High-Throughput Sequencing of mGluR Signaling Pathway Genes Reveals Enrichment of Rare Variants in Autism

Raymond J. Kelleher III1,2,3,*, Ute Geigenmüller4,5, Hayk Hovhannisyan6, Edwin Trautman4, Robert Pinard4, Barbara Rathmell5, Randall Carpenter5, David Margulies4

1 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 2 Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, 3 Program in Neuroscience, Harvard Medical School, Boston, Massachusetts, United States of America, 4 Correlagen Diagnostics, Inc., Waltham, Massachusetts, United States of America, 5 Seaside Therapeutics, Cambridge, Massachusetts, United States of America

Abstract

Identification of common molecular pathways affected by genetic variation in autism is important for understanding disease pathogenesis and devising effective therapies. Here, we test the hypothesis that rare genetic variation in the metabotropic glutamate-receptor (mGluR) signaling pathway contributes to autism susceptibility. Single-nucleotide variants in genes encoding components of the mGluR signaling pathway were identified by high-throughput multiplex sequencing of pooled samples from 290 non-syndromic autism cases and 300 ethnically matched controls on two independent next-generation platforms. This analysis revealed significant enrichment of rare functional variants in the mGluR pathway in autism cases. Higher burdens of rare, potentially deleterious variants were identified in autism cases for three pathway genes previously implicated in syndromic autism spectrum disorder, TSC1, TSC2, and SHANK3, suggesting that genetic variation in these genes also contributes to risk for non-syndromic autism. In addition, our analysis identified HOMER1, which encodes a postsynaptic density-localized scaffolding protein that interacts with Shank3 to regulate mGluR activity, as a novel autism-risk gene. Rare, potentially deleterious HOMER1 variants identified uniquely in the autism population affected functionally important protein regions or regulatory sequences and co-segregated closely with autism among children of affected families. We also identified rare ASD-associated coding variants predicted to have damaging effects on components of the Ras/MAPK cascade. Collectively, these findings suggest that altered signaling downstream of mGlurRs contributes to the pathogenesis of non-syndromic autism.

Introduction

Autism is a brain disorder of early childhood characterized by impaired communication, impaired social interaction and restricted and repetitive patterns of activities and interests. The phenotypic breadth of autism is encompassed by the term autism spectrum disorder (ASD), which collectively affects nearly 1% of children, typically with onset prior to the age of three years. Insights into the genetic landscape of ASDs have indicated substantial heterogeneity. Identification of the genes responsible for recognized neurogenetic syndromes with high prevalence of ASD, such as FMR1 and TSC1/2, and localization of causative genes within microdeletions associated with non-syndromic (idiopathic) autism, such as SHANK3, NLGN3, and NLGN4, have highlighted the role of single-gene mutations in disease pathogenesis [reviewed in [1–4]]. A series of recent genome-wide studies have revealed a significant contribution of rare de novo copy number variants (CNVs) affecting many different loci to ASD susceptibility [5–9]. These known genetic factors are estimated to account for less than 20% of ASD cases, however, and thus much of the genetic basis of ASDs remains unexplained.

Identification of common molecular pathways affected by genetic variation in autism is essential to understand disease pathophysiology and devise effective therapeutic strategies [10]. Increasing evidence suggests a central role for defects in synaptic structure and function in the pathogenesis of autism despite the underlying genetic heterogeneity [3,11–14]. However, the neuronal pathways subserving synapse structure and function which are pathologically altered and which may represent convergence points for genetic lesions in autism remain to be defined. In this
study, we tested the hypothesis that functional genetic variation in components of the signaling network coupling group 1 mGluRs to synaptic protein synthesis contributes to the pathogenesis of non-syndromic autism. Several lines of evidence suggest that the “mGluR pathway” may play an important role in ASD pathophysiology [4]. Group 1 mGluRs (mGluR1/5) signal through the Ras/ERK, and PI3K/mTOR signaling cascades to regulate protein synthesis (Fig. 1). Syndromic disorders with high prevalence of ASD are caused by mutations in pathway components that regulate ERK activity (NP1, RAS isoforms, RAF isoforms, MAP2K1, MAP2K2, mTOR activity (TSC1, TSC2, PTEN)), or mRNA translation directly (FMR1) [4]. Analysis of Fmr1−/− mice, a model of fragile X syndrome, the most common inherited cause of autism, has revealed excessive mGluR5-dependent synaptic protein synthesis and plasticity [15]. An array of phenotypes in these mice could be corrected by attenuation of mGluR5 activity [16]. Similarly, analysis of Nf1−/− and Tsc2−/− mice has suggested treatments to correct behavioral impairments in neurofibromatosis and tuberous sclerosis complex (TSC) [17,18]. Thus, understanding how genetic variation in the mGluR pathway contributes to non-syndromic autism may suggest new therapeutic interventions since this pathway is amenable to pharmacological manipulation.

To test the hypothesis that genetic variation in the mGluR pathway increases risk for non-syndromic autism, we took advantage of the massively parallel sequencing capacity offered by next-generation sequencing (NGS) technologies to interrogate a panel of candidate genes in a cohort of autism cases and controls. Our strategy was designed to enhance high-throughput discovery by performing variant discovery in pools of samples (Fig. 2A). We then resequenced “orthogonal” pools on a second next-generation platform to validate variants and assign them to individual samples. We applied this approach to identify variants associated with autism in coding exons or flanking sequences in a panel of 18 genes encoding key components of the mGluR pathway [4] (Fig. 1; Table S1). Genes selected for analysis encode group 1 mGluRs (mGluR1, mGluR5), postsynaptic density (PSD)-associated scaffolding proteins that regulate mGluR function (Homer1, Shank3), components of the Ras/ERK (H-Ras, Raf, MEK1, MEK2) and PI3K/mTOR (PI3K catalytic and regulatory subunits, PTEN, TSC1, TSC2, Rheb) signaling cascades, mRNA-binding factors that regulate protein synthesis (eIF4E, FMRP), a protein target of mGluR-induced translation that regulates synaptic plasticity (Arc), and an E3 ubiquitin ligase known to regulate Arc degradation (Ube3a) [19]. Although some of the genes studied are known to cause syndrome ASD (e.g. FMR1, TSC1, TSC2, UBE3A), their possible role in non-syndromic autism has been unclear.

Results

We analyzed single-nucleotide variants (SNVs) in mGluR pathway genes by high-throughput sequencing in a cohort of 290 unrelated AGRE cases and 300 ethnically-matched controls (Fig. 2A; Table S2). Inclusion criteria for AGRE cases were a diagnosis of idiopathic (“non-syndromic”) autism and at least one affected sibling. AGRE cases were screened to exclude non-idiopathic (“syndromic”) autism secondary to known neurogenetic disorders such as fragile-X syndrome [20]. Two distinct sets of pools were prepared from genomic DNA samples isolated from these cohorts for sequencing on the Illumina GAII and the Helicos HeliScope. An orthogonal strategy for sample pooling was used in which samples were arrayed in a matrix with 15 rows and 20 columns. Samples were pooled along rows to generate 15 pools of 20 samples each for GAII sequencing and pooled along columns to generate 20 pools of 15 samples each for HeliScope sequencing. Each sample representing a single case or control subject was thus identified by a unique combination of two pools representing its unique position within this matrix. Each genomic DNA pool was used as a template for PCR amplification of all coding exons from our panel of 18 mGluR pathway genes (240 exons comprising a total of 40,473 bases). PCR amplicons from each genomic DNA pool were concatenated and sheared to construct libraries for high-throughput sequencing. The average coverage per exon per pool was 610 for the GAII and 1,688 for the HeliScope. An average coverage of ≥10 per individual was achieved for 87% (GAII) and 97% (HeliScope) of exons across all pools. Sensitivity of variant detection was 610 for the GAII and 1,688 for the HeliScope. An average coverage of ≥10 per individual was achieved for 87% (GAII) and 97% (HeliScope) of exons across all pools. Sensitivity of variant detection was
detection was therefore generally limited by the lower coverage achieved on the GAII in this analysis.

We used confirmation by Sanger sequencing to evaluate the sensitivity and fidelity of SNV detection on each NGS platform and on both platforms combined. The false discovery rate for single variant occurrences (singletons) fell sharply without losing sensitivity when considering only SNVs detected concordantly on both platforms (Fig. S1). We therefore limited our subsequent analysis to SNVs that were concordantly detected on both platforms.

Common SNVs were defined as those with allele frequencies ≥1%, while rare SNVs were defined as having allele frequencies <1%. The number of common SNVs did not differ between the AGRE and control groups (Table 1). Rare SNVs were modestly enriched in AGRE samples compared to controls (302 and 276, respectively). However, when we excluded rare SNVs that are either silent or present in both populations, and therefore presumed to be benign (Fig. 2B), a significant enrichment of genetic variation was detected in the autism population, with 80 and 49 SNVs in AGRE and control groups, respectively \((P = 0.001)\). When we additionally eliminated SNVs located in 5’ UTRs or deep intronic regions, thereby focusing on SNVs with potentially deleterious effects, there emerged a two-fold enrichment of variants in the autism population, with 58 and 29 SNVs in the AGRE and control groups, respectively \((P = 0.0005)\). The two-fold enrichment of SNVs in the autism population persisted if we further excluded SNVs characterized as common in dbsNP (build 132), with 57 and 27 SNVs in the AGRE and control populations, respectively \((P = 0.0002)\).

We applied several commonly used computational tools for pathogenicity prediction to assess the functional impact of the rare, potentially deleterious missense variants identified in the AGRE and control groups. Since the performance and reliability of methods for pathogenicity prediction varies widely, and their results typically correlate poorly \([21]\), we compared the predictions derived from SIFT, PolyPhen2, SNP&GO and MutPred. Although SIFT did not predict a significant difference between groups, PolyPhen2, SNP&GO, and MutPred each predicted a 2- to 3-fold enrichment of damaging missense variants in the AGRE group. Overall, comparably high proportions of the variants classified as potentially deleterious in the AGRE and control groups (68% and 71%, respectively) were predicted to be functionally damaging by at least one of the four prediction tools, supporting the notion that disruptive variants within mGlurR pathway components occur at a higher rate in the autism relative to the control population.

At the level of individual pathway genes, we identified a significant excess of SNVs in the autism population for the TSC1, TSC2, SHANK3, and Homer1 genes \((P < 0.05)\). Causal roles for TSC1 and TSC2 have previously been demonstrated in syndromic autism. TSC1 or TSC2 mutations can cause TSC, a syndromic disorder characterized by tumor growth in multiple organs, including the brain. Although the manifestations of TSC include ASD in up to 50% of cases \([22]\), our findings additionally implicate TSC1 and TSC2 as risk genes for non-syndromic autism independent of their causative role in TSC. Consistent with this view, the majority of the rare, potentially disruptive TSC1/TSC2 SNVs identified in the AGRE population are novel, and none of these SNVs has been identified previously as a cause of TSC.
Our identification of increased genetic variation in SHANK3 in autism cases supports the emerging view of SHANK3 as an important autism-risk gene [23]. One missense variant observed in our study (R300C) was previously identified as a potential risk factor for ASD [23]. In addition, we identified a number of novel rare SHANK3 SNVs in the autism population (Table 2). The over-representation of rare, potentially disruptive variants in genes previously implicated in ASD (TSC1, TSC2, SHANK3) provides validation of this approach to detect genes that contribute genetic risk in autism.

The fourth gene displaying a significant enrichment of autism-associated rare SNVs, HOMER1, has not previously been implicated in autism. Homer1 is a PSD-localized scaffolding protein that interacts with a variety of PSD proteins, including mGluRs and Shank proteins [24]. Binding of Homer1 to mGluRs promotes trafficking of mGluRs to the postsynaptic membrane and couples mGluR5 to the mTOR signaling pathway [25]. Homer and Shank proteins interact to form an extended polymeric platform required for recruitment and assembly of synaptic proteins and structural integrity of dendritic spines [26]. Consistent with this function, the Homer-Shank interaction has been shown to promote morphological and functional maturation of dendritic spines [27]. We identified multiple rare missense variants in HOMER1 in AGRE cases but not in controls. All of the identified missense variants in HOMER1 alter residues that are invariant among mammalian species, and all but one is invariant across vertebrate species (Fig. 3A). Two of these variants (c.195G>T, M65I and c.290C>T, S97L) localize to the EVH1 (Ena/VASP homology 1) domain of Homer1, which binds to Pro-Pro-Ser-Pro-Phe motifs in mGluR1 and mGluR5 and a Pro-Pro-Glu-Glu-Phe motif in Shank3 [24]. A third potentially damaging SNV in HOMER1 (c.425C>T, P142L) affects one of the conserved prolines within the P-motif of the CRH1 (conserved region of Homer 1) domain, which serves as an internal binding site for the EVH1 domain. It has been proposed that the P-motif competes for binding of the Homer1 EVH1 domain to the proline-rich motif in target proteins such as mGluRs, thereby modulating Homer1 homo-multimerization and mGluR interaction. Interestingly, one of the GRM5 variants (c.3503T>C, L1168P) detected in AGRE samples is located in close proximity to the conserved Pro-Pro-Ser-Pro-Phe Homer1-binding motif in mGluR5. In addition, we identified an SNV in the HOMER1 3' UTR (c. 1080 C>T) only in the autism and not in the control population. Growing evidence suggests an important role for 3' untranslated regions (UTRs) as the sites of pathogenic variation due to their diversity and density of cis-acting regulatory elements [28,29]. In particular, genetic variants that alter microRNA-binding sites have been implicated in the pathogenesis of a variety of human diseases, including the neuropsychiatric disorder Tourette’s syndrome [30,31]. The identified HOMER1 3' UTR variant, which is located 15 nucleotides distal to the translation termination codon,
lies within a cluster of predicted microRNA binding sites and alters predicted seed pairing for several microRNAs, including miR-96, miR-182, miR-203, and miR-513a-3p (miRanda and Microcosm algorithms, www.microrna.org; www.ebi.ac.uk/enright-srv/) (Fig. 3B). Based on the predicted effects on miRNA binding, this variant may perturb the efficiency and/or regulate mGluR pathway genes in autism cases. For each variant, the nucleotide substitution is shown, and the corresponding amino acid substitution is indicated parenthetically.

Table 2 summarizes the 58 rare, potentially deleterious SNVs that were identified in mGluR pathway genes only in autism cases. For each variant, the nucleotide substitution is shown, and the corresponding amino acid substitution is indicated parenthetically.

Table 2. Rare, potentially deleterious variants identified in mGluR pathway genes in autism cases.

<table>
<thead>
<tr>
<th>Gene</th>
<th>c.587C&gt;T</th>
<th>c.605A&gt;G</th>
<th>c.829G&gt;A</th>
<th>c.898C&gt;T</th>
<th>c.1080C&gt;T</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARC</td>
<td>c.587C&gt;T</td>
<td>c.605A&gt;G</td>
<td>c.829G&gt;A</td>
<td>c.898C&gt;T</td>
<td>c.1080C&gt;T</td>
</tr>
<tr>
<td>GRM1</td>
<td>c.195G&gt;T</td>
<td>c.2294T&gt;C</td>
<td>c.290C&gt;T</td>
<td>c.968G&gt;T</td>
<td>c.1006C&gt;T</td>
</tr>
<tr>
<td>GRM5</td>
<td>c.383C&gt;T</td>
<td>c.3503T&gt;C</td>
<td>c.1217T&gt;C</td>
<td>c.763T&gt;C</td>
<td>c.1292T&gt;C</td>
</tr>
<tr>
<td>HOMER1</td>
<td>c.898C&gt;T</td>
<td>c.920C&gt;G</td>
<td>c.1254T&gt;A</td>
<td>c.2155T&gt;C</td>
<td>c.3252C&gt;G</td>
</tr>
<tr>
<td>HRAS</td>
<td>c.692C&gt;T</td>
<td>c.763C&gt;T</td>
<td>c.356C&gt;T</td>
<td>c.763C&gt;T</td>
<td>c.1912G&gt;A</td>
</tr>
<tr>
<td>MAP2K2</td>
<td>c.3914G&gt;C</td>
<td>c.1178C&gt;T</td>
<td>c.2294T&gt;C</td>
<td>c.1178C&gt;T</td>
<td>c.3252C&gt;G</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>c.425C&gt;T</td>
<td>c.1217T&gt;C</td>
<td>c.1315C&gt;T</td>
<td>c.1342T&gt;C</td>
<td>c.2621C&gt;T</td>
</tr>
<tr>
<td>PIK3R1</td>
<td>c.968G&gt;A</td>
<td>c.920C&gt;G</td>
<td>c.1315C&gt;T</td>
<td>c.1342T&gt;C</td>
<td>c.2621C&gt;T</td>
</tr>
<tr>
<td>RAF1</td>
<td>c.3503T&gt;C</td>
<td>c.1217T&gt;C</td>
<td>c.1342T&gt;C</td>
<td>c.2621C&gt;T</td>
<td>c.3252C&gt;G</td>
</tr>
<tr>
<td>SHANK3</td>
<td>c.1080C&gt;T</td>
<td>c.1315C&gt;T</td>
<td>c.1342T&gt;C</td>
<td>c.2621C&gt;T</td>
<td>c.3252C&gt;G</td>
</tr>
<tr>
<td>TSC1</td>
<td>c.829G&gt;A</td>
<td>c.898C&gt;T</td>
<td>c.920C&gt;G</td>
<td>c.1254T&gt;A</td>
<td>c.3252C&gt;G</td>
</tr>
<tr>
<td>TSC2</td>
<td>c.605A&gt;G</td>
<td>c.829G&gt;A</td>
<td>c.920C&gt;G</td>
<td>c.1254T&gt;A</td>
<td>c.3252C&gt;G</td>
</tr>
<tr>
<td>UBE3A</td>
<td>c.1080C&gt;T</td>
<td>c.1315C&gt;T</td>
<td>c.1342T&gt;C</td>
<td>c.2621C&gt;T</td>
<td>c.3252C&gt;G</td>
</tr>
</tbody>
</table>

The fifth HOMER1 variant (c.195G>T, the only HOMER1 variant carried by a female proband) was not detected in an affected sibling, suggesting that this variant may modify autism risk. Interestingly, the c.968G>A variant was present in two affected male children but absent in both parents, suggesting that this variant arose de novo in one of the parental germlines. This finding is consistent with increasing evidence that de novo CNVs and SNVs with high penetrance play major roles in autism [6,34,35]. The remaining four variants were transmitted to affected children by unaffected carriers, possibly reflecting incomplete penetrance of pathogenic variants among parents in families with multiplex autism [34,36].
in a non-affected as well as an affected sibling, indicating reduced penetrance. A potentially damaging missense variant was also detected in \( \text{HRAS} \) (c.383G>A, R128Q); this substitution alters a highly conserved basic residue required for interaction of GTP-bound H-Ras with the plasma membrane and Raf [37]. Familial segregation analysis revealed absence of this variant in an affected sibling of the proband, suggesting a modifying rather than causal role for this variant. Mutations in \( \text{MAP2K2} \) and \( \text{HRAS} \) are responsible for cardiofaciocutaneous (CFC) and Costello syndrome, respectively, related monogenic disorders characterized by mental retardation, facial dysmorphism, cardiac defects and a high prevalence of autistic features [38,39]. CFC and Costello syndromes are thought to be caused by gain-of-function mutations that activate the Ras/ERK pathway, whereas the \( \text{MAP2K2} \) and \( \text{HRAS} \) variants that we identified in autism cases are most compatible with loss of protein function. These findings raise the possibility that rare genetic variation within the Ras/ERK cascade may contribute to non-syndromic autism risk independent of this pathway’s role in CFC and Costello syndromes.

**Discussion**

The findings reported here lend strong support to the hypothesis that perturbed function of the signaling pathways coupling mGluRs to synaptic protein synthesis plays an important role in autism pathogenesis. Overall, we identified rare, potentially damaging variants in 20% (58/290) of the autism cases analyzed but in only 10% (29/300) of controls, suggesting a substantial contribution of rare functional genetic variation within the mGluR pathway to autism susceptibility. A series of recent genome-wide studies has highlighted the association of rare CNVs with autism [5–9,36,40]. Our findings complement these studies by pointing to an important role for rare SNVs in pathophysiologically relevant genes in autism. Consistent with this view, exome sequencing of 20 trios with idiopathic ASD revealed potentially causative de novo
SNVs in several genes known to be associated with autism, intellectual disability and epilepsy [35]. Interestingly, our analysis revealed a significant excess of rare, potentially deleterious variants in three known autism genes, SHANK3, TSC1, and TSC2, in individuals with non-syndromic autism. Whereas mutations in the TSC1 and TSC2 genes typically cause syndromic autism in the context of TSC, our results suggest that these genes contribute to non-syndromic autism risk independent of their causal role in TSC. Mutations in SHANK3 have previously been reported as rare, monogenic cause of non-syndromic autism and syndromic autism (in the context of the 22q13.3 microdeletion syndrome) [23]. A recent study also identified a number of SNVs in the TSC1, TSC2, and SHANK3 genes in simplex autism cases (although the frequency of SNVs in these genes in controls was not reported) [41]. Taken together, these findings further the view that non-syndromic and syndromic forms of autism share common pathophysiological mechanisms.

The identification of HOMER1 as an autism risk gene adds an important component to the network of PSD proteins causally implicated in autism, which now includes the products of the SHANK2 [7,42], SHANK3 [23], MLGNI [43], MLGNI4 [13], SYNGAP1 [7], and DLGAP2 [7] genes (Fig. 1). A pathogenic role for HOMER1 variation in autism is supported by several observations: multiple rare, potentially deleterious variants were identified in the AGRE population but not in the control population; these autism-specific HOMER1 missense variants affect functionally significant protein motifs or regulatory sequences; these autism-specific variants also display tight co-segregation with autism in children of affected families, including in one case a de novo HOMER1 variant present in two affected children. The Homer1 protein interacts with mGluRs and Shank proteins and couples mGluR5 to the mTOR signaling pathway [24,25]. Homer1 thus provides a novel link between the autism-associated gene products that operate downstream of mGluRs and those that interact with Shank proteins, most notably neurexins [12]. In addition, the Homer/Shank interaction scaffolds and regulates the function of group 1 mGluRs, NMDA receptors and AMPA receptors [24], key glutamate receptor subtypes that mediate synaptic plasticity. Intriguingly, recent evidence indicates that disrupted mGluR5-Homer1 interactions in a mouse model of FXS (Fmr1−/− mice) underlies the development of phenotypes relevant to autism, including hippocampal protein synthesis, neocortical circuit dysfunction and behavior [44]. The association of rare functional variation in Homer1 with genetic risk for autism identified in this study provides further insight into the emerging role of synaptic dysfunction in autism pathogenesis [3,11–14].

Our identification of rare damaging variants in the HRAS and MAP2K2 genes further suggests that dysregulation of the Ras/ERK pathway may contribute to autism risk. Consistent with this notion, mutations have been identified in SYNGAP1 in both ASDs and non-syndromic mental retardation [7,45]. SynGAP1, which interacts with the PSD scaffold through PSD-95, negatively regulates Ras/ERK pathway activity, synaptic delivery of AMPA receptors and synaptic transmission [46]. Prior studies have shown that ERK activation is required for synaptic activity-induced protein synthesis, mGluR-dependent synaptic plasticity and cognitive function [47–49].

Collectively, our findings support the view that genetic susceptibility in autism is attributable to the cumulative contribution of individually rare variants in components of the signaling network that couples PSD proteins and downstream effector mechanisms to synaptic function. Furthermore, our results suggest that pharmacological modulation of the signaling mechanisms coupling mGluRs to synaptic protein synthesis may be an effective therapeutic strategy in autism, and that treatments developed for syndromic forms of autism may apply more broadly to non-syndromic autism.

Materials and Methods

Sample selection

Genomic DNA prepared from blood samples was obtained for all unrelated individuals in the AGRE, (Autism Genetic Resource Exchange) collection [20] who satisfied the diagnostic criteria of idiopathic (non-syndromic) autism by the Autism Diagnostic Interview, Revised (ADI-R), and at least one sibling with ASD (n = 290 total). Genomic DNA prepared from blood samples was obtained from the Coriell collection for healthy control individuals (n = 300). Relevant demographic and diagnostic features of the AGRE and control cohorts are summarized in Table S2. The sex ratios in the AGRE and control groups approximate the sex ratios in the autism and neurotypical populations (4:1 and 1:1, respectively). Females are less likely than males to express an ASD phenotype, and pathogenic variants may display decreased penetrance in female carriers [34,36]. Since we considered variants detected in both the control and AGRE populations to be non-pathogenic, over-representation of females in the control group relative to the AGRE group could lead to undercalling of variants as pathogenic, but is unlikely to lead to overcalling. We did not observe any significant differences in the rates of detection of rare, potentially deleterious variants between males and females in either the AGRE or control groups (variant rates: AGRE males 19% (44/232), AGRE females 19% (11/58); control males 8.5% (12/141), control females 7.5% (12/159)). These sex-stratified data also demonstrate significantly higher rates of rare, potentially deleterious variants in both AGRE males and AGRE females relative to their control counterparts. The secondary analysis of existing, de-identified samples and data from the AGRE and Coriell repositories conducted in this study was considered exempt from IRB review.

Next-generation sequencing

DNA concentration was determined for all samples using Nanodrop (Thermo Fisher Scientific). Equal amounts of control sample DNAs were combined into orthogonal pools of 20 samples each and 15 samples each. The same pooling strategy was applied to the AGRE and control groups (n = 300). Relevant demographic and diagnostic features of the AGRE and control cohorts are summarized in Table S2. The sex ratios in the AGRE and control groups approximate the sex ratios in the Autism and neurotypical populations (4:1 and 1:1, respectively). Females are less likely than males to express an ASD phenotype, and pathogenic variants may display decreased penetrance in female carriers [34,36]. Since we considered variants detected in both the control and AGRE populations to be non-pathogenic, over-representation of females in the control group relative to the AGRE group could lead to undercalling of variants as pathogenic, but is unlikely to lead to overcalling. We did not observe any significant differences in the rates of detection of rare, potentially deleterious variants between males and females in either the AGRE or control groups (variant rates: AGRE males 19% (44/232), AGRE females 19% (11/58); control males 8.5% (12/141), control females 7.5% (12/159)). These sex-stratified data also demonstrate significantly higher rates of rare, potentially deleterious variants in both AGRE males and AGRE females relative to their control counterparts. The secondary analysis of existing, de-identified samples and data from the AGRE and Coriell repositories conducted in this study was considered exempt from IRB review.

Next-generation sequencing

DNA concentration was determined for all samples using Nanodrop (Thermo Fisher Scientific). Equal amounts of control sample DNAs were combined into orthogonal pools of 20 samples each and 15 samples each. The same pooling strategy was applied to the AGRE and control groups (n = 300). Relevant demographic and diagnostic features of the AGRE and control cohorts are summarized in Table S2. The sex ratios in the AGRE and control groups approximate the sex ratios in the Autism and neurotypical populations (4:1 and 1:1, respectively). Females are less likely than males to express an ASD phenotype, and pathogenic variants may display decreased penetrance in female carriers [34,36]. Since we considered variants detected in both the control and AGRE populations to be non-pathogenic, over-representation of females in the control group relative to the AGRE group could lead to undercalling of variants as pathogenic, but is unlikely to lead to overcalling. We did not observe any significant differences in the rates of detection of rare, potentially deleterious variants between males and females in either the AGRE or control groups (variant rates: AGRE males 19% (44/232), AGRE females 19% (11/58); control males 8.5% (12/141), control females 7.5% (12/159)). These sex-stratified data also demonstrate significantly higher rates of rare, potentially deleterious variants in both AGRE males and AGRE females relative to their control counterparts. The secondary analysis of existing, de-identified samples and data from the AGRE and Coriell repositories conducted in this study was considered exempt from IRB review.

Next-generation sequencing

DNA concentration was determined for all samples using Nanodrop (Thermo Fisher Scientific). Equal amounts of control sample DNAs were combined into orthogonal pools of 20 samples each and 15 samples each. The same pooling strategy was applied to the AGRE and control groups (n = 300). Relevant demographic and diagnostic features of the AGRE and control cohorts are summarized in Table S2. The sex ratios in the AGRE and control groups approximate the sex ratios in the Autism and neurotypical populations (4:1 and 1:1, respectively). Females are less likely than males to express an ASD phenotype, and pathogenic variants may display decreased penetrance in female carriers [34,36]. Since we considered variants detected in both the control and AGRE populations to be non-pathogenic, over-representation of females in the control group relative to the AGRE group could lead to undercalling of variants as pathogenic, but is unlikely to lead to overcalling. We did not observe any significant differences in the rates of detection of rare, potentially deleterious variants between males and females in either the AGRE or control groups (variant rates: AGRE males 19% (44/232), AGRE females 19% (11/58); control males 8.5% (12/141), control females 7.5% (12/159)). These sex-stratified data also demonstrate significantly higher rates of rare, potentially deleterious variants in both AGRE males and AGRE females relative to their control counterparts. The secondary analysis of existing, de-identified samples and data from the AGRE and Coriell repositories conducted in this study was considered exempt from IRB review.
Analysis pipeline for next-generation sequencing data

Reads were aligned to a reference sequence comprising the hg18-derived sequence of each amplified exon with 30 flanking intronic (non-coding) bases on each side. The aligner MOSAIK was used for the GAII reads (https://code.google.com/p/mosaik-aligner/), and the open source aligner IndexDP for the HeliScope reads (http://open.helicosbio.com). Variant calling was performed with GigaBayses for the GAII reads without invoking the Bayesian-based algorithm (http://bioinformatics.bc.edu/marthlab/GigaBayses), and with SNPsniffer for the HeliScope reads (http://open.helicosbio.com). No minimum minor allele frequency threshold was set in GigaBayses, while in SNPsniffer a minimum allele frequency of 1% was used because variant calls became largely unspecific below that threshold. In both cases, variant calls were only accepted if they occurred at least once on each DNA strand. No other filters were used during the initial variant calling. In the subsequent analysis, GAI1 variant calls were compared to HeliScope variant calls, and all variants called on both the GAII (at a frequency of at least 0.5%) and the HeliScope (at a frequency of at least 1%) were considered confirmed. Confirmed variants occurring in only one pool on each platform could be assigned to individual samples, based on the orthogonal construction of the GAII and HeliScope sample pools. All rare variants discussed in the text were additionally confirmed by Sanger sequencing.

Sanger sequencing

Sanger sequencing was performed for selected exons in selected samples to confirm rare variants detected during NGS. PCR primers and conditions were the same as those used for amplification of exons from sample pools, except that individual samples rather than pooled samples served as template. Each PCR product was then cycle-sequenced using BigDyeTerminator v3.1 reagents (Applied Biosystems, Foster City, CA), with the specific PCR primers serving as sequencing primers, and the sequencing products were separated on an 3730xl Genetic Analyzer (Applied Biosystems). Sequencing traces were visualized using Sequence Scanner software (Applied Biosystems), and presence or absence of a given mutation determined by manual comparison to the reference sequence.

Validation of variant confirmation strategy

To determine sensitivity of variant detection and false discovery rate in pools on the individual sequencing platforms, we constructed validation pools from 20 (GAII) and 15 (HeliScope) samples in which Sanger sequencing had previously been performed on all coding exons of the genes encoding mGluR pathway components that were subjected to next-generation sequencing in a cohort of autism cases and control groups.

Table S2 Cohort demographic and diagnostic characteristics. The demographic and diagnostic features of the AGRE and control cohorts analyzed in this study are summarized.

Supporting Information

Figure S1 Receiver-operating characteristic curves for sensitivity of variant detection and false discovery rate. The receiver-operating characteristic (ROC) curves show the sensitivity of detecting singleton variants in 20-sample pools (GAII) and/or 15-sample pools (HeliScope) as a function of the false discovery rate (FDR). Each point represents sensitivity and FDR at a different threshold value for the minimal allowed allele frequency in the pool. For detection on both platforms, the allele frequency threshold was varied only for GAII data and kept constant at 1% for HeliScope data. When allele frequency cut-offs of 0.5% (GAII) and 1% (HeliScope) were applied to detection on both platforms, a sensitivity of 99% was achieved for singleton detection at high coverage with a false discovery rate of 11%, thereby eliminating most false-positive variants.

Table S3 Rare variants detected in AGRE and/or control groups. All SNVs in 18 mGluR pathway genes that were concordantly detected on both NGS platforms are listed and annotated in terms of their location and consequence, occurrence in AGRE and/or control groups, minor allele frequency (MAF), and functional significance as outlined in Fig. 2B.

Acknowledgments

We thank Dr. Mark Bear for his role in conceiving this study and for valuable advice and guidance in all aspects of the project. We acknowledge colleagues who helped to assemble and winnow the list of genes chosen for study, including Drs. Steve Warren, Aileen Healy, Robert Malenka, Eric Klam, Peter Vanderklish, Dilja Krueger, Emily Osterweil, and Asha Bhakar. We also thank Drs. Patrice Milos, John Thompson, Stan Letovsky, and Eldar Giladi at Helicos Biosciences for their help with generation and analysis of the HeliScope data, as well as technical help with library preparation from Martin Boayteng and Joe Russell at Correlagen. We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium and the participating AGRE families. AGRE is a program of Autism Speaks and is supported, in part, by grant 1U24MH081810 from the National Institute of Mental Health to Clara M. Lajonchere (PI).
Author Contributions
Conceived and designed the experiments: RJK UG RC DM. Performed the experiments: UG HH RP BR. Analyzed the data: RJK UG ET DM. Wrote the paper: RJK UG DM.

References