Excess variants in AFF2 detected by massively parallel sequencing of males with autism spectrum disorder

Kajari Mondal, Dhanya Ramachandran, Viren C. Patel, Katie R. Hagen, Promita Bose, David J. Cutler and Michael E. Zwick

Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA

Received February 6, 2012; Revised June 13, 2012; Accepted June 29, 2012

Autism spectrum disorder (ASD) is a heterogeneous disorder with substantial heritability, most of which is unexplained. ASD has a population prevalence of one percent and affects four times as many males as females. Patients with fragile X E (FRAXE) intellectual disability, which is caused by a silencing of the X-linked gene AFF2, display a number of ASD-like phenotypes. Duplications and deletions at the AFF2 locus have also been reported in cases with moderate intellectual disability and ASD. We hypothesized that other rare X-linked sequence variants at the AFF2 locus might contribute to ASD. We sequenced the AFF2 genomic region in 202 male ASD probands and found that 2.5% of males sequenced had missense mutations at highly conserved evolutionary sites. When compared with the frequency of missense mutations in 5545 X chromosomes from unaffected controls, we saw a statistically significant enrichment in patients with ASD (OR: 4.9; \( P < 0.014 \)). In addition, we identified rare AFF2 3′ UTR variants at conserved sites which alter gene expression in a luciferase assay. These data suggest that rare variation in AFF2 may be a previously unrecognized ASD susceptibility locus and may help explain some of the male excess of ASD.

INTRODUCTION

The recent surge of advances in second-generation sequencing technologies and better methods of targeted enrichment have made the detection of a more complete spectrum of genetic variation increasingly feasible (1,2). There is keen interest in applying these approaches to uncover the genetic basis of polygenic complex human disease, including autism (OMIM 209850), a childhood-onset disorder characterized by impaired social interactions, abnormal verbal communication, restricted interests and repetitive behaviors. Autism has an estimated prevalence of 1% (3,4), and one of its most striking epidemiological features is a 4-fold excess of affected males.

Autism, or the broader autism spectrum disorder (ASD) phenotype, is an example of a highly heterogenous, multifactorial disorder with substantial heritability (5–8, reviewed in 9). To date, more than 100 different genes and genomic regions have been linked to this complex trait (reviewed in 10,11). Despite this, most of the genetic risk for ASD remains unexplained. Recent studies exploring ASD genetics generally adopt one of two study designs. The first employs genome-wide association studies, which have identified a few loci of interest, but largely failed to replicate findings between studies (12–14). A meta-analysis of these studies, with over 2500 study subjects, reveals it is extremely unlikely that there is any common variant influencing autism susceptibility with an odds ratio of \( >1.5 \) (15). The second design focuses on large but very rare (frequency usually less than 1 in a thousand in the general population) de novo and inherited copy number variants (CNVs). Numerous studies have now shown convincingly that this class of rare variation makes a significant contribution to autism susceptibility (16–27), explaining up to 15% of all ASD cases. Unfortunately, these studies point to a highly heterogenous allelic architecture, as no single risk variant is found in more than 1% of surveyed cases. Overall, although genetic studies have uncovered a large number of candidate loci, much ASD heritability remains unexplained.

Like other disorders with a male preponderance, there is evidence that mutations at X-linked loci in hemizygous males contribute to the observed ASD sex bias (24,28–30). Here, we...
further explore the hypothesis that rare mutations in X-linked genes contribute to ASD by performing targeted sequencing of the AFF2 (formerly FMR2) locus in two large autism cohorts. Our study is motivated by findings from FRAXE (OMIM #309548), a rare form of mild-to-moderate intellectual disability (ID) affecting 1 in 50 000 newborn males. Hallmarks of the FRAXE phenotype include learning difficulties, communication deficits, attention problems, hyperactivity and autistic behavior. The initial discovery and mapping of the FRAXE locus identified a CGG expansion upstream of AFF2 that silences the gene resulting in a complete loss of function. This simultaneously explained the ID and the presence of a chromosomal fragile site (31–33). Large deletions that include both AFF2 and the adjacent FMR1 gene result in a severe ID (34–36), whereas deletions of the AFF2 locus only present with autism and mild ID (37). A duplication of the AFF2 locus has also been reported in a young male with ID (38). Thus, since complete loss of AFF2 leads to a mild non-syndromic form of ID often presenting with autistic features, we hypothesized that hypomorphic alleles with reduced function might act as autism susceptibility loci. If correct, our hypothesis predicts that: (i) males with autism will have an excess of rare variants, (ii) these rare variants will be found at evolutionarily conserved sites, and (iii) these variants will influence gene function. Because males are hemizygous for AFF2, whereas females are diploid, rare variants at this locus could help explain the preponderance of males with ASD, if those damaging rare variants are largely recessive in nature.

Here, we report the results of a targeted second-generation sequencing study of the AFF2 locus in male ASD patients ascertained using different sampling designs from the Autism Genetic Resource Exchange (AGRE) and the Simons Simplex Collection (SSC). As predicted, we see an overrepresentation of rare variants, and missense mutations in particular, in ASD cases, with a more than 4-fold increase in ASD risk among carriers of rare AFF2 protein-coding changes. We observe both inherited and de novo risk alleles in our sample. Furthermore, we also identify two rare 3′UTR variants that appear to alter the expression of a reporter in a luciferase assay in a tissue specific fashion. Taken together, these data show that AFF2 is a susceptibility locus conferring increased risk for ASD in males. These data also suggest that rare coding sequence variations contribute to susceptibility for complex traits like autism, and such variations may account for some of the high heritability for these disorders.

RESULTS

We sequenced the AFF2 locus in a sample of 202 males with a diagnosis of autism. A total of 127 of the cases were obtained from the multiplex AGRE, while the remaining 75 cases were from the SSC. We identified a total of 286 sites of variation, with 269 single nucleotide variants (SNVs) and 17 insertions or deletions (indels). Overall levels of variation were similar between the two data sets ([39]; AGRE 6.0 × 10^−4, SSC 6.7 × 10^−4], with an excess of rare variants as evidenced by a negative value for the Tajima’s D test statistics for both sets of samples ([40]; AGRE 1.46, SSC 1.41]. For the SNVs, a total of 128, or 48%, had not been reported before. As summarized in Figure 1, almost all common variation (>5% frequency in our population) is contained in dbSNP, whereas most rare variants (<5%) have not been cataloged in dbSNP.

Our study is focused on rare variation, so we did not follow-up the 141 previously cataloged variants. Functional annotation of the remaining 128 SNVs revealed that 16 were located at sites with elevated evolutionary conservation (Supplementary Material, Table S1). Each of these variants was observed in a single case. To arrive at a better estimate of their population frequency, we genotyped 13 of the variants in a collection of 1400 unaffected male controls, obtained from the NIMH Human Genetics Initiative (Supplementary Material, Table S1). Only one intronic variant had a frequency >0.01, suggesting that the variants we identified are very rare in the general population.

We next restricted our attention to missense mutations. In our cases, there were 5 (2.5% of total cases sequenced) singleton non-synonymous variants. We estimated that the population frequency for two of these variants was ~0.001, with the other three having frequencies below 0.0001 (Supplementary Material, Table S1). We further noted that these mutations were all at highly conserved evolutionary sites (PhyloP >1, Table 1). To ask whether this distribution

![Figure 1. Summary of SNV and indel variation discovered at the AFF2 locus in males with ASD. The frequency of SNVs and indels (minor alleles) in cases is plotted against their level of evolutionary conservation. Most common variation has already been discovered and exists in public databases (blue; circles and diamonds). Most of the rare variation at AFF2 was discovered in our study and not contained in public databases (red; circles and diamonds).](https://academic.oup.com/hmg/article-abstract/21/19/4356/586946)
was unusual, we compared our case data to sequence data from 5545 X chromosomes of European American controls, obtained from the Exome Variant Server from the NHLBI Exome Sequencing Project. As shown in Table 2, for variants with PhyloP scores $\geq 1.0$ (hence the top 10% of conserved sites in the human genome), there is a 4.2-fold excess of mutations in ASD cases (95% CI: 1.28–11.09, P-value 0.01). If we further restrict our analysis to sites with PhyloP scores $\geq 2$, there is a 4.9-fold excess of rare variants in ASD cases (95% CI: 1.21–14.37, P-value 0.014, Table 2).

One potential explanation for our finding is that the sequenced cases are poorly matched with the controls, and the cases are systemically more variable at many or all X chromosome loci. To test this hypothesis, we sequenced a subset of our cases (75 SSC males) for virtually all coding loci on the X (1006 loci total). For each locus, we performed the same analysis as we did at AFF2, asking whether the cases were enriched for mutations with PhyloP $\geq 1$, or $\geq 2$. Our analysis rejects the hypothesis that our cases are systematically more variable than the controls (Supplementary Material, Fig. S1). At the vast majority of loci, controls have the same or more variation than cases. Furthermore, AFF2 appears to be especially enriched for variation in our cases. Of the 1006 loci, it exhibits the 77th greatest enrichment for variation relative to controls for sites with PhyloP $>1$ (P < 0.02). Because the 75 SSC males were a subset of the total samples used in this analysis, we also asked whether this subset shows evidence of population stratification relative to the whole case sample. To do so, we obtained whole genome association X chromosome genotype data for the AGRE case samples, and assessed stratification for all of our case samples (41–43). We observed no evidence of systematic stratification between our SSC cases and AGRE cases (Supplementary Material, Fig. S2), suggesting that we can reject the hypothesis that a simple model of stratification can account for our findings.

To confirm the patterns of segregation of putative AFF2 susceptibility variants, we genotyped the mother and available siblings for each of the five missense mutations to determine whether they showed segregation as expected (Fig. 2). For two of the missense mutations (R927C, D714N), we observed transmission from the mother to both affected sons. For another missense mutation (S660T), we observed transmission of the risk allele from the mother to the affected son, and transmission of the reference allele to the unaffected son (Fig. 2). Missense mutation R837C was transmitted from mother to affected son. For the AGRE cases, these patterns of transmission were expected because of the manner in which we selected our cases for sequencing. Interestingly, missense mutation I1030L in a SSC case arose as a de novo mutation in the male patient with autism, providing further evidence that AFF2 is an autism susceptibility locus.

### Table 2. Odds ratio (OR) for missense mutations at AFF2 for males with ASD when compared with unaffected controls

<table>
<thead>
<tr>
<th>Study</th>
<th>PhyloP $&gt;1$ Missense variants (PhyloP $&gt;1$)</th>
<th>Odds ratio versus controls (95% CI)</th>
<th>P-value</th>
<th>PhyloP $&gt;2$ Missense variants (PhyloP $&gt;2$)</th>
<th>Odds ratio versus controls (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present study ($n = 202$)</td>
<td>5</td>
<td>4.2 (1.28–11.09)</td>
<td>0.010</td>
<td>4</td>
<td>4.9 (1.21–14.37)</td>
<td>0.014</td>
</tr>
<tr>
<td>Control X chromosomes ($n = 5545$)</td>
<td>33</td>
<td>Reference</td>
<td>N/A</td>
<td>23</td>
<td>Reference</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Figure 2. Segregation pattern of the non-synonymous AFF2 variants in families from AGRE and the SSC ASD patient collections. Note that the I1030L variant is a de novo variant in the male with ASD.
Given that we observed a statistically significant excess of rare missense mutations in \textit{AFF2} in patients with autism, we next wondered whether, among the collection of rare variants we discovered, non-coding variants were also autism susceptibility variants. One way non-coding alleles could exert influence is by changing the expression of \textit{AFF2}. To test this hypothesis, we performed a luciferase assay using constructs containing two of the three 3′ UTR variants and a third with the human genome reference sequence as a comparison in two different cell lines. No other sequence variants were present in the 3′ UTR. As shown in Figure 3, the 3′ UTR variant (ChrX:148076068[C>T]) showed statistically significant reduction in gene expression in the luciferase assay performed in human HEK293T cells ($P < 0.02$), while the 3′ UTR variant (ChrX:148075200[T>C]) did not ($P < 0.31$). Repeating these experiments using mouse Neuro 2A cells resulted in both 3′ UTR variants showing a strong increase of expression relative to a construct containing the human genome reference sequence (ChrX:148076068[C>T], $P < 0.03$; ChrX:148075200[T>C], $P < 0.02$), suggesting that these variants influence gene expression in a tissue specific manner.

**DISCUSSION**

Uncovering the genetic basis of polygenic disorders remains a challenge for research. The idea that common genetic variation contributes significantly to common diseases has been a leading model in human genetics for the past 15 years (44–46); evaluating this model required the development of high-throughput genotyping platforms and a catalog of common human genetic variation as seen in the HapMap project (47,48). The ‘common-disease-common-variant’ model was bolstered by technology development, since direct sequencing was not a viable strategy. Therefore, systematic genotyping of common variants was perceived as the best way to begin to characterize the allelic architecture of complex human traits (49). In the time elapsed since, genome-wide association studies have shown that common variants with large effects are unlikely to exist in the human population for many disorders, although a large number of loci with alleles with much smaller ORs ($<1.2$) remains plausible (reviewed in 50).

Following Haldane in the 1920s, population biologists have long been aware that deleterious alleles of large effect will be maintained only at very low frequencies in the general population (51). Most quantitative traits, including human diseases, show substantial heritability in most populations (52). This observation suggests a possible paradox arising from the belief that these traits are influenced by stabilizing selection for an optimum, or simple directional selection to remove disease alleles, both of which should act to eliminate genetic variation (52,53). One model to account for the maintenance of variation supposes that variation is maintained by a very large number of loci, each with rare alleles of exceedingly small effect (too small to ever be individually detectable in an experiment), and which are individually at mutation-selection balance frequency (54). A later analysis, in contrast, argued that even if mutation-selection balance maintains much variation, then there are likely to be rare alleles with appreciable effects (55), and/or alleles at intermediate frequency created by balancing selection due to their pleiotropic effects on...
multiple phenotypes (52). In support of the former view, copy number variation studies have identified variants with a large effect size, having odds ratios often greater than 5 (16–27); however, these variants are quite rare, often occurring much less often than 1 in 1000 in the general population, a frequency generally consistent with a large effect locus at mutation selection balance. It has been reported that in aggregate, large CNVs account for 15% of the cases of autism. Since WGA studies generally preclude the existence of rampant intermediate frequency loci of large effect, Turelli’s theory predicts the existence of loci with multiple rare alleles, and intermediate ORs should make a significant contribution to polygenic disorders such as autism. Our study of AFF2 has identified one such locus.

In the 202 autistic males sequenced, we detected five rare replacement variants: S660T, D714N, R837C, R927H and I11030L. These variants appear in ~2.5% of boys with a diagnosis of autism. One variant (I11030L) arose as a de novo event. We have also identified rare 3’UTR sequence variants that appear to alter levels of gene expression in a luciferase assay in a tissue-specific manner. AFF2 missense mutations have also been seen in other males with neurodevelopmental disorders (56,57). The observation of deletions of just the AFF2 locus in a patient with autism and mild ID (37) and a duplication of the AFF2 locus in a young male with ID (38) provide further independent support for our findings.

We do not claim that the variants we have identified are monogenic causes of autism. Instead, our data support the idea that AFF2 harbors susceptibility alleles that make modest contributions to the risk of ASD. Furthermore, because AFF2 is found on the X chromosome, recessive acting effects would be expected to preferentially influence the risk of ASD in males. Even so, the size of the effects we observed could be overestimated for at least two reasons. First, the observed excess of rare variation in the human genome might make statistical significance estimates more challenging than modeled here (58). Second, our sample size rendered this study underpowered for detecting very weak effects. Therefore, there is the real possibility that the true OR of AFF2 mutations might be even smaller than we have estimated due to ‘Winner’s Curse’ (59). Finally, we note that we did not perform a genome-wide study, and had we done so with only our sample size, the magnitude of the effect estimated here would not have survived a multiple test correction. Detecting effect sizes this small would require vastly larger numbers of samples for a whole genome study. While recent exome sequencing studies identified de novo variants that appear enriched in patients with ASD (60–63), it is worth noting that the variant we observed could have arisen by chance. Replication of our finding regarding AFF2 in a larger collection of individuals with ASD could help resolve a number of these potential issues and provide further support for our findings.

Much remains to be discovered concerning the structure and function of the AFF2 protein. We know AFF2 is a large gene with 21 exons and 6 annotated isoforms with alternative splicing among exons 2, 3, 5 and 7 (31–33). The longest of the AFF2 isoforms is composed of 1311 amino acids and contains 2 nuclear localization signal sequences (33). AFF2 belongs to a gene family that includes AF4, LAF4 and AF5q31. The AFF2 gene has a common ancestor in Drosophila melanogaster, the Lilliputian. Inactivation of lilli generates a fly of reduced size. By analogy with AF4, AFF2 has been considered a putative transcription factor (64), with the N-terminal domain of the protein (amino acids 1–541) known to have transactivation activity (65). Four of our mutations (S660T, D714N, R837C, R927H) fall in the C-terminal domain of the protein (amino acids 633–966) shown to be responsible for localization of the protein in nuclear speckles (64); none of our mutations fell into putative RNA-binding domains between amino acids 787–815 and 890–917 (64). The careful functional characterization of these and other alleles of AFF2 should help elucidate novel pathways and candidates that, when disrupted, contribute to autism susceptibility.

**MATERIALS AND METHODS**

**Selection of samples**

We selected 127 males from the Autism Genetic Resource Exchange ( AGRE) multiplex collection and 75 males from the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection, New York, NY, USA (SSC) for target DNA amplification and DNA sequencing. From the AGRE collection, we chose multiplex families with two or more male affected sib-pairs who shared >99% of 76 genotyped single nucleotide polymorphisms (SNPs) in the AFF2 genomic region (66). One male was randomly chosen if both affected siblings were equally affected; otherwise, the male with autism was chosen over those boys with a diagnosis of not quite autism or broad spectrum. From the SSC collection, we chose only those boys who were described as autistic and not reported to have any other syndromes. From the SSC collection, we chose 75 male children from different families with a diagnosis of ASD (67). We required subjects to have at least one unaffected sibling with no history of ASD or ID based on medical history and/or pedigree evaluation. Parents showed no signs of a broader autism phenotype. Prior to processing, DNA samples were quantified by measuring OD260/280 using a NanoDrop instrument. Following quantification, 100 ng of DNA from each sample were run on a 0.8% agarose gel to verify the integrity of the genomic DNA.

**Long PCR-based amplification of AFF2 genomic region from AGRE samples**

We prepared target DNA for sequencing the AGRE samples by performing long PCR (LPCR) amplification of the AFF2 genomic region. Long-range PCR primers were designed using EmPrimer (http://primer.genetics.emory.edu) using the following selection criteria: length, 29–32 bases; GC percentage, 45–60; and Tm ~ 68°C. The LPCR primers used are contained within Supplementary Material, Table S2. Five hundred nanograms of DNA from each of the AGRE samples were aliquoted and PCR master mix (1 × LA Taq buffer (TaKaRa Bio Inc., Otsu Shiga, JP), 250 mM dNTP (TaKaRa Bio Inc.), 400 nM of both forward and reverse LPCR primers and 0.1 U/ml of LA Taq (TaKaRa Bio Inc.) were added. If an amplicon had a high GC content, we used 1 × GC Buffer (TaKaRa Bio Inc.) in place of 1 × LA Taq buffer. PCR was...
performed using the following parameters: denaturation at 94°C for 2 min, 29 cycles of 94°C for 10 s, 68°C for 1 min per kilobase of amplicon size and final extension at 68°C for 5 min. Amplification was confirmed by using 1% agarose 96-well E-Gels (Invitrogen, Carlsbad, CA, USA). After purification, concentration of each of the amplicons was quantitated using PicoGreen dsDNA Quantitation reagent (Invitrogen) and the Tecan Ultra Evolution plate reader. An equimolar amount of each amplicon was then pooled by sample. Pooled amplicons were then purified using the Invitrogen PureLink PCR Purification Kit (Invitrogen) with the HC buffer. LPCR-derived, purified products were then sheared using Covaris E210 (Duty cycle 10%, Intensity cycle 5, cycle/burst: 200, time: 180 s).

RainDance microdroplet-based PCR enrichment of ChrX exome from SSC samples
We prepared target DNA for sequencing the SSC samples by using RainDance Technology’s (RDT) microdroplet-based technology to enrich for the human X chromosome exome as described previously (68). For 15 samples, a slightly modified version of the manufacturer’s protocol was used where the purified PCR product was run in 1% agarose gel, and the bands between 200 and 700 bp were extracted using a Qiagen Gel purification kit. The ends of the DNA fragments were repaired at 25°C for 30 min using the New England Biolabs End Repair Module (New England Biolabs, Ipswich, MA, USA), catalogue #E6050L, followed by purification using the Qiagen MinElute columns (Qiagen, Valencia, CA, USA). The PCR fragments were then concatenated at 20°C for 30 min using NEB Quick Ligation Kit (NEB, Ipswich, MA, USA, catalogue #M2200L). The ligated products were purified using the Qiagen MinElute columns and were made up to 100 µl volume by adding Qiagen elution buffer and then fragmented using a Covaris E210 machine (duty cycle: 10%, intensity cycle: 5, cycle/burst: 200, number of treatments: 4, total time per treatment: 45 s).

Illumina library preparation, sequencing and data analysis
The sheared fragments from the LPCR and RainDance libraries were purified using a Qiagen QIAquick PCR purification column and eluted in 32 µl of elution buffer (Qiagen). The standard Illumina Genome Analyzer (IGA) multiplex library preparation protocol was then used to prepare samples for sequencing with only one modification: while purifying the adaptor-ligated products, we used Invitrogen E-Gel SizeSelect 2% (Invitrogen, catalogue #G6610-02) in place of the gel purification method suggested by Illumina. We confirmed sample enrichment by running an Agilent BioAnalyzer 7500 DNA chip. A quantitative qPCR was performed to quantitate the library using a KAPA Library Quantification Kit (Kapa Biosystems, Woburn, MA, USA, catalogue # KK4824). Enriched DNA was denatured and diluted to a concentration of 8 pM. Cluster generation and 70 bp or 100 bp single-end sequencing was performed using standard IGA or Illumina HiSeq protocols. We performed multiplex single-end sequencing of 12 samples per lane for the AGRE samples and 3–4 samples per Illumina lane for the SSC samples.

After the completion of Illumina sequencing, the filtered reads were mapped and variants called using PEMapper (Cutler et al., submitted). The AFF2 reference sequence used for the AGRE samples consists of 10 discontinuous fragments covering 84.8 kb. For the AGRE samples, 99% of the bases had more than 8× coverage. Median depth of coverage was in the range of 388–1548, while the range of average depth of coverage was 459–1907 (Supplementary Material, Table S3). Since the entire human X chromosome exome was sequenced in the SSC samples, we used a reference sequence consisting of 5748 discontinuous fragments covering 4.7 Mb. For the SSC samples, between 83 and 97% of the targeted reference bases had more than 8× coverage. Median depth of coverage was in the range of 20–607, while the range of average depth of coverage was 21–675 (Supplementary Material, Table S3).

Variant annotation, validation and segregation analysis
All common and rare single base pair and indel sequence variants identified by PEMapper were annotated into functional classes using the sequence annotator program, SeqAnt (69). Selected highly conserved rare variants (phastCon score > 0.65) were independently validated by Sanger sequencing. The primers for validation were designed using Primer 3 software (http://frodo.wi.mit.edu/). After validating replacement and UTR variants in cases, we next sequenced the mothers and affected or unaffected male siblings to verify the segregation pattern of the variant. The de novo variant at I1030L was confirmed by Sanger sequencing using a DNA sample obtained from blood for both the mother and male offspring.

Control genotyping
To aid in the estimation of allele frequencies in matched controls, we performed genotyping of male samples of European ancestry obtained from the NIMH Human Genetics Initiative (https://www.nimhgenetics.org/available_data/controls/). These control samples used were identical to those used in a previous study (70). Genotyping was performed by the iPLEX Gold method (Sequenom, San Diego, CA, USA) per the manufacturer’s instructions, using primers (see Supplementary Material, Table S4) designed with the Sequenom Assay Design 3.1 software. A positive control was included in every plate to confirm the sensitivity of the assay. The genotyping for the replacement variants, S660T and I1030L, could not be performed, due to failure of the primer design assay by both TaqMan and Sequenom.

Statistical analysis
All population genetic analyses were calculated using the popgen_fasta2.0.c code (Cutler, unpublished work) on the fasta files as previously described (71). This code calculated the average number of pairwise differences and Watterson’s estimator of the population mutation rate (θw per site) for the entire sequenced region and different annotated SNV functional classes, while accounting for missing data (39). A point estimate for Tajima’s D was determined for all the data and different SNV functional classes (40). Control data from
5545 X chromosomes were obtained from exome sequencing data of individuals of European ancestry publicly available at the NHLBI Exome Variant Server (72). P-values and odds ratios were calculated with Fisher’s exact test in the R statistical package. PhyloP scores (phyloP46wayPlacental) were obtained using the February 2009, GRCh37(hg19) assembly from the UCSC Genome Browser (73). Affymetrix 5.0 data generated by the BROAD Institute for AGRE samples were downloaded from AGRE website (19). A total of 215 SNPs that spanned the X chromosome were successfully genotyped in the AGRE cases and sequenced in our SSC cases. We used PLINK to perform LD-based pruning (indep-pairwise option) resulting in a total of 179 SNPs that were in approximate linkage equilibrium (74,75). We then used EIGENSOFT 4.2 and GCTA for principal component analysis to determine whether there was any evidence for population stratification between our AGRE and SSC cases (41–43).

Luciferase assays
We performed luciferase assays in HEK293T and in Neuro-2a cell lines to determine whether the AFF2 3’UTR variants we identified reduce gene expression relative to a 3’UTR construct with the reference sequence.

Plasmid construction
Full-length 3’UTR was amplified for two rare UTR variants with low (ChrX: 148075200 [T>C]) and high evolutionary conservation (ChrX: 148076688[C>T]). The resulting products were cloned into the pminiGLO Dual-Luciferase vector (Promega, Madison, WI, USA, Cat #E1330). As a control, a full-length 3’UTR that consisted of the human genome reference sequence was amplified from an unaffected HapMap normal control and also cloned into the same vector. We were unable to successfully generate a construct for one of the variants we discovered (ChrX:148075804[T>A]). Sanger sequencing confirmed that each variant construct contained the correct novel variant site.

Cell culture and transfections
HEK293T and Neuro-2a cells were cultured at 37°C with 5% CO2 in DMEM and RPMI media, respectively, with 10% fetal bovine serum. Twenty-four hours before transfection, 2 × 10^5 cells were plated in 1 ml of media in each well of 12-well cell culture dishes. Transfections were carried out in Opti-MEM (Invitrogen), using 8 μl of Lipofectamine 2000 (Invitrogen) and 0.5 μg of plasmid. Each plasmid was transfected into three separate wells. Just before transfection, the old media replaced with fresh media (DMEM or RPMI with 10% fetal bovine serum). Twenty-four hours after transfection, cells were harvested with 250 μl 1× Passive Lysis Buffer (Promega) by rocking at room temperature for 15 min. Lysates were cleared of cell debris by centrifugation at 14000 rpm for 5 min at 4°C.

Luciferase assay
From each transfection, 20 μl of the lysate was added to a luminometer tube. To each luminometer tube, 100 μl of LAR II was added. A manual-load luminometer was used to measure the luminescence over a 10 s period, following a 2 s premeasurement delay. The luminometer measurement was repeated after the addition of 100 μl of Stop & Glo reagent. For each lysate, the firefly luciferase values were divided by the Renilla luciferase values. The results of the independent transfections were averaged and the standard deviation was calculated for each plasmid. We performed three independent transformations of each construct, with three replications of each transformation. The raw data were normalized relative to the reference sequence construct and a two-tailed, unequal variance Student’s t-test was performed to determine whether the 3’UTR containing vectors reduced gene expression relative to a construct with the human genome reference sequence.

SUPPLEMENTARY MATERIAL
Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS
The Emory Custom Cloning Core facility generated constructs to our specifications for our expression analysis. We thank members of the Cutler and Zwick labs for comments on the manuscript, Jennifer Mulle and Stephen T. Warren for discussion, Cheryl T. Strauss for editing and the Emory-Georgia Research Alliance Genome Center (EGC), supported in part by PHS Grant UL1 RR025008 from the Clinical and Translational Science Award program, National Institutes of Health, National Center for Research Resources, for performing the Illumina sequencing runs. The ELLIPSE Emory High Performance Computing Cluster was used for this project. The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926), and the Heart GO Sequencing Project (HL-103010).

Conflict of Interest statement. None declared.

FUNDING
This work was supported by the National Institutes of Health/ National Institutes of Mental Health (NIH/NIMH) and Gift Fund (grant number: MH076439, M.E.Z.); the Simons Foundation Autism Research Initiative (M.E.Z.); and the Training Program in Human Disease Genetics (grant number: 1T32MH087977, D.R.).

REFERENCES


