Walters JT, Chambert KD, O'Dushlaine C, Szatkiewicz J, Richards AL, Georgieva L, Mahoney-Davies G, Legge SE, Moran JL, Genovese G, Levinson D, Morris DW, Cormican P, Kendler KS, O'Neill FA, Riley B, Gill M, Corvin A, Sklar P, Hultman C, Pato C, Pato M, Sullivan PF, Gejman PV, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G. 2014. CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. Hum Mol Genet 23(6):1669–1676.

Reilly MT, Faulkner GJ, Dubnau J, Ponomarev I, Gage FH. 2013. The role of transposable elements in health and diseases of the central nervous system. J Neurosci 33(45):17577–17586.

Singer T, McConnell MJ, Marchetto MC, Coufal NG, Gage FH. 2010. LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes? Trends Neurosci 33(8):345–354.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Smit AF, Toth G, Riggs AD, Jurka J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246(3):401–417.

Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010. http://www. repeatmasker. org. Accessed 11 December 2014.

Stephens SH, Franks A, Berger R, Palionyte M, Fingerlin TE, Wagner B, Logel J, Olincy A, Ross RG, Freedman R, Leonard S. 2012. Multiple genes in the 15q13-q14 chromosomal region are associated with schizophrenia. Psychiatr Genet 22(1):1–14.

Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet 7(8):e1002236.

Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL. 2008. Genomewide association for schizophrenia in the CATIE study: Results of stage 1. Mol Psychiatry 13(6):570–584.

Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, Koyama H, Takeda S, Okada N. 2009. Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. PLoS Genet 5(4):e1000461.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays 29(3): 288–299.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 13(9):2129–2141.

Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, Menzies A, Roman-Garcia P, Fullam A, Gerstung M, Shlien A, Tarpey PS, Papaemmanuil E, Knappskog S, Van Loo P, Ramakrishna M, Davies HR, Marshall J, Wedge DC, Teague JW, Butler AP, Nik-Zainal S, Alexandrov L, Behjati S, Yates LR, Bolli N, Mudie L, Hardy C, Martin S, McLaren S, O'Meara S, Anderson E, Maddison M, Gamble S, Foster C, Warren AY, Whitaker H, Brewer D, Eeles R, Cooper C, Neal D, Lynch AG, Visakorpi T, Isaacs WB, van't Veer L, Caldas C, Desmedt C, Sotiriou C, Aparicio S, Foekens JA, Eyfjord JE, Lakhani SR, Thomas G, Myklebost O, Span PN, Borresen-Dale AL, Richardson AL, Van de Vijver M, Vincent-Salomon A, Van den Eynden GG, Flanagan AM, Futreal PA, Janes SM, Bova GS, Stratton MR, McDermott U, Campbell PJ. 2014. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science 345(6196):1251343.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320(5875):539–543.

Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. 2006. DbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat 27(4):323–329.

Wong K, Keane TM, Stalker J, Adams DJ. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol 11(12): R128.

Wu J, Lee WP, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. 2014. Tangram: A comprehensive toolbox for mobile element insertion detection. BMC Genomics 15:795.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB. 2009. Mobile elements create structural variation: Analysis of a complete human genome. Genome Res 19(9): 1516–1526.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18:821–829.

Zingler N, Willhoeft U, Brose HP, Schoder V, Jahns T, Hanschmann KM, Morrish TA, Lower J, Schumann GG. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. Genome Res 15(6):780–789.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.