# COMPOSITIONS AND METHODS FOR DIAGNOSING AUTISM SPECTRUM DISORDERS

## Related Applications

The present invention claims priority under 35 USC 119(e) to U.S. Provisional Patent Application No. 61/240,469, filed September 8, 2009. The disclosure of U.S. Provisional Patent Application No. 61/240,469 is incorporated by reference in its entirety herein.

## Field of the Invention

The invention generally relates to compositions and methods for diagnosing autism spectrum disorders.

## Background

Autism is a complex developmental disability that interferes with normal development of the brain in the areas of social interaction and communication skills. Typically, autistic children and adults have difficulties with verbal and non-verbal communication, social interactions, and leisure or play activities.

Autism generally is characterized as one of five disorders coming under the umbrella of Pervasive Developmental Disorders (PDD), a category of neurological disorders characterized by severe and pervasive impairment in several areas of development, including social interaction and communications skills. The five disorders under PDD include Autistic Disorder, Asperger's Disorder, Childhood Disintegrative Disorder (CDD), Rett's Disorder, and PDD-Not Otherwise Specified (PDD-NOS). Specific diagnostic criteria for each of these disorders can be found in the American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders,* Fourth Edition, Text Revision. Washington, DC, American Psychiatric Association, 2000, as distributed by the American Psychiatric Association.

There is no definitive diagnostic test for biological manifestations of autism, and thus it remains one of the only neurological disorders that must be diagnosed almost entirely through behavioral symptoms. The DSM-IV classifies autism as a Pervasive Developmental Disorder (PDD) characterized by twelve diagnostic criteria. Those criteria fall into three categories: impairments in social interaction; impairments in communication; and a restricted repertoire of activities and interests. A diagnosis of autism requires that a child display at least six of the twelve symptoms.

If a child does not fit the definition of autism given above, he/she may be diagnosed with a condition called Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS). Such a diagnosis of non-specific forms of Pervasive Developmental Disorder (PDD)

may include atypical types of autism that do not fall into the above categories because of late age of onset, for example, or sub-threshold or atypical symptoms. According to the DSM-IV, this diagnosis is to be used when autistic-like behaviors are present, in particular when there is severe impairment in the development of social and verbal communication skills, but the child does not meet the criteria for classic autism or any other specific Pervasive Developmental Disorder, Schizophrenia, Schizotypal Personality Disorder or Avoidant Personality Disorder.

A variety of agents have been postulated to be associated with the development of autism including, but not limited to, exposure to pesticides and/or agents that can cause birth defects. In at least some cases, it appears that autism may have a genetic basis. The genetics of autism appear to be complex. For example, copy number variation and chromosomal structural abnormalities (both large and small) have been shown be present in particular genomic regions in patients with autism or syndromes in which autistic behavior is common (Abrahams and Geschwind, Nature Reviews Genetics, 2008, 9:341-355). DNA hybridization studies have shown structural abnormalities in autistic populations. A causal role for genetic variation in many different genes has been suggested based on evidence from association or linkage studies. Still, genome wide association studies have failed to link specific common variants, acting singly or in combination, though such studies have identified association peaks that may point to other causative genes or pathways. There is some evidence that genetic variation may be the cause of at least non-syndromic autism.

Evaluations to diagnose a child are made by a team typically including doctors and the child's parents. Because diagnosis of autism spectrum disorders is subjective, misdiagnosis of a child can frequently occur. Thus, there is an unmet need for diagnostic tests that can provide an objective determination of whether a subject suffers from an autism spectrum disorder.

## Summary

The invention generally relates to compositions and methods for diagnosing the presence or an increased risk of developing autism spectrum disorders. The methods and compositions of the present invention may be used to obtain or provide genetic information from a subject in order to objectively diagnose the presence of an autism spectrum disorder (ASD), or an increased risk for that subject, or other subjects, to develop an autism spectrum disorder.

In one embodiment, the invention comprises methods for diagnosing the presence or an increased risk of developing an autism spectrum disorder in a subject. The method may

comprise the steps of obtaining a nucleic acid from a biological sample (e.g., a tissue or body fluid sample) from a subject and conducting an assay to identify whether there is a variant sequence in the subject's nucleic acid. In certain embodiments, the method may comprise comparing the variant to known variants associated with an autism spectrum disorder and determining whether the variant is a variant that has been previously identified as being associated with autism. Or, the method may comprise identifying the variant as a new, previously uncharacterized or previously not described variant. If the variant is a new variant, the method may further comprise performing an analysis to determine whether the mutation is expected to be deleterious to expression of the gene and/or the function of the protein encoded by the gene. The method may further comprise using the variant profile (i.e., the compilation of mutations identified in the subject) to diagnose the presence of an autism spectrum disorder or an increased risk of developing an autism spectrum disorder. In some embodiments, the method may comprise obtaining a nucleic acid from a tissue or body fluid sample from a subject and sequencing at least a portion of a nucleic acid in order to obtain a sample nucleic acid sequence for at least one gene.

Yet other embodiments of the invention may comprise methods for identifying mutations (i.e., variants) correlated with the presence or increased risk of developing an autism spectrum disorder. The method may comprise the step of identifying a nucleic acid to be evaluated as having a sequence that if mutated may be associated with the development of autism. Also, the method may comprise obtaining a nucleic acid sample from a biological sample (e.g., a tissue or body fluid sample) from a subject having an autism spectrum disorder; and conducting an assay to identify whether there is a mutation in the nucleic acid sequence in the subject having autism as compared to the nucleic acid sequence in individuals who do not have an autism spectrum disorder, wherein the presence of the mutation in a subject with an autism spectrum disorder indicates that the mutation may be associated with the development of the autism spectrum disorder. If the variant is a new variant, the method may further comprise performing an analysis to determine whether the mutation is expected to be deleterious to expression of the gene and/or the function of the protein encoded by the gene. The method may further comprise compiling a panel of variant mutations that can be used to diagnose the presence of an autism spectrum disorder or an increased risk of developing an autism spectrum disorder.

In yet other embodiments, the invention comprises an isolated nucleic acid comprising a nucleic acid of at least one of the following genes or genomic regions: *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1,*

*MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* or *UBE3A*, wherein the sequence comprises a variant that is indicative of or associated with an autism spectrum disorder.

There are additional features of the invention which will be described hereinafter. It is to be understood that the invention is not limited in its application to the details set forth in the following claims, description and figures. The invention is capable of other embodiments and of being practiced or carried out in various ways.

## Brief Description of the Figures

Various features, aspects and advantages of the present invention will become more apparent with reference to the following figures.

**FIG. 1** shows genes that are involved in mGluR signaling in accordance with an embodiment of the present invention.

**FIG. 2** shows a method for variant classification in accordance with an embodiment of the present invention.

**FIG. 3, Panels A-LL,** depicts the DNA sequences for the *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* and *UBE3A* genes and protein sequences encoded by these genes as SEQ ID NO: 1-38.

**FIG. 4** depicts exon and flanking sequences used as described in the examples to identify mutations in the *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* and *UBE3A* genes as SEQ ID NOs: 39-271 as well as the chromosomal location of the sequences.

## Detailed Description

Notwithstanding that the numerical ranges and parameters setting forth the broad scope of the invention are approximations, the numerical values set forth in the specific examples are reported as precisely as possible. Any numerical value, however, inherently contains certain errors necessarily resulting from the standard deviation found in their respective testing measurements. Moreover, all ranges disclosed herein are to be understood to encompass any and all subranges subsumed therein. For example, a stated range of "1 to 10" should be considered to include any and all subranges between (and inclusive of) the minimum value of 1 and the maximum value of 10; that is, all subranges beginning with a minimum value of 1 or more, e.g. 1 to 6.1, and ending with a maximum value of 10 or less, e.g., 5.5 to 10. Additionally, any reference referred to as being "incorporated herein" is to be understood as being incorporated in its entirety.

It is further noted that, as used in this specification, the singular forms "a," "an," and "the" include plural referents unless expressly and unequivocally limited to one referent. The term "and/or" generally is used to refer to at least one or the other. In some case the term "and/or" is used interchangeably with the term "or".

Also, the terms "portion" and "fragment" are used interchangeably to refer to parts of a polypeptide, nucleic acid, or other molecular construct.

"Polypeptide" and "protein" are used interchangeably herein to describe protein molecules that may comprise either partial or full-length proteins. The term "peptide" is used to denote a less than full-length protein or a very short protein unless the context indicates otherwise.

As is known in the art, "proteins", "peptides," "polypeptides" and "oligopeptides" are chains of amino acids (typically L-amino acids) whose alpha carbons are linked through peptide bonds formed by a condensation reaction between the carboxyl group of the alpha carbon of one amino acid and the amino group of the alpha carbon of another amino acid. Typically, the amino acids making up a protein are numbered in order, starting at the amino terminal residue and increasing in the direction toward the carboxy terminal residue of the protein.

As is known in the art, conditions for hybridizing nucleic acid sequences to each other can be described as ranging from low to high stringency. Generally, highly stringent hybridization conditions refer to washing hybrids in low salt buffer at high temperatures. Hybridization may be to filter bound DNA using hybridization solutions standard in the art such as 0.5M $NaHPO_4$, 7% sodium dodecyl sulfate (SDS), at 65°C, and washing in 0.25 M $NaHPO_4$, 3.5% SDS followed by washing 0.1 x SSC/0.1% SDS at a temperature ranging from room temperature to 68°C depending on the length of the probe (*see e.g.* Ausubel, F.M. *et al., Short Protocols in Molecular Biology*, 4th Ed., Chapter 2, John Wiley & Sons, N.Y). For example, a high stringency wash comprises washing in 6x SSC/0.05% sodium pyrophosphate at 37°C for a 14 base oligonucleotide probe, or at 48°C for a 17 base oligonucleotide probe, or at 55°C for a 20 base oligonucleotide probe, or at 60°C for a 25 base oligonucleotide probe, or at 65°C for a nucleotide probe about 250 nucleotides in length. Nucleic acid probes may be labeled with radionucleotides by end-labeling with, for example, [$\gamma$-$^{32}$P]ATP, or incorporation of radiolabeled nucleotides such as [$\alpha$-$^{32}$P]dCTP by random primer labeling. Alternatively, probes may be labeled by incorporation of biotinylated or fluorescein labeled nucleotides, and the probe detected using Streptavidin or anti-fluorescein antibodies.

As used herein, the term "upstream" refers to a residue that is N-terminal to a second residue where the molecule is a protein, or 5' to a second residue where the molecule is a nucleic acid. Also as used herein, the term "downstream" refers to a residue that is C-terminal to a second residue where the molecule is a protein, or 3' to a second residue where the molecule is a nucleic acid. The protein, polypeptide and peptide sequences disclosed herein are all listed from N-terminal amino acid to C-terminal acid and the nucleic acid sequences disclosed herein are all listed from the 5' end of the molecule to the 3' end of the molecule.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art. Practitioners are particularly directed to Current Protocols in Molecular Biology (Ausubel) for definitions and terms of the art. Abbreviations for amino acid residues are the standard 3-letter and/or 1-letter codes used in the art to refer to one of the 20 common L-amino acids.

A "nucleic acid" is a polynucleotide such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). The term is used to include single-stranded nucleic acids, double-stranded nucleic acids, and RNA and DNA made from nucleotide or nucleoside analogues.

The terms "identity" or "percent identical" refers to sequence identity between two amino acid sequences or between two nucleic acid sequences. Percent identity can be determined by aligning two sequences and refers to the number of identical residues (*i.e.*, amino acid or nucleotide) at positions shared by the compared sequences. Sequence alignment and comparison may be conducted using the algorithms standard in the art (*e.g.* Smith and Waterman, 1981, *Adv. Appl. Math.* 2:482; Needleman and Wunsch, 1970, *J. Mol. Biol.* 48:443; Pearson and Lipman, 1988, *Proc. Natl. Acad. Sci., USA*, 85:2444) or by computerized versions of these algorithms (Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Drive, Madison, WI) publicly available as BLAST and FASTA. Also, ENTREZ, available through the National Institutes of Health, Bethesda MD, may be used for sequence comparison. In other cases, commercially available software, such as GenomeQuest, may be used to determine percent identity. When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (e.g., BLASTN; available at the Internet site for the National Center for Biotechnology Information) may be used. In one embodiment, the percent identity of two sequences may be determined using GCG with a gap weight of 1, such that each amino acid gap is weighted as if it were a single amino acid mismatch between the two sequences. Or, the ALIGN program

(version 2.0), which is part of the GCG (Accelrys, San Diego, CA) sequence alignment software package may be used.

As used herein, the term "conserved residues" refers to amino acids that are the same among a plurality of proteins having the same structure and/or function. A region of conserved residues may be important for protein structure or function. Thus, contiguous conserved residues as identified in a three-dimensional protein may be important for protein structure or function. To find conserved residues, or conserved regions of 3-D structure, a comparison of sequences for the same or similar proteins from different species, or of individuals of the same species, may be made.

As used herein, the term "similar" or "homologue" when referring to amino acid or nucleotide sequences means a polypeptide having a degree of homology or identity with the wild-type amino acid sequence. Homology comparisons can be conducted by eye, or more usually, with the aid of readily available sequence comparison programs. These commercially available computer programs can calculate percent homology between two or more sequences (e.g. Wilbur, W. J. and Lipman, D. J., 1983, Proc. Natl. Acad. Sci. USA, 80:726-730). For example, homologous sequences may be taken to include an amino acid sequences which in alternate embodiments are at least 70% identical, 75% identical, 80% identical, 85% identical, 90% identical, 95% identical, 97% identical, or 98% identical to each other.

As used herein, the term at least 90% identical thereto includes sequences that range from 90 to 100% identity to the indicated sequences and includes all ranges in between. Thus, the term at least 90% identical thereto includes sequences that are 91, 91.5, 92, 92.5, 93, 93.5. 94, 94.5, 95, 95.5, 96, 96.5, 97, 97.5, 98, 98.5, 99, 99.5 percent identical to the indicated sequence. Similarly the term "at least 70% identical includes sequences that range from 70 to 100% identical, with all ranges in between. The determination of percent identity is determined using the algorithms described herein.

As used herein, a polypeptide or protein "domain" comprises a region along a polypeptide or protein that comprises an independent unit. Domains may be defined in terms of structure, sequence and/or biological activity. In one embodiment, a polypeptide domain may comprise a region of a protein that folds in a manner that is substantially independent from the rest of the protein. Domains may be identified using domain databases such as, but not limited to PFAM, PRODOM, PROSITE, BLOCKS, PRINTS, SBASE, ISREC PROFILES, SAMRT, and PROCLASS.

As used herein a gene is a unit of heredity. Generally, a gene is a portion of DNA that encodes a protein or a functional RNA. A modern working definition of a gene is is a

locatable region of genomic sequence corresponding to a unit of inheritance. A gene may associated with regulatory regions, transcribed regions, and or other functional sequence regions.

As used herein a gene regulatory element or regulatory sequence is a segment of DNA where regulatory proteins, such as transcription factors, bind to regulate gene expression. Such regulatory regions are often upstream of the gene being regulated.

As used herein an exon is a nucleic acid sequence that is found in mature or processed RNA after other portions of the RNA (e.g., intervening regions known as introns) have been removed by RNA splicing. As such, exon sequences generally encode for proteins or portions of proteins. An intron is the portion of the RNA that is removed from surrounding exon sequences by RNA splicing.

As used herein expressed RNA is an RNA that encodes for a protein or polypeptide ("coding RNA"), and any other RNA that is transcribed but not translated ("non-coding RNA").

As used herein micro RNA is microRNAs (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression. microRNA can affect both the stability and translation of mRNAs. For example, microRNAs can bind to complementary sequences in the 3'UTR of target mRNAs and cause gene silencing. miRNAs are transcribed by RNA polymerase II as part of capped and polyadenylated primary transcripts (pri-miRNAs) that can be either protein-coding or non-coding. The primary transcript can be cleaved by the Drosha ribonuclease III enzyme to produce an approximately 70-nt stem-loop precursor miRNA (pre-miRNA), which can further be cleaved by the cytoplasmic Dicer ribonuclease to generate the mature miRNA and antisense miRNA star (miRNA*) products. The mature miRNA can be incorporated into a RNA-induced silencing complex (RISC), which can recognize target mRNAs through imperfect base pairing with the miRNA and most commonly results in translational inhibition or destabilization of the target mRNA.

As used herein, siRNA is essentially a double-stranded RNA molecule composed of about 20 complementary nucleotides. siRNA is created by the breakdown of larger double-stranded (ds) RNA molecules. siRNA can suppress gene expression by inherently splitting its corresponding mRNA in two by way of the interaction of the siRNA with the mRNA, leading to degradation of the mRNA. siRNAs can also interact with DNA to facilitate chromating silencing and the expansion of heterochromatin.

As used herein, an epigenetic element can change gene expression by a mechanism

other than a change in the underlying DNA sequences. Such elements may include elements that regulate paramutation, imprinting, gene silencing, X chromosome inactivation, position effect, reprogramming, transvection, maternal effects, histone modification, and heterochromatin.

As used herein, the terms mutation and variant are used interchangeably to describe a nucleic acid or protein sequence change.

As used herein, "associated with an autism spectrum disorder" means that the variant is found with in patients with autism more than in non-autistic controls. Generally, the statistical significance of such association can be determined by assaying a plurality of patients.

As used herein, a region of interest is a portion of the chromosome that is being targeted for assaying for variants in the DNA sequence.

Methods and compositions for diagnosing autism spectrum disorders

Embodiments of the present invention comprise compositions and methods for diagnosing presence or increased risk of developing autism spectrum disorders. The methods and compositions of the present invention may be used to obtain or provide genetic information from a subject in order to objectively diagnose the presence or increased risk for that subject, or other subjects to develop an autism spectrum disorder.

In one embodiment, the invention comprises methods for diagnosing the presence or an increased risk of developing an autism spectrum disorder in a subject. The method may comprise the steps of obtaining a nucleic acid from a tissue or body fluid sample from a subject and conducting an assay to identify whether there is a variant sequence (i.e., a mutation) in the subject's nucleic acid. In certain embodiments, the method may comprise comparing the variant to known variants associated with an autism spectrum disorder and determining whether the variant is a variant that has been previously identified as being associated with autism. Or, the method may comprise identifying the variant as a new, previously uncharacterized variant. If the variant is a new variant, the method may further comprise performing an analysis to determine whether the mutation is expected to be deleterious to expression of the gene and/or the function of the protein encoded by the gene. The method may further comprise using the variant profile (i.e., the compilation of mutations identified in the subject) to diagnose the presence of an autism spectrum disorder or an increased risk of developing an autism spectrum disorder.

In certain embodiments, the invention comprises a method for diagnosing the presence or an increased risk of developing an autism spectrum disorder in a subject, the

method comprising: obtaining a nucleic acid from a tissue or body fluid sample from a subject; conducting an assay to identify whether there is a variant sequence, or a plurality of variant sequences, in the subject's nucleic acid; for each variant detected, determining if the variant is a known variant associated with an autism spectrum disorder or a previously undescribed variant; if the variant is a previously undescribed variant, determining if the variant is expected to have a deleterious effect on at least one of gene expression and/or protein function; and diagnosing the presence or an increased risk of developing the autism spectrum disorder based on the variant sequence or the plurality of variant sequences detected.

In some embodiments, the method may comprise obtaining a nucleic acid from a tissue or body fluid sample from a subject and sequencing at least a portion of a nucleic acid in order to obtain a sample nucleic acid sequence for at least one gene. In certain embodiments, the method may comprise comparing the variant to known variants associated with an autism spectrum disorder and determining whether the variant is a variant that has been previously identified as being associated with autism. Or, the method may comprise identifying the variant as a new, previously uncharacterized variant. If the variant is a new variant, or in some cases for previously characterized (i.e., identified) variants, the method may further comprise performing an analysis to determine whether the mutation is expected to be deleterious to expression of the gene and/or the function of the protein encoded by the gene. The method may further comprise using the variant profile (i.e., a compilation of variants identified in the subject) to diagnose the presence of an autism spectrum disorder or an increased risk of developing an autism spectrum disorder.

In embodiments of each of the methods of the invention, the method may comprise performing the assay (e.g., sequencing) in a plurality of individuals to determine the statistical significance of the association.

In various embodiments of the methods of the invention and as described in more detail herein, the assay comprises at least one of nucleic acid sequencing, hybrid capture, and/or epigenetic analysis. For example, in certain embodiments, next generation (massively-parallel sequencing) may be used. Or, Sanger sequencing may be used. Or, a combination of next generation (massively-parallel sequencing) and Sanger sequencing may be used. Additionally and/or alternatively, the sequencing comprises at least one of single-molecule sequencing-by-synthesis. Thus, in certain embodiments, a plurality of DNA samples are analyzed in a pool to identify samples that show a variation. Additionally or

alternatively, in certain embodiments, a plurality of DNA samples are analyzed in a plurality of pools to identify an individual sample that shows the same variation in at least two pools.

Also, in various embodiments, the nucleic acid in the conducting step comprises a gene, an RNA, an exon, an intron, a gene regulatory element, an expressed RNA, an siRNA, or an epigenetic element. Also, regulatory elements, including splice sites, transcription factor binding, A-I editing sites, microRNA binding sites, and functional RNA structure sites may be evaluated for mutations (i.e., variants).

In certain embodiments, the nucleic acid selected for analyzing for a variant comprises a sequence selected from a sequence known or suspected to be associated with one or more autism spectrum disorders. For, example, the nucleic acid comprises at least a portion of one of the genes in Table 1. Or, the nucleic acid may comprise a gene that encodes for a protein involved in a biochemical pathway that can be important in the develomement of an austism spectrum disorder (ASD). For example, in certain embodiments, the nucleic acid is derived from a gene that encodes a protein in the metabotropic glutamate receptor signaling pathway. For example, in certain embodiments, the variant comprises at least one of the variants in Table 2. Thus, in certain embodiments of the methods of the invention, the nucleic acid comprises at least a portion of a gene for at least one of *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* or *UBE3A*. In some embodiments, the nucleic acid comprises at least a portion of a gene for at least one of *TSC1, TSC2, SHANK3,* or *HOMER1*. In certain embodiments, the variant comprises at least one of the following mutations: *HOMER 1* c.195G>T, M65I; *HOMER 1* c.290C>T, S97L; *HOMER 1* c.425C>T, P142L; *GRM5* c.3503T>C, L1168P; *MAPK2* c.581-1G>T; *HRAS* c.383G>A, R128Q; a *MECP2* c.1477G>T, E483X.

In the various embodiments of the methods of the invention, the autism spectrum disorder may be at least one of non-syndromic autism, classical autism, Asperger's syndrome, Rett's syndrome, childhood disintegrative disorder, or pervasive developmental disorder not otherwise specified (PDD-NOS). In certain embodiments, the autism spectrum disorder comprises non-syndromic autism (i.e., patients who display symptoms of autism but who do not exhibit physical manifestations often found with autism).

The methods of the invention may further comprise diagnosing a the presence of, or an increased risk of developing, a genetic syndrome linked to autism, wherein the genetic syndrome comprises a manifesting phenotype. For example, in certain embodiments, the genetic syndrome comprises at least one of Angelman syndrome, Prader-Willi syndrome,

15q11-q13 duplication, fragile X syndrome, fragile X premutation, deletion of chromosome 2q, XYY syndrome, Smith-Lemli-Opitz syndrome, Apert syndrome, mutations in the ARX gene, De Lange syndrome, Smith-Magenis syndrome, Williams syndrome, Noonan syndrome, Down syndrome, velo-cardio-facial syndrome, myotonic dystrophy, Steinert disease, tuberous sclerosis, Duchenne's disease, Timothy syndrome, 10p terminal deletion, Cowden syndrome, 45,X/46,XY mosaicism, Myhre syndrome, Sotos syndrome, Cohen syndrome, Goldenhar syndrome, Joubert syndrome, Lujan-Fryns syndrome, Moebius syndrome, hypomelanosis of Ito, neurofibromatosis type 1, CHARGE syndrome, and/or HEADD syndrome.

The methods may be used to assist in the diagnosis of individuals who do not yet display symptoms of an ASD, or for whom, the diagnosis is equivocal. For example, the subject may be a child or a fetus.

The techniques for sequencing nucleic acids (both DNA and RNA) are highly sensitive and therefore, can be used almost any biological sample (i.e., tissue or body fluid) taken from subject. For example, in alternate embodiments, the body fluid comprises at least one of cerebrospinal fluid, blood, amniotic fluid, maternal blood, or urine.

As noted above, in certain embodiments, the genes for which mutations are evaluated are genes that encode proteins in biochemical pathway or pathways that are relevant to the development of autism. For example, in certain embodiments, the genes are involved in the metabotropic glutamate receptor pathway. In one embodiment, the pathway is the mGluR5 signaling pathway and/or includes genes important to the activity of the mGluR5 receptor. Or, other biochemical pathways that are related to certain types of autism syndromes may be evaluated. For example, in certain embodiments, at least one of the genes and/or genomic regions in Table 1 may be evaluated.

Where the pathway is the mGluR5 signaling pathway and/or includes genes important to the activity of the mGluR5 receptor, the DNA sequences may be derived from genes or genomic regions comprising the genes shown in Table 2. In certain embodiments of the methods, the genes and/or genomic regions being evaluated for mutations that may be indicative of the presence or an increase risk of an ASD are *ARC, EIF4E, FMR1, GRM1, GRM5, HOMER1, HRAS, MAP2K1, MAP2K2, MECP2, PIK3CA, PIK3R1, PTEN, RAF1, RHEB, SHANK3, TSC1, TSC2,* and/or *UBE3A*. In certain embodiments, the native or non-variant sequence used in the assay comprises an exon sequence from at least one of the following genes: *ARC, EIF4E, FMR1, GRM1, GRM5, HOMER1, HRAS, MAP2K1, MAP2K2, MECP2, PIK3CA, PIK3R1, PTEN, RAF1, RHEB, SHANK3, TSC1, TSC2,* and/or *UBE3A*.

For example, in certain embodiments, the gene sequence being evaluated for a variant comprises the exon sequences. Or, intron sequences or other non-coding regions may be evaluated for potentially deleterious mutations. In certain embodiments, the exon sequence and additional flanking sequence (e.g., about 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 or more nucleotides of UTR and/or intron sequence) is analyzed in the assay. Or portions of these sequences may be used. In certain embodiments, the gene sequence being evaluated comprises an exon sequence and/or flanking intron or UTR sequence from at least one of the following genes: *HOMER1, SHANK3, TSC1,* and/or *TSC2*. In certain embodiments, the gene sequence being evaluated comprises an exon sequence from the *HOMER1* gene. Such variant gene sequences may include sequences having at least one of the mutations as shown in Table 2.

Yet other embodiments of the invention may comprise methods for identifying mutations correlated with the presence or increased risk of developing an autism spectrum disorder. The method may comprise the step of identifying a nucleic acid sequence, such as a gene or a genomic region, that if mutated may be associated with the development of autism. Also, the method may comprise obtaining a nucleic acid sample from a tissue or body fluid sample from a subject having an autism spectrum disorder; and conducting an assay to identify whether there is a mutation in the nucleic acid sequence in the subject having autism as compared to the nucleic acid sequence in individuals who do not have an autism spectrum disorder, wherein the presence of the mutation in a subject with an autism spectrum disorder indicates that the mutation may be associated with the development of the autism spectrum disorder. Or, the method may comprise analyzing the sequence of the selected gene or genomic region for new variants (i.e., previously undiscovered mutations). If the variant is a new variant, or in some cases for a previously identified variant, the method may further comprise performing an analysis to determine whether the mutation is expected to be deleterious to expression of the gene and/or the function of the protein encoded by the gene. The method may further comprise compiling a panel of variant mutations that can be used to diagnose the presence of an autism spectrum disorder or an increased risk of developing an autism spectrum disorder.

Thus, the method may comprise a method for identifying mutations correlated with the presence or increased risk of developing an autism spectrum disorder, comprising: identifying a nucleic acid to be evaluated as having a sequence that if mutated may be or is associated with the development of autism; obtaining a nucleic acid sample from a tissue or body fluid sample from a subject having an autism spectrum disorder; and conducting an

assay to identify whether there is a mutation in the nucleic acid sequence in the subject having autism as compared to the nucleic acid sequence in individuals who do not have an autism spectrum disorder, wherein the presence of the mutation in a subject with an autism spectrum disorder indicates that the mutation may be associated with the development of the autism spectrum disorder.

In embodiments the methods of the invention for identifying new mutations, the method may comprise performing the assay (e.g., sequencing) in a plurality of individuals to determine the statistical significance of the association.

In certain embodiments, the mutation is a variant that has been previously associated with the development of an autism spectrum disorder. Or, the the mutation may be a previously undescribed variant. The method may additionally comprise determining if the mutation is expected to have a deleterious effect on at least one of gene expression and/or protein function.

In certain embodiments, the nucleic acid selected for analyzing for a variant comprises a sequence selected from a sequence known or suspected to be associated with one or more autism spectrum disorders. For, example, the nucleic acid comprises at least a portion of one of the genes in Table 1. Or, the nucleic acid may comprise a gene that encodes for a protein involved in a biochemical pathway that can be important in the develomement of an austism spectrum disorder (ASD). For example, in certain embodiments, the nucleic acid is derived from a gene that encodes a protein in the metabotropic glutamate receptor signaling pathway. For example, in certain embodiments, the variant comprises at least one of the variants in Table 2. Thus, in certain embodiments of the methods of the invention, the nucleic acid comprises at least a portion of a gene for at least one of *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* or *UBE3A*. In some embodiments, the nucleic acid comprises at least a portion of a gene for at least one of *TSC1, TSC2, SHANK3,* or *HOMER1*.

In the various embodiments of the methods of the invention, the autism spectrum disorder may be at least one of non-syndromic autism, classical autism, Asperger's syndrome, Rett's syndrome, childhood disintegrative disorder, or pervasive developmental disorder not otherwise specified (PDD-NOS). In certain embodiments, the autism spectrum disorder comprises non-syndromic autism.

Or, the association of variants with other syndromes that are associated (e.g., genetically linked to) with autism, such as at least one of Angelman syndrome, Prader-Willi syndrome, 15q11-q13 duplication, fragile X syndrome, fragile X premutation, deletion of

chromosome 2q, XYY syndrome, Smith-Lemli-Opitz syndrome, Apert syndrome, mutations in the ARX gene, De Lange syndrome, Smith-Magenis syndrome, Williams syndrome, Noonan syndrome, Down syndrome, velo-cardio-facial syndrome, myotonic dystrophy, Steinert disease, tuberous sclerosis, Duchenne's disease, Timothy syndrome, 10p terminal deletion, Cowden syndrome, 45,X/46,XY mosaicism, Myhre syndrome, Sotos syndrome, Cohen syndrome, Goldenhar syndrome, Joubert syndrome, Lujan-Fryns syndrome, Moebius syndrome, hypomelanosis of Ito, neurofibromatosis type 1, CHARGE syndrome, and/or HEADD syndrome.

In various embodiments, of the methods of the invention and as described in more detail herein, the assay comprises at least one of nucleic acid sequencing, hybrid capture, and epigenetic analysis. For example, in certain embodiments, next generation (massively-parallel sequencing) may be used. Or, Sanger sequencing may be used. Or, a combination of next generation (massively-parallel sequencing) and Sanger sequencing may be used. Additionally and/or alternatively, the sequencing comprises at least one of single-molecule sequencing-by-synthesis. Thus, in certain embodiments, a plurality of DNA samples are analyzed in a pool to identify samples that show a variation. Additionally or alternatively, in certain embodiments, a plurality of DNA samples are analyzed in a plurality of pools to identify an individual sample that shows the same variation in at least two pools.

Also, in various embodiments, the nucleic acid in the conducting step comprises a gene, an RNA, an exon, an intron, a gene regulatory element, an expressed RNA, an siRNA, or an epigenetic element. Also, regulatory elements, including splice sites, transcription factor binding, A-I editing sites, microRNA binding sites, and functional RNA structure sites may be evaluated for mutations (i.e., variants).

The methods may be used to assist in the diagnosis of individuals who do not yet display symptoms of an ASD, or for whom, the diagnosis is equivocal. For example, the subject may be a child or a fetus.

The techniques for sequencing nucleic acids (both DNA and RNA) are highly sensitive and therefore, can be used almost any biological sample (i.e., tissue or body fluid) taken from subject. For example, in alternate embodiments, the body fluid comprises at least one of cerebrospinal fluid, blood, amniotic fluid, maternal blood, or urine.

Again, in certain embodiments, the genes for which new mutations are evaluated are genes that encode proteins in biochemical pathway or pathways that are relevant to the development of autism. For example, in certain embodiments, the genes are involved in the metabotropic glutamate receptor pathway. In one embodiment, the pathway is the mGluR5

signaling pathway and/or includes genes important to the activity of the mGluR5 receptor. Or, other biochemical pathways that are related to certain types of autism syndromes may be evaluated. For example, in certain embodiments, at least one of the genes and/or genomic regions in Table 1 may be evaluated.

Where the pathway is the mGluR5 signaling pathway and/or includes genes important to the activity of the mGluR5 receptor the DNA sequences may be derived from genes or genomic regions comprising the genes shown in Table 2. In certain embodiments of the methods the genes and/or genomic regions being evaluated for new mutations that may be indicative of the presence or an increase risk of an ASD are *ARC, EIF4E, FMR1, GRM1, GRM5, HOMER1, HRAS, MAP2K1, MAP2K2, MECP2, PIK3CA, PIK3R1, PTEN, RAF1, RHEB, SHANK3, TSC1, TSC2,* and/or *UBE3A*. In certain embodiments, the native or non-variant sequence comprises an exon sequence from at least one of the following genes: *ARC, EIF4E, FMR1, GRM1, GRM5, HOMER1, HRAS, MAP2K1, MAP2K2, MECP2, PIK3CA, PIK3R1, PTEN, RAF1, RHEB, SHANK3, TSC1, TSC2,* and/or *UBE3A*. For example, in certain embodiments, the gene sequence being evaluated for a variant comprises the exon sequences. In certain embodiments, the exon sequence and additional flanking sequence (e.g., about 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 or more nucleotides of UTR and/or intron sequence) is analyzed in the assay. Or, intron sequences or other non-coding regions may be evaluated for potentially deleterious mutations. Or, portions of these sequences may be used. Such variant gene sequences may include sequences having at least one of the mutations as shown in Table 2.

Other embodiments of the invention provide isolated gene sequences containing mutations that relate to autism spectrum disorders. Such gene sequences may be used to objectively diagnose the presence or increased risk for a subject to develop an autism spectrum disorder. In certain embodiments, the isolated nucleic acid may contain a non-variant sequence or a variant sequence of any one or combination of *ARC, EIF4E, FMR1, GRM1, GRM5, HOMER1, HRAS, MAP2K1, MAP2K2, MECP2, PIK3CA, PIK3R1, PTEN, RAF1, RHEB, SHANK3, TSC1, TSC2,* and/or *UBE3A*. For example, in certain embodiments, the gene sequence comprises the exon sequences. In certain embodiments, the exon sequence and additional flanking sequence (e.g., about 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 or more nucleotides of UTR and/or intron sequence) is analyzed in the assay. Or, intron sequences or other non-coding regions may be used. Or, portions of these sequences may be used. In certain embodiments, the gene sequence comprises an exon sequence from at least one of the following genes: *HOMER1, SHANK3, TSC1,* and/or *TSC2*. In certain embodiments, the gene

sequence comprises an exon sequence from the *HOMER1* gene. Such variant gene sequences include sequences having at least one of the mutations as shown in Table 2. In an embodiment, the isolated nucleic acid may comprise at least one of the following variants: *HOMER 1* c.195G>T, M65I; *HOMER 1* c.290C>T, S97L; *HOMER 1* c.425C>T, P142L; *GRM5* c.3503T>C, L1168P; *MAPK2* c.581-1G>T; *HRAS* c.383G>A, R128Q; a *MECP2* c.1477G>T, E483X.

Autism spectrum disorders are generally characterized as one of five disorders coming under the umbrella of Pervasive Developmental Disorders (PDD). The five disorders under PDD include autism (classical autism), Asperger's Syndrome, Rett's Syndrome, childhood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS). According to the invention, one may analyze a panel of genes known or suspected to be associated with one of the five disorders and/or an autism spectrum disorder. In certain embodiments, the autism is non-syndromic autism. Or, the presence or increased risk of developing other types of autism spectrum disorders may be characterized.

The methods and compositions of the invention may further be used for diagnosing or predicting increased risk of developing a genetic syndrome linked to autism, thereby determining whether the subject is affected with, or at increased risk of developing, syndromic autism or non-syndromic autism or another autism spectrum disorder. Genetic disorders that are generally linked to autism include, for example, Angelman syndrome, Prader-Willi syndrome, 15q11-q13 duplication, fragile X syndrome, fragile X premutation, deletion of chromosome 2q, XYY syndrome, Smith-Lemli-Opitz syndrome, Apert syndrome, mutations in the ARX gene, De Lange syndrome, Smith-Magenis syndrome, Williams syndrome, Noonan syndrome, Down syndrome, velo-cardio-facial syndrome, myotonic dystrophy, Steinert disease, tuberous sclerosis, Duchenne's disease, Timothy syndrome, 10p terminal deletion, Cowden syndrome, 45,X/46,XY mosaicism, Myhre syndrome, Sotos syndrome, Cohen syndrome, Goldenhar syndrome, Joubert syndrome, Lujan-Fryns syndrome, Moebius syndrome, hypomelanosis of Ito, neurofibromatosis type 1, CHARGE syndrome, and HEADD syndrome.

The methods of the invention may utilize nucleic acid sequencing, hybridization, quantitative PCR or other techniques known in the art to identify variants associated with autism spectrum disorder. A description of such techniques may be found in textbooks used by those in the art. Or, newer sequencing technologies, such as those described in more detail herein may be used (see e.g., Bowers et al., 2009, Nature Methods, 6:593-595; Ozsolak et al., Nature, 2009, 461: 814-818. By utilizing an objective diagnostic test, methods of the

invention greatly reduce and/or eliminate misdiagnoses associated with subjective methods of diagnosing an autism spectrum disorder.

For example, in certain embodiments, the invention provides methods for diagnosing presence or increased risk of developing an autism spectrum disorder in a subject (e.g., a child or a fetus) by obtaining a nucleic acid sample from the subject and identifying a sequence variant, rearrangement, copy number variant and the like that is indicative of an autism spectrum disorder. The sequence variant may be one that has been previously identified in a subject or subjects having an ASD. Or, the sequence variant may be new (i.e., previously undescribed). The identification of the variant may be empirical or may be made by comparison to known sequence alterations associated with one or more autism spectrum disorders as taught herein.

The nucleic acid source material may be obtained from a body fluid or tissue, such as cerebrospinal fluid, blood, amniotic fluid, maternal blood, buccal swab, sputum, or urine. Diagnosis may be made by analysis of any genetic element, such as, but not limited to, genes, exons, introns, gene regulatory elements, introns, expressed RNA, micro RNA, siRNA, and epigenetic elements. Sequencing methods sensitive enough to detect single copies of a gene may be used.

Yet other elements in the genome may be important to gene expression and as such, are contemplated as variants that may be used in the diagnostics of an ASD. For example, for the *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* and *UBE3A* genes, regulatory elements, including splice sites, transcription factor binding, A-I editing sites, microRNA binding sites, functional RNA structure sites, have been mapped and can be evaluated for mutations (variants) as described herein.

Thus, for each of the methods and compositions of the invention, the variant may comprise a nucleic acid sequence that encompasses at least one of the following: (1) A-to-I editing sites - adenosine-to-inosine (A-to-I) RNA editing exhibits precise regional specificity in the brain and is essential for normal behavior, and alterations in specific editing sites have been associated with a range of neuropathologies, including epilepsy and schizophrenia; (2) splice sites - it is estimated that nearly half of the causative mutations affect pre-mRNA splicing, and that many neurological diseases are caused by a splicing defect, including myotonic dystrophy and Parkinsonism linked to chromosome 17; (3) conserved functional RNA structures - single-stranded RNA-mediated regulation is structure dependent, and several core secondary structures are repeatedly used, such as hairpins and stem-loops, and

alteration of these structures may affect their function to cause disease, as in the classical example of SEPN1-related myopathy; (4) validated transcription factor binding sites (TFBS) - the Encyclopedia of DNA Elements (ENCODE) project has validated the binding of several transcription factors to predicted transcription factor binding sites (TFBS) using CHiP-seq, and mutations in TFBS are associated with several psychiatric disorders, including schizophrenia and bipolar disorder; (5) microRNA (miRNA) binding sites - miRNAs are increasingly recognized as key regulators of brain development, inducing global shifts in gene expression programs by silencing target mRNAs, and mutations in microRNA binding sites have been implicated in Tourette Syndrome and TDP43-positive frontotemporal dementia; (6) polyadenylation sites -3 polyadenylation is necessary for mRNA stabilization, and polyadenylation defects may indirectly lead to altered expression of their mRNA, or, rarely have a direct gain of function effect, such as in oculopharyngeal muscular dystrophy; (7) known regulatory elements - the Open REGulatory ANNOtation database (ORegAnno) is a database for the curation of known regulatory elements from scientific literature; (8) miRNA genes encoded in the region of interest (ROI) as several miRNA genes are embedded within protein coding genes, and polymorphisms in miRNA genes are associated with Alzheimers disease and schizophrenia; (9) small nucleolar RNA genes encoded in the ROIs - several snoRNA genes are hosted in protein coding genes, and alterations in brain specific snoRNAs have been associated with certain diseases e.g., Prader-Willi Syndrome; (10) ultraconserved elements across placental mammals - ultraconserved elements have been under tremendous evolutionary pressure to prevent any sequence changes over millions of years, and as such are thought to carry a key functional role.

For example, embodiments of the invention provide methods for diagnosing the presence or an increased risk of developing an autism spectrum disorder in a subject, e.g., a child or a fetus. Such methods may include obtaining a nucleic acid from a tissue or body fluid sample from a subject, or, in the case of a fetus, from its mother. The method may further include the steps of sequencing the nucleic acid or determining the genomic arrangement or copy number of the nucleic acid to detect whether there is a variant or variants in the nucleic acid sequence or genomic arrangement or copy number. The method may further include the steps of assessing the clinical significance of a variant or variants in the nucleic acid sequence or genomic arrangement or copy number for autism spectrum disorders. Such analysis may include an evaluation of the extent of association of the variant sequence in affected populations (i.e., subjects having the disease). Such analysis may also include an analysis of the extent of effect the mutation may have on gene expression and/or

protein function. The method may also include diagnosing presence or increased risk of developing the autism spectrum disorder based on results of this assessment.

Many different genomic analysis techniques can be used in order to make the assessments taught herein. For example, target resequencing, whole genome sequencing, single nucleotide polymorphism (SNP) analysis, copy number, epigenetic comparisons, rearrangements, deletions, and identification/analysis of other variants can be used to make the comparisons and identifications taught herein. The exemplification below is intended as illustrative and the skilled artisan understands that any available genomic analysis technique can be used in order to achieve the results specified herein.

Nucleic acid for analysis according to the invention may be obtained from a human sample, e.g. a human tissue or body fluid in any clinically acceptable manner. Nucleic acid can be obtained from adults or children or can be fetal material (e.g., fetal chromosomal materials from maternal serum or amniotic fluid). Any tissue or body fluid source is acceptable, including cellular material from tissue or fluids, such as mucous, blood, plasma, serum, serum derivatives, bile, blood, maternal blood, phlegm, saliva, sweat, amniotic fluid, mammary fluid, urine, and cerebrospinal fluid (CSF). A sample may also be a swab or a fine needle aspirate or biopsied tissue. A sample also may be media containing cells or biological material. In embodiments in which the subject is a fetus, the liquid sample can be obtained from either the amniotic fluid or the maternal blood.

The nucleic acid may be sequenced and/or its genomic arrangement and/or copy number is determined in order to detect variants (i.e., mutations) in the nucleic acid compared to a reference sequence derived from one or more individuals not known to suffer from an autism spectrum disorder at the time of sampling. As noted above, sequence variants may also be obtained empirically. The nucleic acid can include a plurality of nucleic acids derived from a plurality of genetic elements. Methods of detecting sequence variants or genomic arrangement or copy number are known in the art, and sequence variants or genomic arrangement or copy numbers can be detected by any sequencing method known in the art e.g., ensemble sequencing or single molecule sequencing, and by any method for detecting genomic arrangement or copy number known in the art, e.g., array comparative genomic hybridization.

One conventional method to perform sequencing is by chain termination and gel separation, as described by Sanger et al., 1977, Proc Natl Acad Sci U S A, 74:5463-67. Another conventional sequencing method involves chemical degradation of nucleic acid fragments. See, Maxam et al., 1977, Proc. Natl. Acad. Sci., 74:560-564. Finally, methods

have been developed based upon sequencing by hybridization. See, e.g., Harris et al., U.S. Patent Application Publication No. 20090156412. Each of these references are incorporated by reference in there entireties herein.

In certain embodiments, sequencing is performed by the Sanger sequencing technique. Classical Sanger sequencing involves a single-stranded DNA template, a DNA primer, a DNA polymerase, radioactively or fluorescently labeled nucleotides, and modified nucleotides that terminate DNA strand elongation. If the label is not attached to the dideoxynucleotide terminator (e.g., labeled primer), or is a monochromatic label (e.g., radioisotope), then the DNA sample is divided into four separate sequencing reactions, containing four standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP). These dideoxynucleotides are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides during DNA strand elongation. If each of the dideoxynucleotides carries a different label, however, (e.g., 4 different fluorescent dyes), then all the sequencing reactions can be carried out together without the need for separate reactions.

Incorporation of a dideoxynucleotide into the nascent, i.e., elongating, DNA strand terminates DNA strand extension, resulting in a nested set of DNA fragments of varying length. The newly synthesized and labeled DNA fragments are thenn denatured, and separated by size using gel electrophoresis on a denaturing polyacrylamide-urea gel capable of resolving single-base differences in chain length. If each of the four DNA synthesis reactions was labeled with the same, monochromatic label (e.g., radioisotope), then they are separated in one of four individual, adjacent lanes in the gel, in which each lane in the gel is designated according to the dideoxynucleotide used in the respective reaction, i.e., gel lanes A, T, G, C. If four different labels were utilized, then the reactions can be combined in a single lane on the gel. DNA bands are then visualized by autoradiography or fluorescence, and the DNA sequence can be directly read from the X-ray film or gel image or a continuous monitoring of fluorescence as the reaction products pass by a certain point in the gel.

The terminal nucleotide base is identified according to the dideoxynucleotide that was added in the reaction resulting in that band or its corresponding direct label. The relative positions of the different bands in the gel are then used to read (from shortest to longest) the DNA sequence as indicated. The Sanger sequencing process can be automated using a DNA sequencer, such as those commercially available from PerkinElmer, Beckman Coulter, Life Technologies, and others.

In other embodiments, sequencing of the nucleic acid is accomplished by massively parallel sequencing (also known as "next generation sequencing") of single-molecules or groups of largely identical molecules derived from single molecules by amplification through a method such as PCR. Massively parallel sequencing is shown for example in Lapidus et al., U.S. patent number 7,169,560, Quake et al. U.S. patent number 6,818,395, Harris U.S. patent number 7,282,337 and Braslavsky, et al., PNAS (USA), 100: 3960-3964 (2003), the contents of each of which are incorporated by reference herein.

In next generation sequencing, PCR or whole genome amplification can be performed on the nucleic acid in order to obtain a sufficient amount of nucleic acid for analysis. In some forms of next generation sequencing, no amplification is required because the method is capable of evaluating DNA sequences from unamplified DNA. Once determined, the sequence and/or genomic arrangement and/or genomic copy number of the nucleic acid from the test sample is compared to a standard reference derived from one or more individuals not known to suffer from an autism spectrum disorder at the time their sample was taken. All differences between the sequence and/or genomic arrangement and/or genomic arrangement and/or copy number of the nucleic acid from the test sample and the standard reference are considered variants.

In next generation (massively parallel sequencing), all regions of interest are sequenced together, and the origin of each sequence read is determined by comparison (alignment) to a reference sequence. The regions of interest can be enriched together in one reaction, or they can be enriched separately and then combined before sequencing. In certain embodiments, and as described in more detail in the examples herein, the DNA sequences derived from coding exons of genes included in the assay are enriched by bulk hybridization of randomly fragmented genomic DNA to specific RNA probes. The same adapter sequences are attached to the ends of all fragments, allowing enrichment of all hybridization-captured fragments by PCR with one primer pair in one reaction. Regions that are less efficiently captured by hybridization are amplified by PCR with specific primers. In addition, PCR with specific primers is may be used to amplify exons for which similar sequences ("pseudo exons") exist elsewhere in the genome.

In certain embodiments where massively parallel sequencing is used, PCR products are concatenated to form long stretches of DNA, which are sheared into short fragments (e.g., by accoustic energy). This step ensures that the fragment ends are distributed throughout the regions of interest. Subsequently, a stretch of dA nucleotides is added to the 3' end of each fragment, which allows the fragments to bind to a planar surface coated with oligo(dT)

primers (the "flow cell"). Each fragment may then be sequenced by extending the oligo(dT) primer with fluorescently-labeled nucleotides. During each sequencing cycle, only one type of nucleotide (A, G, T, or C) is added, and only one nucleotide is allowed to be incorporated through use of chain terminating nucleotides. For example, during the 1st sequencing cycle, a fluorescently labeled dCTP could be added. This nucleotide will only be incorporated into those growing complementary DNA strands that need a C as the next nucleotide. After each sequencing cycle, an image of the flow cell is taken to determine which fragment was extended. DNA strands that have incorporated a C will emit light, while DNA strands that have not incorporated a C will appear dark. Chain termination is reversed to make the growing DNA strands extendible again, and the process is repeated for a total of 120 cycles.

The images are converted into strings of bases, commonly referred to as "reads," which recapitulate the 3' terminal 25 to 60 bases of each fragment. The reads are then compared to the reference sequence for the DNA that was analyzed. Since any given string of 25 bases typically only occurs once in the human genome, most reads can be "aligned" to one specific place in the human genome. Finally, a consensus sequence of each genomic region may be built from the available reads and compared to the exact sequence of the reference at that position. Any differences between the consensus sequence and the reference are called as sequence variants.

Methods to Identify Autism Markers

In certain embodiments, the invention comprises methods to identify autism markers (i.e., variants in nucleic acid sequence that are associated with autism in a statistically significant manner). The genes and/or genomic regions assayed for new markers may be selected based upon their importance in biochemical pathways that show linkage and/or causation to autism. Or, the genes and/or genomic regions assayed for markers may be selected based on genetic linkage to DNA regions that are genetically linked to the inheritance of autism in families (e.g., Abrahams and Geschwind, 2008). Or, the genes and/or genomic regions assayed for markers may be evaluated systematically to cover certain regions of chromosomes not yet evaluated.

As discussed herein, autism spectrum disorders are generally characterized as one of five disorders coming under the umbrella of Pervasive Developmental Disorders (PDD). The five disorders under PDD include Autistic Disorder, Asperger's Disorder, Childhood Disintegrative Disorder (CDD), Rett's Disorder, and PDD-Not Otherwise Specified (PDD-NOS). In certain cases, the autism may be non-syndromic. Table 1 below provides a panel of

genes or genomic regions that may be evaluated for new markers to diagnose an autism spectrum disorder according to the methods of the invention.

Table 1

| Gene | Protein encoded |
|---|---|
| EIF4E | Eukaryotic translation initiation factor 4E |
| EBP1 | Eukaryotic translation initiation factor 4E-binding protein 1 |
| EBP2 | Eukaryotic translation initiation factor 4E-binding protein 2 |
| AKT1 | RAC-alpha serine/threonine-protein kinase |
| AKT2 | RAC-beta serine/threonine-protein kinase |
| AKT3 | RAC-gamma serine/threonine-protein kinase |
| PRKAA1 | 5'-AMP-activated protein kinase catalytic subunit alpha-1 |
| APP | Amyloid precursor protein |
| ARC | activity-regulated cytoskeleton-associated |
| ARX | Aristaless related homeobox |
| CACNA1C | calcium channel, voltage-dependent, L type, alpha 1C subunit |
| CAMK2G | Calcium/calmodulin-dependent protein kinase type II gamma chain |
| CDKL5 | cyclin-dependent kinase-like 5 |
| MET | MNNG (N-Methyl-N'-nitro-N-nitroso-guanidine) HOS transforming |
| CNTNAP2 | Contactin-associated protein-like 2 |
| DHCR7 | 7-dehydrocholesterol reductase |
| DRD3 | D(3) dopamine receptor |
| MAPK3 | Mitogen-activated protein kinase 3 |
| MAPK1 | Mitogen-activated protein kinase 1 |
| FKBP1A | Peptidyl-prolyl cis-trans isomerase FKBP1A |
| FMR1 | fragile X mental retardation 1 protein (FMRP) |
| AFF2 | AF4/FMR2 family member 2 |
| FOXP2 | Forkhead box protein P2 |
| FXR1 | Fragile X mental retardation syndrome-related protein 1 |
| FXR2 | Fragile X mental retardation syndrome-related protein 2 |
| GCH1 | GTP cyclohydrolase 1 |
| Gq-alpha | $G_q$ protein or $G_{q/11}$ |
| HLA-A | human leukocyte antigen |
| HOMER1 | Homer protein |
| HOXA1 | Homeobox protein Hox-A1 |
| HRAS | A ras oncogene |
| HTR3A | 5-hydroxytryptamine receptor 3A |
| HTR3C | 5-hydroxytryptamine receptor 3C |
| IGF1R | insulin-like growth factor 1 receptor |
| IGFBP1 | Insulin-like growth factor-binding protein 1 |
| MIRLET7B microRNA let-7b | A micro RNA (no protein) |
| MAP1B | Microtubule-associated protein 1B |
| MECP2 | Methyl CpG binding protein 2 |
| MAP2K1 | mitogen-activated protein kinase kinase 1 |
| MAP2K2 | mitogen-activated protein kinase kinase 1 |
| GRM1 | glutamate receptor, metabotropic 1 |
| GRM5 | glutamate receptor, metabotropic 5 |
| MKNK1 | MAP kinase-interacting serine/threonine-protein kinase 1 |

| | |
|---|---|
| MTOR | mammalian target of rapamycin (mTOR) |
| NF1 | Neurofibromatosis type I (NF-1) |
| NLGN3 | Neuroligin-3 |
| NLGN4 | Neuroligin-4 |
| NLGN4X | Neuroligin-4, X-linked |
| NLGN4Y | Neuroligin-4, X-linked |
| NRXN1 | Neurexin-1-alpha |
| OXTR | oxytocin receptor |
| PAK1 | Serine/threonine-protein kinase PAK 1 |
| PAK2 | Serine/threonine-protein kinase PAK 2 |
| PDPK1 | 3-phosphoinositide dependent protein kinase-1 |
| PDK1 | Pyruvate dehydrogenase [lipoamide] kinase isozyme 1, mitochondrial |
| PDK2 | Pyruvate dehydrogenase [lipoamide] kinase isozyme 2, mitochondrial |
| PIK3CA | Phosphatidylinositol 3-kinase, catalytic subunit |
| PIK3R1 | Phosphatidylinositol 3-kinase, catalytic subunit |
| PPP2CA | Protein phosphatase 2 (PP2) |
| PPP1CA | Serine/threonine-protein phosphatase PP1-alpha catalytic subunit |
| PPP1CC | Serine/threonine-protein phosphatase PP1-gamma catalytic subunit |
| PPP2R2B | Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B beta isoform |
| PPP2R3B | Serine/threonine-protein phosphatase 2A regulatory subunit B subunit beta |
| PPP3CA | Calcineurin |
| PPP3CB | Serine/threonine-protein phosphatase 2B catalytic subunit beta isoform |
| PPP3CC | Serine/threonine-protein phosphatase 2B catalytic subunit gamma isoform |
| PRKCB | Protein kinase C beta type |
| DLG4 | Disks large homolog 4 |
| PTEN | Phosphatase and tensin homolog |
| PTPRD | Receptor-type tyrosine-protein phosphatase delta |
| PTPRF | Receptor-type tyrosine-protein phosphatase F |
| PTPRM | Receptor-type tyrosine-protein phosphatase mu |
| PTPRZ1 | Receptor-type tyrosine-protein phosphatase zeta |
| RAC1 | Ras-related C3 botulinum toxin substrate 1 |
| RAF1 | Map kinase |
| RPTOR | regulatory associated protein of MTOR complex |
| RELN | RELN protein |
| RGS4 | Regulator of G protein signaling 4 |
| RHEB | Ras homolog enriched in brain |
| RPS6KB1 | Ribosomal protein S6 kinase beta-1 |
| SHANK1 | Shank protein 1 |
| SHANK3 | Shank protein 3 |
| SLC6A4 | Solute carrier family 6 (neurotransmitter transporter, serotonin) member |
| SNRPN | Small nuclear ribonucleoprotein-associated protein N |
| TSC1 | Tuberous sclerosis 1 |
| TSC2 | Tuberous sclerosis 2 |
| UBE3A | Ubiquitin protein ligase E3A |

In other embodiments, the genes or genomic regions evaluated for new markers may be part of a biochemical pathway that may be linked to the development of autism. For example, in certain embodiments, the genes and/or genomic regions are involved in the metabotropic glutamate receptor pathway. In one embodiment, the pathway is the mGluR5

signaling pathway and/or includes genes important to the activity of the mGluR5 receptor as the mGluR5 receptor. Thus, the mGluR5 receptor signaling pathway may, based upon its apparent importance in the development of fragile X syndrome and the association of several other components within the broadly defined mGluR5 signaling pathway with ASD, provide markers predictive of ASD. Cumulative contribution of individually rare sequence variants within multiple components of a given pathway to the same phenotype has been shown to occur for other genetic diseases. Or, other biochemical pathways that are related to certain types of autism syndromes may be evaluated.

For example, FIG. 1 provides a depiction of genes that are involved in the mGluR5 signaling pathway and that may be evaluated according to the invention to determine if mutations in such genes are linked to the development of autism. Where evidence indicates that such sequence variations may be linked to the development of autism, isolated sequences may be provided for use in DNA sequencing of patient samples to provide an indication of the presence and/or increased risk of developing autism in the subject. For example, and as described in more detail herein, Table 2 provides a subset of genes and/or genomic regions that may be evaluated, as well as mutations found in autistic subjects (i.e., patients diagnosed with non-sydromic autism).

As depicted in FIG. 2, the variants and/or variant combinations may be assessed for their clinical significance for autism spectrum disorders based on one or more of the following methods. If a variant or a variant combination is reported or known to occur more often in nucleic acid from subjects with, than in subjects without, autism spectrum disorders, it is considered to be at least potentially predisposing to autism spectrum disorders. If a variant or a variant combination is reported or known to be transmitted exclusively or preferentially to individuals with an autism spectrum disorder, it is considered to be at least potentially predisposing to autism spectrum disorders. Conversely, if a variant is found in both populations at a similar frequency, it is less likely to be associated with the development of an autism syndrome disorder (ASD) (see FIG. 2, right hand side).

If a variant or a variant combination is reported or known to have an overall deleterious effect on the function of a protein or a biological system in an experimental model system appropriate for measuring the function of this protein or this biological system, and if this variant or variant combination affects a gene or genes known to be associated with autism spectrum disorders, it is considered to be at least potentially predisposing to autism spectrum disorders. For example, if a variant or a variant combination is predicted to have an overall deleterious effect on a protein or gene expression (i.e., resulting in a nonsense

mutation, a frameshift mutation, or a splice site mutation, or even a missense mutation), based on the predicted effect on the sequence and/or the structure of a protein or a nucleic acid, and if this variant or variant combination affects a gene or genes known to be associated with autism spectrum disorders, it is considered to be at least potentially predisposing to autism spectrum disorders (see FIG. 2, left-hand side).

Also, in certain embodiments, the overall number of variants may be important. If, in the test sample, a variant or several variants are detected that are, individually or in combination, assessed as at least probably associated with an autism spectrum disorder, then the individual in whose genetic material this variant or these variants were detected can be diagnosed as being affected with or at high risk of developing an autism spectrum disorder.

Methods and Compositions for Diagnosing An Autism Spectrum Disorder

In certain embodiments, diagnosis of the autism spectrum disorder is carried out by detecting variation in the sequence, genomic location or arrangement, and/or genomic copy number of a nucleic acid or a panel of nucleic acids. For example, in some embodiments, the gene or genomic regions assessed for variants is selected from the genes in Table 1. The panel can include at least 5, 10, 20, 30, 40, 50, 60, 70, 80, or 90 of the genes in Table 1. In other embodiments, the diagnosis is made with less than 5 genes from Table 1, and in certain embodiments, with only 1 gene from Table 1.

For example, Table 2 below provides a subset of genes from Table 1, at least some of which are involved in mGluR5 receptor signaling. Table 2 also provides variants for these genes that may be detected in subjects with autism. These variants may, in certain embodiments of the methods and compositions of the invention, be indicative of an autism spectrum disorder in a subject.

Table 2

Variants for detection in ASD from mGluR5 pathway

| Gene | Position | Variant | AA change | Type | Major allele | Minor allele | Variant type |
|---|---|---|---|---|---|---|---|
| ARC | 95 | c.65T>G | p.Val22Gly | missense | T | G | T>G |
| ARC | 155 | c.125T>G | p.Val42Gly | missense | T | G | T>G |
| ARC | 167 | c.137A>C | p.His46Pro | missense | A | C | A>C |
| ARC | 173 | c.143A>C | p.His48Pro | missense | A | C | A>C |
| ARC | 188 | c.158T>G | p.Val53Gly | missense | T | G | T>G |
| ARC | 200 | c.170T>G | p.Val57Gly | missense | T | G | T>G |
| ARC | 229 | c.199T>G | p.Ser67Ala | missense | T | G | T>G |
| ARC | 266 | c.236T>G | p.Val79Gly | missense | T | G | T>G |
| ARC | 341 | c.311A>C | p.Asn104Thr | missense | A | C | A>C |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ARC | 376 | c.346T>G | p.Trp116Gly | missense | T | G | T>G |
| ARC | 413 | c.383A>C | p.Asp128Ala | missense | A | C | A>C |
| ARC | 469 | c.439A>C | p.Thr147Pro | missense | A | C | A>C |
| ARC | 479 | c.449T>G | p.Val150Gly | missense | T | G | T>G |
| ARC | 485 | c.455T>G | p.Val152Gly | missense | T | G | T>G |
| ARC | 503 | c.473A>C | p.Tyr158Ser | missense | A | C | A>C |
| ARC | 619 | c.589T>G | p.Trp197Gly | missense | T | G | T>G |
| ARC | 632 | c.602A>G | p.Glu201Gly | missense | A | G | A>G |
| ARC | 656 | c.626T>G | p.Val209Gly | missense | T | G | T>G |
| ARC | 698 | c.668A>C | p.His223Pro | missense | A | C | A>C |
| ARC | 722 | c.692T>-invalid | | deletion | T | - | T>- invalid |
| ARC | 722 | c.692T>G | p.Val231Gly | missense | T | G | T>G |
| ARC | 726 | c.696C>G | p.Gly232Gly | silent | C | G | C>G |
| ARC | 739 | c.709T>G | p.Tyr237Asp | missense | T | G | T>G |
| ARC | 748 | c.718T>G | p.Ser240Pro | missense | T | G | T>G |
| ARC | 787 | c.757T>G | p.Trp253Gly | missense | T | G | T>G |
| ARC | 790 | c.760T>G | p.Trp254Gly | missense | T | G | T>G |
| ARC | 859 | c.829G>A | p.Gly277Ser | missense | G | A | G>A |
| ARC | 1012 | c.982A>C | p.Thr328Pro | missense | A | C | A>C |
| ARC | 1043 | c.1013G>C | p.Arg338Pro | missense | G | C | G>C |
| ARC | 1046 | c.1016A>C | p.His339Pro | missense | A | C | A>C |
| ARC | 1060 | c.1030A>C | p.Thr344Pro | missense | A | C | A>C |
| ARC | 1094 | c.1064T>G | p.Val355Gly | missense | T | G | T>G |
| ARC | 1136 | c.1106A>C | | missense | A | C | A>C |
| ARC | 1139 | c.1109T>C | p.Leu370Pro | missense | T | C | T>C |
| ARC | 1145 | c.1115T>G | p.Val372Gly | missense | T | G | T>G |
| ARC | 1162 | c.1132A>C | p.Thr378Pro | missense | A | C | A>C |
| ARC | 1166 | c.1136T>C | p.Leu379Pro | missense | T | C | T>C |
| ARC | 1168 | c.1138A>C | p.Thr380Pro | missense | A | C | A>C |
| ARC | 1181 | c.1151A>C | p.Asn384Thr | missense | A | C | A>C |
| ARC | 1200 | c.1170T>G | p.Ser390Arg | missense | T | G | T>G |
| ARC | 1202 | c.1172A>C | p.Asp391Ala | missense | A | C | A>C |
| ARC | 1219 | c.1189T>G | p.X397Glu | missense | T | G | T>G |
| EIF4E | 113 | c.622G>A | p.Gly208Ser | missense | G | A | G>A |
| GRM1 | 56 | c.26T>-invalid | | deletion | T | - | T>-invalid |
| GRM1 | 443 | c.413A>C | p.Asn138Thr | missense | A | C | A>C |
| GRM1 | 654 | c.624C>T | p.Asp208Asp | silent | C | T | C>T |
| GRM1 | 71 | c.1643C>T | p.Thr548Met | missense | C | T | C>T |
| GRM1 | 35 | c.1734T>G | p.Cys578Trp | missense | T | G | T>G |
| GRM1 | 63 | c.1762T>G | p.Trp588Gly | missense | T | G | T>G |
| GRM1 | 183 | c.1882C>G | p.Arg628Gly | missense | C | G | C>G |
| GRM1 | 240 | c.1939A>C | p.Thr647Pro | missense | A | C | A>C |
| GRM1 | 261 | c.1960A>C | p.Thr654Pro | missense | A | C | A>C |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| GRM1 | 284 | c.1983C>T | p.Arg661Arg | silent | C | T | C>T |
| GRM1 | 399 | c.2098T>G | p.Phe700Val | missense | T | G | T>G |
| GRM1 | 486 | c.2185C>A | p.Pro729Thr | missense | C | A | C>A |
| GRM1 | 603 | c.2302A>C | p.Thr768Pro | missense | A | C | A>C |
| GRM1 | 641 | c.2340C>T | p.Asn780Asn | silent | C | T | C>T |
| GRM1 | 882 | c.2581G>A | p.Gly861Ser | missense | G | A | G>A |
| GRM1 | 95 | c.2725A>C | p.Met909Leu | missense | A | C | A>C |
| GRM1 | 163 | c.2793G>- invalid | | deletion | G | - | G>- invalid |
| GRM1 | 229 | c.2859C>T | p.Thr953Thr | silent | C | T | C>T |
| GRM1 | 332 | c.2962A>C | p.Thr988Pro | missense | A | C | A>C |
| GRM1 | 538 | c.3168T>- invalid | | deletion | T | - | T>- invalid |
| GRM1 | 580 | c.3210A>C | p.Pro1070Pro | silent | A | C | A>C |
| GRM1 | 583 | c.3213T>G | p.Pro1071Pro | silent | T | G | T>G |
| GRM1 | 620 | c.3250A>C | p.Thr1084Pro | missense | A | C | A>C |
| GRM1 | 727 | c.3357G>C | p.Thr1119Thr | silent | G | C | G>C |
| HOMER 1 | 234 | c.1080C>T | | 3'UTR | C | T | C>T |
| HRAS | 21 | c.-10C>T | | 5'UTR | C | T | -C>T |
| HRAS | 56 | c.26T>G | p.Val9Gly | missense | T | G | T>G |
| HRAS | 69 | c.39T>G | p.Gly13Gly | silent | T | G | T>G |
| HRAS | 50 | c.131T>G | p.Val44Gly | missense | T | G | T>G |
| HRAS | 70 | c.151T>G | p.Cys51Gly | missense | T | G | T>G |
| GRM5 | 23 | c.-8T>G | | 5'UTR | T | G | -T>G |
| GRM5 | 117 | c.87T>C | p.Ala29Ala | silent | T | C | T>C |
| GRM5 | 126 | c.96G>A | p.Pro32Pro | silent | G | A | G>A |
| GRM5 | 390 | c.360A>G | p.Ser120Ser | silent | A | G | A>G |
| GRM5 | 96 | c.727G>T | p.Ala243Ser | missense | G | T | G>T |
| GRM5 | 50 | c.1167A>G | p.Thr389Thr | silent | A | G | A>G |
| GRM5 | 207 | c.1563+8G>A | | intronic | G | A | +G>A |
| GRM5 | 27 | c.1691-4G>T | | intronic | G | T | -G>T |
| GRM5 | 120 | c.1780A>C | p.Thr594Pro | missense | A | C | A>C |
| GRM5 | 467 | c.2127T>A | p.Val709Val | silent | T | A | T>A |
| GRM5 | 719 | c.2379T>C | p.Phe793Phe | silent | T | C | T>C |
| GRM5 | 805 | c.2465T>G | p.Val822Gly | missense | T | G | T>G |
| GRM5 | 838 | c.2498T>G | p.Val833Gly | missense | T | G | T>G |
| GRM5 | 894 | c.2554T>G | p.Ser852Ala | missense | T | G | T>G |
| GRM5 | 52 | c.2652G>A | p.Thr884Thr | silent | G | A | G>A |
| GRM5 | 53 | c.2653T>G | p.Trp885Gly | missense | T | G | T>G |
| GRM5 | 63 | c.2663A>C | p.Asn888Thr | missense | A | C | A>C |
| GRM5 | 111 | c.2711A>C | p.His904Pro | missense | A | C | A>C |
| GRM5 | 147 | c.2747T>G | p.Val916Gly | missense | T | G | T>G |
| GRM5 | 344 | c.2944G>A | p.Ala982Thr | missense | G | A | G>A |
| GRM5 | 344 | c.2944G>T | p.Ala982Ser | missense | G | T | G>T |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| GRM5 | 345 | c.2945C>-invalid | | deletion | C | - | C>- invalid |
| GRM5 | 354 | c.2954G>A | p.Arg985His | missense | G | A | G>A |
| GRM5 | 354 | c.2954G>C | p.Arg985Pro | missense | G | C | G>C |
| GRM5 | 354 | c.2954G>T | p.Arg985Leu | missense | G | T | G>T |
| GRM5 | 355 | c.2955C>G | p.Arg985Arg | silent | C | G | C>G |
| GRM5 | 356 | c.2956T>A | p.Ser986Thr | missense | T | A | T>A |
| GRM5 | 356 | c.2956T>C | p.Ser986Pro | missense | T | C | T>C |
| GRM5 | 432 | c.3032A>C | p.His1011Pro | missense | A | C | A>C |
| GRM5 | 500 | c.3100A>C | p.Thr1034Pro | missense | A | C | A>C |
| GRM5 | 509 | c.3109A>C | p.Thr1037Pro | missense | A | C | A>C |
| GRM5 | 523 | c.3123C>T | p.Ser1041Ser | silent | C | T | C>T |
| GRM5 | 533 | c.3133T>C | p.Ser1045Pro | missense | T | C | T>C |
| GRM5 | 548 | c.3148A>C | p.Thr1050Pro | missense | A | C | A>C |
| GRM5 | 570 | c.3170T>G | p.Val1057Gly | missense | T | G | T>G |
| GRM5 | 626 | c.3226A>C | p.Thr1076Pro | missense | A | C | A>C |
| GRM5 | 754 | c.3354T>-invalid | | deletion | T | - | T>- invalid |
| GRM5 | 754 | c.3354T>G | p.Ala1118Ala | silent | T | G | T>G |
| GRM5 | 755 | c.3355G>-invalid | | deletion | G | - | G>- invalid |
| GRM5 | 756 | c.3356C>G | p.Ala1119Gly | missense | C | G | C>G |
| GRM5 | 763 | c.3363C>A | p.Ala1121Ala | silent | C | A | C>A |
| GRM5 | 786 | c.3386T>G | p.Val1129Gly | missense | T | G | T>G |
| GRM5 | 794 | c.3394A>C | p.Thr1132Pro | missense | A | C | A>C |
| GRM5 | 822 | c.3422T>G | p.Val1141Gly | missense | T | G | T>G |
| GRM5 | 851 | c.3451C>A | p.Pro1151Thr | missense | C | A | C>A |
| GRM5 | 884 | c.3484T>G | p.Ser1162Ala | missense | T | G | T>G |
| GRM5 | 899 | c.3499A>C | p.Thr1167Pro | missense | A | C | A>C |
| GRM5 | 903 | c.3503T>C | p.Leu1168Pro | missense | T | C | T>C |
| GRM5 | 920 | c.3520A>C | p.Thr1174Pro | missense | A | C | A>C |
| GRM5 | 920 | c.3520A>G | p.Thr1174Ala | missense | A | G | A>G |
| GRM5 | 946 | c.3546G>T | | 3'UTR | G | T | G>T |
| MAP2K1 | 54 | c.315C>T | p.Pro105Pro | silent | C | T | C>T |
| RAF1 | 152 | c.122G>A | p.Arg41Gln | missense | G | A | G>A |
| RAF1 | 66 | c.356C>T | p.Ala119Val | missense | C | T | C>T |
| RAF1 | 19 | c.1537-12T>G | | intronic | T | G | -T>G |
| RAF1 | 181 | c.1668+19G>T | | intronic | G | T | +G>T |
| RAF1 | 18 | c.1669-13T>C | | intronic | T | C | -T>C |
| RAF1 | 168 | c.1941C>T | p.Val647Val | silent | C | T | C>T |
| SHANK3 | 106 | c.524A>C | p.His175Pro | missense | A | C | A>C |
| SHANK3 | 120 | c.538A>C | p.Thr180Pro | missense | A | C | A>C |
| SHANK3 | 135 | c.553A>C | p.Thr185Pro | missense | A | C | A>C |
| SHANK3 | 54 | c.624A>C | p.Ser208Ser | silent | A | C | A>C |
| SHANK3 | 24 | c.769-7C>G | | intronic | C | G | -C>G |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SHANK3 | 125 | c.863A>C | p.His288Pro | missense | A | C | A>C |
| SHANK3 | 43 | c.898C>T | p.Arg300Cys | missense | C | T | C>T |
| SHANK3 | 254 | c.1254G>A | p.Glu418Glu | silent | G | A | G>A |
| SHANK3 | 44 | c.2091C>G | p.Pro697Pro | silent | C | G | C>G |
| SHANK3 | 1217 | c.3585G>A | p.Lys1195Lys | silent | G | A | G>A |
| SHANK3 | 1559 | c.3927C>T | p.Ser1309Ser | silent | C | T | C>T |
| SHANK3 | 1781 | c.4149C>T | p.Asp1383Asp | silent | C | T | C>T |
| SHANK3 | 2000 | c.4368C>T | p.Ser1456Ser | silent | C | T | C>T |
| SHANK3 | 468 | c.5090A>C | p.His1697Pro | missense | A | C | A>C |
| MAP2K2 | 132 | c.405G>C | p.Gly135Gly | silent | G | C | G>C |
| MAP2K2 | 147 | c.420C>T | p.Asp140Asp | silent | C | T | C>T |
| MAP2K2 | 108 | c.528G>A | p.Ala176Ala | silent | G | A | G>A |
| MAP2K2 | 171 | c.846C>T | p.Pro282Pro | silent | C | T | C>T |
| MAP2K2 | 58 | c.1074G>A | p.Ala358Ala | silent | G | A | G>A |
| MECP2 | 27 | c.378-4A>C | | intronic | A | C | -A>C |
| MECP2 | 87 | c.434T>G | p.Val145Gly | missense | T | G | T>G |
| MECP2 | 90 | c.437G>T | p.Gly146Val | missense | G | T | G>T |
| MECP2 | 93 | c.440A>C | p.Asp147Ala | missense | A | C | A>C |
| MECP2 | 105 | c.452A>C | p.Asp151Ala | missense | A | C | A>C |
| MECP2 | 138 | c.485G>T | p.Arg162Ile | missense | G | T | G>T |
| MECP2 | 220 | c.567A>C | p.Gly189Gly | silent | A | C | A>C |
| MECP2 | 235 | c.582C>T | p.Ser194Ser | silent | C | T | C>T |
| MECP2 | 242 | c.589A>C | p.Thr197Pro | missense | A | C | A>C |
| MECP2 | 271 | c.618T>G | p.Gly206Gly | silent | T | G | T>G |
| MECP2 | 338 | c.685T>G | p.Ser229Ala | missense | T | G | T>G |
| MECP2 | 355 | c.702T>G | p.Ala234Ala | silent | T | G | T>G |
| MECP2 | 364 | c.711T>- invalid | | deletion | T | - | T>- invalid |
| MECP2 | 364 | c.711T>G | p.Gly237Gly | silent | T | G | T>G |
| MECP2 | 387 | c.734T>G | p.Val245Gly | missense | T | G | T>G |
| MECP2 | 393 | c.740T>G | p.Val247Gly | missense | T | G | T>G |
| MECP2 | 403 | c.750C>T | p.Arg250Arg | silent | C | T | C>T |
| MECP2 | 406 | c.753C>T | p.Pro251Pro | silent | C | T | C>T |
| MECP2 | 436 | c.783T>G | p.Pro261Pro | silent | T | G | T>G |
| MECP2 | 516 | c.863T>G | p.Val288Gly | missense | T | G | T>G |
| MECP2 | 552 | c.899T>G | p.Val300Gly | missense | T | G | T>G |
| MECP2 | 555 | c.902T>C | p.Leu301Pro | missense | T | C | T>C |
| MECP2 | 555 | c.902T>G | p.Leu301Arg | missense | T | G | T>G |
| MECP2 | 609 | c.956T>G | p.Val319Gly | missense | T | G | T>G |
| MECP2 | 612 | c.959T>G | p.Val320Gly | missense | T | G | T>G |
| MECP2 | 627 | c.974T>G | p.Val325Gly | missense | T | G | T>G |
| MECP2 | 632 | c.979A>C | p.Thr327Pro | missense | A | C | A>C |
| MECP2 | 640 | c.987T>G | p.Gly329Gly | silent | T | G | T>G |
| MECP2 | 649 | c.996C>A | p.Ser332Arg | missense | C | A | C>A |
| MECP2 | 649 | c.996C>T | p.Ser332Ser | silent | C | T | C>T |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| MECP2 | 733 | c.1080A>C | p.Ser360Ser | silent | A | C | A>C |
| MECP2 | 805 | c.1152A>C | p.Pro384Pro | silent | A | C | A>C |
| MECP2 | 815 | c.1162C>T | p.Pro388Ser | missense | C | T | C>T |
| MECP2 | 817 | c.1164A>C | p.Pro388Pro | silent | A | C | A>C |
| MECP2 | 823 | c.1170A>C | p.Pro390Pro | silent | A | C | A>C |
| MECP2 | 842 | c.1189G>A | p.Glu397Lys | missense | G | A | G>A |
| MECP2 | 851 | c.1198A>C | p.Thr400Pro | missense | A | C | A>C |
| MECP2 | 882 | c.1229G>T | p.Ser410Ile | missense | G | T | G>T |
| MECP2 | 910 | c.1257C>T | p.Pro419Pro | silent | C | T | C>T |
| MECP2 | 933 | c.1280A>C | p.Asp427Ala | missense | A | C | A>C |
| MECP2 | 977 | c.1324A>C | p.Thr442Pro | missense | A | C | A>C |
| MECP2 | 986 | c.1333A>C | p.Thr445Pro | missense | A | C | A>C |
| MECP2 | 1088 | c.1435A>C | p.Thr479Pro | missense | A | C | A>C |
| MECP2 | 1090 | c.1437G>A | p.Thr479Thr | silent | G | A | G>A |
| MECP2 | 1095 | c.1442T>G | p.Val481Gly | missense | T | G | T>G |
| PIK3CA | 114 | c.1143C>G | p.Pro381Pro | silent | C | G | C>G |
| PIK3CA | 76 | c.1297A>C | p.Thr433Pro | missense | A | C | A>C |
| PIK3CA | 155 | c.1529A>C | p.His510Pro | missense | A | C | A>C |
| PIK3CA | 35 | c.1544A>G | p.Asn515Ser | missense | A | G | A>G |
| PIK3CA | 72 | c.1788A>G | p.Glu596Glu | silent | A | G | A>G |
| PIK3CA | 53 | c.2439A>G | p.Thr813Thr | silent | A | G | A>G |
| PIK3CA | 154 | c.3060A>G | p.Ala1020Ala | silent | A | G | A>G |
| PIK3CA | 169 | c.3075C>T | p.Thr1025Thr | silent | C | T | C>T |
| PIK3R1 | 18 | c.837-13C>T | | intronic | C | T | -C>T |
| TSC1 | 166 | c.346T>G | p.Leu116Val | missense | T | G | T>G |
| TSC1 | 52 | c.935A>C | p.Tyr312Ser | missense | A | C | A>C |
| TSC1 | 123 | c.1006C>T | p.Arg336Trp | missense | C | T | C>T |
| TSC1 | 67 | c.1178C>T | p.Thr393Ile | missense | C | T | C>T |
| TSC1 | 115 | c.1523A>C | p.Tyr508Ser | missense | A | C | A>C |
| TSC1 | 151 | c.1559A>C | p.His520Pro | missense | A | C | A>C |
| TSC1 | 172 | c.1580A>G | p.Gln527Arg | missense | A | G | A>G |
| TSC1 | 200 | c.1608A>C | p.Leu536Phe | missense | A | C | A>C |
| TSC1 | 202 | c.1610A>C | p.His537Pro | missense | A | C | A>C |
| TSC1 | 275 | c.1683T>G | p.Ser561Arg | missense | T | G | T>G |
| TSC1 | 373 | c.1781T>G | p.Val594Gly | missense | T | G | T>G |
| TSC1 | 391 | c.1799A>C | p.Gln600Pro | missense | A | C | A>C |
| TSC1 | 421 | c.1829T>G | p.Val610Gly | missense | T | G | T>G |
| TSC1 | 435 | c.1843A>C | p.Thr615Pro | missense | A | C | A>C |
| TSC1 | 436 | c.1844C>A | p.Thr615Lys | missense | C | A | C>A |
| TSC1 | 509 | c.1917T>G | p.Gly639Gly | silent | T | G | T>G |
| TSC1 | 535 | c.1943T>G | p.Val648Gly | missense | T | G | T>G |
| TSC1 | 550 | c.1958T>G | p.Ile653Arg | missense | T | G | T>G |
| TSC1 | 552 | c.1960C>A | p.Gln654Lys | missense | C | A | C>A |
| TSC1 | 552 | c.1960C>G | p.Gln654Glu | missense | C | G | C>G |
| TSC1 | 555 | c.1963C>A | p.Gln655Lys | missense | C | A | C>A |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TSC1 | 591 | c.1997+2T>G | | splice site | T | G | +T>G |
| TSC1 | 183 | c.2194C>T | p.His732Tyr | missense | C | T | C>T |
| TSC1 | 82 | c.2865C>T | p.Thr955Thr | silent | C | T | C>T |
| TSC1 | 97 | c.3042C>T | p.His1014His | silent | C | T | C>T |
| TSC1 | 114 | c.3059C>T | p.Thr1020Ile | missense | C | T | C>T |
| TSC1 | 157 | c.3102T>G | p.Gly1034Gly | silent | T | G | T>G |
| TSC1 | 160 | c.3105T>G | p.Gly1035Gly | silent | T | G | T>G |
| UBE3A | 301 | c.333C>G | p.Asn111Lys | missense | C | G | C>G |
| UBE3A | 126 | c.457G>A | p.Val153Ile | missense | G | A | G>A |
| UBE3A | 261 | c.592G>A | p.Ala198Thr | missense | G | A | G>A |
| UBE3A | 287 | c.618A>T | p.Ala206Ala | silent | A | T | A>T |
| UBE3A | 823 | c.1154T>G | p.Val385Gly | missense | T | G | T>G |
| UBE3A | 1007 | c.1338T>C | p.Phe446Phe | silent | T | C | T>C |
| UBE3A | 1097 | c.1428A>G | p.Thr476Thr | silent | A | G | A>G |
| TSC2 | 80 | c.275A>T | p.Glu92Val | missense | A | T | A>T |
| TSC2 | 127 | c.433G>A | p.Ala145Thr | missense | G | A | G>A |
| TSC2 | 26 | c.649-5A>C | | intronic | A | C | -A>C |
| TSC2 | 118 | c.736A>C | p.Thr246Pro | missense | A | C | A>C |
| TSC2 | 52 | c.796A>C | p.Thr266Pro | missense | A | C | A>C |
| TSC2 | 119 | c.848+15T>G | | intronic | T | G | +T>G |
| TSC2 | 65 | c.1292C>T | p.Ala431Val | missense | C | T | C>T |
| TSC2 | 66 | c.1875A>C | p.Ser625Ser | silent | A | C | A>C |
| TSC2 | 190 | c.3126G>T | p.Pro1042Pro | silent | G | T | G>T |
| TSC2 | 45 | c.3299T>G | p.Val1100Gly | missense | T | G | T>G |
| TSC2 | 198 | c.3778A>C | p.Thr1260Pro | missense | A | C | A>C |
| TSC2 | 43 | c.3827C>T | p.Ser1276Phe | missense | C | T | C>T |
| TSC2 | 61 | c.3914C>T | p.Pro1305Leu | missense | C | T | C>T |
| TSC2 | 133 | c.3986G>A | p.Arg1329His | missense | G | A | G>A |
| TSC2 | 23 | c.4006-8C>T | | intronic | C | T | -C>T |
| TSC2 | 76 | c.4051G>A | p.Glu1351Lys | missense | G | A | G>A |
| TSC2 | 294 | c.4269G>A | p.Leu1423Leu | silent | G | A | G>A |
| TSC2 | 310 | c.4285G>T | p.Ala1429Ser | missense | G | T | G>T |
| TSC2 | 24 | c.4990-7C>T | | intronic | C | T | -C>T |
| TSC2 | 69 | c.5028G>A | p.Leu1676Leu | silent | G | A | G>A |
| TSC2 | 23 | c.5069-8C>T | | intronic | C | T | -C>T |
| TSC2 | 130 | c.5359G>A | p.Gly1787Ser | missense | G | A | G>A |
| TSC2 | 200 | c.5429G>A | | 3'UTR | G | A | G>A |

In Table 2, all numbers and names for variants are relative to the the human reference sequence as published at the genome.ucsc.edu web-site March of 2006 (hg18) and according to the system suggested by the Human Genome Variation Society. According to the HGVS system, the start of the coding sequence (ie, the "A" of the start codon ATG) is designated as +1. All coding nucleotides, i.e., all exonic nucleotides, in the designated mRNA isoform are

numbered consecutively. Intronic nucleotides are numbered relative to the nearest exonic nucleotide. For example, the first three nucleotides of a gene (atg) would be numbered 1, 2, and 3 respectively with non-exon elements numbered as shown below (see e.g., Correlagen web-site for a discussion).

| Variant Numbering | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exon 1 | | | Intron 1 | | | | Exon 2 | | | | | | | | Intron 2 | | | | Exon 3 | | | | | | | | | | | | |
| 5'UTR | | | | | | | 5'UTR | | Met | | | Glu | | | | | | | Val | | stop | | | 3'UTR | | | | | | | |
| G | A | G | T | | T | A | G | G | T | A | T | G | G | A | G | G | T | A | G | G | T | A | T | G | A | G | A | | | | |
| -5 | -4 | -3 | -3+1 | | -3+2 | -2-2 | -2-1 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 6+1 | 6+2 | 7-2 | 7-1 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | | | | |

As shown in Table 2, sequence variants are named according to the change they cause in the DNA sequence and the resultant change (if any) to the peptide sequence. The most common types of changes are substitutions of one nucleotide for another nucleotide (e.g., c.3G>T). Other types of variants include deletions of one or more nucleotides (eg, c.4_6delGAA), insertions of one or more nucleotides (eg, c.4_5insT), or substitutions of a group of nucleotides for a group of different nucleotides, where the number of deleted and inserted nucleotides can be different (eg, c.4_6delinsT).

Mutations, even a single nucleotide substitution, can have very different results. Splice site mutations destroy an existing splice site or create a new splice site. Both types of variations can lead to altered mRNA processing and thus, a dramatically different mature mRNA and different protein.

Nonsense mutations introduce a stop codon in the middle of a coding region, which leads to truncation of the protein. Missense mutations change one amino acid in the protein into another. Synonymous mutations are mutations that do not change the amino acid sequence.

Frameshift mutations cause a shift in reading frame leading to a complete change of the amino acid sequence downstream of the mutation (i.e., the frameshift site). A frameshift mutation is caused by a net deletion or insertion of a number of nucleotides not divisible by 3. In-frame deletions and/or insertions lead to deletion or insertion of one or more amino acids in the protein, but do not alter the reading frame and so, do not change the amino acid sequence downstream of the deletion or insertion site.

The variants in Table 2 have been detected in subjects with non-syndromic autism using the methods described herein. In certain embodiments, diagnosis of the autism spectrum disorder can be carried out by comparing a sample nucleic acid including the

variant(s) to a panel of nucleic acids including the nucleic acid variants selected from the genes in Table 2. Or, novel variants may be included in the panel. The panel can include at least 1, 2, 3, 5, 10, 15, 16 or all of the genes in Table 2. In other embodiments, the diagnosis is made with less than 3 genes from Table 2, and in certain embodiments, with only 1 gene from Table 2.

Thus, ARC (activity-regulated cytoskeleton-associated) encodes a protein that is important for consolidation of synaptic plasticity as well as formation of long-term memory. ARC also regulates endocytosis of AMPA receptors in response to synaptic activity and is involved in homeostatic synaptic scaling of AMPA receptors. The ARC gene is located on chromosome 8 at 8q24.3, starting 143,689,412 bp from the p-terminus and ending 143,692,835 bp from the p-terminus (3,424 bases; orientation: minus strand). The genomic sequence of ARC is found in GenBank at accession number NC_000008. The gene sequence (NM_015193) is shown in FIG. 3A as SEQ ID NO: 1 (coding sequence from 202-1392); the protein sequence is shown in FIG. 3B as SEQ ID NO: 2. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

EIF4E (eukaryotic translation initiation factor 4E) encodes the eukaryotic translation initiation factor 4E. EIF4E is a eukaryotic translation initiation factor involved in directing ribosomes to the 7-methyl-gaunosine cap structure of mRNA. EI4FE is part of the EIF4E pre-initiation complex. The genomic sequence of EIF4E is found in GenBank at accession number NC_000004. The gene sequence (NM_001968) is shown in FIG. 3C as SEQ ID NO: 3 (coding sequence from 1524-2177); the protein sequence is shown in FIG. 3D as SEQ ID NO: 4. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

FMR1 (fragile X mental retardation 1) encodes fragile X mental retardation protein (FMRP). This protein is normally made in many tissues and may play a role in the development of synaptic connections between nerve cells in the brain. FMRP may be involved in the regulation of synaptic plasticity, which can be important in memory and learning. The FMR1 gene is located on the long arm of the X chromosome at position 27.3, from base pair 146,699,054 to base pair 146,736,156. A genomic sequence of FMR1 is found in GenBank at accession number NC_000023. The gene sequence (NM_002024) is shown in

FIG. 3E as SEQ ID NO: 5 (coding sequence from 230-2128); the protein sequence is shown in FIG. 3F as SEQ ID NO: 6. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

GRM1 (glutamate receptor, metabotropic 1) encodes the metabotropic glutamate receptor 1 (mGluR1) protein. GRM5 (glutamate receptor, metabotropic 5) encodes the metabotropic glutamate receptor 5 (mGluR5) protein. L-glutamate is the major excitatory neurotransmitter in the central nervous system and activates both ionotropic and metabotropic glutamate receptors. Glutamatergic neurotransmission is involved in most aspects of normal brain function and can be perturbed in many neuropathologic conditions. The metabotropic glutamate receptors are a family of G protein-coupled receptors, that have been divided into 3 groups on the basis of sequence homology, putative signal transduction mechanisms, and pharmacologic properties. Group I includes GRM1 and GRM5 and these receptors have been shown to activate phospholipase C. Group II includes GRM2 and GRM3 while Group III includes GRM4, GRM6, GRM7 and GRM8. Group II and III receptors are linked to the inhibition of the cyclic AMP cascade but differ in their agonist selectivities.

The GRM1 gene is located on chromosome 6 at 6q24, starting 146,390,611 bp from the p-terminus and ending 146,800,427 bp from the p-terminus (409,817 bases; orientation: plus strand). The genomic sequence of GRM1 is found in GenBank at accession number NC_000006. The gene sequence (NM_000838) is shown in FIG. 3G as SEQ ID NO: 7 (coding sequence from 471-4055); the protein sequence is shown in FIG. 3H as SEQ ID NO: 8. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

The GRM5 gene is located on chromosome 11 at 11q14.2-q14.3, starting 87,880,626 bp from the p-terminus and ending 88,438,761 bp from the p-terminus (558,136 bases; orientation: minus strand). The genomic sequence of GRM5 is found in GenBank at accession number NC_000011. The gene sequence (NM_000842) is shown in FIG. 3I as SEQ ID NO: 9 (coding sequence from 369-3911); the protein sequence is shown in FIG. 3J as SEQ ID NO: 10. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be

associated with non-syndromic autism or syndromic autism.

HOMER1 encodes a member of the homer family of dendritic proteins. Members of this family regulate group 1 metabotrophic glutamate receptor function. The HOMER1 gene is located on chromosome 5 at 5q14.2, starting 78,704,215 bp from the p-terminus and ending 78,845,796 bp from the p-terminus (141,582 bases; orientation: minus strand). The genomic sequence of HOMER1 is found in GenBank at accession number NC_000005. The gene sequence (NM_004272) is shown in FIG. 3K as SEQ ID NO: 11 (coding sequence from 1104-2168); the protein sequence is shown in FIG. 3L as SEQ ID NO: 12. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

HRAS belongs to the Ras oncogene family, whose members are related to the transforming genes of mammalian sarcoma retroviruses. The products encoded by these genes function in signal transduction pathways. These proteins can bind GTP and GDP, and they have intrinsic GTPase activity. The HRAS gene is located on chromosome 11 at 11p15.5, starting 522,242 bp from the p-terminus and ending 525,591 bp from the p-terminus (3,350 bases; orientation: minus strand). The genomic sequence of HRAS is found in GenBank at accession number NC_000011. The gene sequence (NM_176795) is shown in FIG. 3M as SEQ ID NO: 13 (coding sequence from 189-701); the protein sequence is shown in FIG. 3N as SEQ ID NO: 14. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

MAP2K1 (mitogen-activated protein kinase kinase 1) encodes a protein known as MEK1 protein kinase. MAP2K2 (mitogen-activated protein kinase kinase 2) encodes a protein known as MEK2 protein kinase. These proteins are part of a signaling pathway called the RAS/MAPK pathway, which transmits chemical signals from outside the cell to the cell's nucleus. RAS/MAPK signaling helps control the growth and division (proliferation) of cells, the process by which cells mature to carry out specific functions (differentiation), cell movement, and the self-destruction of cells (apoptosis).

The MAP2K1 gene is located on chromosome 15 at 15q22.1-q22.33, starting 64,466,674 bp from the p-terminus and ending 64,570,936 bp from the p-terminus (104,263 bases; orientation: plus strand). The genomic sequence of MAP2K1 is found in GenBank at

accession number NC_000015. The gene sequence (NM_002755) is shown in FIG. 3O as SEQ ID NO: 15 (coding sequence from 476-1657); the protein sequence is shown in FIG. 3P as SEQ ID NO: 17. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

The MAP2K2 gene is located on chromosome 19 at 19p13.3, starting 4,041,319 bp from the p-terminus and ending 4,075,126 bp from the p-terminus (33,808 bases; orientation: minus strand). The genomic sequence of MAP2K2 is found in GenBank at accession number NC_000019. The gene sequence (NM_030662) is shown in FIG. 3Q as SEQ ID NO: 17 (coding sequence from 255-1457); the protein sequence is shown in FIG. 3R as SEQ ID NO: 18. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

The MECP2 gene (methyl CpG binding protein 2) encodes a protein (MeCP2) that is essential for normal brain development. This protein seems to be important for the function of nerve cells in the brain and is present in high levels in mature nerve cells. Studies suggest that the MeCP2 protein plays a role in forming synapses between nerve cells, where cell-to-cell communication occurs. This protein silences several other genes, preventing them from making proteins. The MECP2 gene is located on chromosome X at Xq28, starting 152,940,218 bp from the p-terminus and ending 153,016,406 bp from the p-terminus (76,189 bases; orientation: minus strand). The genomic sequence of MECP2 is found in GenBank at accession number NC_000023. The gene sequence (NM_004992) is shown in FIG. 3S as SEQ ID NO: 19 (coding sequence from 227-1687); the protein sequence is shown in FIG. 3R as SEQ ID NO: 20. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

PIK3CA encodes a protein that represents the catalytic subunit of Phosphatidylinositol 3-kinase, which uses ATP to phosphorylate PtdIns, PtdIns4P and PtdIns(4,5)P2. The gene is located on chromosome 3 at 3q26.3 , starting 180,349,005 bp from the p-terminus and ending 180,435,194 bp from the p-terminus (86,190 bases; orientation: plus strand). The genomic sequence of is found in GenBank at accession

number NC_000003. The gene sequence (NM_006218) is shown in FIG. 3U as SEQ ID NO: 21 (coding sequence from 158-3364); the protein sequence is shown in FIG. 3V as SEQ ID NO: 22. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

PIK3R1 encodes a protein that represents the 85kD regulatory unit of Phosphatidylinositol 3-kinase. The gene is located on chromosome 5 at 5q13.1, starting 67,558,218 bp from the p-terminus and ending 67,633,405 bp from the p-terminus (75,188 bases; orientation: plus strand). The genomic sequence of is found in GenBank at accession number NC_000005. The gene sequence (NM_181523) is shown in FIG. 3W as SEQ ID NO: 23 (coding sequence from 43-2217); the protein sequence is shown in FIG. 3X as SEQ ID NO: 24. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

PTEN encodes the phosphatase and tensin homology protein, is a 3,4,5-triphosphate 3-phosphatse that contains a tensin like domain as well as a catalytic domain similar to that of the dual specificity protein tyrosine phosphatases. The PTEN protein preferentially dephosphorylates phosphoinositide substrates, and negatively regulate intracellular levels of phosphatidylinosito-3,4,5-triphosphate in cells. The PTEN protein is involved in the regulation of the cell cycle, preventing cells from growing too rapidly. The genomic sequence of is found in GenBank at accession number NC_007466. The gene sequence (NM_000314) is shown in FIG. 3Y as SEQ ID NO: 25 (coding sequence from 1032-2243); the protein sequence is shown in FIG. 3Z as SEQ ID NO: 26. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

RAF1 encodes a MAP kinase that functions downstream of the Ras family of membrane associated GTPases to which it binds directly. Once activated, the cellular RAF1 protein can phosphorylate to activate the dual specificity protein kinases MEK1 and MEK2, which in turn phosphorylate to activate the serine/threonine specific protein kinases, ERK1 and ERK2. Activated ERKs are pleiotropic effectors of cell physiology and play an important role in the control of gene expression involved in the cell division cycle, apoptosis, cell

differentiation and cell migration. The RAF1 gene is located on chromosome 3 at 3p25, starting 12,600,108 bp from the p-terminus and ending 12,680,678 bp from the p-terminus (80,571 bases; orientation: minus strand). The genomic sequence of RAF1 is found in GenBank at accession number NC_000003. The gene sequence (NM_002880) is shown in FIG. 3AA as SEQ ID NO: 27 (coding sequence from 416-2362); the protein sequence is shown in FIG. 3BB as SEQ ID NO: 28. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

RHEB encodes the GTP-binding protein known as Ras Homology Enriched in Brain. Rheb is a member of the Ras superfamily and may be involved in neural plasticity. The protein is a member of the small GTPase superfamily and encodes a lipid-anchored cell membrane protein with five repeats of the Ras-related GTP-binding region. A genomic sequence of RHEB is found in GenBank at accession number NC_000007. The gene sequence (NM_005614) is shown in FIG. 3CC as SEQ ID NO: 29 (coding sequence from 414-968); the protein sequence is shown in FIG. 3DD as SEQ ID NO: 30. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

SHANK3 encodes proteins necessary to construct synapses in the brain. Shank proteins are multidomain scaffold proteins of the postsynaptic density that connect neurotransmitter receptors, ion channels, and other membrane proteins to the actin cytoskeleton and G-protein-coupled signaling pathways. Shank proteins also play a role in synapse formation and dendritic spine maturation. The gene is located on chromosome 22 at 22q13.3, starting 49,459,936 bp from the p-terminus and ending 49,518,507 bp from the p-terminus (58,572 bases; orientation: plus strand). A genomic sequence of SHANK3 is found in GenBank at accession number NC_000022. The gene sequence (NM_001080420) is shown in FIG. 3EE as SEQ ID NO: 31 (coding sequence from 1-5244); the protein sequence is shown in FIG. 3FF as SEQ ID NO: 32. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

TSC1 (Tuberous sclerosis 1) encodes a peripheral membrane protein that has been implicated as a tumor suppressor. TSC1 is also involved in vesicular transport and docking, in complex with TSC2. The TSC1 gene is located on chromosome 9 at 9q34, starting 134,756,557 bp from the p-terminus and ending 134,809,841 bp from the p-terminus (53,285 bases; orientation: minus strand). The gene sequence of TSC1 is found in GenBank at accession number NC_000009. The gene sequence (NM_000368) is shown in FIG. 3GG as SEQ ID NO: 33 (coding sequence from 235-3729); the protein sequence is shown in FIG. 3HH as SEQ ID NO: 34. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

The gene TSC2 (Tuberous sclerosis 2) encodes a protein called tuberin and has been implicated as a tumor suppressor. Its gene product associates with hamartin in a cytosolic complex, acting as a chaperone for hamartin. TSC2 has a function in vesicular transport, and interaction between TSC1 and TSC2 facilitates vesicular docking. Gene products of TSC1 and TSC2 work together to help control cell growth and size. The TSC2 gene is located on chromosome 16 at 16p13.3, starting 2,037,991 bp from the p-terminus and ending 2,078,714 bp from the p-terminus (40,724 bases; orientation: plus strand). A genomic sequence of TSC2 is found in GenBank at accession number NC_000016. The gene sequence (NM_000548) is shown in FIG. 3II as SEQ ID NO: 35 (coding sequence from 107-5530); the protein sequence is shown in FIG. 3JJ as SEQ ID NO: 36. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g., non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

UBE3A (ubiquitin protein ligase E3A) encodes an enzyme called ubiquitin protein ligase E3A. This enzyme is involved in targeting proteins to be broken down (degraded) within cells. The gene is located on chromosome 15 at 15q11-q13, starting 23,133,489 bp from the p-terminus and ending 23,235,221 bp from the p-terminus (101,733 bases; orientation: minus strand). A genomic sequence of is found in GenBank at accession number NC_000015. The gene sequence (NM_130839) is shown in FIG. 3KK as SEQ ID NO: 37 (coding sequence from 658-3276); the protein sequence is shown in FIG. 3LL as SEQ ID NO: 38. Except as noted herein, variants in this sequence are believed to have not previously been shown to be associated with at least some autism spectrum disorders (e.g.,

non-syndromic autism), and variants in Table 2 have not previously been shown to be associated with non-syndromic autism or syndromic autism.

For example, for the first ARC variant in Table 2, one may use the DNA sequence of SEQ ID NO: 1 to determine the coding sequence of the gene encoding the variant: c65T>G, and the protein sequence of the protein having the variant p.Val22Gly to determine that this variant comprises a missense mutation. The nature of the mutation may further be assessed on the protein sequence (SEQ ID NO: 2) and function. For example, this mutation may be expected to have a moderate to minor effect as the amino acid substitution of Gly for Val may be considered to be a conservative substitution. Or, a more detailed analysis of three-dimensional conformational data for the protein may indicate that the mutation may be deleterious to protein function. A similar analysis may be performed for each of the variants described in Table 2, using the sequences provided in Figure 4, panes A-LL.

Thus, in certain embodiments, the number and nature of DNA sequence variants in the coding regions and contiguous RNA regulatory regions in 19 genes that code for a number of the key proteins involved in the mGluR5 pathway in samples from patients with an ASD (e.g., non-syndromic autism) and from controls (i.e., non-autistic individuals) may be examined. Target regions may be enriched in pools of samples and sequenced by next-generation technology to enable the detection of rare variants. An embodiment of such methods are discussed in more detail in the Examples herein. Such sequencing is generally performed under conditions of high sensitivity and low false discovery rates at acceptable cost using the methods described herein. Sensitivity of variant detection may generally depend on the number of reads covering this position (known as "depth" or "coverage"), i.e., the amount of sequence information available for that particular position. Since both the enrichment methods and the sequencing step are influenced by the sequence context, coverage may vary from region to region. In addition, sensitivity of variant detection also differs by type of variant (substitution versus deletion and/or insertion). At high coverages (i.e., sequencing a region $\geq$30x), sensitivity is in certain embodiments, about 99% for detecting substitution variants, 90% for detecting deletions and/or insertions spanning $\leq 5$ bases, and approximately 30% for detecting deletions and/or insertions spanning from 6 to about 40 bases. Also in certain embodiments, insertions and/or deletions spanning $\leq$5 bases or $\geq$6 bases account for about 10% and 1%, respectively, of all variant occurrences and for about 16% and 2.6%, respectively, of all pathogenic variant occurrences. Taking into account coverage at each base position within the sequenced regions, length of the sequenced regions, and variant-type specific sensitivity, an overall sensitivity of variant detection for each gene

included in the assay may be provided. For example, if 80% of the analyzed bases in a gene have a coverage corresponding to 97% sensitivity, 15% have a coverage corresponding to 92% sensitivity, and 5% have a coverage corresponding to 80% sensitivity, the overall sensitivity for that gene would be calculated as 95%. Generally, exons with a sensitivity of less than 50% are not included in the overall sensitivity estimate per gene, but are reported separately as segments not sequenced.

In certain embodiments, all sequence variants detected by next-generation sequencing that are known or predicted to be pathogenic as well as all sequence variants that are novel (i.e., not previously described in the literature or a database) are confirmed by uni-directional Sanger sequencing. Therefore, the false positive rate of reported variants is generally very low. Using this method, a statistically significant increase in the number of rare variants overall as well as rare, potentially disruptive variants in cases compared to controls in several different genes may be detected.

For example, in an embodiment, the variants depicted in Table 2 were found in samples from individuals with a type of autism syndrome (i.e., non-syndromic autism) that are not seen in unaffected individuals.

For example, Tables 3 - 5 provide an analysis of variants associated with mGluR5 signaling as found in individuals that either did not display the symptoms of autism (i.e., controls), or that were diagnosed with non-syndromic autism. As described in more detail in the Examples herein, the variants in Table 2 were found in samples from 290 individuals with an autism-spectrum disorder (non-syndromic autism).

As shown in Tables 3-5, at least 4 of these genes (*SHANK 3, TSC1, TSC2* and *HOMER 1*) had mutations that based upon an increased detection in autism patients. Also, for these genes, and at least some of the other genes assayed, the severity of the mutation with respect to gene expression or protein function indicated that the variants may be associated with the development of non-syndromic autism. These mutations may also be involved in other types of autism syndromes. At least one of these genes (*HOMER1*) has not previously been associated with autism either functionally, or genetically.

Table 3 shows a comparison of the number of common and rare variants in samples from patients with non-syndromic autism as provided by the AGRE sample database, as compared to controls (i.e., individuals who do not have autism or ASD). It can be seen that for some of the genes, there is a distinct increase in the number of rare variants in the individuals from the patient pool, whereas the more common variants exhibit similar frequencies in both groups.

**Table 3**

| | | Number of rare variants | | Number of common variants | |
|---|---|---|---|---|---|
| **Gene** | **Total # variants** | **AGRE** | **Controls** | **AGRE** | **Controls** |
| *ARC* | 14 | 6 | 9 | 2 | 2 |
| *EIF4E* | 7 | 4 | 5 | 1 | 1 |
| *FMR1* | 10 | 3 | 7 | 2 | 3 |
| *GRM1* | 41 | 23 | 28 | 7 | 7 |
| *GRM5* | 71 | 38 | 36 | 22 | 21 |
| *HOMER1* | 13 | 8 | 2 | 4 | 4 |
| *HRAS* | 7 | 3 | 2 | 3 | 3 |
| *MAP2K1* | 8 | 5 | 4 | 2 | 3 |
| *MAP2K2* | 32 | 21 | 19 | 5 | 6 |
| *MECP2* | 21 | 15 | 15 | | 1 |
| *PIK3CA* | 27 | 9 | 4 | 16 | 16 |
| *PIK3R1* | 18 | 9 | 8 | 6 | 6 |
| *PTEN* | 6 | 5 | 5 | 1 | 1 |
| *RAF1* | 10 | 7 | 8 | 1 | |
| *RHEB* | 3 | 1 | 2 | 1 | 1 |
| *SHANK3* | 98 | 81 | 61 | 7 | 6 |
| *TSC1* | 31 | 26 | 13 | 3 | 5 |
| *TSC2* | 107 | 64 | 73 | 14 | 12 |
| *UBE3A* | 12 | 8 | 9 | 2 | 1 |
| **Grand Total** | **536** | **336** | **310** | **99** | **99** |

Table 4 shows a comparison of the number of rare and potentially disruptive mutations,(i.e., based on the nature of the mutation, these mutations are expected to disrupt gene expression or protein function, in patients with non-syndromic autism as compared to controls. It can be seen that for some of the genes, there is a distinct increase in the number of potentially disruptive variants in the AGRE population as compared to the controls.

**Table 4**

| | Number of rare, potentially disruptive variants | | |
|---|---|---|---|
| **Gene** | **Total** | **AGRE** | **Controls** |
| *ARC* | 6 | 5 | 3 |
| *EIF4E* | 2 | 2 | 2 |
| *FMR1* | 3 | | 2 |

| | | | |
|---|---|---|---|
| *GRM1* | 19 | 10 | 14 |
| *GRM5* | 31 | 13 | 17 |
| *HOMER1* | 6 | 6 | 1 |
| *HRAS* | 2 | 1 | |
| *MAP2K1* | 3 | 2 | 3 |
| *MAP2K2* | 8 | 6 | 4 |
| *MECP2* | 6 | 4 | 6 |
| *PIK3CA* | 13 | 2 | 3 |
| *PIK3R1* | 4 | 2 | 1 |
| *PTEN* | 1 | 1 | 1 |
| *RAF1* | 4 | 2 | 2 |
| *RHEB* | 1 | 1 | 1 |
| *SHANK3* | 42 | 37 | 26 |
| *TSC1* | 18 | 17 | 8 |
| *TSC2* | 44 | 35 | 27 |
| *UBE3A* | 3 | 1 | 1 |
| **Total** | **216** | **147** | **122** |

Table 5 shows the number of rare, potentially disruptive variants that were found in only one sample source. It can be seen that for at least four genes (*HOMER, SHANK 3, TSC1* and *TSC2*, there are mutations that are found in the patient (AGRE) population, but are not found in the controls. This demonstrates a statistically significant difference between rare (i.e., novel in that these variants have never been reported before) variants in specific genes in the mGluR5 pathway and non-syndromic autism.

**Table 5**

**Number of rare, potentially disruptive variants found in only one sample source**

| Gene | Total | AGRE | Control | Fisher's 2 tailed test |
|---|---|---|---|---|
| *ARC* | 4 | 3 | 1 | 0.3654 |
| *FMR1* | 3 | | 2 | 0.2412 |
| *GRM1* | 8 | 2 | 6 | 0.2863 |
| *GRM5* | 9 | 3 | 6 | 0.505 |
| *HOMER1* | 5 | 5 | | **0.0282** |
| *HRAS* | 1 | 1 | | 0.4915 |
| *MAP2K1* | 1 | | 1 | 0.4915 |
| *MAP2K2* | 4 | 3 | 1 | 0.3654 |

| | | | | |
|---|---|---|---|---|
| *MECP2* | 2 | | 2 | 0.4994 |
| *PIK3CA* | 3 | 1 | 2 | 1 |
| *PIK3R1* | 1 | 1 | | 0.4915 |
| *RAF1* | 4 | 2 | 2 | 1 |
| *SHANK3* | 17 | 14 | 3 | **0.006** |
| *TSC1* | 8 | 8 | | **0.0032** |
| *TSC2* | 19 | 14 | 5 | **0.0356** |
| *UBE3A* | 1 | 1 | | 0.4915 |
| **Grand Total** | **90** | **58** | **31** | **0.0012** |

Thus, in certain embodiments, the present invention provides methods and or nucleic acid sequences that can be used to determine if a subject has, or is at increased risk for developing an ASD. As noted above, in some cases, the variant nucleic acid may be a novel (i.e., not previously reported) variant, or it may be a variant that has previously been found to be associated with an ASD. In certain embodiments, the variant may be a novel variant, or a previously reported variant in one of the genes that is important to the metabotropic glutamate receptor pathway. Or, genes from other biochemical pathways may be analyzed. For example, in at least one embodiment, at least four genes (*HOMER, SHANK 3, TSC1* and *TSC2*), there are mutations that are found in patients with non-syndromic autism but that are not found in the controls.

In an embodiment, the variant (mutation) may be one of the variants listed in Table 2. Or, the variant may be at least one of a *HOMER 1* such as, but not limited to: a c.195G>T, M65I; a c.290C>T, S97L mutation; or a c.425C>T, P142L mutation. Additionally or alternatively, the mutation may comprise a *GRM5* c.3503T>C, L1168P mutation. Additionally or alternatively, the mutation may comprise a *MAPK2* c.581-1G>T mutation and/or a *HRAS* c.383G>A, R128Q mutation. Additionally or alternatively, the mutation may comprise a *MECP2* c.1477G>T, E483X mutation.

For example, two of *HOMER1* variants (c.195G>T, M65I and c.290C>T, S97L) are located in the EVH1 domain in Homer1, which interacts with the Pro-Pro-Ser-Pro-Phe motifs in mGluR1 and mGluR5. A third potentially damaging variant in *HOMER1* (c.425C>T, P142L) affects one of the conserved prolines within the P-motif of the CRH1 domain, which serves as an internal binding site for the EVH1 domain. It has been proposed that EVH1 binding to mGluR induces homo-multimerization of Homer1, while EVH1 binding to the internal P-motif in Homer1 arrests this homo-multimerization. Interestingly, one of the *GRM5* variants (c.3503T>C, L1168P) detected in AGRE samples is located relatively close to the conserved Pro-Pro-Ser-Pro-Phe Homer1 binding motif in mGluR5.

In other embodiments, the mutations detected are in either the TSC1 or the TSC2 genes (see Table 2). In yet other embodiments, the mutations detected are in either the SHANK 3 gene (see Table 2).

In other embodiments, AGRE samples may have a variant in *MAP2K2* that affects a conserved splice-site and is thus highly likely to be damaging (c.581-1G>T). In yet other embodiments, a potentially damaging variant was also detected in *HRAS*, another gene in the RAS/MAPK signalling pathway. This *HRAS* variant (c.383G>A, R128Q) disrupts an arginine at position 128 that plays an important role in membrane binding and function of GTP-bound H-ras.

In yet other embodiments, the method may further detect a nonsense mutation (c.1477G>T, E483X) in *MECP2* in a single AGRE sample, a gene known to be associated with Rett's syndrome, another syndromic form of ASD.

## Examples

The method is exemplified by the following non-limiting examples.

### Example 1: Variant Discovery in Autism Candidate Genes

All coding exons of 19 candidate genes hypothesized to be associated with autism spectrum disorders and of 4 control genes known to be associated with hypertrophic cardiomyopathy in 290 samples from the AGRE collection and 290 ethnically matched samples from the Coriell collection were amplified. Before amplification, the DNA concentration in each sample was determined by measurement on a NANODROP spectrophotometer, and equal amounts of DNA were then used to generate 15 pools of 20 AGRE samples each and 15 pools of 20 Coriell samples each.

Ten samples of each collection were represented in two pools, allowing for independent replication of variant detection in those samples. For each pool, a total of 293 PCR products were generated, encompassing about 116,000 bases in total. PCR products covered all coding regions of every mRNA isoform as well as flanking intronic regions. A high-fidelity polymerase was used for PCR amplification, to minimize introduction of errors during PCR. PCR primers were tailed with sequences containing a NotI restriction site. Following PCR amplification, PCR products were pooled and subjected to digestion with NotI restriction enzyme. The NotI-digested PCR products were ligated to generate concatemers of several kb in length. Concatemers were then randomly sheared into fragments 200 to 250 bp in length. Following ILLUMINA's protocol, the fragments were end-repaired, A-tailed, and ligated to forked adapter molecules. Adapter-ligated fragments were selectively enriched by PCR. During the enrichment step, a 6-bp index was added to the fragments.

Indexing of the fragments allowed sequencing of fragments from different sample pools on the same lane of the Illumina GA2 instrument.

Sequencing was performed for 50 cycles on the ILLUMINA GA2. Minimal yield per lane was 5 million reads. Fragment libraries from two different sample pools were sequenced per ILLUMINA GA2 lane, for an average target coverage of 800 fold per base and sample pool, or 40 fold per individual (20 fold per individual chromosome). This average coverage was sufficient to detect occurrence of a single heterozygous variant in a pool of 20 samples. It was found that coverage and thus detection sensitivity varied within and between amplified regions, as well as between fragment libraries.

Sequence data derived from each ILLUMINA GA2 lane were processed through BUSTARD for base calling, and the output data then separated into different files based on index. Only index reads differing by 1 or less bases from the actual index sequence were used. After index-splitting, sequence data were analyzed using the pipeline developed at Boston College by Dr. Gabor Marth, which is comprised of an aligner (MOSAIK) and a variant caller (GIGABAYES). Sequence reads were aligned to a reference sequence assembled from the hg18-derived sequence for all coding exons represented in the original PCR-product library plus about 30 nucleotides of flanking non-coding sequence. For a read to be considered aligned, at least 60% of the bases had to be aligned with a maximum of 1 mismatch. Variant calling in the pooled data was based on the successive application of three types of filters followed by use of the Bayesian-based variant calling algorithm employed by GIGABAYES. The filters were designed to reduce the false-positive rate while maintaining 80-90% sensitivity for detecting single heterozygous variant occurrences in a pool of 20 samples by adhering to the following experimental conditions: (1) The QV value of the base call had to be at least 20; (2) a minimum number of minor allele calls had to be derived from each DNA strand (coding and non-coding); and (3) the minor allele frequency had to reach a certain value.

Filters were applied as follows. Minor alleles that occurred at least 4 times on each DNA strand at a QV of at least 20 were kept in consideration as potential variants. At positions where the total coverage (i.e., the total number of base calls of any QV value) was below 1200, minor alleles that occurred at least 3 times on each DNA strand at a QV of at least 20 were kept in consideration as potential variants. At positions where the total coverage was below 900, minor alleles that occurred at least 2 times on each DNA strand at a QV of at least 20 were kept in consideration as potential variants. At all positions where a potential variant was called based on the criteria described above, all variant calls with any filter in any

other sample pools were then kept in consideration as potential variants. All potential variant calls were then subjected to the GIGABAYES variant calling algorithm with the appropriate settings for variant calling in pooled samples. Of the resulting variant calls, only those with a minor allele frequency of 1.5% or more were accepted.

Example 2- Number of variants found in AGRE and Control Samples

A total of 536 variants were detected in the AGRE and/or the control samples on both the GA2 and the HELISCOPE platform (Table 3). These variants were called 'common' if found with an allele frequency $\geq$ 1% and rare if found with an allele frequency of <1%. 336 and 310 of the variants detected in AGRE samples and in controls, respectively, were rare. The number of both common and rare variants differed between individual genes. Table 2 shows at least some of the variants that were detected.

The method may include selecting for rare variants with a potentially disruptive effect. In this group, variants which create a missense or nonsense change on the protein level, affect a conserved splice-site, or are located in the 3'UTR or 5' UTR and may thus impact mRNA transcription or processing were included. Of a total of 216 rare potentially disruptive variants, 147 were found in AGRE and 122 in controls (Table 4). Of those, 58 were only found in AGRE samples and 31 only in controls (Table 5), indicating as statistically significant enrichment of rare, potentially disruptive variants in AGRE samples. On the level of the individual genes, the enrichment reached statistical significance for the genes *HOMER1*, *SHANK3*, *TSC1*, and *TSC2* (Table 5).

Three of these genes (*SHANK3*, *TSC1*, and *TSC2*) have previously demonstrated causal roles in autism. Notably, however, autism due to variants in *TSC1* or *TSC2* is typically seen in the context of tuberous sclerosis, while in the current study, samples from individuals with syndromic forms of ASD were excluded. The fourth gene (*HOMER1*) has not previously been causally related to autism. Two of *HOMER1* variants (c.195G>T, M65I and c.290C>T, S97L) are located in the EVH1 domain in Homer1, which has been shown to interact with the Pro-Pro-Ser-Pro-Phe motifs in mGluR1 and mGluR5. A third potentially damaging variant in *HOMER1* (c.425C>T, P142L) affects one of the conserved prolines within the P-motif of the CRH1 domain, which serves as an internal binding site for the EVH1 domain. It has been proposed that EVH1 binding to mGluR induces homo-multimerization of Homer1, while EVH1 binding to the internal P-motif in Homer1 arrests this homo-multimerization. Interestingly, one of the *GRM5* variants (c.3503T>C, L1168P) detected in AGRE samples is located relatively close to the conserved Pro-Pro-Ser-Pro-Phe Homer1 binding motif in mGluR5.

Several of the rare, potentially disruptive *TSC1* and *TSC2* variants observed only in AGRE samples have been classified by others as rare polymorphism because they were seen together with clear disease variants and/or did not clearly segregate with a tuberous sclerosis phenotype. These variants may thus represent hypomorphic variants with regard to tuberous sclerosis and act as modifiers when occurring together with other variants in *TSC1* and *TSC2*. The pleomorphic nature of monogenic disorders and the role of hypomorphic variants in milder forms of monogenic disease is increasingly well recognized.

While the enrichment of rare, potentially disruptive variants in AGRE samples reached statistical significance for four of the genes with this initial sampling, specific single variants suggest causal relationship of additional genes to ASD. Specifically, one AGRE sample harbored a variant in *MAP2K2* that affects a conserved splice-site and is thus highly likely to be damaging (c.581-1G>T). A potentially damaging variant was also detected in *HRAS*, another gene in the RAS/MAPK signalling pathway. This *HRAS* variant (c.383G>A, R128Q) disrupts an arginine at position 128 that has been shown to play an important role in membrane binding and function of GTP-bound H-ras. *MAP2K2* and *HRAS* are known to be associated cardiofaciocutaneous and Costello syndrome, respectively, both monogenic disorders associated with mental delay and retardation. However, *MAP2K2* has not previously been linked to autism, while early association studies did suggest a link between *HRAS* and ASD.

The method further detected one nonsense mutation (c.1477G>T, E483X) in *MECP2* in a single AGRE sample, a gene known to be associated with Rett's syndrome, another syndromic form of ASD. Interestingly, this nonsense mutation causes a deletion of only the 3 C-terminal amino acids of MECP2, and may thus also represent a hypomorphic variant.

The average coverage for each gene in each pool at all positions where a variant was detected in any or the pools was determined. Low coverage in one or a few pools has little effect on detection of common variants, since the variant will be found in many different pools. However, rare variants may be missed if they occur only in a pool of lower coverage. To appreciate these effects two measures were evaluated: (1) the number of pools with coverage below a certain cut-off value (e.g., 160 for the 20-specimen pools and 120 for the 15-specimen pools); and (2) the relative frequency of common and rare variants. Both presence of low coverage in several pools and an unbalanced ratio of common and rare variants between populations lower the detection confidence for rare variants in a given gene.

As the methods of the invention provide for the discovery of rare variants in candidate genes, assays were performed to determine that that the sample-pool size did not limit sensitivity of variant detection.

To validate sensitivity of variant detection in the larger (20-sample) pools, a validation pool from 20 samples that had previously been Sanger sequenced for all coding exons of the genes *MYBPC3*, *MHY7*, *TNNT2*, and *TNNI3* was constructed and enriched for these targets using PCR. The PCR products were concatenated, sheared, and sequenced on the GA2 sequencer under conditions of high coverage. Sequencing detected all of the 46 single-nucleotide variants previously detected by Sanger sequencing, including 20 variants that were heterozygously present in only 1 of the 20 samples (singletons), demonstrating the high sensitivity of variant detection in such pools (e.g., 20-sample pools on the GA2 under conditions of high coverage). Although for some of the singletons, the allele frequency detected in the pool deviated from the theoretical value of 0.025, all singletons were detected at an allele frequency of $\geq 0.012$, or half the theoretical value. At this allele-frequency cut-off, however, an additional 82 variants were detected that had not been found by Sanger sequencing and are thus likely to be false positives, for a false-discovery rate (FDR) of 64%.

Example 3: Methods

**Sample Selection**

DNA samples (n=290) from individuals with an autism-spectrum disorder (ASD) were obtained from the Autism Genetic Research Exchange (AGRE) collection, based on the following inclusion criteria: diagnosis of autism by Autism Diagnostic Interview, Revised (ADI-R) and Autism Diagnostic Schedule (ADOS); idiopathic (i.e., non-syndromic) autism; at least one affected family member; and availability of complete data for RAVEN, Peobody, and SRS. Sample ethnicity was given as white, not Hispanic or Latino, and not more than one race for 221 individuals; as white, not Hispanic or Latino, and more than one race for 11 individuals; as Hispanic or Latino for 53 individuals; and as Asian for 5 individuals. Three hundred (300) control DNA samples were obtained from the Coriell collection and consisted of 248 samples with Caucasian or European ethnicities and 52 with Hispanic or Latino ethnicities.

**Next-generation sequencing**

DNA concentration was determined for all samples using a NANODROP analysis system, and equal amounts of control sample DNAs were combined into orthogonal pools of either 20 samples each and 15 samples. Each pool then served as a one DNA template for PCR amplification of all coding exons of the longest isoform of each of the 19 candidate

genes, using specific PCR primers tailed at the 5' end with a 14-bp sequence containing a Not1 restriction site. All PCR products derived from the same template (i.e., sample pool) were pooled, digested with Not1, and ligated to form concatemers, which were subsequently randomly sheared into fragments with a mean size of 150 to 300 bp, using a COVARIS S2 instrument. These fragments were prepared for sequencing on either an ILLUMINA GA2 (20-sample pools) or a HELICOS HELISCOPE (15-sample pools) according to the manufacturers' instructions. ILLUMINA sequencing was performed for 50 cycles, resulting in a read length of up close to 50 bases, and HELISCOPE sequencing was performed for 120 cycles or 30 quads, resulting in an average read length of about 32 bases.

**Analysis of next-generation sequencing data**

Reads were aligned to a reference sequence that included the hg18-derived sequence of each amplified exon "padded" on each side with 30 flanking non-coding bases. The aligner MOSAIK was used for the GA2 reads, and the aligner INDEXDP for the HELISCOPE reads. Variant calling was performed with GIGABAYES for the GA2 reads, but without invoking the Bayesian-based algorithm, and with SNPSNIFFER for the HELISCOPE reads. SNPSNIFFER required a minimum minor allele frequency threshold of 1%. No minimum minor allele frequency threshold was set in GIGABAYES. 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.

**Sanger sequencing**

Sanger sequencing was performed for selected gene regions and selected samples, to confirm variants detected during next-generation sequencing. PCR primers and conditions were the same as before, except that individual samples were used as template instead of sample pools. Each PCR product was then cycle-sequenced using ABI BIGDYE reagents, with the specific PCR primers serving as sequencing primers, and the sequencing products were separated on an ABI3730exl. Sequencing traces were visualized using SEQUENCESCANNER (ABI), and presence or absence of a given mutation determined by manual comparison to the reference sequence.

References and citations to other documents, such as patents, patent applications, patent publications, journals, books, papers, web contents, have been made throughout this disclosure. All such documents are hereby incorporated herein by reference in their entirety for all purposes.

Various modifications and equivalents of the invention and many further embodiments thereof, in addition to those shown and described herein, will become apparent

to those skilled in the art from the full contents of this document, including references to the scientific and patent literature cited herein. The subject matter herein contains information, exemplification and guidance that can be adapted to the practice of this invention in its various embodiments and equivalents thereof.

What is claimed is:

1. A method for diagnosing the presence or an increased risk of developing an autism spectrum disorder in a subject, the method comprising:

obtaining a nucleic acid from a tissue or body fluid sample from a subject;

conducting an assay to identify whether there is a variant sequence, or a plurality of variant sequences, in the subject's nucleic acid;

for each variant detected, determining if the variant is a known variant associated with an autism spectrum disorder or a previously undescribed variant;

if the variant is a previously undescribed variant, determining if the variant is expected to have a deleterious effect on at least one of gene expression and/or protein function; and

diagnosing the presence or an increased risk of developing the autism spectrum disorder based on the variant sequence or the plurality of variant sequences detected.

2. The method according to claim 1, wherein the assay comprises at least one of nucleic acid sequencing, hybrid capture, and epigenetic analysis.

3. The method of claim 1, wherein the nucleic acid in the conducting step comprises a gene, an exon, an intron, a gene regulatory element, an expressed RNA, an siRNA, or an epigenetic element.

4. The method of claim 1, wherein the nucleic acid comprises a sequence selected from a sequence known or suspected to be associated with one or more autism spectrum disorders.

5. The method according to claim 4, wherein the nucleic acid comprises at least a portion of one of the genes in Table 1.

6. The method of claim 1, wherein the nucleic acid is derived from a gene that encodes a protein in the metabotropic glutamate receptor signaling pathway.

7. The method according to claim 1, wherein the nucleic acid comprises at least a portion of a gene for at least one of *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* or *UBE3A*.

8. The method according to claim 1, wherein the nucleic acid comprises at least a portion of a gene for at least one of *TSC1, TSC2, SHANK3,* or *HOMER1*.

9. The method according to claim 2, wherein the sequencing comprises at least one of single-molecule sequencing-by-synthesis or massively parallel sequencing.

10. The method according to claim 2, wherein a plurality of DNA samples are analyzed in a pool to identify samples that show a variation.

11. The method according to claim 10, wherein a plurality of DNA samples are analyzed in a plurality of pools to identify an individual sample that shows the same variation in at least two pools.

12. The method according to claim 1, wherein the autism spectrum disorder comprises at least one of non-syndromic autism, classical autism, Asperger's syndrome, Rett's syndrome, childhood disintegrative disorder, or pervasive developmental disorder not otherwise specified (PDD-NOS).

13. The method according to claim 1, wherein the autism spectrum disorder comprises non-syndromic autism.

14. The method according to claim 1, further comprising diagnosing a the presence of, or an increased risk of developing a genetic syndrome linked to autism, wherein the genetic syndrome comprises a manifesting phenotype.

15. The method according to claim 14, wherein the genetic syndrome comprises at least one of Angelman syndrome, Prader-Willi syndrome, 15q11-q13 duplication, fragile X syndrome, fragile X premutation, deletion of chromosome 2q, XYY syndrome, Smith-Lemli-Opitz syndrome, Apert syndrome, mutations in the ARX gene, De Lange syndrome, Smith-Magenis syndrome, Williams syndrome, Noonan syndrome, Down syndrome, velo-cardio-facial syndrome, myotonic dystrophy, Steinert disease, tuberous sclerosis, Duchenne's disease, Timothy syndrome, 10p terminal deletion, Cowden syndrome, 45,X/46,XY mosaicism, Myhre syndrome, Sotos syndrome, Cohen syndrome, Goldenhar syndrome, Joubert syndrome, Lujan-Fryns syndrome, Moebius syndrome, hypomelanosis of Ito, neurofibromatosis type 1, CHARGE syndrome, or HEADD syndrome.

16. The method according to claim 1, wherein the subject is a child.

17. The method according to claim 1, wherein the subject is a fetus.

18. The method according to claim 1, wherein the body fluid comprise at least one of cerebrospinal fluid, blood, amniotic fluid, maternal blood, and urine.

19. The method according to claim 1, wherein the variant comprises at least one of the variants in Table 2.

20. The method according to claim 1, wherein the variant comprises at least one of the following mutations: *HOMER 1* c.195G>T, M65I; *HOMER 1* c.290C>T, S97L; *HOMER 1* c.425C>T, P142L; *GRM5* c.3503T>C, L1168P; *MAPK2* c.581-1G>T; *HRAS* c.383G>A, R128Q; a *MECP2* c.1477G>T, E483X.

21. A method for identifying mutations correlated with the presence or increased risk of developing an autism spectrum disorder, the method comprising:

identifying a nucleic acid to be evaluated as having a sequence that if mutated may be or is associated with the development of autism;

obtaining a nucleic acid sample from a tissue or body fluid sample from a subject having an autism spectrum disorder; and

conducting an assay to identify whether there is a mutation in the nucleic acid sequence in the subject having autism as compared to the nucleic acid sequence in individuals who do not have an autism spectrum disorder, wherein the presence of the mutation in a subject with an autism spectrum disorder indicates that the mutation may be associated with the development of the autism spectrum disorder.

22. The method of claim 21, wherein the mutation is a variant that has been previously associated with the development of an autism spectrum disorder.

23. The method of claim 21, wherein the mutation is a previously undescribed variant.

24. The method of claim 21, further comprising determining if the mutation is expected to have a deleterious effect on at least one of gene expression and/or protein function.

25. The method of claim 21, wherein the nucleic acid sequence for which the presence or absence of a mutation is evaluated is at least a portion of a gene that encodes a protein in the metabotropic glutamate receptor signaling pathway.

26. The method of claim 21, wherein the autism spectrum disorder is non-syndromic autism.

27. An isolated nucleic acid comprising a variant sequence from at least one of the following genes, *TSC1, TSC2, MECP2, SHANK3, GRM1, GRM5, ARC, EIF4E, HOMER1, HRAS, MAP2K1, MAP2K2, RAF1, PIK3CA, PIK3R1, FMR1, PTEN, RHEB* or *UBE3A*, wherein the sequence comprises a variant that is indicative of an autism spectrum disorder.

28. The isolated nucleic acid of claim 27, comprising a variant sequence as described in Table 2.

29. The isolated nucleic acid of claim 27, comprising at least one of the following variants: *HOMER 1* c.195G>T, M65I; *HOMER 1* c.290C>T, S97L; *HOMER 1* c.425C>T, P142L; *GRM5* c.3503T>C, L1168P; *MAPK2* c.581-1G>T; *HRAS* c.383G>A, R128Q; a *MECP2* c.1477G>T, E483X.