(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 April 2016 (14.04.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/057810 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2015/054726

(22) International Filing Date:
8 October 2015 (08.10.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/061,656 8 October 2014 (08.10.2014) US

(71) Applicant: **THE BROAD INSTITUTE, INC.** [US/US]; 415 Main Street, Cambridge, MA 02142 (US).

(72) Inventors; and
(71) Applicants : **LINDBLAD-TOH, Kerstin** [US/US]; 90 Cherry Street, Malden, MA 02148-1714 (US). **NOH, Hyun, Ji** [KR/US]; 375A Harvard Street, Apt. 8A, Cambridge, MA 02111 (US). **KARLSSON, Elinor** [US/US]; 35 Bigelow Street, Apt. 3, Cambridge, MA 02139 (US). **TANG, Ruqi** [CN/CN]; Building 104, Baoshan er cun, Baoshan District, Shanghai, 200940 (CN). **FENG, Guoping** [US/US]; 242 Homer Street, Newton, MA 02459 (US).

(74) Agent: **COLANTONIO, Jessica, R.**; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210-2206 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

[Continued on next page]

(54) Title: MARKERS FOR ASSESSING RISK OF DEVELOPING OR HAVING OBSESSIVE COMPULSIVE DISORDER

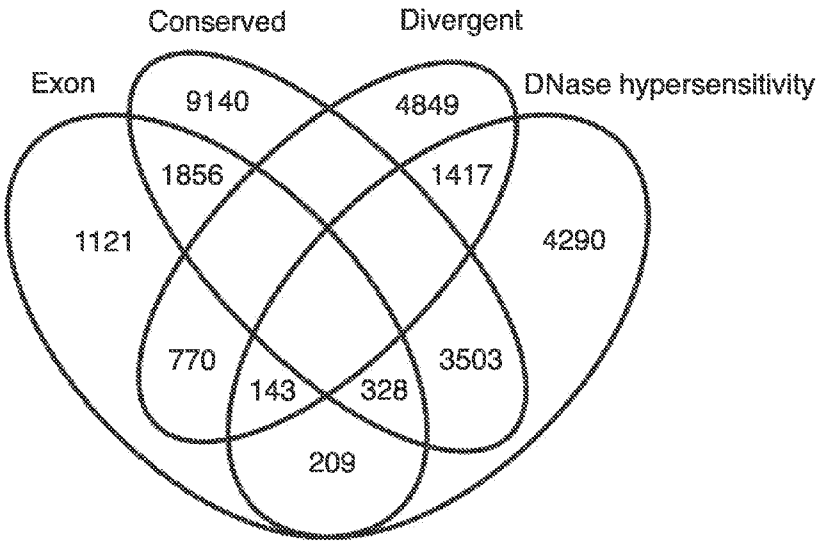


FIG. 1A

(57) Abstract: Provided herein are methods and compositions for identifying subjects as having elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder. These subjects are identified based on the presence of one or more mutations.

WO 2016/057810 A1

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

MARKERS FOR ASSESSING RISK OF DEVELOPING OR HAVING OBSESSIVE COMPULSIVE DISORDER

RELATED APPLICATIONS

This application claims the benefit under 35 U.S.C. § 119(e) of U.S. provisional application number 62/061,656, filed October 8, 2014, the contents of which are incorporated by reference herein in their entirety.

SEQUENCE LISTING

The Sequence Listing filed on October 8, 2015 as an ASCII text file is incorporated by reference herein. The ASCII text file is named B1195.70033WO00-SEQ, was created on October 8, 2015, and is 4,685 bytes in size.

BACKGROUND OF INVENTION

Neuropsychiatric disorders, such as obsessive-compulsive disorder, autism spectrum disorder, and Tourette syndrome, affect millions of people world-wide. Such neuropsychiatric disorders can hamper the quality of life of affected individuals. Such disorders are often inherited, but the genetic factors are not well-understood.

Obsessive-compulsive disorder (OCD), a severe neuropsychiatric disorder manifested in time-consuming repetition of behaviors, affects 1-3% of the human population. While highly heritable, complex genetics have hampered attempts to elucidate OCD etiology.

SUMMARY OF INVENTION

The invention is premised, in part, on a study involving targeted sequencing of both coding and functional non-coding regions indicated by evolutionary conservation, for a heritable psychiatric disease, obsessive-compulsive disorder (OCD). Six hundred and eight (608) genes were surveyed in a small cohort to identify genes and variants that correlated with OCD. Five genes were identified with significant mutation loads in OCD after multiple testing: *NRXN1*, *HTR2A*, *CTTNBP2*, *REEP3,* and *LIPH*. Two hundred and eighteen (218) variants associated with these genes were identified and regulatory functions were validated for a subset of the variants. Exemplary variants were found in coding regions as well as in or near functional non-coding regions including DNase hypersensitivity sites and untranslated regions (UTRs).

Accordingly, aspects of the invention relate to methods for identifying subjects at elevated risk of developing a neuropsychiatric disorder (e.g., OCD) or to identify subjects having a neuropsychiatric disorder (e.g., OCD). Other aspects of the invention relate to methods of treatment of subjects identified according to the methods set forth herein as having a neuropsychiatric disorder (e.g., OCD) or having an elevated risk of developing a neuropsychiatric disorder (e.g., OCD).

In some aspects, the invention provides a method, comprising:

(a) analyzing genomic DNA from a subject for the presence of a mutation within or near a gene selected from NRXN1, CTTNBP2, HTR2A, REEP3, or LIPH; and

(b) identifying a subject having the mutation as a subject at elevated risk of developing a neuropsychiatric disorder or as a subject having a neuropsychiatric disorder,

wherein if the gene is HTR2A, the mutation is within an exon or is a SNP provided in Table 2.

In some embodiments, the mutation is within 100 kb, upstream or downstream, of the chromosomal boundaries/co-ordinates provided in Table 1. In some embodiments, the gene is selected from CTTNBP2 or REEP3. In some embodiments, the mutation is within an untranslated region (UTR), exon, or DNase1 hypersensitivity site of the gene.

In some embodiments of any one of the methods described, the gene is NRXN1 and the mutation is within the isoform AK093260. In some embodiments of any one of the methods described, the gene is NRXN1 and the mutation is a SNP located at a chromosome location selected from chr2:51256161, chr2:50762143, chr2:51153020, chr2:50733992, chr2:51000979, chr2:50842279, chr2:51067620, chr2:50606585, chr2:50927403, chr2:50703012, chr2:50956849, chr2:50606521, chr2:50733958, chr2:50733841, chr2:50755127, chr2:50542571, chr2:50619213, chr2:50699305, chr2:50927347, chr2:50448284, chr2:50400991, chr2:50924511, chr2:50850245, chr2:50570754, chr2:50171755, chr2:50343508, chr2:50762346, chr2:51148378, chr2:51244839, chr2:50607797, chr2:50941362, chr2:51236675, chr2:50187207, chr2:50848555, chr2:50198372, chr2:50981817, chr2:50693162, chr2:50570237, chr2:50683609, chr2:50849551, chr2:50735998, chr2:51246886, chr2:50200776, chr2:50323549, chr2:50675110, chr2:50922035, chr2:51057550, chr2:50386107, chr2:50386080, chr2:50847195, chr2:51252712, chr2:51245440, chr2:50354237, chr2:50719598,

chr2:50952610, chr2:50792080, chr2:50542527, chr2:50750575, chr2:50155007, chr2:50779791, chr2:50733581, chr2:50400809, chr2:50201255, chr2:50178130, chr2:51146148, chr2:50575137, chr2:51148372, chr2:51171979, chr2:50779943, chr2:50848551, chr2:50165016, chr2:51149889, chr2:50774153, chr2:50389636, chr2:50434866, chr2:50724642, chr2:50981813, chr2:51085557, chr2:50463984, chr2:50724745, chr2:50981807, chr2:50598207, chr2:50675639, chr2:50653833, chr2:51145459, chr2:50542372, chr2:50952571, chr2:50548140, chr2:50765412, chr2:50850686, chr2:50934666, chr2:50682914, chr2:50709350, chr2:50979527, chr2:50386109, chr2:50542308, chr2:50607943, chr2:50735814, chr2:50981815, chr2:50155737, chr2:50683701, chr2:50842256, chr2:50148728, chr2:50952482, chr2:51153206, chr2:50560998, chr2:50996952, chr2:50458593, chr2:50924466, chr2:51005207, chr2:50602031, chr2:50178059, chr2:50850340, chr2:51016384, chr2:50175865, chr2:50571910, chr2:50570602, chr2:50548103, chr2:50518040, chr2:50236859, chr2:50464065, chr2:50598321, chr2:50282777, chr2:51245472, chr2:50735943, chr2:50927534, chr2:50941367, chr2:50952709, chr2:51067726, chr2:51079254, chr2:50277539, chr2:50424938, chr2:50765589, chr2:50699377, chr2:51149368, chr2:50723068, chr2:50723000, chr2:51245656, chr2:50571784, chr2:50148783, chr2:50598280, chr2:50850307, chr2:50850394, chr2:50563875, chr2:50614848, chr2:50531295, chr2:50877741, chr2:50733745, chr2:50919652, chr2:50570601, chr2:50981811, or chr2:51021463. In some embodiments of any one of the methods described, the mutation is a SNP located at a chromosome location selected from chr2:50847195, chr2:50779791, chr2:50779943, chr2:50724642, chr2:50463984, chr2:50724745, chr2:50765412, chr2:50850686, chr2:50464065, chr2:50765589, chr2:50723068, or chr2:50733745.

In some embodiments of any one of the methods described, the gene is CTTNBP2 and the mutation is within or near a DNase1 hypersensitivity site or within an exon of CTTNBP2. In some embodiments of any one of the methods described, the gene is CTTNBP2 and the mutation is a SNP located at a chromosome location selected from chr7:117430669, chr7:117358107, chr7:117431704, chr7:117396664, chr7:117374935, chr7:117391129, chr7:117368123, chr7:117446174, chr7:117456904, chr7:117356081, chr7:117427551, chr7:117354909, chr7:117452215, chr7:117431202, chr7:117358129,

chr7:117359713, chr7:117457141, chr7:117450810, chr7:117431879, chr7:117386178, chr7:117385978, chr7:117468334, chr7:117396706, chr7:117501314, chr7:117390966, chr7:117354258, chr7:117352306, chr7:117351979, chr7:117431079, chr7:117417559, chr7:117427686, chr7:117421141, or chr7:117468056. In some embodiments, the mutation is a SNP located at a chromosome location selected from chr7:117456904, chr7:117356081, chr7:117450810, chr7:117390966, chr7:117417559, chr7:117421141, or chr7:117468056.

In some embodiments of any one of the methods described, the gene is HTR2A and the mutation is within an exon of HTR2A. In some embodiments, the mutation is within the last exon of HTR2A. In some embodiments of any one of the methods described, the gene is HTR2A and the mutation is a SNP located at a chromosome location selected from chr13:47454997, chr13:47440198, chr13:47409048, chr13:47440301, chr13:47466592, chr13:47418543, chr13:4743474, chr13:47448370, chr13:47466622, chr13:47409701, chr13:47440209, chr13:47421746, chr13:47418629, chr13:47408946, chr13:47455071, or chr13:47469335. In some embodiments, the mutation is a SNP located at a chromosome location selected from chr13:47409048, chr13:47466622, or chr13:47409701.

In some embodiments of any one of the methods described, the gene is REEP3 and the mutation is within or near a DNase1 hypersensitivity site of REEP3. In some embodiments of any one of the methods described, the gene is REEP3 and the mutation is a SNP located at a chromosome location selected from chr10:65339450, chr10:65368263, chr10:65358911, chr10:65326034, chr10:65359513, chr10:65332906, chr10:65354650, chr10:65357754, chr10:65287863, chr10: 65387644, chr10: 65387722, chr10: 65388750, chr10: 65384621, or chr10:65307923. In some embodiments, the mutation is a SNP located at a chromosome location selected from chr10:65332906, chr10: 65387644, chr10: 65387722, chr10: 65388750, chr10: 65384621, or chr10:65307923.

In some embodiments of any one of the methods described, the gene is LIPH and the mutation is within an untranslated region (UTR), intron, or exon of LIPH. In some embodiments of any one of the methods described, the mutation is within the 3'UTR of LIPH or near a splice site of LIPH. In some embodiments of any one of the methods described, the gene is LIPH and the mutation is a SNP located at a chromosome location selected from chr3:185241792, chr3:185229283, chr3:185226492, chr3:185229464, chr3:185226396, chr3:185225638, or chr3:185225644. In some embodiments, the mutation is a SNP located

at a chromosome location selected from chr3:185226492, chr3:185226396, chr3:185225638, or chr3:185225644.

In some embodiments of any one of the methods described, the genomic DNA is obtained from a bodily fluid or tissue sample of the subject. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a bead array. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a nucleic acid sequencing assay.

In some embodiments of any one of the methods described, the subject is a human subject.

In some embodiments of any one of the methods described, the method further comprises:

(c) administering a therapeutic agent to the subject identified as at elevated risk of developing a neuropsychiatric disorder or as a subject having a neuropsychiatric disorder.

In some embodiments of any one of the methods described, the method further comprises:

(c) performing behavioral therapy on the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

In some embodiments of any one of the methods described, the neuropsychiatric disorder is obsessive-compulsive disorder.

In some embodiments of any one of the methods described, the mutation is at least two mutations. In some embodiments of any one of the methods described, the gene is at least two genes.

In another aspect, the invention provides a method, comprising:

(a) analyzing genomic DNA from a subject for the presence of a SNP in Table 2; and

(b) identifying the subject having the SNP as a subject at elevated risk of developing or as a subject having a neuropsychiatric disorder. In some embodiments, the SNP is a SNP located at a chromosome location selected from chr2:50847195, chr2:50779791, chr2:50779943, chr2:50724642, chr2:50463984, chr2:50724745, chr2:50765412, chr2:50850686, chr2:50464065, chr2:50765589, chr2:50723068, chr2:50733745, chr7:117456904, chr7:117356081, chr7:117450810, chr7:117390966, chr7:117417559,

chr7:117421141, chr7:117468056, chr13:47409048, chr13:47466622, chr13:47409701, chr10:65332906, chr10: 65387644, chr10: 65387722, chr10: 65388750, chr10: 65384621, chr10:65307923, chr3:185226492, chr3:185226396, chr3:185225638, or chr3:185225644.

In some embodiments of any one of the methods described, the genomic DNA is obtained from a bodily fluid or tissue sample of the subject. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a bead array. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a nucleic acid sequencing assay.

In some embodiments of any one of the methods described, the method further comprises:

(c) administering a therapeutic agent to the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

In some embodiments of any one of the methods described, the method further comprises:

(c) performing behavioral therapy on the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

In some embodiments of any one of the methods described, the neuropsychiatric disorder is obsessive-compulsive disorder. In some embodiments of any one of the methods described, the mutation or SNP is two mutations or SNPs.

Yet other aspects of the invention provide a method, comprising:

(a) analyzing genomic DNA from a subject for the presence of a first mutation within or near HTR2A and a second mutation within a gene selected from LIPH, NRXN1, CTTNBP2, or REEP3; and

(b) identifying a subject having the first and second mutation as a subject at elevated risk of developing or as a subject having a neuropsychiatric disorder. In some embodiments, the first mutation and the second mutation are each independently within 100 kb, upstream or downstream, of the chromosomal boundaries/co-ordinates provided in Table 1. In some embodiments, the mutation is within an untranslated region (UTR), exon, or DNase1 hypersensitivity site of the gene.

In some embodiments of any one of the methods described, the first mutation is within an exon of HTR2A. In some embodiments of any one of the methods described, the first mutation is within the last exon of HTR2A. In some embodiments of any one of the methods described, the gene is HTR2A and the second mutation is a SNP provided in Table 2.

In some embodiments of any one of the methods described, the gene is LIPH and the second mutation is within an untranslated region (UTR), intron, or exon of LIPH. In some embodiments of any one of the methods described, the second mutation is within the 3'UTR of LIPH or near a splice site of LIPH. In some embodiments of any one of the methods described, the gene is LIPH and the second mutation is a SNP provided in Table 2.

In some embodiments of any one of the methods described, the gene is NRXN1 and the second mutation is within the isoform AK093260. In some embodiments of any one of the methods described, the gene is NRXN1 and the second mutation is a SNP provided in Table 2.

In some embodiments of any one of the methods described, the gene is CTTNBP2 and the second mutation is within or near a DNase1 hypersensitivity site or within an exon of CTTNBP2. In some embodiments of any one of the methods described, the gene is CTTNBP2 and the second mutation is a SNP provided in Table 2.

In some embodiments of any one of the methods described, the gene is REEP3 and the second mutation is within or near a DNase1 hypersensitivity site of REEP3. In some embodiments of any one of the methods described, the gene is REEP3 and the second mutation is a SNP provided in Table 2.

In some embodiments of any one of the methods described, the genomic DNA is obtained from a bodily fluid or tissue sample of the subject. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a bead array. In some embodiments of any one of the methods described, the genomic DNA is analyzed using a nucleic acid sequencing assay.

In some embodiments of any one of the methods described, the subject is a human subject.

In some embodiments of any one of the methods described, the method further comprises:

(c) administering a therapeutic agent to the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

In some embodiments of any one of the methods described, the method further comprises:

(c) performing behavioral therapy on the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

In some embodiments of any one of the methods described, the neuropsychiatric disorder is obsessive-compulsive disorder.

BRIEF DESCRIPTION OF DRAWINGS

The following drawings form part of the present specification and are included to further demonstrate certain aspects of the present disclosure. These aspects can be better understood by reference to one or more of these drawings in combination with the detailed description of specific embodiments presented herein.

FIGs. 1A and 1B show a diagram and a graph showing gene burden analyses. **FIG. 1A** shows the number and the overlaps of the detected variants in four categories, i.e. Exon, DNase hypersensitivity, (evolutionarily) Conserved (GERP>2) and (evolutionarily) Divergent (GERP<-2) categories. **FIG. 1B** shows P-value distribution of 608 genes from 12 burden tests stratified by variant types. 'Overall' category is for any variant types, 'Exon' category is for coding variants, and 'DHS' category is for regulatory variants indicated by DNase hypersensitivity. The three categories are further stratified by the evolutionary status of a variant site, where 'Cons.' means evolutionarily conserved sites, 'Div.' means evolutionarily divergent sites, and 'Evo.' means 'Cons.' and 'Div.' combined. Each circle represents a gene, whose location is determined by the p-values indicated on the y-axis. The horizontal red line indicates the threshold at p-value<0.05 after jointly corrected for the number of genes and stratified tests performed. Five significantly associated genes are labeled, with red color-code for the gene with 'Overall' mutation burden, blue colors for the

genes with 'Exon' mutation burden, and green colors for the genes with 'DHS' mutation burden.

FIGs. 2A-E are a series of graphs showing candidate variants of the associated genes. **FIG. 2A** shows LIPH (44.8Kb). **FIG. 2B** shows NRXN1 (1.1 Mb). **FIG. 2C** shows HTR2A (65.5Kb). **FIG. 2D** shows CTTNBP2 (163Kb). **FIG. 2E** shows REEP3 (104Kb). '-$Log_{10}P_{single}$' shows the negative log value of the t-test's p-value for individual variants. 'AF ca/co ratio' means the ratio of case allele frequency over control allele frequency. In the 'Candidate variant' track, each bar represents the location of candidate variants, and red represents missense, green represents synonymous coding variant, blue represents DHS variant, and black represents no functional annotation. The red numbers at the right end shows the number of [candidate variants] ([of which genotyped in independent set]) / [all detected variants]. The 'Conserved elements' track shows the mammalian conserved elements (PhastCons), from which we determined the regions sequenced (grey blocks above the green peaks).

FIGs. 3A and **3B** are a diagram and a photograph of a gel shift showing examples of candidate 'divergent' sites within mammalian constraints and functional validation of candidate regulatory mutations. **FIG. 3A** shows sequences for all primates and placental mammals that have alignments for both example positions, based on UCSC Multiz Alignments of 46 Vertebrates. **FIG. 3B** shows electrophoretic mobility shift assay (EMSA) with nuclear extracts from human neuroblasts for candidate regulatory variants in REEP3 and CTTNBP2. Arrows indicate allele-specific gel shifts. WT, wild-type allele; MT, candidate variant. Excessive cold probes outcompete biotin-labeled probes, verifying specific DNA-protein bindings.

FIG. 4 is a drawing showing the proposed mechanism at synapses for genes associated with OCD. The disclosure and the claims are not bound to or limited by this proposed mechanism.

FIG. 5A shows mean read depth coverage of the targeted regions in each pool. Red bars indicate case pools and blue bars indicate control pools.

FIG.5B shows the strong correlation of allele frequency calls for 41504 SNPs high-confidence variants between two different variant calling algorithms, Syzygy and SNVer.

FIGs. 6A and B show the strong correlation of allele frequency calls on a subset of detected variants between sequencing experiment and Sequenom genotyping experiment for case (A) and control (B) samples. SEQ, sequencing; GENO, genotyping.

FIGs. 7A-E show the concordance between sequencing and genotyping experiments in various variant categories (e.g. allele frequency, call confidence) and predicted concordance with simulated data for missing genotypes.

FIGs. 8A-E show quantile-quantile plots for gene-based burden tests (A-C, and E) and summary results for gene-based burden test considering rare variants only (D). The x-axis for each plot in FIGs. 8A-C and E is Empirical $-\log_{10}(p)$. The y-axis for each plot in FIGs. 8A-C and E is Observed $-\log_{10}(p)$.

FIG. 9 shows the distribution of minimum P values generated from gene-based burden test on permuted data (case-control label permutation) that was used for multiple testing correction.

FIG. 10 is a heatmap showing the genes' p-values for 16 gene-based burden tests performed. Y-axis shows individual 608 genes and X-axis shows individual 16 tests performed.

FIG. 11A is a quantile-quantile plot for polygenic burden test on 989 candidate GO terms.

FIG. 11B shows the minimum P distribution generated from permutation for multiple testing correction.

FIGs. 12A-C show diagrams of functional or bioinformatic evidences for a subset of candidate variants, such as DNase hypersensitivity, transcription factor binding sites, histone marks and mammalian conservations.

FIG. 13 is a graph showing the comparison of p-values between 5 genes detected from OCD animal models (*CDH2, PGCP, ATXN1, CTTNBP2, DLGAP3*) and the rest of the targeted genes.

FIGs. 14A-B are photographs of electrophoretic mobility shift assays (EMSA) results, using nuclear extracts from human neuroblasts, for candidate regulatory variants in *REEP3* and *CTTNBP2*. WT, wildtype (reference allele); MT, mutant (candidate mutation). Excessive cold probes outcompete biotin-labeled probes, verifying specific DNA-protein bindings.

DETAILED DESCRIPTION OF INVENTION

Aspects of the invention relate to mutations (such as single nucleotide polymorphisms (SNPs) and other mutations, in or near genes) and various methods of use and/or detection thereof. The invention is premised, in part, on results from targeted sequencing of both coding and functional noncoding regions indicated by evolutionary conservation, for a heritable psychiatric disease, obsessive-compulsive disorder (OCD). Six hundred and eight (608) genes were surveyed in a small cohort to identify genes and variants that correlated with OCD. Five genes were identified with significant mutation loads in OCD after multiple testing correction: *NRXN1*, *HTR2A*, *CTTNBP2*, *REEP3,* and *LIPH*. Two hundred and eighteen (218) candidate variants associated with these genes were identified and regulatory functions were validated for a subset of the variants.

Accordingly, aspects of the invention provide methods that involve detecting a mutation (e.g., one or more mutations) within a region surrounding a gene (e.g., within 100 kilobases (kb) on either side of a gene) and using such detection to identify subjects having an elevated risk of developing a neuropsychiatric disorder (e.g., OCD) or to identify subjects having a neuropsychiatric disorder (e.g., OCD).

Identifying subjects having an elevated risk of developing a neuropsychiatric disorder and identifying subjects having a neuropsychiatric disorder is useful in a number of applications. For example, the methods can be used for prognostic purposes and for diagnostic purposes. Accordingly, the invention provides diagnostic and prognostic methods for use in subjects, such as human subjects. In some embodiments, such diagnostic or prognostic methods can be paired with a treatment (e.g., a therapeutic agent or behavioral therapy). Subjects identified as at elevated risk may be monitored, including monitored more regularly, for the appearance of disorder-like symptoms and/or may be treated prophylactically (e.g., prior to the development of the symptoms) or therapeutically (e.g., after the appearance of symptoms or after the diagnosis is made). Animal subjects carrying one or more of the mutations described herein may also be used to further study the neuropsychiatric disorders and optionally to study the efficacy of various treatments.

Elevated risk of developing a neuropsychiatric disorder a neuropsychiatric disorder

The mutations of the invention can be used to identify subjects at elevated risk of developing a neuropsychiatric disorder or to identify subjects having a neuropsychiatric disorder. An elevated risk means a lifetime risk, or a risk within a certain amount of time during the lifespan of a subject, of developing or having such a disorder that is higher than the risk of developing or having the same disorder in (a) a population that is unselected for the presence or absence of the mutation (i.e., the general population) or (b) a population that does not carry the mutation.

Neuropsychiatric disorder and diagnostic/prognostic methods

Aspects of the invention include various methods, such as prognostic and diagnostic methods, related to neuropsychiatric disorders. Non-limiting examples of neuropsychiatric disorders include obsessive-compulsive disorder, autism spectrum disorder, Tourette syndrome, and obsessive-compulsive spectrum such as dermatillomania, trichotillomania, and onychophagia.

Obsessive-compulsive disorder (OCD) is disorder characterized by intrusive, persistent thoughts (obsessions) and/or repetitive, intentional behaviours (compulsions) that result in significant distress or dysfunction. It affects 1 to 3% of the general population. In humans, symptoms of the disorder include excessive washing or cleaning; repeated checking; extreme hoarding; preoccupation with sexual, violent or religious thoughts; relationship-related obsessions; aversion to particular numbers; and nervous rituals, such as opening and closing a door a certain number of times before entering or leaving a room.

Diagnosis of OCD generally involves identifying obsessions, compulsions, or both that are "fixed" (e.g., present for a certain length of time) in a subject. Diagnosis of human subjects may be made according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases, 10th Edition (ICD). Obsessions include distressing ideas, images, or impulses that enter a subject's mind repeatedly. The obsessions are often violent, obscene, or perceived to be senseless and the subject finds these ideas difficult to resist. Compulsions include stereotyped behaviors that are not enjoyable that are repeated over and over and are perceived to prevent an unlikely event that is in reality unlikely to occur. The subject often recognizes that the behavior is ineffectual and makes

attempts to resist it, but is unable to. Compulsions may also include repetitive behaviours or mental acts that are carried out to reduce or prevent anxiety or distress and are perceived to prevent a dreaded event or situation.

The diagnostic criteria for OCD, according to the DSM, are as follows:

1. Obsessional symptoms or compulsive acts or both must be present on most days for at least 2 successive weeks and be a source of distress or interference with activities.

2. Obsessional symptoms should have the following characteristics:

a. they must be recognized as the individual's own thoughts or impulses.

b. there must be at least one thought or act that is still resisted unsuccessfully, even though others may be present which the sufferer no longer resists.

c. the thought of carrying out the act must not in itself be pleasurable (simple relief of tension or anxiety is not regarded as pleasure in this sense).

d. the thoughts, images, or impulses must be unpleasantly repetitive.

Autism Spectrum Disorder (ASD) is a developmental disorder characterized by abnormalities in social interactions and communication, as well as restricted interests and repetitive behaviours. ASD may be diagnosed using the DSM, which provides diagnostic criteria for identifying ASD. The criteria include persistent deficits in social communication and social interaction combined with restricted, repetitive patterns of behavior, interests, or activities.

Tourette syndrome is a disorder generally having onset in childhood, characterized by multiple physical (motor) tics and at least one vocal (phonic) tic. Tourette's may be diagnosed using the DSM. The diagnostic criteria include that a person exhibits both multiple motor and one or more vocal tics (although these do not need to be concurrent) over the period of a year, with no more than three consecutive tic-free months.

Dermatillomania is characterized by the repeated urge to pick at one's own skin, often to the extent that damage is caused. Dermatillomania may be classified as an impulse control disorder by DSM-IV. Trichotillomania is characterized by compulsive urge to pull out one's own hair leading to noticeable hair loss, distress, and social or functional impairment. Trichotillomania may be classified as an impulse control disorder by DSM-IV. Onychophagia is an oral compulsive habit characterized by nail biting. Nail biting is considered an impulse

control disorder in the DSM-IV-R, and is classified under obsessive-compulsive and related disorders in the DSM-5.

In some embodiments, diagnostic methods include measuring a mutation as described herein in combination with a known diagnostic method (e.g., a behavioral test or use of a questionnaire or assessment provided in DSM IV, DSM IV-R, or DSM 5).

Mutations

Aspects of the invention relate to a (i.e., at least one) mutation and uses and detection of such mutation(s) in various methods. As used herein, a mutation is one or more changes in the nucleotide sequence of the genome of the subject. As used herein, mutations include, but are not limited to, point mutations (e.g., SNPs), insertions, deletions, rearrangements, inversions and duplications. Mutations also include, but are not limited to, silent mutations, missense mutations, and nonsense mutations. In some embodiments, the mutation is a SNP. SNPs are further described herein.

The mutation can be a germ-line mutation or a somatic mutation. In some embodiments, the mutation is a germ-line mutation. A germ-line mutation is generally found in the majority, if not all, of the cells in a subject. Germ-line mutations are generally inherited from one or both parents of the subject (i.e., were present in the germ cells of one or both parents). Germ-line mutations as used herein also include *de* novo germ-line mutations, which are spontaneous mutations that occur at single-cell stage level during development. A somatic mutation occurs after the single-cell stage during development. Somatic mutations are considered to be spontaneous mutations. Somatic mutations generally originate in a single cell or subset of cells in the subject.

A mutation as described herein may be found within a gene described herein or within a region encompassing such a gene (e.g., a region that encompasses the gene as well as 100 kb or more upstream and 100 kb or more downstream of the gene).

A mutation as described herein may be a private mutation. As used herein, the term "private mutation" generally refers to a rare gene mutation that is usually found only in a single family or a small population.

A mutation as described herein may be a synonymous mutation. A codon in RNA is a set of three nucleotides that encode a specific amino acid. Most amino acids have several

RNA codons that are translated into that particular amino acid. As used herein, the term "synonymous mutation" refers to a mutation which causes no change to the amino acid that is encoded by the mutated RNA codon as compared to the non-mutated RNA codon. In some embodiments, a synonymous mutation changes only a single base pair in the RNA copy of the DNA.

A mutation as described herein may be a non-synonymous mutation. As used herein, the term "non-synonymous mutation" refers to a mutation which does cause a change to one or more amino acids that are encoded by the mutated RNA codon(s) as compared to the non-mutated codon. In some embodiments, a non-synonymous mutation is an insertion or a deletion of one or more nucleotides, e.g., nucleotides that encode a specific amino acid. In some embodiments, a non-synonymous mutation is a frame shift mutation. In some embodiments, a non-synonymous mutation leads to the introduction of an early stop codon, e.g., nonsense mutation. In some embodiments, a non-synonymous mutation leads to a stop codon becoming a codon for an amino acid, e.g., run-on mutation.

Genes

In some embodiments, a mutation provided herein is a mutation within or near a gene. In some embodiments, the gene is a gene provided in Table 1. The boundaries of each gene are defined using the "start" and "end" coordinates provided in columns 3 and 4 respectively of Table 1 for human subjects. It is to be understood that these coordinates are inclusive (i.e., including the boundaries).

The start and end coordinates (i.e., the chromosome coordinates) and the Ensembl Gene IDs in Table 1 are based on the 19th human genome assembly (Hg19, see, e.g., UCSC Genome Browser). With regard to the coordinates, the first base pair in each chromosome is labeled 0 and the position of the start and end is then the number of base pairs from the first base pair.

A gene may include regulatory sequences (e.g., promoters, enhancers, suppressors, or sequences associated with DNase1 hypersensitivity, either adjacent to or far from the coding sequence) and coding sequences. As used herein, a coding sequence includes the first DNA nucleotide to the last DNA nucleotide that is transcribed into an mRNA that includes the untranslated regions (UTRs), exons, and introns. The coding sequence for each gene can be

obtained using the Ensembl database by entering the Ensembl gene IDs provided in Table 1, or by other methods known in the art. In some embodiments, the mutation is within or near (e.g., within 100 kb of) the coding sequence of a gene. Thus, it is to be understood that this disclosure provides for detecting mutations within or near "genes" or within or near coding sequence, and that although many embodiments are described relative to "gene" co-ordinates this is for the sake of brevity only and the disclosure contemplates and provides parallel embodiments relative to coding sequence (and its co-ordinates) as well.

In some embodiments, a mutation, such as a SNP, is contained within or near the gene, such as in a DNase1 hypersensitivity site (DHS), an exon, or UTR. DHSes can be identified, e.g., using the ENCODE database (e.g., the DNaseI Hypersensitivity Uniform Peaks from ENCODE) or performing DNase1 hypersensitivity assays known in the art, such as DNAse-seq/DNAase-ChIP assays (see, e.g., Song and Crawford, 2010; Boyle et al., 2008a; Crawford et al., 2006). In an exemplary method, cells are lysed with NP40, and intact nuclei are digested with optimal levels of DNaseI enzyme. DNaseI-digested ends are captured from three different DNase concentrations, and material is sequenced. Alternatively, for example, material is hybridized to a NimbleGen Human ENCODE tiling arrays (1% of the genome) along with the input DNA as reference.

In some embodiments, a mutation is near a gene if the mutation is within 5000 kb, 2500 kb, 1000 kb, 900 kb, 800 kb, 700 kb, 600 kb, 500 kb, 400 kb, 300 kb, 200 kb, 150 kb, 100 kb, 50 kb, 25 kb, 10 kb, or 5 kb of a gene or of the coding sequence of the gene, as described herein, such as an UTR, exon, or DHS of the gene. In some embodiments, a mutation is contained within the boundaries provided in the "start - 100 kb (or more)" column and the "end + 100 kb (or more)" column of Table 1.

In some embodiments, a mutation may be in a functional non-coding region, such a region containing a DNase1 hypersensitivity site.

Table 1. Human Genes

Gene	HG19 Chr	HG19 start (bp)	HG19 end (bp)	HG19 start - 100 kb	HG19 end +100 kb	Ensembl ID
LIPH	3	185,225,570	185,270,369	185,125,570	185,370,369	ENSG00000163898
NRXN1	2	50,145,643	51,259,674	50,045,643	51,359,674	ENSG00000179915
HTR2A	13	47,405,677	47,471,211	47,305,677	47,571,211	ENSG00000102468

CTTNBP2	7	117,350,706	117,513,561	117,250,706	117,613,561	ENSG00000077063
REEP3	10	65,281,306	65,380,629	65,181,306	65,480,629	ENSG00000165476

In some embodiments, the gene is a gene identified in a pathway described herein or a gene as described in Table 4 or Table 5. In some embodiments, the pathway is 5-HT signaling or LPA signaling or a pathway involved in cell adhesion, programmed cell death, or cell migration including telencephalic tangential migration.

In some embodiments, a mutation in a gene or within a region encompassing a gene (e.g., a region that includes the gene plus 100 kb or 150 kb upstream and 100 kb or 150 kb downstream of the gene) is used in the methods described herein. In some embodiments, the method comprises:

(a) analyzing genomic DNA from a subject for the presence of a mutation (i) within a gene (e.g., within and including the start and end coordinates provided in columns 3 and 4 of Table 1) and/or (ii) near a gene (e.g., within 150 kb, 100 kb, 50 kb, 25 kb, 10 kb, or 5 kb of the start and end coordinates provided in columns 3 and 4 of Table 1) and/or (iii) within and including the coordinates provided in columns 5 and 6 of Table 1); and

(b) identifying a subject having the mutation as a subject at elevated risk of developing a neuropsychiatric disorder or as a subject having a neuropsychiatric disorder. It is to be understood that the start and end coordinates in Table 1 are coordinates on the chromosome number provided in column 2.

It is to be understood that any number of mutations (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 or more mutations) in or near any number of genes (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 or more genes) are contemplated. Any mutation of any size located within or near a gene is contemplated herein, e.g., a SNP, a deletion, an inversion, a translocation, or a duplication. In some embodiments, the mutation is a SNP. In some embodiments, the gene is two genes, the first gene being HTR2A and the second gene being selected from LIPH, NRXN1, CTTNBP2, and REEP3.

In some embodiments, the mutation is within or near a gene, wherein the gene is selected from LIPH, NRXN1, HTR2A, CTTNBP2, and REEP3. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10 (or more) mutations are within or near 1, 2, 3, 4, or 5 genes, wherein

the genes are 1, 2, 3, 4, or all 5 of LIPH, NRXN1, HTR2A, CTTNBP2, and REEP3. In some embodiments, LIPH and/or HTR2A are excluded.

In some embodiments, the gene is LIPH and the mutation is within an untranslated region (UTR), intron, or exon of LIPH. In some embodiments, the mutation is within the 3'UTR of LIPH or near a splice site of LIPH. Exemplary mutations that fall within such regions are provided, e.g., in Table 2.

In some embodiments, the gene is NRXN1 and the mutation is within the regions of NRXN1 that encode the isoform AK093260. The sequence of AK093260 is provided below. Exemplary mutations that fall within such regions are provided, e.g., in Table 2.

ATTTGTGACTTCTTCAGACCACTTTCATGACTTCTGGAAGCAAATGTCCTAGTACAGACTCAGGAAAAGA
AGAATATATTGCCACGTTCAAAGGATCTGAATACTTCTGCTACGACTTGTCTCAAAACCCCATTCAAAGC
AGCAGTGATGAAATAACTCTGTCATTTAAAACCCTTCAGAGGAATGGACTGATGCTTCACACTGGGAAAT
CGGCTGATTATGTCAATCTTGCCCTGAAAAATGGAGCTGTCTCTCTGGTCATTAATTTGGGATCAGGGGC
CTTTGAAGCACTAGTGGAGCCTGTGAATGGAAAGTTTAATGATAATGCCTGGCATGATGTGAAAGTCACC
AGGAATCTGCGTCAGGTGACAATATCAGTGGATGGGATTCTTACCACAACGGGCTACACGCAAGAAGATT
ATACCATGCTGGGGTCTGATGACTTTTTCTATGTTGGAGGCAGTCCCAGCACAGCCGACCTTCCAGGGTC
ACCAGTCAGTAACAACTTTATGGGCTGTCTCAAAGAGGTTGTATATAAAAATAATGATGTGAGGCTGGAA
TTATCTCGACTTGCCAAGCAAGGAGATCCTAAGATGAAGATCCATGGAGTGGTGGCATTTAAATGTGAGA
ATGTTGCAACTTTAGACCCAATCACCTTTGAAACCCCAGAGTCTTTCATCTCTTTGCCTAAATGGAATGC
AAAGAAAACTGGCTCCATATCATTTGATTTCCGTACAACAGAGCCAAATGGCCTCATCTTATTTAGCCAT
GGCAAGCCAAGACATCAGAAAGATGCCAAGCACCCACAGATGATAAAGGTGGACTTCTTTGCTATTGAGA
TGCTAGATGGCCACCTCTACCTCCTCCTGGACATGGGGTCAGGTACTATAAAAATAAAAGCCCTGTTGAA
GAAAGTGAATGATGGAGAATGGTATCATGTGGACTTCCAGAGAGACGGACGGTCAGGTACCATTTCTGTC
AACACGTTGCGTACTCCCTACACTGCTCCTGGTGAGAGTGAGATTCTGGACCTGGATGATGAGTTGTACC
TGGGGGGGCTGCCAGAAAATAAAGCTGGCCTTGTCTTCCCCACCGAGGTGTGGACTGCTCTGCTCAACTA
TGGCTACGTGGGCTGCATCAGGGATTTGTTCATCGATGGCCAAAGCAAAGATATCCGGCAAATGGCTGAA
GTTCAAAGTACTGCTGGAGTGAAGCCTTCCTGCTCAAAGGAAACAGCAAAACCGTGCCTTAGCAACCCTT
GCAAAAACAATGGCATGTGCAGGGATGGGTGGAACAGATATGTCTGTGATTGTTCCGGAACAGGCTATCT
TGGCAGGTCCTGTGAGAGAGAGGCAACGGTTTTGAGCTATGATGGGAGCATGTTTATGAAAATTCAGCTC
CCCGTAGTCATGCATACGGAGGCTGAGGATGTTTCCTTACGGTTCCGATCCCAGCGTGCATATGGCATTC
TGATGGCAACCACTTCTAGAGACTCTGCTGACACCCTCCGCCTGGAGCTAGACGCAGGACGTGTGAAACT
GACGGTCAATCTAGGCAAAGGTCCCGAGACTCTTTTTGCTGGCTATAACCTCAATGATAACGAGTGGCAC
ACAGTGCGTGTAGTTCGGCGTGGAAAAAGTTTAAAGTTAACAGTGGATGACCAACAGGCCATGACAGGTC
AAATGGCAGGTGATCATACTAGGCTGGAGTTCCATAACATAGAGACTGGCATCATCACAGAACGACGGTA
TCTTTCTTCTGTCCCCTCCAACTTCATTGGACACCTGCAGAGCTTGACATTTAATGGAATGGCATACATT
GACCTGTGTAAAAATGGCGACATAGATTACTGTGAGCTTAATGCCAGATTTGGCTTCAGGAACATCATAG
CAGATCCTGTCACCTTCAAGACCAAATCGAGCTATGTTGCCTTAGCTACCTTGCAAGCCTACACTTCTAT
GCATCTTTTTTTCCAGTTCAAGACAACATCCCTAGATGGATTAATTCTATATAACAGTGGGGATGGAAAT
GACTTTATTGTGGTTGAATTAGTTAAAGGGTACTTACATTACGTGTTTGATTTGGGAAATGGTGCTAACC
TCATCAAAGGAAGCTCAAATAAACCTCTCAATGACAATCAGTGGCACAACGTGATGATATCAAGGGACAC
CAGCAACCTCCACACTGTAAAGATTGACACAAAAATCACAACGCAAATCACCGCCGGAGCCAGGAACTTA
GACCTCAAGAGTGACTTATATATAGGAGGAGTAGCTAAAGAAACATACAAATCCTTACCAAAACTTGTAC
ATGCCAAAGAAGGCTTTCAAGGCTGCCTGGCATCAGTTGATTTAAATGGACGGCTTCCGGACCTCATCTC
CGATGCTCTTTTCTGCAACGGACAGATCGAGAGAGGATGTGAAGGGCCCAGCACAACCTGCCAAGAGGAC
TCATGTTCCAATCAAGGTGTGTGCTTGCAACAATGGGATGGCTTCAGCTGTGACTGTAGTATGACTTCCT
TCAGTGGACCACTCTGCAATGACCCTGGGACGACATATATCTTTAGCAAAGGTGGTGGACAAATCACGTA
TAAGTGGCCTCCTAATGACCGACCCAGTACACGAGCAGACAGACTGGCCATAGGTTTTAGCACTGTTCAG
AAAGAAGCCGTATTGGTGCGAGTGGACAGTTCTTCAGGCTTGGGTGACTACCTAGAACTGCATATAACTC
TTCAATGATCAGAATGCTTTTGTGGAGGCAACTGCTATGCTTGAAAGAACATAGATGGCCTTTGAAGTAT

AATATTTCATTCAAAGTCAAGTTCCGTCACATAGTAGCTACGGAATCATAAGCAAATTATCTAACATCTG
CAACCTTCTGTTTCCTTATTGGAATCCTTGACCTGCAGTCATTATAATGAAATCTGGAGCCAATGATGGA
TATATGAACCTTTGGATTCAATCTTTCCAAACATATTTCTGACTAGAAGGAATTTGGTTTTATTTTCATC
CAGCTAACGTGAAATAAAATTGTCCTTTCAAAAT

In some embodiments, the gene is HTR2A and the mutation is within an exon of HTR2A. In some embodiments, the exon is the last coding exon of HTR2A. In some embodiments, the last coding exon of HTR2A is exon 3. In some embodiments, the exon is the last exon of HTR2A. In some embodiments, the last exon of HTR2A is exon 4. Exemplary mutations that fall within such regions are provided, e.g., in Table 2.

In some embodiments, the gene is CTTNBP2 and the mutation is within or near a DNase1 hypersensitivity site or within an exon of CTTNBP2. Exemplary mutations that fall within or near such regions are provided, e.g., in Table 2.

In some embodiments, the gene is REEP3 and the mutation is within or near a DNase1 hypersensitivity site or within an exon of REEP3. Exemplary mutations that fall within or near such regions are provided, e.g., in Table 2.

SNPs

In some embodiments, a mutation provided herein is a single nucleotide polymorphism (SNP). A SNP is a mutation that occurs at a single nucleotide location on a chromosome. The nucleotide located at that position may differ between individuals in a population and/or between paired chromosomes in an individual.

In some embodiments, the subject is a human subject and the mutation is a (i.e., at least one) SNP selected from Table 2. The risk nucleotide ("Risk allele" in Table 2) is the nucleotide identity that is associated with having a neuropsychiatric disorder or having an elevated risk of developing a neuropsychiatric disorder. In some embodiments, the risk nucleotide is a nucleotide identity other than a reference nucleotide ("Ref allele" in Table 2). The positions (i.e., the chromosome coordinates) in Table 2 are based on the 19th human genome assembly (Hg19, see, e.g., UCSC Genome Browser). The first base pair in each chromosome is labeled 0 and the position of the SNP is then the number of base pairs from the first base pair.

Table 2. SNPs

Chromosome	Position	Ref allele	Risk allele	Alle Freq. Cases	Alle Freq. Ctrls	Associated Gene
chr7	117430669	G	A	0.00260532	0.00118267	CTTNBP2
chr7	117358107	T	C	0.00350794	0.00095125	CTTNBP2
chr7	117431704	C	T	0.00088087	1.42E-05	CTTNBP2
chr7	117396664	C	A	0.00088924	1.45E-05	CTTNBP2
chr7	117374935	C	T	0.00088429	5.27E-05	CTTNBP2
chr7	117391129	A	G	0.00361388	0.00012395	CTTNBP2
chr7	117368123	T	C	0.00127929	8.36E-05	CTTNBP2
chr7	117446174	A	G	0.00166889	0.00094875	CTTNBP2
chr7	117456904	C	T	0.6498415	0.5937507	CTTNBP2
chr7	117356081	T	G	0.00087716	2.99E-05	CTTNBP2
chr7	117427551	T	C	0.00087421	3.17E-05	CTTNBP2
chr7	117354909	A	G	0.00117812	0.00099729	CTTNBP2
chr7	117452215	C	A	0.5441493	0.5061831	CTTNBP2
chr7	117431202	C	A	0.00096297	3.93E-05	CTTNBP2
chr7	117358129	T	A	0.00085854	1.44E-05	CTTNBP2
chr7	117359713	G	A	0.00087283	3.70E-05	CTTNBP2
chr7	117457141	G	C	0.6593968	0.617696	CTTNBP2
chr7	117450810	C	T	0.00099473	4.05E-05	CTTNBP2
chr7	117431879	G	A	0.00087544	2.46E-05	CTTNBP2
chr7	117386178	T	C	0.00089977	0.00053898	CTTNBP2
chr7	117385978	G	T	0.0017548	5.38E-05	CTTNBP2
chr7	117468334	T	C	0.6548906	0.6101791	CTTNBP2
chr7	117396706	C	T	0.00177246	0.00012335	CTTNBP2
chr7	117501314	G	A	0.00089729	4.85E-05	CTTNBP2
chr7	117390966	T	D	0.00088245	2.89E-05	CTTNBP2
chr7	117354258	G	A	0.00097471	0.000125	CTTNBP2
chr7	117352306	C	T	0.00086396	2.20E-05	CTTNBP2
chr7	117351979	G	A	0.00092798	7.15E-05	CTTNBP2
chr7	117431079	T	G	0.01548976	0.00350967	CTTNBP2
chr7	117417559	A	G	0.00364527	0.00162631	CTTNBP2
chr7	117427686	A	G	0.0033584	0.00214344	CTTNBP2
chr7	117421141	C	A	0.00097602	0.00046574	CTTNBP2
chr7	117468056	C	T	0.569068	0.5062561	CTTNBP2
chr13	47454997	G	T	0.00113719	5.49E-05	HTR2A
chr13	47440198	G	A	0.0023679	3.66E-05	HTR2A
chr13	47409048	G	A	0.00358304	0.00123287	HTR2A

chr13	47440301	T	C	0.00094808	2.75E-05	HTR2A
chr13	47466592	G	D	0.00085616	1.16E-05	HTR2A
chr13	47418543	T	C	0.00253451	0.00104533	HTR2A
chr13	47434747	A	C	0.00088582	3.29E-05	HTR2A
chr13	47448370	A	G	0.00092162	6.44E-05	HTR2A
chr13	47466622	G	A	0.03858394	0.022473	HTR2A
chr13	47409701	G	A	0.00197878	0.00120059	HTR2A
chr13	47440209	A	G	0.00108192	2.09E-05	HTR2A
chr13	47421746	G	T	0.00086305	1.62E-05	HTR2A
chr13	47418629	T	A	0.00085409	7.90E-06	HTR2A
chr13	47408946	C	A	0.00098246	3.55E-05	HTR2A
chr13	47455071	A	C	0.00088107	5.01E-06	HTR2A
chr13	47469335	T	C	0.9543982	0.9390308	HTR2A
chr3	185241792	G	C	0.0009132	5.44E-05	LIPH
chr3	185229283	C	T	0.0042522	0.00222155	LIPH
chr3	185226492	C	T	0.2717248	0.2349959	LIPH
chr3	185229464	T	C	0.4683532	0.4386795	LIPH
chr3	185226396	A	G	0.4670401	0.4330542	LIPH
chr3	185225638	A	T	0.02462703	0.0146825	LIPH
chr3	185225644	A	C	0.00087868	3.69E-05	LIPH
chr2	51256161	T	C	0.00090636	5.91E-05	NRXN1
chr2	50762143	C	T	0.00103184	0.00032744	NRXN1
chr2	51153020	G	A	0.00091575	3.75E-05	NRXN1
chr2	50733992	T	C	0.00085497	9.52E-06	NRXN1
chr2	51000979	G	I	0.00084935	5.07E-06	NRXN1
chr2	50842279	C	T	0.00091102	2.11E-05	NRXN1
chr2	51067620	G	C	0.00085562	9.72E-06	NRXN1
chr2	50606585	C	A	0.00087202	0.00014205	NRXN1
chr2	50927403	C	A	0.00169499	0.00023584	NRXN1
chr2	50703012	T	I	0.2621787	0.2424891	NRXN1
chr2	50956849	G	C	0.00085956	7.22E-06	NRXN1
chr2	50606521	G	A	0.07435779	0.05919093	NRXN1
chr2	50733958	T	C	0.429355	0.3910412	NRXN1
chr2	50733841	G	A	0.0008674	3.83E-05	NRXN1
chr2	50755127	T	A	0.00090075	3.99E-05	NRXN1
chr2	50542571	T	C	0.0008589	4.61E-06	NRXN1
chr2	50619213	C	T	0.0009083	9.50E-05	NRXN1
chr2	50699305	T	A	0.6060599	0.5651899	NRXN1
chr2	50927347	A	C	0.0008774	2.15E-05	NRXN1

chr2	50448284	T	C	0.00855504	0.00521072	NRXN1
chr2	50400991	T	C	0.00154868	0.00123783	NRXN1
chr2	50924511	C	G	0.00327626	5.88E-05	NRXN1
chr2	50850245	T	C	0.00180565	3.32E-05	NRXN1
chr2	50570754	G	C	0.00091261	0.00076225	NRXN1
chr2	50171755	G	A	0.00831418	0.00491064	NRXN1
chr2	50343508	C	G	0.00087991	3.36E-05	NRXN1
chr2	50762346	T	C	0.00093353	4.76E-05	NRXN1
chr2	51148378	A	T	0.000854	1.03E-05	NRXN1
chr2	51244839	T	C	0.06029333	0.04490447	NRXN1
chr2	50607797	G	A	0.0092228	0.00603568	NRXN1
chr2	50941362	C	T	0.00703044	0.0027606	NRXN1
chr2	50570601	A	ND	0.01920468	0.0102919	NRXN1
chr2	51236675	T	C	0.06442485	0.04570236	NRXN1
chr2	50187207	C	D	0.00154891	0.00107621	NRXN1
chr2	50848555	T	C	0.00129122	0.00092416	NRXN1
chr2	50198372	C	T	0.00090593	1.90E-05	NRXN1
chr2	50981817	T	C	0.0034023	0.00206087	NRXN1
chr2	50693162	A	G	0.616713	0.5631101	NRXN1
chr2	50570237	C	A	0.00261337	0.0012821	NRXN1
chr2	50683609	C	A	0.00089837	5.22E-05	NRXN1
chr2	50849551	T	C	0.00088462	3.71E-05	NRXN1
chr2	50735998	G	C	0.4071833	0.3689128	NRXN1
chr2	51246886	A	T	0.00200887	0.00098123	NRXN1
chr2	50200776	C	G	0.01047886	0.00580885	NRXN1
chr2	50323549	G	A	0.00089426	4.72E-05	NRXN1
chr2	50675110	T	C	0.0165951	0.00514289	NRXN1
chr2	50922035	G	A	0.00701069	0.0025022	NRXN1
chr2	51057550	G	A	0.00089328	9.57E-05	NRXN1
chr2	50386107	C	G/T	0.00086118	0.0001051	NRXN1
chr2	50386080	C	T	0.00085941	3.01E-05	NRXN1
chr2	50847195	G	A	0.00770192	0.00362506	NRXN1
chr2	51252712	C	T	0.00208087	0.0001516	NRXN1
chr2	51245440	A	T	0.00085531	9.40E-06	NRXN1
chr2	50354237	C	G	0.00091993	7.38E-05	NRXN1
chr2	50719598	C	T	0.00138124	0.00110033	NRXN1
chr2	50952610	G	T	0.00262365	7.26E-05	NRXN1
chr2	50792080	G	A	0.00099	0.000137	NRXN1
chr2	50542527	T	C	0.00165871	1.17E-05	NRXN1

chr2	50750575	C	G	0.00091001	5.41E-05	NRXN1
chr2	50155007	T	C	0.00120133	0.00116964	NRXN1
chr2	50779791	C	T	0.0008481	4.60E-06	NRXN1
chr2	50733581	T	D	0.00132345	0.00101267	NRXN1
chr2	50400809	T	C	0.01732912	0.00833458	NRXN1
chr2	50201255	G	A	0.00357087	0.00194928	NRXN1
chr2	50178130	A	C	0.00095967	4.23E-05	NRXN1
chr2	51146148	T	G	0.00091513	5.22E-05	NRXN1
chr2	50575137	T	A	0.0008994	2.38E-05	NRXN1
chr2	51148372	G	A	0.00097176	1.27E-05	NRXN1
chr2	51171979	G	C	0.00088798	4.81E-05	NRXN1
chr2	50779943	T	C	0.00086198	6.72E-06	NRXN1
chr2	50848551	T	C	0.00193478	0.00015711	NRXN1
chr2	50165016	A	C	0.00219682	0.00104104	NRXN1
chr2	51149889	C	G	0.00088559	0.00061487	NRXN1
chr2	50774153	G	T	0.00091927	4.62E-05	NRXN1
chr2	50389636	T	G	0.00567169	0.00318092	NRXN1
chr2	50434866	C	T	0.00094894	7.53E-05	NRXN1
chr2	50724642	A	G	0.00085032	2.60E-06	NRXN1
chr2	50981813	C	ND	0.00235601	0.00105149	NRXN1
chr2	51085557	G	D	0.02479016	0.01799434	NRXN1
chr2	50463984	G	A	0.00086916	0.00059959	NRXN1
chr2	50724745	G	T	0.00173593	0.0009164	NRXN1
chr2	50981807	A	D	0.00221101	0.00106565	NRXN1
chr2	50598207	A	G	0.00090418	5.73E-05	NRXN1
chr2	50675639	G	A	0.00176738	4.44E-05	NRXN1
chr2	50653833	G	A	0.00089544	4.24E-05	NRXN1
chr2	51145459	G	A	0.00166583	0.00087083	NRXN1
chr2	50542372	C	T	0.00086827	1.97E-05	NRXN1
chr2	50952571	T	C	0.00174664	4.07E-05	NRXN1
chr2	50548140	G	C	0.00084975	1.05E-05	NRXN1
chr2	50765412	G	T	0.00894677	0.00282258	NRXN1
chr2	50850686	G	A	0.01074276	0.00132826	NRXN1
chr2	50934666	T	C	0.00094389	6.75E-05	NRXN1
chr2	50682914	T	C	0.0935983	0.07463179	NRXN1
chr2	50709350	T	G	0.00087111	1.64E-05	NRXN1
chr2	50979527	T	C	0.00178027	8.23E-05	NRXN1
chr2	50386109	C	T	0.0009151	1.09E-05	NRXN1
chr2	50542308	C	D	0.02148694	0.01503297	NRXN1

chr2	50607943	G	C	0.00086566	1.88E-05	NRXN1
chr2	50735814	C	T	0.00182226	0.00099818	NRXN1
chr2	50981815	G	A	0.00215973	0.00097383	NRXN1
chr2	50155737	T	C	0.00448853	0.00110602	NRXN1
chr2	50683701	G	A	0.02152826	0.00996433	NRXN1
chr2	50842256	C	G	0.00087221	1.92E-05	NRXN1
chr2	50148728	G	I	0.00086118	1.51E-05	NRXN1
chr2	50952482	T	D	0.1064954	0.09330609	NRXN1
chr2	51153206	C	T	0.00088029	3.36E-05	NRXN1
chr2	50560998	C	G	0.00308847	0.00184582	NRXN1
chr2	50996952	C	T	0.2170998	0.1858741	NRXN1
chr2	50458593	T	C	0.00156192	7.95E-05	NRXN1
chr2	50924466	A	T	0.00096108	3.25E-05	NRXN1
chr2	51005207	C	T	0.00148103	8.60E-05	NRXN1
chr2	50602031	A	T	0.00102945	7.00E-05	NRXN1
chr2	50178059	C	A	0.00132313	0.00095255	NRXN1
chr2	50850340	T	A	0.00155961	0.00096682	NRXN1
chr2	51016384	T	C	0.00235489	0.0001423	NRXN1
chr2	50175865	A	G	0.01599345	0.00465917	NRXN1
chr2	50571910	A	T	0.00494906	0.00251935	NRXN1
chr2	50570602	G	ND	0.06257605	0.05465273	NRXN1
chr2	50548103	G	C	0.00084982	9.27E-06	NRXN1
chr2	50518040	G	A	0.00099883	0.00091739	NRXN1
chr2	50236859	G	A	0.00116077	0.00099526	NRXN1
chr2	50464065	C	T	0.00089161	0.00030345	NRXN1
chr2	50598321	G	A	0.00126708	2.42E-05	NRXN1
chr2	50282777	T	C	0.0011488	0.00084684	NRXN1
chr2	51245472	G	A	0.00141775	2.17E-05	NRXN1
chr2	50735943	C	G	0.4921696	0.4449057	NRXN1
chr2	50927534	C	G	0.03215577	0.01906242	NRXN1
chr2	50941367	G	T	0.00088856	2.77E-05	NRXN1
chr2	50952709	G	A	0.00091409	2.82E-05	NRXN1
chr2	51067726	G	C	0.00174473	9.57E-05	NRXN1
chr2	51079254	C	G	0.00086376	2.40E-05	NRXN1
chr2	50277539	C	A	0.00095583	0.00030738	NRXN1
chr2	50424938	C	T	0.00086417	4.48E-05	NRXN1
chr2	50765589	T	C	0.00160868	9.11E-06	NRXN1
chr2	50699377	C	T	0.00101147	0.00098491	NRXN1
chr2	51149368	G	T	0.00200754	0.00057659	NRXN1

chr2	50723068	G	A	0.00344394	6.56E-05	NRXN1
chr2	50981811	T	ND	0.001479825	0.000823127	NRXN1
chr2	50723000	C	A	0.00186579	0.00104014	NRXN1
chr2	51245656	C	ND	0.00707575	0.0024419	NRXN1
chr2	50571784	G	C	0.00528244	0.00206069	NRXN1
chr2	50148783	C	T	0.00087211	1.48E-05	NRXN1
chr2	50598280	A	C	0.00085786	1.24E-05	NRXN1
chr2	50850307	G	A	0.0012766	3.03E-05	NRXN1
chr2	50850394	C	A	0.00086764	4.52E-05	NRXN1
chr2	50563875	C	T	0.02105221	0.00729743	NRXN1
chr2	50614848	C	T	0.00092797	4.98E-05	NRXN1
chr2	50531295	A	T	0.00098164	6.54E-05	NRXN1
chr2	50877741	C	T	0.00704379	0.00255214	NRXN1
chr2	50733745	G	C	0.00172869	1.54E-05	NRXN1
chr2	50919652	T	A	0.00204463	0.00011175	NRXN1
chr2	51021463	A	G	0.00100263	0.0001034	NRXN1
chr10	65339450	C	A	0.00148089	8.31E-05	REEP3
chr10	65368263	T	A	0.00097042	0.00011515	REEP3
chr10	65358911	C	A	0.00580936	0.00110724	REEP3
chr10	65326034	A	G	0.006148	0.00115539	REEP3
chr10	65359513	T	C	0.00086289	1.85E-05	REEP3
chr10	65332906	T	C	0.00636706	0.00204213	REEP3
chr10	65354650	A	G	0.00086983	1.13E-05	REEP3
chr10	65357754	C	A	0.00106936	0.00028568	REEP3
chr10	65287863	G	C	0.00659347	0.00184873	REEP3
chr10	65307923	A	G	0.05183556	0.03045778	REEP3
chr10	65387644	C	G	8.60E-04	2.60E-05	REEP3
chr10	65387722	C	D	1.09E-03	1.52E-05	REEP3
chr10	65388750	G	A	9.00E-04	9.01E-05	REEP3
chr10	65384621	C	T	1.87E-03	3.03E-04	REEP3

Legend: ND = Not determined, D = deletion , I = insertion, Alle Freq. Cases = Allele frequency in cases, Alle Freq. Ctrls = allele frequency in controls

In some embodiments, a SNP can be used in the methods described herein. In some embodiments, the method comprises:

(a) analyzing genomic DNA from a subject for the presence of a SNP (e.g., a SNP in Table 2); and

(b) identifying a subject having the SNP as a subject at elevated risk of developing a neuropsychiatric disorder or as a subject having a neuropsychiatric disorder.

Any number of SNPs are contemplated herein for use, e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more SNPs.

Genome analysis methods

Methods provided herein comprise analyzing genomic DNA. In some embodiments, analyzing genomic DNA comprises carrying out a nucleic acid-based assay, such as a sequencing-based assay or a hybridization-based assay. In some embodiments, the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array. In some embodiments, the genomic DNA is analyzed using a bead array. Methods of genetic analysis are known in the art. Examples of genetic analysis methods and commercially available tools are described below.

Affymetrix: The Affymetrix SNP 6.0 array contains over 1.8 million SNP and copy number probes on a single array. The method utilizes at a simple restriction enzyme digestion of 250 ng of genomic DNA, followed by linker-ligation of a common adaptor sequence to every fragment, a tactic that allows multiple loci to be amplified using a single primer complementary to this adaptor. Standard PCR then amplifies a predictable size range of fragments, which converts the genomic DNA into a sample of reduced complexity as well as increases the concentration of the fragments that reside within this predicted size range. The target is fragmented, labeled with biotin, hybridized to microarrays, stained with streptavidin-phycoerythrin and scanned. To support this method, Affymetrix Fluidics Stations and integrated GS-3000 Scanners can be used.

Illumina Infinium: Examples of commercially available Infinium array options include the 660W-Quad (>660,000 probes), the 1MDuo (over 1 million probes), and the custom iSelect (up to 200,000 SNPs selected by user). Samples begin the process with a whole genome amplification step, then 200 ng is transferred to a plate to be denatured and neutralized, and finally plates are incubated overnight to amplify. After amplification the samples are enzymatically fragmented using end-point fragmentation. Precipitation and resuspension clean up the DNA before hybridization onto the chips. The fragmented, resuspended DNA samples are then dispensed onto the appropriate BeadChips and placed in

the hybridization oven to incubate overnight. After hybridization the chips are washed and labeled nucleotides are added to extend the primers by one base. The chips are immediately stained and coated for protection before scanning. Scanning is done with one of the two Illumina iScan™ Readers, which use a laser to excite the fluorophore of the single-base extension product on the beads. The scanner records high-resolution images of the light emitted from the fluorophores. All plates and chips are barcoded and tracked with an internally derived laboratory information management system. The data from these images are analyzed to determine SNP genotypes using Illumina's BeadStudio. To support this process, Biomek F/X, three Tecan Freedom Evos, and two Tecan Genesis Workstation 150s can be used to automate all liquid handling steps throughout the sample and chip prep process.

Illumina BeadArray: The Illumina Bead Lab system is a multiplexed array-based format. Illumina's BeadArray Technology is based on 3-micron silica beads that self-assemble in microwells on either of two substrates: fiber optic bundles or planar silica slides. When randomly assembled on one of these two substrates, the beads have a uniform spacing of ~5.7 microns. Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays. BeadArray technology is utilized in Illumina's iScan System.

Sequenom: During pre-PCR, either of two Packard Multiprobes is used to pool oligonucleotides, and a Tomtec Quadra 384 is used to transfer DNA. A Cartesian nanodispenser is used for small-volume transfer in pre-PCR, and another in post-PCR. Beckman Multimeks, equipped with either a 96-tip head or a 384-tip head, are used for more substantial liquid handling of mixes. Two Sequenom pin-tool are used to dispense nanoliter volumes of analytes onto target chips for detection by mass spectrometry. Sequenom Compact mass spectrometers can be used for genotype detection.

In some embodiments, methods provided herein comprise analyzing genomic DNA using a nucleic acid sequencing assay. Methods of genome sequencing are known in the art. Examples of genome sequencing methods and commercially available tools are described below.

Illumina Sequencing: 89 GAIIx Sequencers are used for sequencing of samples. Library construction is supported with 6 Agilent Bravo plate-based automation, Stratagene

MX3005p qPCR machines, Matrix 2-D barcode scanners on all automation decks and 2 Multimek Automated Pipettors for library normalization.

454 Sequencing: Roche® 454 FLX-Titanium instruments are used for sequencing of samples. Library construction capacity is supported by Agilent Bravo automation deck, Biomek FX and Janus PCR normalization.

SOLiD Sequencing: SOLiD v3.0 instruments are used for sequencing of samples. Sequencing set-up is supported by a Stratagene MX3005p qPCR machine and a Beckman SC Quanter for bead counting.

ABI Prism® 3730 XL Sequencing: ABI Prism® 3730 XL machines are used for sequencing samples. Automated Sequencing reaction set-up is supported by 2 Multimek Automated Pipettors and 2 Deerac Fluidics - Equator systems. PCR is performed on 60 Thermo-Hybaid 384-well systems.

Ion Torrent: Ion PGM™ or Ion Proton™ machines are used for sequencing samples. Ion library kits (Invitrogen) can be used to prepare samples for sequencing.

Other Technologies: Examples of other commercially available platforms include Helicos Heliscope Single-Molecule Sequencer, Polonator G.007, and Raindance RDT 1000 Rainstorm.

Devices and Kits

Any of the methods provided herein can be performed on a device, e.g., an array. Suitable arrays are described herein and known in the art. Accordingly, a device, e.g., an array, for detecting any of the mutations (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations, or at least 10, at least 20, at least 30, at least 40, at least 50, or more mutations, or up to 5, up to 10, up to 15, up to 20, up to 25, up to 30, up to 35, up to 40, up to 45, up to 50, up to 75 or up to 100 mutations) described herein is also contemplated.

Reagents for use in any of the methods provided herein can be in the form of a kit. Accordingly, a kit for detecting any of the mutations (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations, or at least 10, at least 20, at least 30, at least 40, at least 50, or more mutations, or up to 5, up to 10, up to 15, up to 20, up to 25, up to 30, up to 35, up to 40, up to 45, up to 50, up to 75 or up to 100 mutations) described herein is also contemplated. In some embodiments, the kit comprises reagents for detecting any of the mutations described herein,

e.g., reagents for use in a method described herein. Suitable reagents are described herein and are known in the art.

Controls

Some of the methods provided herein involve determining the presence or absence a mutation in a biological sample and then comparing that presence or absence to a control in order to identify a subject having an elevated risk of developing a neuropsychiatric disorder or to identify a subject having a neuropsychiatric disorder. The control may be the identity of the nucleic acid(s) at the corresponding location in a control tissue, control subject, or a population of control subjects.

The control may be (or may be derived from) a normal subject (or normal subjects). A normal subject, as used herein, refers to a subject that is healthy, such a subject experiencing none of the symptoms associate with a neuropsychiatric disorder. The control population may be a population of normal subjects.

In other instances, the control may be (or may be derived from) a subject (a) having a similar neuropsychiatric disorder to that of the subject being tested and (b) who is negative for the mutation.

It is to be understood that the methods provided herein do not require that a control identity be measured every time a subject is tested. Rather, it is contemplated that control identities are obtained and recorded and that any test identity is compared to such a pre-determined identity.

In some embodiments, the mutation is a SNP described in Table 2 and the control is a nucleotide other than the risk nucleotide as described in Table 2 ("Risk allele" in Table 2). In some embodiments, the mutation is a SNP described in Table 2 and the control is a reference allele nucleotide as described in Table 2 ("Ref allele" in Table 2).

Samples

The methods provided herein detect and optionally measure (and thus analyze) particular mutations in biological samples. Biological samples, as used herein, refer to samples taken or obtained from a subject. These biological samples may be tissue samples or they may be fluid samples (e.g., bodily fluid). Examples of biological fluid samples are

whole blood, plasma, serum, urine, sputum, phlegm, saliva, tears, and other bodily fluids. In some embodiments, the biological sample is a whole blood or saliva sample. In some embodiments, the biological sample is a biopsy sample, e.g., a central nervous system biopsy sample.

In some embodiments, the biological sample may comprise a polynucleotide (e.g., genomic DNA or mRNA) derived from a tissue sample or fluid sample of the subject. In some embodiments, the biological sample may be manipulated to extract a polynucleotide. In some embodiments, the biological sample may be manipulated to amplify a polynucleotide sample. Methods for extraction and amplification (e.g., PCR) are well known in the art.

Subjects

Subjects are preferably human subjects. In some embodiments, a subject (e.g., a subject identified in a method herein) is at elevated risk of developing a neuropsychiatric disorder or has a neuropsychiatric disorder. In some embodiments, a subject (e.g., a subject identified in a method herein) has a neuropsychiatric disorder.

It is to be understood that methods of the invention may be used in a variety of other subjects including but not limited to mammals such as humans, canines, felines, mice, rats, rabbits, and apes.

Computational analysis

Methods of computation analysis of genomic and expression data are known in the art. Examples of available computational programs are: Genome Analysis Toolkit (GATK, Broad Institute, Cambridge, MA), Expressionist Refiner module (Genedata AG, Basel, Switzerland), GeneChip - Robust Multichip Averaging (CG-RMA) algorithm, PLINK (Purcell et al, 2007), GCTA (Yang et al, 2011), the EIGENSTRAT method (Price et al 2006), EMMAX (Kang et al, 2010). In some embodiments, methods described herein include a step comprising computational analysis.

Treatment

Other aspects of the invention relate to diagnostic or prognostic methods that comprise a treatment step (also referred to as "theranostic" methods due to the inclusion of

the treatment step). Any treatment for a neuropsychiatric disorder is contemplated. In some embodiments, treatment comprises behavioral therapy and/or one or more therapeutic agents.

In some embodiments, treatment comprises administration of an effective amount of an appropriate therapeutic agent for the particular neuropsychiatric disorder, e.g., an antidepressant, a stimulant, an antidopaminergic, or a central adrenergic inhibitor. Non-limiting examples of antidepressants include ariprazole, doxepin, clomipramine, bupropion, amoxapine, nortriptyline, citalopram, duloxetine, trazodone, venlafaxine, selegiline, amitriptyline, escitalopram, isocarboxazid, phenelzine, desipramine, trazodone, nortriptyline, tranylcypromine, paroxetine, paroxetine, fluoxetine, desvenlafaxine, mirtazapine, fluoxetine, quetiapine, nefazodone, doxepin, trimipramine, imipramine, vilazodone, protriptyline, bupropion, sertraline, and olanzapine. Non-limiting examples of antidopaminergics include domperidone, haloperidol, chlorpromazine and alizapride. Non-limiting examples of stimulants include Adderall, Adderall XR, Concerta, Dexedrine, Dexedrine spansule, Daytrana, Metadate CD, Metadate ER, Methylin ER, Ritalin, Ritalin LA, Ritalin SR, Vyvanse, and Quillivant XR. Non-limiting examples of central adrenergic inhibitors include clonidine and guanfacine.

In some embodiments, the neuropsychiatric disorder is OCD. Non-limiting examples of therapeutic agents for OCD include anti-depressants such as selective serotonin reuptake inhibitors (SSRIs) (e.g., paroxetine, sertraline, fluoxetine, escitalopram and fluvoxamine) and tricyclic antidepressants (e.g., clomipramine). Other non-limiting examples of therapeutic agents for OCD include riluzole, memantine, gabapentin, N-Acetylcysteine, lamotrigine, and atypical antipsychotics, such as olanzapine, quetiapine, and risperidone.

In some embodiments, treatment comprises behavioral therapy. Non-limiting examples of behavioral therapy include exposure and response prevention (ERP) and habit-reversal training.

In some embodiments, treatment comprises electroconvulsive therapy. In some embodiments, treatment comprises deep brain stimulation (DBS).

It is to be understood that any treatment described herein may be used alone or may be used in combination with any other treatment described herein.

In some embodiments, a subject identified as being at elevated risk of developing or as having a neuropsychiatric disorder is treated. In some embodiments, the method

comprises selecting a subject for treatment on the basis of the presence of one or more mutations as described herein. In some embodiments, the method comprises treating a subject having a neuropsychiatric disorder characterized by the presence of one or more mutations as defined herein.

As used herein, "treat" or "treatment" includes, but is not limited to, preventing or reducing the development of a neuropsychiatric disorder or reducing or eliminating the symptoms of a neuropsychiatric.

An effective amount is a dosage of a therapy sufficient to provide a medically desirable result, such as treatment of a neuropsychiatric disorder. The effective amount will vary with the disorder to be treated, the age and physical condition of the subject being treated, the severity of the disorder, the duration of the treatment, the nature of any concurrent therapy, the specific route of administration and the like factors within the knowledge and expertise of the health practitioner.

Administration of a treatment may be accomplished by any method known in the art (see, e.g., Harrison's Principle of Internal Medicine, McGraw Hill Inc.). Administration may be local or systemic. Administration may be parenteral (e.g., intravenous, subcutaneous, or intradermal) or oral (e.g., sublingual or buccal). Compositions for different routes of administration are well known in the art (see, e.g., Remington's Pharmaceutical Sciences by E. W. Martin). Dosage will depend on the subject and the route of administration. Dosage can be determined by the skilled artisan.

EXAMPLES

Example 1: Integrating evolutionary and regulatory evidence identifies genes for Obsessive-Compulsive Disorder

Whole exome sequencing has proven successful in identifying variants for single gene disorders, but not in polygenic disorders yet. >90% genome-wide association loci for common diseases reside in noncoding regions, suggesting that current sequencing studies for polygenic disorders are hindered by neglecting noncoding disease variants. As described below, targeted sequencing of both coding and functional noncoding regions indicated by evolutionary conservation for a heritable psychiatric disease, obsessive-compulsive disorder

(OCD), was performed. Combining the genetic and neurobiological understanding of OCD, 608 genes (13Mb) in a small cohort (592 cases and 560 controls) were studied to identify genes and variants for OCD. Genic burden tests stratified by coding, regulatory and evolutionary status identified five genes with significant mutation loads in OCD after multiple testing correction: *NRXN1* and *HTR2A* with coding, *CTTNBP2* and *REEP3* with regulatory, and *LIPH* with all mutation types. 218 candidate variants were identified, some of which were experimentally validated for regulatory functions. In conclusion, this study demonstrates that integrating evolutionary and biochemical evidence facilitates the detection of disease genes and variants and that both regulatory and coding regions are required to be examined for polygenic disorders.

INTRODUCTION

Sequencing of genomic regions yields a nearly complete catalog of all genetic variation in the study population and can identify the causal variants underlying traits and diseases. One popular strategy, whole exome sequencing, targets the protein coding portion of the genome and has successfully identified variants for single gene Mendelian disorders. Studies of polygenic diseases have been less successful, potentially because the causal variation lies in noncoding regulatory regions not captured by whole exome sequencing. This hypothesis is supported by the observation that genomic regions significantly associated with complex diseases in genome-wide association studies (GWAS) are enriched for non-coding regions [ref. 1], with over 90% of the GWAS variants residing within noncoding regions [ref. 2].

Functional noncoding elements comprise a much larger proportion of the genome than protein coding sequence, and play an important role in human phenotypic variation and disease risk that is just beginning to be elucidated. The ENCODE project, which aims to identify all functional elements in the human genome, has compiled an extensive catalog of noncoding, biochemically active elements, as indicated by biochemical signatures such as DNase hypersensitivity, transcription factor occupancy, and histone marks in different cell types. These elements are enriched both in genomic regions highly conserved or highly diverged across mammals, suggesting they are subject to evolutionary pressure. Massively parallel reporter assays that systematically disrupt and induce noncoding enhancers clearly

demonstrate that these elements have regulatory functions, and that the effect of such perturbations varies by cell type.

Here, a targeted sequencing approach was used that captured both coding and functional non-coding elements to identify genes and variants associated with obsessive-compulsive disorder (OCD), a common, complex and heritable psychiatric disease affecting 1-3% of the population. As with other complex diseases, both common and rare variation likely contribute to OCD risk, and genome-wide complex trait analysis (GCTA) suggests an unusually large proportion of OCD heritability is due to common variation. In addition, the genomic search space for OCD can be intelligently reduced by the well-established corticostriato-thalamo-cortical (CSTC) neurocircuitry model, as well as the significant GWAS results from canine OCD that remarkably resembles human OCD. The current level of genetic and neurobiological understanding of OCD, and the role of common variation in heritability, make OCD particularly well-suited to a targeted sequencing strategy, as hundreds of candidate regions can be identified and statistical power can be achieved even with relatively small sample sizes.

Coding and noncoding evolutionarily conserved elements in or near 608 genes implicated in OCD (13Mb) were sequenced in 592 cases and 560 controls. It was found that the resulting sequence dataset captured both rare and common variations, and included dramatically more potentially functional variants compared to a whole exome approach that focused only on rare, protein damaging mutations. This allows for use of sensitive association tests that evaluate the collective effect of rare and common variations. Five genes were identified as being significantly associated with OCD after multiple testing corrections. Four of the five genes were identified in analyses stratified based on sequence annotations (evolutionary conservation, protein-coding status and regulatory activity), demonstrating the power of integrating evolutionary and biochemical evidence. Associated Gene Ontology (GO) sets were also identified that highlight the role of synaptic cell adhesion molecules in the neurobiology of OCD.

RESULTS

Design of targeted regions

A custom NimbleGen hybrid capture array was designed targeting 82,723 evolutionarily constrained regions in and around 608 genes (total 13Mb, 58bp-16kb size range, median size 237 bases). This included all regions within 1kb of the start and end of each of 608 targeted gene with SiPhy evolutionarily constraint score >7, a set that included all exons [ref. 3]. In the intergenic region up and downstream of each gene, regions were selected for sequencing using a constraint score threshold that became more stringent with distance from the gene. The targeted regions were sequenced in 592 European DSM-IV OCD cases and 560 ancestry matched controls using a pooled sequencing strategy. Each uniquely barcoded pool included 16 phenotype matched individuals, for a total of 37 "case" pools and 35 "control" pools.

95% of the target regions at >30x read depth per pool were captured, with a median of 112x (approximately 7x per individual) (FIG. 5). Two different algorithms, Syzygy and SNVer, were used for variant detection and 41,504 'high-confidence' single nucleotide variants (SNVs) detected by both methods became the focus of the study. Another 83,037 'low-confidence' SNVs were detected by only one algorithm (42,712 by Syzygy and 40,325 by SNVer). Supporting the accuracy of the variant calling of 'high-confidence' SNVs, the allele frequencies determined by Syzygy and SNVer are strongly correlated (AF_{syzygy} vs. AF_{snver}, $\rho = 0.999$, $p < 2.2 \times 10-16$; FIG. 5). Genotyping a random subset of 22 'high-confidence' and 43 'low-confidence' variants in 1,085 samples (544 cases and 541 controls) using Sequenom MassArray validated the allele frequencies measured in the pooled sequencing data ($\rho = 0.999$, $p < 2.2 \times 10-16$; FIGs. 6 and 7). Similar results were obtained in a separate analysis that extended the sample cohort to 2986 individuals (1,321 OCD cases and 1,665 controls) that included the original sample cohort.

Subsequently, the detected variants were annotated based on whether they resided within coding, regulatory, or evolutionary regions as defined below. Variants overlapping with exons were annotated as coding variants. For regulatory variants, ENCODE DNase1 hypersensitivity sites (DHS) were used, which indicate genomic regions where transcription factors can access to the open chromatin. For the evolutionary signatures potentially driven by positive and negative selection left on the variant sites, the base-specific mammalian constraint GERP++ score was applied, where a zero indicated a given position has the neutral rate of base substitution across multiple species, a positive score indicated a slow rate of base

substitution, and a negative score indicated a fast rate of base substitution. The sites with GERP>2 were termed as 'evolutionarily conserved', with GERP<-2 were termed as 'evolutionarily divergent', and both combined were termed as 'evolutionary' sites, based on the threshold used by the 1000 genomes project. The annotation resulted in 4,427 exonic variants, 9,890 DHS variants, 14,827 variants in conserved sites, and 7,179 variants in divergent sites, assigning functional annotations to 67% (27,626/41,504) of the detected variants (FIG. 1A).

Examining the variants with multiple annotations revealed complex overlaps between the functional classes. 49.3% (2,184) exonic variants and 38.7% (3,831) DHS variants were at conserved sites, and 20.6% (913) exonic variants and 15.8% (1,560) DHS variants were at divergent sites. 680 variants overlapped with both exons and DHS (i.e. 15.4% of exonic and 6.88% of DHS variants), of which 328 (49.2%) were at conserved sites and 143 (21%) at divergent sites, leaving 25.3% (1,121) exonic variants and 43.4% (4,290) DHS variants to be annotated by a single category. As these overlaps potentially indicated multiple functions of a locus, the overlapping annotation structure was maintained to incorporate into association analyses.

Stratified analyses identify associated genes with coding and regulatory variants separately

Using the annotated variants and their case-control frequencies in the data set, genic burden tests of excessive non-reference allele rates in cases was performed. The burden tests were stratified into three categories, where the test was performed on (i) all detected variants ('Overall'), (ii) coding variants ('Exon'), and (iii) DHS variants ('DHS'). Each category had three additional sub-categories stratified by evolutionary status: variants at the (i) 'evolutionary' sites ('Evo.'), (ii) 'evolutionarily conserved' sites ('Cons.') and (iii) 'evolutionarily divergent' sites ('Div.') (FIGs. 1B and 8A). Multiple testing was rigorously controlled for by employing the empirical 'minP' procedure that jointly corrected for all tested stratifications and for all sequenced genes (Methods and FIG. 9).

The large majority (99%) of the 608 sequenced genes showed no evidence of association in any of these twelve tests (FIG. 8A). However, it should be noted that more genes may become significant with a larger sample size. Five genes were identified with a

significant excess of variants in cases after multiple testing correction (corr. $p < 0.0487$; FIG. 1B and FIG. 9): 'Overall' burden test identified *LIPH* ($p = 4 \times 10^{-4}$) to have a significant excess of all mutation types in cases and 'Overall-Div.' burden test identified *REEP3* ($p = 10^{-4}$) to have significant excess of 'divergent' site mutations in cases; 'Exon' tests identified *NRXN1* ($p = 2 \times 10^{-4} \sim 3 \times 10^{-4}$ in [Exon-All,-Evo., and -Cons.]) and *HTR2A* ($p = 3 \times 10^{-4}$ in [Exon-Cons.]); DHS tests identified *CTTNBP2* ($p = 5 \times 10^{-4} \sim 9 \times 10^{-4}$ in [DHS-All, -Cons.]) and *REEP3* again ($p = 6 \times 10^{-4} \sim 8 \times 10^{-4}$ in [DHS-All, -Evo.]). Stratified burden tests using rare variants (AF_{1kg} <0.01) did not result in any genes with significant burden of mutations after multiple testing corrections (FIG. 8B). It is notable that *NRXN1* and *HTR2A*, the genes that have excessive coding mutations showed no association signals in DHS tests, and the opposite is the case for *CTTNBP2* and *REEP3* (FIG. 1B). This result implies distinct patterns for Exon and DHS mutations for OCD-associated genes, and the observation is not by random chance from variant sampling, as the general test results of the eleven stratified burden analyses correlated with the 'overall' burden test despite the variable signals among the top genes (FIG. 10 and Table 3).

Table 3. Genic Burden Correlation

Number	Test	Rho*	P-value*	No. variants tested
1	Overall-All	NA	NA	41504
2	Overall-Evo	0.847	2.20E-16	22006
3	Overall-Cons	0.647	2.20E-16	14827
4	Overall-Div	ome	2.20E-16	7179
5	Exon-All	0.384	2.20E-16	4427
6	Exon-Evo	0.357	4.36E-14	3097
7	Exon-Cons	0.238	2.28E-06	2184
8	Exon-Div	0.305	5.22E-08	913
9	DHS-All	0.622	2.20E-16	9890
10	DHS-Evo	0.532	2.20E-16	5391
11	DHS-Cons	0.399	1.49E-14	3831
12	DHS-Div	0.432	3.08E-16	1560
13	Rare-All	0.106	0.020	13589
14	Rare-Evo	0.060	0.194	7806
15	Rare-Cons	0.052	0.281	5767
16	Rare-Div	0.008	0.881	2039

*Spearman's rank correlation coefficient and p-value for comparison with Overall-All test result

Polygenic burden overlaid on GO network pins down disease-related pathways

In order to gain pathway-level insights, a polygenic burden of non-reference alleles in cases for the gene sets was evaluated that may be relevant to OCD. Gene Ontology (GO) categories that were weakly enriched by the 608 targeted genes (enrichment p <0.1) were used to obtain comprehensive gene sets of biological relevance that represented the search space (Table 4). To this end, 989 GO sets were obtained that covered a range of brain-related functions from synaptic transmission and ion channel activity to glutamate and dopamine signaling, as well as non-brain-specific terms such as regulation of metabolic processes and cytoskeleton organization. In order to understand the relationships between these GO sets, a network generation algorithm was employed that was optimally designed for visualizing many highly-related gene sets [ref. 4]. This automatically placed 415 less-redundant GO sets as nodes and created 1,942 connecting edges, determined by genetic overlaps between two GO sets.

Table 4. Polygenic Burden

GOID	Description	Enrichment P	Burden P	# Total Genes	# Target Genes	Targeted Genes
GO:0010942	positive regulation of cell death	1.58E-03	3.00E-04	435	30	CCK, CADM1, PTGS2, ADORA2A, GRIK2, NELL1, PPP3R1, PTEN, GPX1, AGTR2, CD44, APOE, RAC1, PPP3CC, BCL6, APC, KNG1, KCNMA1, ARHGEF2, GRIN1, NF1, YWHAB, LGALS12, MBD4, ADIPOQ, DAPK1, ADRB2, TNFSF10, SST, PRODH
GO:0043065	positive regulation of apoptosis	2.69E-03	5.00E-04	430	29	CCK, CADM1, PTGS2, ADORA2A, NELL1, PPP3R1, PTEN, GPX1,

						AGTR2, CD44, APOE, RAC1, PPP3CC, BCL6, APC, KNG1, KCNMA1, ARHGEF2, GRIN1, NF1, YWHAB, LGALS12, MBD4, ADIPOQ, DAPK1, ADRB2, TNFSF10, SST, PRODH
GO:0043068	positive regulation of programmed cell death	1.49E-03	5.00E-04	433	30	CCK, CADM1, PTGS2, ADORA2A, GRIK2, NELL1, PPP3R1, PTEN, GPX1, AGTR2, CD44, APOE, RAC1, PPP3CC, BCL6, APC, KNG1, KCNMA1, ARHGEF2, GRIN1, NF1, YWHAB, LGALS12, MBD4, ADIPOQ, DAPK1, ADRB2, TNFSF10, SST, PRODH
GO:0031334	positive regulation of protein complex assembly	3.21E-05	7.00E-04	35	9	WNT2, ACTR3, CTTNBP2, CCK, ARPC2, GSK3B, MAP1B, RAC1, APC
GO:0060249	anatomical structure homeostasis	4.25E-02	1.30E-03	106	9	ALDOA, ADRB2, RAC2, DMD, ANKRD11, RAC1, ADD1, CTNNB1, APC
GO:0032273	positive regulation of protein polymerization	2.40E-04	3.70E-03	25	7	WNT2, ACTR3, CTTNBP2, ARPC2, MAP1B, RAC1, APC
GO:0070830	tight junction assembly	7.26E-02	5.10E-03	2	2	STRN, APC
GO:0019900	kinase binding	3.44E-02	5.20E-03	179	13	NELL1, NBEA, CDH2, FER, CTNNB1, WNT2, CTTNBP2, GSK3B, AVPR1A, AKAP1, MAP3K13, APC, DLG1
GO:0016023	cytoplasmic membrane-bounded vesicle	9.39E-05	5.30E-03	550	40	SEPT5, CADM1, DRD3, GRIP1, F13A1, OXT, TH, BCAN, CNP, ITGB3, WNT2, CTTNBP2, ECE2, BDNF,

						GRIN2B, GP1BB, RAC1, DLG4, HRG, EHD3, MLANA, AP2M1, STX1A, AVP, TAOK2, GRIN1, YWHAB, GRIA3, GRIA4, CACNG2, DBH, PCLO, AP2A2, DOC2A, GRIA2, AP2A1, GRIA1, SYT17, CLTCL1, MAP6D1
GO:0031410	cytoplasmic vesicle	1.48E-04	5.80E-03	642	44	SEPT5, CADM1, DRD3, GRIP1, OXT, F13A1, TH, BCAN, CNP, ITGB3, WNT2, CTTNBP2, BDNF, ECE2, GRIN2B, GP1BB, RAC1, DLG4, HRG, EHD3, PRSS12, MLANA, AP2M1, STX1A, AVP, TAOK2, ZDHHC8, GRIN1, YWHAB, GRIA3, GRIA4, CACNG2, DBH, PCLO, AP2A2, DOC2A, GRIA2, CADPS2, AP2A1, GRIA1, SLC18A1, SYT17, CLTCL1, MAP6D1
GO:0031988	membrane-bounded vesicle	8.74E-05	5.90E-03	568	41	SEPT5, ALDOA, CADM1, DRD3, GRIP1, F13A1, OXT, TH, BCAN, CNP, ITGB3, WNT2, CTTNBP2, ECE2, BDNF, GRIN2B, GP1BB, RAC1, DLG4, HRG, EHD3, MLANA, AP2M1, STX1A, AVP, TAOK2, GRIN1, YWHAB, GRIA3, GRIA4, CACNG2, DBH, PCLO, AP2A2, DOC2A, GRIA2, AP2A1, GRIA1, SYT17, CLTCL1, MAP6D1

GO:0006897	endocytosis	3.35E-03	6.50E-03	220	18	HRAS, DRD3, ADORA2A, AHSG, CORO1C, ADRB2, AP2A2, GRIA2, AP2A1, GRIA1, APOE, RAC1, TSC2, DLG4, GRK4, CLTCL1, DNM1, VLDLR
GO:0010324	membrane invagination	3.35E-03	6.60E-03	220	18	HRAS, DRD3, ADORA2A, AHSG, CORO1C, ADRB2, AP2A2, GRIA2, AP2A1, GRIA1, APOE, RAC1, TSC2, DLG4, GRK4, CLTCL1, DNM1, VLDLR
GO:0032880	regulation of protein localization	7.01E-02	6.60E-03	138	10	CADM1, DRD3, ADORA2A, DRD2, GSK3B, NF1, DRD4, YWHAB, CDH2, APC
GO:0019901	protein kinase binding	9.59E-02	8.40E-03	147	10	NELL1, GSK3B, AVPR1A, NBEA, FER, CDH2, AKAP1, MAP3K13, APC, DLG1
GO:0007623	circadian rhythm	2.49E-04	1.00E-02	46	9	KCNMA1, NPAS2, DRD1, EGR3, DRD3, DRD2, HTR7, PER1, ADA
GO:0006970	response to osmotic stress	5.96E-03	1.12E-02	32	6	KCNMA1, AVP, OXT, RAC1, BDKRB2, SST
GO:0021826	substrate-independent telencephalic tangential migration	4.75E-04	1.32E-02	5	4	ARX, DRD1, DRD2, RAC1
GO:0044433	cytoplasmic vesicle part	2.31E-05	1.37E-02	187	21	STX1A, F13A1, TH, BCAN, GRIA3, CACNG2, ITGB3, GRIA4, DBH, WNT2, CTTNBP2, ECE2, AP2A2, GRIA2, AP2A1, GP1BB, GRIA1, DLG4, HRG, CLTCL1, AP2M1
GO:0048871	multicellular organismal homeostasis	1.13E-03	1.37E-02	85	11	GPX1, ADRB2, DRD1, RAC2, DRD2, ANKRD11, RAC1, DBH, ADD1,

						CTNNB1, HTR2A
GO:0007622	rhythmic behavior	2.10E-04	1.44E-02	16	6	KCNMA1, NPAS2, DRD1, EGR2, DRD3, ADA
GO:0009968	negative regulation of signal transduction	3.51E-03	1.44E-02	221	18	DRD3, DRD2, STRN3, NF1, PTEN, ADIPOQ, ADA, AHSG, ATXN1, AGTR2, ADRB2, TSC1, TSC2, DGKZ, GRK4, BCL6, CHRD, APC
GO:0046716	muscle maintenance	9.26E-02	1.47E-02	14	3	ALDOA, DMD, APC
GO:0021843	substrate-independent telencephalic tangential interneuron migration	4.75E-04	1.52E-02	5	4	ARX, DRD1, DRD2, RAC1
GO:0048512	circadian behavior	1.37E-03	1.54E-02	14	5	KCNMA1, NPAS2, DRD1, DRD3, ADA
GO:0012506	vesicle membrane	4.91E-05	1.66E-02	151	18	STX1A, TH, BCAN, GRIA3, CACNG2, ITGB3, GRIA4, WNT2, CTTNBP2, ECE2, AP2A2, GRIA2, AP2A1, GP1BB, GRIA1, DLG4, CLTCL1, AP2M1
GO:0030659	cytoplasmic vesicle membrane	1.67E-05	1.68E-02	139	18	STX1A, TH, BCAN, GRIA3, CACNG2, ITGB3, GRIA4, WNT2, CTTNBP2, ECE2, AP2A2, GRIA2, AP2A1, GP1BB, GRIA1, DLG4, CLTCL1, AP2M1
GO:0019216	regulation of lipid metabolic process	5.56E-02	1.71E-02	112	9	KCNMA1, AVP, DRD3, APOE, DHCR7, RAC1, AVPR1A, GHSR, ADIPOQ
GO:0030838	positive regulation of actin filament polymerization	7.04E-02	1.74E-02	12	3	ACTR3, ARPC2, RAC1
GO:0010648	negative regulation of cell	5.37E-06	1.75E-02	248	26	DRD1, DRD3, PTGS2, GRIK2,

	communication					DRD2, GRIK3, PTEN, ADA, AHSG, HTR1B, AGTR2, BCL6, APC, AVP, STRN3, NF1, ADIPOQ, ATXN1, ADRB2, TSC1, TSC2, AVPR1A, DGKZ, GRK4, CHRD, HTR2A
GO:0019905	syntaxin binding	2.65E-02	1.79E-02	31	5	WNT2, SEPT5, CTTNBP2, CPLX3, NSF
GO:0006898	receptor-mediated endocytosis	3.19E-03	1.80E-02	53	8	ADRB2, GRIA2, DRD3, APOE, GRIA1, GRK4, CLTCL1, DNM1
GO:0010741	negative regulation of protein kinase cascade	1.87E-03	1.85E-02	36	7	DRD3, DRD2, NF1, TSC2, ADIPOQ, PTEN, APC
GO:0010638	positive regulation of organelle organization	2.29E-04	1.86E-02	83	12	WNT2, ACTR3, CTTNBP2, DRD3, ARPC2, MAP1B, RAC1, TAC1, RANBP1, TPM1, SYNPO, APC
GO:0051495	positive regulation of cytoskeleton organization	3.24E-05	1.86E-02	45	10	WNT2, ACTR3, CTTNBP2, ARPC2, MAP1B, RAC1, TAC1, TPM1, SYNPO, APC
GO:0007588	excretion	6.25E-02	2.13E-02	58	6	KCNMA1, KNG1, AVP, ADORA2A, DRD2, CACNA1C
GO:0021831	embryonic olfactory bulb interneuron precursor migration	7.26E-02	2.16E-02	2	2	ARX, RAC1
GO:0019228	regulation of action potential in neuron	1.45E-04	2.19E-02	54	10	KCNMA1, SCN1A, DRD1, EGR2, TSC1, GRIK2, NF1, CLDN1, OLIG2, EIF2B5
GO:0048511	rhythmic process	9.01E-04	2.26E-02	128	14	KCNMA1, ERMP1, EGR3, DRD1, EGR2, DRD3, DRD2, OXTR, ADA, NPAS2, GRIN2B, HTR7, PER1, EIF2B5
GO:0031982	vesicle	4.69E-05	2.31E-02	670	47	SEPT5, ALDOA, CADM1, DRD3, GRIP1, OXT, F13A1,

						TH, BCAN, CNP, ITGB3, WNT2, CTTNBP2, BDNF, ECE2, GRIN2B, GP1BB, RAC1, DLG4, HRG, EHD3, PRSS12, MLANA, AP2M1, STX1A, AVP, PLD1, TAOK2, ZDHHC8, GRIN1, YWHAB, GRIA3, GRIA4, CACNG2, DBH, PCLO, AP2A2, DOC2A, GRIA2, CADPS2, GRIA1, AP2A1, PLN, SLC18A1, SYT17, CLTCL1, MAP6D1
GO:0003014	renal system process	4.66E-04	2.36E-02	28	7	KCNMA1, KNG1, AVP, AGTR2, ADORA2A, DRD2, CACNA1C
GO:0006972	hyperosmotic response	1.09E-02	2.37E-02	13	4	AVP, OXT, RAC1, SST
GO:0022028	tangential migration from the subventricular zone to the olfactory bulb	7.26E-02	2.37E-02	2	2	ARX, RAC1
GO:0005829	cytosol	1.50E-02	2.39E-02	1330	66	FHIT, GRIP1, SLC6A4, NBEA, PTEN, CTNNB1, WNT2, PRKAR2B, CTTNBP2, GPX1, GSN, RPL39L, PSMD2, RPL10, RAPGEF4, NOS2, NSF, DLG1, PIK3CG, ARHGEF2, UFD1L, PRKCG, PCM1, PRKCB, EIF4G1, CAMK4, ADK, EIF4A2, ADSL, TPH1, MAP3K13, ADD1, CPLX3, UBE3A, TH, PPP3R1, COMT, ADA, TUBGCP5, RAC1, PPP3CC, NEFL, APC, AP2M1, ITK, MAP1A, UPB1, ASMT, NAT2,

						MAP1B, YWHAB, CAMK2N2, AP2A2, EIF4E, TSC1, AP2A1, GSK3B, NTRK2, TSC2, MYO16, SPTBN1, GRK4, QPRT, DPYD, SPTAN1, CBS
GO:0021772	olfactory bulb development	3.29E-02	2.42E-02	8	3	ARX, DLX2, RAC1
GO:0042592	homeostatic process	8.63E-07	2.43E-02	751	56	SLC9A9, CYP11B1, GRIK2, GRIK3, CTNNB1, DMPK, GPX1, GRIN2B, APOE, ANKRD11, GRID2, OLIG2, EIF2B5, KCNMA1, KNG1, AVP, EGR2, CACNG2, PRKCB, ADRB2, CLDN1, TXNRD2, ADD1, ALDOA, DRD1, SCN1A, CCK, DRD3, ADORA2A, DRD2, OXT, DRD4, GPR6, TAC1, OXTR, SFXN1, BDKRB2, RAC2, DMD, RAC1, BCL6, APC, NF1, GRIN1, NLGN3, DBH, ADIPOQ, ATXN1, TSC1, PLN, AVPR1A, CACNA1G, CACNA1E, CACNA1F, CACNA1C, HTR2A
GO:0021988	olfactory lobe development	4.13E-02	2.60E-02	9	3	ARX, DLX2, RAC1
GO:0000149	SNARE binding	1.12E-02	2.62E-02	37	6	WNT2, SEPT5, CTTNBP2, CPLX3, STX1A, NSF
GO:0030665	clathrin coated vesicle membrane	4.73E-02	2.67E-02	53	6	STX1A, AP2A2, GRIA2, AP2A1, BCAN, CLTCL1
GO:0070160	occluding junction	1.91E-02	2.68E-02	73	8	ARHGEF2, CLDN5, LIN7B, CLDN1, STRN, LIN7C, SYNPO, APC
GO:0050998	nitric-oxide synthase binding	4.15E-02	2.75E-02	9	3	WNT2, CTTNBP2, DMD
GO:0001508	regulation of action potential	3.58E-05	2.79E-02	68	12	KCNMA1, SCN1A, DRD1, EGR2,

						GRIN2B, TSC1, GRIK2, NF1, CLDN1, TAC1, OLIG2, EIF2B5
GO:0042311	vasodilation	2.86E-06	2.80E-02	26	9	WNT2, KCNMA1, KNG1, CTTNBP2, GPX1, ADRB2, AGTR2, ADORA2A, APOE
GO:0005923	tight junction	1.91E-02	2.95E-02	73	8	ARHGEF2, CLDN5, LIN7B, CLDN1, STRN, LIN7C, SYNPO, APC
GO:0048839	inner ear development	7.17E-02	2.97E-02	79	7	KCNMA1, HOXA1, EYA1, BDNF, DLX6, TBX1, FZD3
GO:0030135	coated vesicle	6.70E-03	3.04E-02	159	14	SEPT5, STX1A, CADM1, TH, GRIN1, BCAN, PCLO, AP2A2, DOC2A, GRIA2, GRIN2B, AP2A1, SYT17, CLTCL1
GO:0051345	positive regulation of hydrolase activity	6.84E-02	3.13E-02	179	12	DRD1, AGTR2, CCK, TSC1, DRD2, NF1, TSC2, AVPR1A, HOMER1, HTR2C, TPM1, HTR2A
GO:0001666	response to hypoxia	2.69E-02	3.15E-02	134	11	KCNMA1, FLT1, NF1, TH, ARNT2, NR4A2, ABAT, NOS2, ADIPOQ, ADA, VLDLR
GO:0006940	regulation of smooth muscle contraction	4.21E-04	3.15E-02	38	8	KCNMA1, ADRB2, FLT1, PTGS2, OXT, OXTR, GHSR, ADA
GO:0070482	response to oxygen levels	1.55E-02	3.26E-02	141	12	KCNMA1, FLT1, NF1, TH, ARNT2, NR4A2, ABAT, OXTR, NOS2, ADIPOQ, ADA, VLDLR
GO:0030877	beta-catenin destruction complex	7.96E-03	3.28E-02	4	3	GSK3B, CTNNB1, APC
GO:0030136	clathrin-coated vesicle	1.33E-03	3.42E-02	132	14	SEPT5, STX1A, CADM1, TH, GRIN1, BCAN, PCLO, AP2A2, DOC2A, GRIA2, GRIN2B, AP2A1, SYT17, CLTCL1

GO:0034747	Axin-APC-beta-catenin-GSK3B complex	4.08E-03	3.43E-02	3	3	GSK3B, CTNNB1, APC
GO:0019899	enzyme binding	1.04E-03	3.50E-02	523	35	GRIK2, NELL1, STRN, NBEA, FER, BDKRB2, CDH2, AKAP11, CTNNB1, WNT2, CTTNBP2, RANBP9, ANK1, DMD, RAC1, RANBP1, LAMB1, NEFL, DLG1, APC, ARHGEF2, STRN3, STRN4, YWHAB, DOCK8, PPP1R9B, ADRB2, GSK3B, AVPR1A, ABAT, CYFIP1, CACNA1C, AKAP1, MAP3K13, CBS
GO:0001894	tissue homeostasis	8.29E-02	3.55E-02	63	6	ADRB2, RAC2, ANKRD11, RAC1, ADD1, CTNNB1
GO:0045744	negative regulation of G-protein coupled receptor protein signaling pathway	1.37E-03	3.68E-02	14	5	ADRB2, DRD3, DRD2, GRK4, ADA
GO:0046033	AMP metabolic process	2.53E-02	3.95E-02	7	3	AK3, ADSL, NT5E
GO:0008013	beta-catenin binding	9.82E-02	3.98E-02	30	4	DVL3, GSK3B, CDH2, APC
GO:0032844	regulation of homeostatic process	1.03E-03	4.09E-02	114	13	DRD1, AVP, CCK, DRD2, OXT, NF1, TAC1, CDH2, ADA, AHSG, CD44, CACNA1G, AVPR1A
GO:0043297	apical junction assembly	1.27E-02	4.12E-02	5	3	STRN, CTNNA1, APC
GO:0007589	body fluid secretion	2.78E-03	4.32E-02	27	6	VIP, KCNMA1, KNG1, AVP, ADORA2A, DRD2
GO:0030139	endocytic vesicle	5.89E-05	4.36E-02	59	11	AP2A2, GRIA2, DRD3, AP2A1, GRIA1, DLG4, GRIA3, CACNG2, GRIA4, EHD3, AP2M1
GO:0045667	regulation of osteoblast differentiation	7.34E-02	4.49E-02	43	5	NELL1, OSTN, CHRD, CTNNB1, APC

GO:0016281	eukaryotic translation initiation factor 4F complex	4.22E-02	4.50E-02	9	3	EIF4G1, EIF4E, EIF4A2
GO:0005244	voltage-gated ion channel activity	5.91E-02	4.52E-02	195	13	KCNMA1, SCN1A, CLCN2, SCN2A, CACNG2, KCNV2, CACNA1G, CACNA1H, CACNA1E, CACNA1F, KCTD13, CACNA1C, CACNA1D
GO:0022832	voltage-gated channel activity	5.91E-02	4.57E-02	195	13	KCNMA1, SCN1A, CLCN2, SCN2A, CACNG2, KCNV2, CACNA1G, CACNA1H, CACNA1E, CACNA1F, KCTD13, CACNA1C, CACNA1D
GO:0005885	Arp2/3 protein complex	6.18E-05	4.58E-02	7	5	ACTR3, ACTR2, ARPC3, ARPC2, ARPC4
GO:0031112	positive regulation of microtubule polymerization or depolymerization	1.09E-02	4.71E-02	13	4	WNT2, CTTNBP2, MAP1B, APC
GO:0009167	purine ribonucleoside monophosphate metabolic process	3.58E-02	4.82E-02	20	4	AK3, ADSL, NT5E, ADA
GO:0009126	purine nucleoside monophosphate metabolic process	3.58E-02	4.83E-02	20	4	AK3, ADSL, NT5E, ADA
GO:0031116	positive regulation of microtubule polymerization	8.60E-03	4.97E-02	12	4	WNT2, CTTNBP2, MAP1B, APC
GO:0005184	neuropeptide hormone activity	9.41E-03	4.98E-02	23	5	VIP, AVP, CCK, PENK, OXT
GO:0019725	cellular homeostasis	2.94E-08	5.01E-02	466	44	ALDOA, DRD1, SCN1A, CCK, DRD3, ADORA2A, DRD2, GRIK2, GRIK3, OXT, DRD4, GPR6, TAC1, OXTR, BDKRB2, DMPK, GPX1, GRIN2B, APOE, DMD, GRID2,

						OLIG2, EIF2B5, APC, KCNMA1, KNG1, AVP, EGR2, NF1, GRIN1, NLGN3, CACNG2, ADIPOQ, PRKCB, ATXN1, TSC1, PLN, AVPR1A, CLDN1, CACNA1G, TXNRD2, CACNA1F, CACNA1C, ADD1
GO:0030666	endocytic vesicle membrane	2.19E-05	5.15E-02	33	9	AP2A2, GRIA2, AP2A1, GRIA1, DLG4, GRIA3, CACNG2, GRIA4, AP2M1
GO:0031113	regulation of microtubule polymerization	1.09E-02	5.28E-02	13	4	WNT2, CTTNBP2, MAP1B, APC
GO:0050878	regulation of body fluid levels	3.63E-02	5.34E-02	141	11	VIP, KCNMA1, KNG1, AVP, ADORA2A, GP1BB, DRD2, OXT, F13A1, ITGB3, DTNBP1
GO:0043296	apical junction complex	3.19E-02	5.49E-02	99	9	ARHGEF2, CLDN5, LIN7B, CLDN1, STRN, LIN7C, CTNNB1, SYNPO, APC
GO:0003779	actin binding	4.90E-08	5.55E-02	326	35	MYO5A, ALDOA, ARPC4, FER, TPM2, CAPZB, TPM1, TPM3, CORO2B, ACTR3, ACTR2, MACF1, ARPC3, ARPC2, GSN, DMD, SYNPO, KCNMA1, PHACTR1, MAP1A, MAP1B, CORO1C, ARPC1A, PPP1R9B, CORO1B, PPP1R9A, SYNE1, MYO16, CYFIP1, SPTBN1, DBN1, ADD2, SNTG2, ADD1, SPTAN1
GO:0016327	apicolateral plasma membrane	3.71E-02	5.64E-02	102	9	ARHGEF2, CLDN5, LIN7B, CLDN1, STRN, LIN7C, CTNNB1, SYNPO, APC
GO:0046578	regulation of Ras protein signal	4.83E-02	5.66E-02	210	14	GIT1, ARFGAP1, HRAS, ARHGEF2,

	transduction					NF1, CDH2, ARHGEF10, MCF2L2, TSC1, RAC1, DGKZ, BCL6, TBC1D7, AGAP1
GO:0006937	regulation of muscle contraction	1.16E-05	5.97E-02	72	13	KCNMA1, FLT1, PTGS2, OXT, OXTR, TPM1, ADA, TPM3, DMPK, SSTR2, ADRB2, CACNA1G, GHSR
GO:0007628	adult walking behavior	1.11E-04	6.02E-02	22	7	KCNMA1, EPHA4, SCN1A, DRD1, EFNB3, DRD2, CACNA1C
GO:0042755	eating behavior	1.31E-03	6.11E-02	23	6	DRD1, CCK, ADORA2A, OXT, TH, OXTR
GO:0051130	positive regulation of cellular component organization	1.22E-05	6.14E-02	181	21	CCK, DRD3, OXT, MAP1B, OXTR, TAC1, TPM1, AHSG, WNT2, NTRK3, ACTR3, CTTNBP2, TSC1, ROBO1, ARPC2, GSK3B, RAC1, RANBP1, NEFL, SYNPO, APC
GO:0001656	metanephros development	7.85E-02	6.19E-02	44	5	EYA1, BDNF, CD44, NF1, PBX1
GO:0016328	lateral plasma membrane	9.44E-02	6.39E-02	14	3	CLDN1, CTNNB1, APC
GO:0009898	internal side of plasma membrane	8.02E-04	6.44E-02	316	25	HRAS, TH, DMD, RAC1, DLG4, AP3D1, GNG2, EHD3, AP2M1, PLD1, GNB1L, RAB39B, STMN2, MAOA, CTNNA1, CTNNA2, SYNE1, AP2A2, CADPS2, AP2A1, SPTBN1, CLTCL1, ADD2, ADD1, SNTG2
GO:0050880	regulation of blood vessel size	2.03E-05	6.44E-02	53	11	WNT2, KCNMA1, KNG1, CTTNBP2, GPX1, ADRB2, AVP, AGTR2, ADORA2A, APOE, CACNA1G
GO:0035150	regulation of tube size	2.03E-05	6.62E-02	53	11	WNT2, KCNMA1, KNG1, CTTNBP2, GPX1, ADRB2, AVP, AGTR2, ADORA2A,

						APOE, CACNA1G

Polygenic burden tests were performed on the 989 GO gene sets, using the same test as the genic burden test above, but looking at the collective allele effects at the polygenic level, instead of individual gene level. The overall test results were moderately inflated compared to the theoretical null, possibly due to the functional grouping of genes (GO sets) that are relevant to OCD (FIG. 11). The top five GO sets, which deviated even more from the null, included three GO sets related to regulation of cell death (GO:0010942, GO:0043065, and GO:0043068, $p = 3 \times 10^{-4} \sim 5 \times 10^{-4}$, corr. $p < 0.03 \sim 0.05$), positive regulation of protein complex assembly (GO:0031334, $p = 7 \times 10^{-4}$, corr. $p < 0.06$), and anatomical structure homeostasis (GO:006024, $p = 1.3 \times 10^{-3}$, corr. $p < 0.1$) (Tables 4 and 5). Additional functional themes were identified among the top 82 GO sets with nominal burden ($p < 0.05$), i.e. endocytosis, rhythmic process, cytoplasmic membrane-bounded vesicle, tight junction, and protein kinase binding (Table 5). Additionally, overlaying the polygenic burden test results onto the GO genetic networks allowed for identification of clusters of GO sets with strong p-values, such as regulation of protein polymerization and cytoskeleton organization, regulation of action potential, telencepahlic tangential migration, and membrane-bounded vesicle (Table 5).

Table 5.

GOID	GO description	Burden P	No. Total Genes	No. Target Genes	Target Genes
GO:0010942	positive regulation of cell death	3.00E-04	435	30	CCK, CADM1, PTGS2, ADORA2A, GRIK2, NELL1, PPP3R1, PTEN, GPX1, AGTR2, CD44, APOE, RAC1, PPP3CC, BCL6, APC, KNG1, KCNMA1, ARHGEF2, GRIN1, NF1, YWHAB, LGALS12, MBD4, ADIPOQ, DAPK1, ADRB2, TNFSF10, SST, PRODH
GO:0031334	positive regulation of protein	7.00E-04	35	9	WNT2, ACTR3, CTTNBP2, CCK, ARPC2, GSK3B, MAP1B, RAC1, APC

	complex assembly				
GO:0060249	anatomical structure homeostasis	1.30E-03	106	9	ALDOA, ADRB2, RAC2, DMD, ANKRD11, RAC1, ADD1, CTNNB1, APC
GO:0070830	tight junction assembly	5.10E-03	2	2	STRN, APC
GO:0019900	kinase binding	5.20E-03	179	13	NELL1, NBEA, CDH2, FER, CTNNB1, WNT2, CTTNBP2, GSK3B, AVPR1A, AKAP1, MAP3K13, APC, DLG1
GO:0016023	cytoplasmic membrane-bounded vesicle	5.30E-03	550	40	SEPT5, CADM1, DRD3, GRIP1, F13A1, OXT, TH, BCAN, CNP, ITGB3, WNT2, CTTNBP2, ECE2, BDNF, GRIN2B, GP1BB, RAC1, DLG4, HRG, EHD3, MLANA, AP2M1, STX1A, AVP, TAOK2, GRIN1, YWHAB, GRIA3, GRIA4, CACNG2, DBH, PCLO, AP2A2, DOC2A, GRIA2, AP2A1, GRIA1, SYT17, CLTCL1, MAP6D1
GO:0021826	substrate-independent telencephalic tangential migration	1.32E-02	5	4	ARX, DRD1, DRD2, RAC1
GO:0007622	rhythmic behavior	1.44E-02	16	6	KCNMA1, NPAS2, DRD1, EGR2, DRD3, ADA
GO:0019228	regulation of action potential in neuron	2.19E-02	54	10	KCNMA1, SCN1A, DRD1, EGR2, TSC1, GRIK2, NF1, CLDN1, OLIG2, EIF2B5

GOID = Gene Ontology ID

Evaluation of individual variants leads to experimentally tractable candidate variants

Individual variants contributing to the significant genic burdens that were identified were further investigated. A list of candidate variants was compiled based on the p-values produced by comparing case-control frequency of a single variant (*psingle*), ratio between the case and control allele frequencies, and novelty. 897 (low and high confidence) variants were

identified in the five candidate genes, of which 218 variants met the selection criteria (*LIPH*, 7/19 [36.8%]; *NRXN1*, 152/651 [23.3%]; *HTR2A* 16/71 [22.5%]; *CTTNBP2*, 33/114 [28.9%]; *REEP3*, 11/43 [25.6%]; FIG. 2).

Of the seven *LIPH* candidate variants, four variants, i.e. three variants (rs55854644, Chr3:185226492 [C>T], MAF 0.143, *psingle* = 0.031, OCD AF=0.27, control AF=0.23; rs6788865, Chr3:185226396 [A>G], MAF 0.37, *psingle* = 0.030, OCD AF=0.47, control AF=0.43; rs141497229, Chr3:185225638 [A>T], MAF 0.0092, *psingle* = 0.066, OCD AF=0.025, control AF=0.015) and one private (estimated allele count=1) novel variant (Chr3:185225644 [A>C]), were located in the 3'UTR (FIGs. 2 and 5A, and Table 6, which includes Table 6A-6C). Three miRNAs (miR-23ab, let-7/98, and miR-874) have been predicted to bind near these four variants in the 3'UTR [ref. 5], suggesting that these candidate mutations might be affecting the binding affinity of miRNAs that regulate *LIPH* expression (FIG. 12A). Additionally, a synonymous (rs61730237, Chr3:185229464 [T>C], MAF 0.403, *psingle* = 0.05, OCD AF=0.47, control AF=0.44) and a private noncoding variant (Chr3:185241792 [G>C]) were located near (~20bp) the both splice sites of the second last exon of *LIPH* (366 to 423 amino acids of UniProt protein Q8WWY8) that contains a transmembrane domain surrounded by a glycosylated residue (357aa) and a disulfide bond (427↔446aa), raising a possibility for a role for these variants in splicing out the transmembrane domain of *LIPH* (FIG. 2).

Table 6A.

Chr	Position	Ref	Alt	rsID	OCD AF	Control AF	MAF (UCSC)	P single	Gene	Potential Function	Amino acid change	GERP score
3	185241792	G	C	N/A	9.13E-04	5.44E-05	N/A	3.23E-01	LIPH	~70bp Away Exon	N/A	-0.838
3	185229283	C	T	rs200468410	4.25E-03	2.22E-03	3.80E-03	3.55E-01	LIPH	~30 bp Away Exon		-0.579
3	185226492	C	T	rs55854644	2.72E-01	2.35E-01	1.43E-01	3.12E-02	LIPH	3'UTR		3.77
3	185229464	T	C	rs61730237	4.68E-01	4.39E-01	4.03E-01	5.00E-02	LIPH	syn, cluster with Chr3:185241792		1.42

3	185226396	A	G	rs6788865	4.67E-01	4.33E-01	3.70E-01	3.04E-02	LIPH	3'UTR		-1.64
3	185225638	A	T	rs141497229	2.46E-02	1.47E-02	9.18E-03	6.62E-02	LIPH	3'UTR		2.51
3	185225644	A	C	N/A	8.79E-04	3.69E-05	N/A	3.28E-01	LIPH	3'UTR		0.784
2	50850686	G	A	rs2303298	1.07E-02	1.33E-03	5.00E-02	5.30E-04	NRXN1	syn, 5'UTR, nc	N/A	3.51
2	50723068	G	A	rs56402642	3.44E-03	6.56E-05	2.30E-03	4.64E-02	NRXN1	syn, nc	N/A	-11.2
2	50765412	G	T	rs56086732	8.95E-03	2.82E-03	3.20E-03	5.55E-02	NRXN1	mis, nc	L[CTT]>I[ATT]	5.11
2	50765589	T	C	rs200074974	1.61E-03	9.11E-06	N/A	1.41E-01	NRXN1	mis, nc	I[ATC]>V[GTC]	5.16
2	50733745	G	C	rs147984237	1.73E-03	1.54E-05	2.70E-03	1.58E-01	NRXN1	syn, nc	N/A	3.41
2	50847195	G	A	rs78540316	7.70E-03	3.63E-03	2.00E-03	1.61E-01	NRXN1	mis, nc	P[CCA]>S[TCA]	6.16
2	50779943	T	C	novel	8.62E-04	6.72E-06	N/A	3.18E-01	NRXN1	mis, nc	N[TTA]>S[TCA]	5.92
2	50724642	A	G	novel	8.50E-04	2.60E-06	N/A	3.22E-01	NRXN1	mis, nc	I[TAT]>T[TGT]	5.58
2	50779791	C	T	novel	8.48E-04	4.60E-06	N/A	3.25E-01	NRXN1	mis, nc	A[CGG]>T[TGG]	5.93
2	50464065	C	T	rs80094872	8.92E-04	3.03E-04	8.00E-03	5.12E-01	NRXN1	syn, nc	N/A	-10.8
2	50724745	G	T	rs201818223	1.74E-03	9.16E-04	6.00E-03	5.84E-01	NRXN1	mis, nc	L[CTG]>M[ATG]	4.18
2	50463984	G	A	rs147580960	8.69E-04	6.00E-04	2.00E-04	7.88E-01	NRXN1	syn, nc	N/A	-11.2
13	47466622	G	A	rs6305	3.86E-02	2.25E-02	2.00E-02	1.88E-02	HTR2A	syn	N/A	2.58
13	47409048	G	A	rs6308	3.58E-03	1.23E-03	2.60E-03	2.26E-01	HTR2A	mis	A[GCT]>V[GTT]	3.9
13	47409701	G	A	rs141413930	1.98E-03	1.20E-03	1.00E-04	6.27E-01	HTR2A	syn	N/A	-11.9
13	47408946	C	A	rs370829834	9.82E-04	3.55E-05	N/A	2.76E-01	HTR2A	3'UTR	N/A	2.63

Chr= Chromosome, Pos= Position, Ref = reference allele, Alt= alternative allele, rsID = rs identifier, OCD AF= allele frequency in OCD cases, Control AF = allele frequency in controls MAF = minor allele frequency, P single = p-values from single variant association test, syn – synonymous, nc = non-coding, mis = missense.

Table 6B.

Chr	Pos	Ref	Alt	rsID	OCD AF	Control AF	MAF(UCSC)	P single	Gene	Functional Annotation	GERP score
7	117468056	C	T	rs2067080	5.69E-01	5.06E-01	5.30E-01	1.23E-03	CTTNBP2	dhs	4.11
7	117456904	C	T	rs12706157	6.50E-01	5.94E-01	6.20E-01	1.30E-03	CTTNBP2	dhs	0.0946
7	117450810	C	T	rs34868515	9.95E-04	4.05E-05	2.00E-02	2.72E-01	CTTNBP2	dhs	-0.41
7	117417559	A	G	rs75322384	3.65E-03	1.63E-03	2.70E-03	2.89E-01	CTTNBP2	dhs	5.63
7	117390966	T	D	N/A	8.82E-04	2.89E-05	N/A	3.21E-01	CTTNBP2	dhs	-0.477
7	117356081	T	G	N/A	8.77E-04	2.99E-05	N/A	3.23E-01	CTTNBP2	dhs	-0.1
7	117421141	C	A	N/A	9.76E-04	4.66E-04	N/A	5.64E-01	CTTNBP2	dhs	2.07
10	65307923	A	G	rs78109635	5.18E-02	3.05E-02	2.00E-02	2.49E-03	REEP3	dhs	-5.76
10	65332906	T	C	rs76646063	6.37E-03	2.04E-03	1.00E-02	8.34E-02	REEP3	dhs	2.87
7	117431079	T	G	rs113126080	1.55E-02	3.51E-03	4.13E-03	2.81E-03	CTTNBP2	N/A, cluster	1.25
7	117457141	G	C	rs13242822	6.59E-01	6.18E-01	6.47E-01	3.87E-03	CTTNBP2	N/A	0
7	11746833	T	C	rs2111209	6.55E-01	6.10E-01	7.20E-01	6.36E-03	CTTNBP2	N/A	-0.708
7	117385978	G	T	rs10274022	1.27E-03	5.38E-05	1.00E-02	1.59E-01	CTTNBP2	mis: Q[CAA]>K[AAA]	4.98
7	117358107	T	C	rs142089340	3.51E-03	9.51E-04	1.37E-03	1.74E-01	CTTNBP2	mis: T[ACT]>A[GCT], nc	-2.34
7	117431202	C	A	N/A	9.63E-04	3.93E-05	N/A	2.84E-01	CTTNBP2	mis: S[AGC]>I[ATC], cluster	5.54
7	117431704	C	T	N/A	8.81E-04	1.42E-05	N/A	3.13E-01	CTTNBP2	mis: G[GGG]>R[AGG], cluster	5.6
7	117396664	C	A	rs200975491	8.89E-04	1.45E-05	9.00E-04	3.14E-01	CTTNBP2	mis: S[AGT]>I[ATT]	5.43
7	117431879	G	A	N/A	8.75E-04	2.46E-05	N/A	3.21E-01	CTTNBP2	syn, cluster	-0.221
7	117501314	G	A	N/A	8.97E-04	4.85E-05	N/A	3.25E-01	CTTNBP2	syn	-4.7
10	65387644	C	G	rs56311840	8.60E-04	2.60E-05	4.13E-03	3.32E-01	REEP3	dhs	-1.96
10	65387722	C	D	N/A	1.09E-03	1.52E-05	N/A	2.25E-01	REEP3	dhs	1.4

10	65388750	G	A	N/A	9.00E-04	9.01E-05	N/A	3.57E-01	REEP3	dhs	5.79
10	65384621	C	T	rs7089835	1.87E-03	3.03E-04	8.89E-03	1.96E-01	REEP3	3'UTR	2.8

Table 6C.

Chr	Pos	Ref	Alt	rsID	DHS/TF	EMSA
7	117468056	C	T	rs2067080	64DHS,10TF	
7	117456904	C	T	rs12706157	11DHS,1TF	
7	117450810	C	T	rs34868515	5DHS,9TF	
7	117417559	A	G	rs75322384	16DHS,H3K36me3	
7	117390966	T	D	N/A	62DHS,3TF	
7	117356081	T	G	N/A	106DHS,7TF	++
7	117421141	C	A	N/A	31DHS,17TF,H3K27me3	
10	65307923	A	G	rs78109635	33DHS,2TF	+
10	65332906	T	C	rs76646063	72DHS,4TF	++
7	117431079	T	G	rs113126080	2DHS(~10bp away)	
7	117457141	G	C	rs13242822	same as rs12706157	
7	11746833	T	C	rs2111209	same as rs2067080	
7	117385978	G	T	rs10274022	NoEvidence	
7	117358107	T	C	rs142089340	3DHS (~200bp away),1TF	
7	117431202	C	A	N/A	2DHS(~100bp away)	
7	117431704	C	T	N/A	2TF	
7	117396664	C	A	rs200975491	NoEvidence	
7	117431879	G	A	N/A	1TF	
7	117501314	G	A	N/A	4DHS(~450bp away)	
10	65387644	C	G	rs56311840	45DHS,1TF	
10	65387722	C	D	N/A	same as rs56311840	
10	65388750	G	A	N/A	91DHS	
10	65384621	C	T	rs7089835	NoEvidence	

DHS/TF = DNase Hypersensitivity site/Transcription Factor, - = no change; + = shift; ++ = difference in shift between ref and alt.

Focusing on the coding variants for *NRXN1* and *HTR2A* that showed significant burdens of exonic variants in cases, fifteen coding variants were identified (twelve in *NRXN1* and three in *HTR2A*) that passed the selection criteria. Interestingly, all twelve *NRXN1* candidate coding variants were found within a specific isoform (UCSC gene uc002rxc.1 at chr2:50,426,275- 50,883,552) among hundreds of possible alternative transcripts [ref. 6],

suggesting a potential role for this isoform in OCD (FIG. 2). Of the twelve coding variants, seven are missense, of which three are novel private variants (chr2:50779943 [T>C], N>S; chr2:50724642 [A>G], I>T; and chr2:50779791 [C>T], A>T; Table 6 and Table 7) and four have been reported in dbSNP 138 previously (rs56086732, L>I, *psingle* = 0.056, OCD AF=0.009, control AF=0.003 ; rs200074974, I>V, *psingle* = 0.14, OCD AF=0.0016, control AF~0; rs78540316, P>S, *psingle* = 0.16, OCD AF=0.008, control AF=0.004; and rs201818223, L>M, *psingle* = 0.58, OCD AF=0.002, control AF=0.0009; Table 6 and Table 7). All of these seven missense mutations change amino acids of laminin G or EGF-like domains that are important in binding other synaptic adhesion molecules, hence potentially affecting NRXN1's role in synapse formation and/or maintenance.

Two of the three HTR2A candidate coding variants, a missense and a synonymous variant, were located in the last coding exon (205 to 471aa in the UniProt protein P28223). Of note, the missense variant (rs6308, A>V, *psingle* = 0.627, OCD AF=0.002, control AF=0.001) was located in the cytoplasmic domain which contains the PDZ-binding motif where protein interacting partners bind. Previous mutagenesis experiments, which happened to test several missense variants within the 102bp window between the missense and a private 3'UTR candidate mutation (rs370829834), have shown these mutations to cause a loss of interaction with PDZ-based scaffolding proteins at synapse such as CASK, DLG1, DLG4, and APBA1, or alternatively, to acquire HTR2C binding properties instead of HTR2A7. This suggests that the identified *HTR2A* candidate mutations might affect the binding affinity or specificity to HTR2A's interacting partners.

Candidate DHS variants for *CTTNBP2* and *REEP3* that showed significant burdens of DHS variants in cases were also analyzed. In *CTTNBP2*, seven DHS variants passed the candidate selection criteria (FIG. 2). All of the seven variants overlap with DHS in neural stem cells (SK-N-MC) or neuroblasts (SK-N-SH/BE2 C/SH-SY5Y/SK-N-SH RA) as well as several other ENCODE cell lines [ref. 8]. The *CTTNBP2* DHS candidate variants also overlapped with binding sites of several transcription factors including FOXP2, CTCF, and RAD21 in the brain-derived cell lines, suggesting that the candidate mutations may be disrupting transcription factor binding sites for *CTTNBP2* (Table 6). Additionally, five of the seven coding variants in *CTTNBP2* that passed the selection criteria resided within or near regulatory marks, despite the lack of DHS annotation to those variants: Three private coding

variants, i.e. two missense variants (Chr7:117431704 [C>T], G>R and Chr7:117431202 [C>A], S>I) and one synonymous (Chr7:117431879 [G>A]) variant, congregated together with an additional noncoding candidate variant (rs113126080, Chr7:117431079 [T>G], MAF 0.0041, *psingle* = 0.0028, OCD AF=0.016, control AF=0.0035), within a ~750bp regulatory region that is indicated by H3K27Ac and/or H3K4Me1 marks in several cell lines including differentiated neuroblasts SK-N-SH RA (FIG. 12B). The remaining two coding variants were located 200~450bp away from two different DHS in neuroblasts (Table 6). Additionally, two noncoding candidate variants without Exon or DHS annotations (rs13242822, Chr7:117457141 [G>C], AF 0.647, *psingle* = 0.0039, OCD AF=0.66, control AF=0.62 and rs2111209, Chr7:11746833 [T>C], AF 0.647, *psingle* = 0.0064, OCD AF=0.66, control AF=0.61) resided within the same regulatory regions (indicated by transcription factor binding sites and histone marks) for the top two DHS variants (rs12706157 and rs2067080, respectively; FIG. 12B), proposing additional regulatory candidate variants for *CTTNBP2*.

In *REEP3*, two DHS variants passed the selection criteria for candidacy (FIG. 2): One candidate variant was rs78109635 (Chr10:65307923 [A>G], MAF 0.017, *psingle* = 0.0025, OCD AF=0.052, control AF=0.030) that overlapped with DHS marks and GATA2 transcription factor binding sites in neuroblasts. The other variant is rs76646063 (Chr10:65332906 [T>C], MAF 0.007, *psingle* = 0.083, OCD AF=0.006, control AF=0.002), which overlaps with DHS marks in neural stem cells and with binding sites for GATA2, GATA3, and EP300 in neuroblasts and POLR2A in neural stem cells (Table 6). In addition to the two DHS candidate variants within *REEP3*, there are three private DHS candidate variants clustering ~3Kb upstream of *REEP3* (FIG. 2). These variants overlapped with DHS and binding sites for POLR2A and EP300 in neuroblasts and/or neural stem cells (FIG. 12C, Table 6), providing additional regulatory candidate variants for *REEP3*.

Gel shift assay validates functions of candidate regulatory variants

All nine candidate DHS variants and nine additional regulatory candidate variants identified in *CTTNBP2* (seven DHS and six additional variants) and *REEP3* (two genic and three intergenic DHS variants) were tested as to whether these variants affected protein bindings using gel shift assays (FIGs. 2 and 3). While both of the candidate DHS sites in REEP3 showed protein bindings to the reference alleles, a clear reduction of the protein

binding was shown for rs76646063, but no apparent difference for rs78109635. Of the seven DHS candidate variants for CTTNBP2, at least three variants showed differential protein bindings between reference and risk alleles. The variant Chr7:117356081 [T>G] clearly reduced the protein binding that existed to the reference allele T. These results validate that at least multiple candidate regulatory mutations indeed affected protein binding to their regulatory regions *in vitro*. It may well be the case that the remaining variants will also be validated when the assay conditions are optimized for individual test sites accounting for different transcription factor binding conditions, length of binding motifs, etc.

In further studies, we performed electrophoretic mobility shift assays (EMSA) in a human neuroblastoma cell line, SK-N-SH, for a subset of 18 candidate regulatory variants in or near *CTTNBP2* and *REEP3* (12 DHS and 6 non-DHS; Table 7), and found evidence for at least six variants that cause clear changes in transcription factor binding to these DNA sequences (FIGs. 14A and 14B).

In *REEP3,* the two tested DHS candidate variants, chr10:65307923A>G and chr10:65332906T>C showed a clear reduction of specific DNA-protein binding for the alternate allele relative to its reference allele (FIG. 14A). Both variants were validated for the same directional effect in the genotyping cohort (Table 7). Of the three candidate DHS variants that share the same regulatory element upstream of *REEP3*, chr10:65387722C>D showed evidence of differential DNA-protein binding.

Of the seven DHS candidate variants tested for *CTTNBP2*, three variants (chr7:117456904C>T, chr7:117417559A>G, and chr7:117356081T>G) showed clear differential DNA-protein bindings (FIG. 14B). Chr7:117456904C>T and chr7:117417559A>G were validated for the same directional effect in the genotyping cohort, and chr7:117356081T>G was private to an OCD patient in the sequencing cohort (Table 7). The other two DHS variants that were private to OCD patients in the sequencing cohort (chr7:117390966T>D and chr7:117421141C>A) showed less clear patterns (data not shown). Of the six *CTTNBP2* candidate regulatory variants without DHS annotation, chr7:117431879G>A, a private variant to an OCD patient in the sequencing cohort, showed clear changes to specific DNA-protein binding relative to the reference allele (FIG. 14B and Table 7), while chr7:117431704C>T, another private variant to OCD patient in the sequencing cohort, showed a less clear pattern (data not shown).

Overall, the results confirmed regulatory activities of four *CTTNBP2* candidate mutations (three noncoding and one synonymous) and two *REEP3* noncoding candidate mutations in a human neuroblastoma cell line. Of the six, all the four genotyped variants were validated for the same directional effect in the genotyping cohort (FIGs. 14A and 14B and Table 7).

In the EMSA experiments, 8/12 (67%) tested DHS candidate variants were confirmed or showed some evidence of regulatory activity. This result supports the predictability of regulatory activities by DHS annotations.

Table 7. Missense and regulatory variants for OCD

Chr:pos	Ref	Alt	rsID	OCD AF	Ctrl AF	Gene	Function
chr2:50779943	T	C	NA	AC=1	0	*NRXN1*	Missense (N>S)
chr2:50724642	A	G	NA	AC=1	0		Missense (I>T)
chr2:50779791	C	T	NA	AC=1	0		Missense (A>T)
chr2:50765412	G	T	rs56086732	0.0075	0.0046		Missense (L>I)
chr2:50765589	T	C	rs200074974	0.0016*	0*		Missense (I>V)
chr2:50847195	G	A	rs78540316	0.0031	0.0043		Missense (P>S)
chr2:50724745	G	T	rs201818223	0.002*	0.0009*		Missense (L>M)
chr10:65307923	A	G	rs78109635	0.0404	0.0400	*REEP3*	Regulatory (DHS)
chr10:65332906	T	C	rs76646063	0.0067	0.0018		Regulatory (DHS)
chr7:117456904	C	T	rs12706157	0.629	0.616	*CTTNBP2*	Regulatory (DHS)
chr7:117417559	A	G	rs75322384	0.0054	0.0025		Regulatory (DHS)
chr7:117356081	T	G	NA	AC=1*	0*		Regulatory (DHS)
chr7:117431879	G	A	NA	AC=1*	0*		Regulatory (nonDHS)

* Data from sequencing cohort
AF=Allele Frequency; AC=Allele Count

Correlation to canine and mouse OCD genes

Additionally, we examined whether the genes implicated in OCD animal models could be reidentified in the human data set. Five genes were analyzed: *CDH2* (neural cadherin), *CTNNA2* (catenin alpha2), *ATXN1* (ataxin-1), and *PGCP/CPQ* (plasma glutamate carboxypeptidase/ carboxypeptidase Q) that were identified from a GWAS followed by targeted resequencing in dog OCD, and *SAPAP3/ DLGAP3*, which, when deleted, causes OCD-like behavior and abnormal activities of the CSTC neurocircuit in mice. Among the sixteen genic burden tests performed, *CDH2* showed the strongest association in [DHS-Cons.] test ($p = 0.03$), *CTNNA2* in Rare tests ($p = 0.02 \sim 0.05$; [Rare-All], [Rare-Evo.],

[Rare-Cons.] and [Rare-Div.]), *ATXN1* in Exon tests ($p = 0.027 \sim 0.076$; [Exon-All], [Exon-Evo.], [Exon-Cons.]), *PGCP/ CPQ* in DHS tests ($p = 0.072 \sim 0.078$; [DHS-Evo.] and [DHS-Div.]), and *SAPAP3/ DLGAP3* in [Rare-Div.] test ($p = 0.007$). None of the individual associations of the five OCD model genes held up after multiple testing corrections. However, these five genes have significantly lower p-values than the remaining sequenced genes ($pwilcox = 0.027$ from the comparison of p-values from the twelve genic burden tests excluding rare variant tests). The significance is much stronger when evaluating rare variant tests ($pwilcox = 6.43 \times 10^{-5}$; FIG. 14).

Comparison of sequencing and genotyping data

In order to confirm the validity of allele frequency calls estimated from the pooled sequencing data, a subset of sequence variants (66 variants) were genotyped, covering different allele frequencies and call confidence, in 547 (of 592) sequenced cases and 542 (of 560) sequenced controls that were available for genotyping. After filtering out the variants and individuals with missing genotype rate>0.1, genotyping data for 65 variants in 544 cases and 541 controls were retained for analysis.

Overall, the allele frequency estimates obtained from the sequencing data ('AFseq') and the allele frequency calculated from the genotyping data ('AFgeno') showed almost perfect correlation (OCD AFseq vs. OCD AFgeno, $\rho \sim 0.999$, $p < 2.2 \times 10^{-16}$; control AFseq vs. control AFgeno, $\rho \sim 0.999$, $p < 2.2 \times 10^{-16}$; FIG. 6).

The directionality of allele frequencies for a given variant ('case-control AF directionality'; e.g. whether the variant is more common in cases than in controls) was also compared between sequencing and genotyping data. Of the 65 genotyped variants, 24 variants (37%) were concordant for their case-control AF directionality between sequencing and genotyping data, and 41 variants (63%) were discordant (FIG. 7). This high discordance was affected by call confidence, as expected; 71% (29/41) of discordant variants were low confidence calls that were supported by one algorithm rather than two, and the proportion was higher than the expected (66%, 43/65; FIG. 7). Another major source of discordance arose from rare variants; 93% (38/41) of the discordant variants are rare (i.e. OCD AFseq< 0.01), while the expected proportion was 81.5% (53/65; FIG. 7). Taking into account both call confidence and rare variant biases, the concordance rate reached 86% (6/7) for the high

confidence (i.e. supported by two algorithms), not rare (i.e. OCD AFseq≥ 0.01) variants (FIG. 7).

Two possible explanations for the discordance bias towards rare variants were: (i) a low accuracy of AFseq in rare variants, and (ii) a presence of rare case-excessive alleles in the missing individuals (48 cases and 19 controls) in the genotyping data. As a simple scenario for the situation of missing rare case-excessive alleles, suppose for a given variant one or two additional alleles are present in the missing cases. Combining the genotyping data with simulated genotypes based on the above assumption, OCD AFgeno for a given variant is expected to increase by +0.0007 and +0.0016 for the presence of one and two additional alleles, respectively (Table 3). Adjusting the OCD AFgeno with the expected increase greatly improved the concordance rate in rare variants, from 28% (15/53) to 62% (33/53) and 85% (45/53), for one and two additional alleles, respectively (FIG. 7). The adjusted concordance rates for rare variants (62-85%) are comparable to that for not rare variants (75%).

CTTNBP2 and REEP3 candidate variants

Lastly, we describe the individual candidate variants that passed the candidate selection criteria for *CTTNBP2* and *REEP3* in greater detail to supplement the above.

CTTNBP2-variant 1: The top DHS candidate variant was rs2067080 (Chr7:117468056 [C>T], AF (T allele) 0.53, *psingle* = 0.001, OCD AF=0.57, control AF=0.51). The T allele was a derived allele and the ancestral allele (C) frequency for rs2067080 varied widely between different populations with C allele being major in African populations while T allele being major in East Asian populations [ref. 62]. rs2067080 overlapped with DHS in 64 ENCODE cell lines (among the ENCODE 125 cell types [ref. 63]) including SK-N-MC (neuroepithelioma cell line), HBMEC (brain microvascular endothelial cell), HAc (astrocytes-cerebellar), and Gliobla (glioblastoma), and at least ten different transcription factors (TF) are bound in this region in multiple cell lines (among the 161 factors from ENCODE [ref. 64]) including EP300 in SK-N-SH RA (neuroblastoma cell line), FOXP2 in SK-N- MC, GATA2 in HUVEC, JUND in H1-hESC. ~300bp away from this SNP, there is a nonDHS candidate variant rs2111209 (Chr7:117468334 [T>C], AF (C allele) 0.72, *psingle* = 0.006), which is also within the same regulatory region indicated by

the DHS and TF ChIP-seq data. rs2111209 also has evidence for a derived allele, i.e. the reference allele in chimp/orangutan/macaque is C, while the human reference allele is T.

CTTNBP2-variant 2: The second variant was rs12706157 (Chr7:117456904[C>T], AF (T allele) 0.62; *psingle* = 0.001, OCD AF=0.65, control AF=0.59). Interestingly, the chimp allele for this position is C, but orangutan and macaque alleles are T, which may indicate the derived status of this site. rs12706157 overlapped with 11 DHS marks including HBMEC and transcription factor FOS ChIP-seq mark in HUVEC (umbilical vein endothelial cells). ~100bp away from this SNP, a nonDHS candidate variant rs13242822 (Chr7:117457141 [G>C], AF[C] = 0.647, *psingle* = 0.0039) exists, overlapping with the same DHS/TF cluster. rs13242822 also had evidence for derived status that chimp/orangutan/macaque's reference was a C allele while human's was G at this position.

CTTNBP2-variant 3: The third variant was rs34868515 (Chr7:117450810 [C>T], MAF 0.026, *psingle* = 0.27, OCD AF=0.001, control AF ~ 0). This SNP also had evidence for being a derived allele, with chimp/orangutan/macaque's reference allele being T while human's was a C allele. The variant overlapped with DHS marks in five cell lines including three embryonic stem cells (H1- hESC, H7-hESC, and H9ES). It also overlapped with at least nine transcription factors (SP1, YY1, EP300, JUND, TCF12, HDAC2, NANOG, BCL11A, and TEAD4) ChIP-seq marks in H1-hESC.

CTTNBP2-variant 4: The fourth variant was rs75322384 (Chr7:117417559 [A>G], MAF 0.003, *psingle* = 0.29, OCD AF=0.003, control AF=0.0016). The variant was predicted as splice region variant based on UCSC gene uc003vjf.3. It overlapped with DHS marks in at least 16 cell lines including BE2 C (neuroblastoma) and HAc. Histone (H3K36me3) modification mark for H1- and H7-hESC was found in this region. No ENCODE ChIP-seq signal was found.

CTTNBP2-variant 5: The fifth variant was a deletion (Chr7:117390966 [T>D], *psingle* = 0.32, OCD AF=0.0009, control AF ~ 0). This deletion overlapped with 62 DHS marks including BE2 C, H1-hESC, HAc, HBMEC, and SK-N-MC. It also overlapped with three transcription factors' (CTCF, SMC3, and RAD21) ChIP-seq marks in numerous cell lines, which included CTCF signals in 78 cell lines including BE2 C, Gliobla, H1-hESC, HBMEC, and SK-N-SH RA, and RAD21 signals in H1-hESC and SK-N-SH RA.

CTTNBP2-variant 6: The sixth variant was a private variant (Chr7:117356081 [T>G], *psingle* = 0.32, OCD AF=0.0009, control AF ~ 0). This variant overlapped with DHS in 106 cell lines including BE2 C, Gliobla, H1-/H7-hESC, HAc, HBMEC, and SK-N-MC. It also overlaps at least seven transcription factors' (TCF12, ELF1, CTCF, EBF1, SMC3, ZNF143, and RAD21) ChIP-seq marks in multiple cell lines, which include CTCF signals in 95 cell lines including BE2 C, Gliobla, H1- hESC, HAc, HBMEC, and SK-N-SH RA, and RAD21 in 11 cell lines including H1-hESC and SK-N-SH RA. An 16bp insertion has been reported (rs199826384, MAF=0.022) at this position.

CTTNBP2-variant 7: The seventh variant was a private variant (Chr7:117421141 [C>A], *psingle* = 0.56, OCD AF=0.001, control AF=0.0005). This variant overlapped with 31 DHS marks including HBMEC and SK-N-MC and with 17 transcription factors' binding sites in GM12878 (B- lymphocyte). It also overlapped with ENCODE histone H3K27me3 mark in SK-N-SH RA, and with ENCODE ChIP-seq peaks of CTCF in Gliobla, HBMEC, SK-N-SH RA; POL2 in Gliobla; NRSF in PFSK-1 (neuroectodermal cell derived from cerebral brain tumor) and U87 (glioblastoma cell); GATA-2 in SH-SY5Y (SK-N-SH clone); P300 in SK-N-SH RA; RAD21 in SK-N-SH RA; USF1 in SK-N-SH RA, and YY1 in SK-N-SH RA.

CTTNBP2-coding variants with regulatory evidences: There were seven coding variants that passed the candidate selection criteria, of which three private coding variants found in cases cluster within a ~750bp regulatory region, together with an additional nonDHS variant. A missense (Chr7:117431202 [C>A], S>I) and a nonDHS variant (rs113126080, Chr7:117431079 [T>G], MAF 0.004, *psingle* = 0.0028, OCD AF=0.016, control AF=0.0035) were located ~10 and ~100bp away from DHS in two cell lines, i.e. SK-N-SH and K562 (leukemia cell line). Another missense (Chr7:117431704 [C>T], G>R) within this cluster overlapped with at least two transcription factor binding sites, i.e. RAD21 in h1-hESC and CTCF in K562. The remaining coding variant was synonymous (Chr7:117431879 [G>A]), which overlapped with binding site of RBBP5 in K562. The G allele may have been subject to positive selection during primate speciation, considering that the G allele was present across human/chimp/rhesus/baboon while the absolute majority of the remaining available vertebrate genomes had the A allele. Two of the remaining four coding variants were also located near regulatory marks: a missense variant rs142089340

(Chr7:117358107 [T>C], T>A, MAF 0.0014, *psingle* = 0.174, OCD AF=0.0035, control AF=0.001) was located ~200bp away from DHS in at least three cell lines including SK-N-SH, and overlapped with a binding site for CEBPB in IMR90 (fetal lung cells). A private case-only synonymous variant (Chr7:117501314 [G>A]) was located ~450bp away from DHS in at least four cell lines including SK-N-MC.

REEP3-intragenic: In REEP3, two DHS variants passed the selection criteria for candidate variants: The top variant was rs78109635 (Chr10:65307923 [A>G], MAF 0.017, *psingle* = 0.0025, OCD AF=0.052, control AF=0.030), which overlapped with DHS marks in 33 cell lines including BE2 C, HAc, HBMEC, and SK-N-SH RA. It also overlapped with ChIP-seq marks for GATA2 in SH-SY5Y and EGR1 in K562. The second variant was rs76646063 (Chr10:65332906 [T>C], MAF 0.007, *psingle* = 0.083, OCD AF=0.006, control AF=0.002), which overlapped with DHS marks in 72 cell lines including H7-hESC, H9ES, HAc, HBMEC, and SK-N-MC. It also overlapped with four transcription factors' ChIP-seq data, i.e. GATA3 in SH-SY5Y and T-47D, GATA2 in HUVEC and SH-SY5Y, POLR2A in HUVEC and SK-N-MC, and EP300 in SK-N-SH RA.

REEP3-upstream: ~3kb upstream of REEP3, there were three private DHS variants found in cases that met the candidate selection criteria. Two variants (rs56311840, Chr10:65387644 [C>G], MAF 0.0041 and Chr10:65387722 [C>D]) overlapped with DHS in 45 cell lines including SK-N-MC, and with a binding site for POLR2A in SK-N-MC. The other variant (Chr10:65388750) overlapped with DHS in 91 cell lines including BE2_C, SK-N-MC, SK-N-SH RA and binding sites for 21 transcription factors including SIN3AK20, POLR2A, and REST in SK-N-SH, and USF1, EP300, and RAD21 (40bp away) in SK-N-SH RA.

DISCUSSION

Functions of the five associated genes

The *LIPH* gene, located within an OCD genome-wide linkage interval 3q27-289, encodes the protein lipase member H that catalyzes the production of lysophosphatidic acid (LPA) [ref. 10]. LPA signaling is essential for the development of the nervous system and neuronal differentiation and growth. Lack of LPA production (by the absence of autotaxin) leads to the defects in neural tube closure during the development, resulting in embryonic

lethality in mice [ref. 11]. After development, LPA signaling also regulates the rearrangement of the synaptic connections and the cytoskeleton [ref. 12]. A role for LPA signaling in psychiatric disorder has been merging [ref. 13]. Multiple strains of LPA1 receptor knockout mice exhibit behavioral and/or the nervous system problems including schizophrenia- like phenotypes and abnormal prepulse inhibition. Furthermore, all the three miRNAs that were predicted to bind near the candidate mutations are expressed in human neuroblasts [ref. 14, 15] and at least two are known to promote neuronal differentiation (miR-23ab)16 and neural tube closure (let-7) [ref. 16]. Without wishing to be bound by theory, considering that the *LIPH* candidate mutations are likely to regulate *LIPH* expression via altered miRNAs binding rather than directly impacting the protein sequence and that other LPA-producing enzymes such as autotaxin exist at least during the developmental stage, a subtle change in the LPA production rate, rather than a complete knockout of LPA production, may be affecting the LPA signaling that plays a role in the synaptic connectivity and cytoskeleton in OCD (FIG. 4).

NRXN1 encodes neurexin 1α protein, a synapse organizing cell adhesion molecule and a receptor [ref. 6]. Deletions in *NRXN1* have been implicated in autism, schizophrenia and mental retardation [ref. 17]. As the role of neurexins in synaptic function and cognitive diseases have been extensively documented and reviewed elsewhere [ref. 18], the focus will be on the isoforms of *NRXN1* here. Neurexin 1α splice-forms interact with a range of different synaptic proteins including neuroligins, neurexophilins, and leucine-rich repeat transmembrane neuronal proteins (LRRTMs), to promote the adhesion between two neurons, providing specificity to the synaptic connections (FIG. 4). All of the NRXN1 candidate coding mutations were located within an isoform AK093260 (GenBank ID), suggesting that this particular variant may be relevant to OCD. This isoform produces lowly-expressed transcripts that are predominantly found in the brain, and contains four laminin G domains that have cell attachment activity [ref. 19] and two EGF domains that are found in the extracellular domain of membrane-bound or secreted proteins [ref. 20]. It does not contain a signal peptide or transmembrane domain, unlike other longer isoforms [ref. 21]. Without wishing to be bound by theory, mutations in this isoform may have a potential to exert a dominant negative effect by inaccurate cellular localization or by competing for neurexin 1α interacting partners, particularly through neurexophilins [ref. 21], suggesting possible

dysfunctions in the synapse in OCD. Notably, a similar gene *CDH2*, which encodes neural cadherin that also mediates the pre- and post-synaptic adhesion, has been confidently implicated in dog OCD [refs. 22, 23]. In addition, the second strongest association in the most recent GWAS for human OCD (OCGAS-GWAS) [ref. 24] was located in the cadherin cluster containing *CDH9* and *CDH10*, corroborating the roles for synaptic adhesion molecules in OCD (FIG. 4).

HTR2A encodes a G-protein coupled receptor, serotonin (5-HT) receptor 2A that is expressed widely throughout the central nervous system, including the prefrontal cortex. Several allelic variants have been reported to increase the susceptibility to cognitive disorders [ref. 25]. While only partially effective, selective serotonin reuptake inhibitors (SSRI) are a standard medication for OCD. Multiple clinical reports showed that an *HTR2A* antagonist, risperidone showed clinically significant improvement in patients with OCD who did not respond to SSRIs [ref. 26]. PET studies suggest a reduction in the *HTR2A* in the cortex in drug-naive OCD patients [ref. 27]. Activation of post-synaptic *HTR2A* triggers IP3 mediated calcium release from the endoplasmic reticulum (ER), regulating a range of calcium-dependent processes including synaptic transmission and neuronal excitation and differentiation [ref. 28] (FIG. 4).

CTTNBP2, located within an autism susceptibility locus at 7q3129, encodes cortactin binding protein 2, which regulates postsynaptic excitatory synapse formation and maintenance by modulating cortactin (cortical actin binding protein) mobility and distribution [ref. 30] (FIG. 4). *CTTNBP2* interacts with *CDH2* a gene implicated in dog OCD [ref. 31]. RAD21 and CTCF, the transcription factors that bind to at least 4/7 of the candidate DHS sites in brain-derived or embryonic cells, play important roles in brain development and neuronal migration. RAD21, one of the four core subunits of cohesin that connects replicated DNA molecules during cell division, is highly expressed in the human fetal cerebral cortex. The coordinated expression of RAD21 along with other key mitotic regulators is fundamental for the developing brain [ref. 32]. Apart from its role in cell division, RAD21 has also demonstrated a role in axon pruning and locomotion of neurons, presumably driven by its transcriptional activity [ref. 33]. In fact, the binding sites of cohesin and CTCF, a well-known transcription insulator, colocalize in the human genome, and it has been demonstrated both that CTCF is required for the positioning of cohesion and cohesin is required for the

insulator function of CTCF [ref. 34]. Without wishing to be bound by theory, both RAD21 and CTCF may be required to regulate the transcription of CTTNBP2, and the candidate mutations in their binding sites may dysregulate the CTTNBP2 expression, hence affecting protein stoichiometry, which is vital for synaptic formation and maintenance.

REEP3, a positional candidate for autism [ref. 35], encodes receptor expression enhancing protein 3. While the function of the protein is not well characterized, it may regulate cellular vesicle trafficking between the ER and the Golgi [ref. 35] and may transport G-protein coupled receptors to the cell surface membrane considering its sequence similarity to the related proteins [ref. 36]. Recent work showed that REEP proteins are important in shaping tubular ER membranes that extend throughout the highly polarized cells like neurons: REEP1-4 interact with microtubules through an extended C- terminal cytoplasmic domain, possibly mediating the formation and stabilization of the tubular ER network [ref. 37]. Although the functional connection between membrane shaping and neuronal function is unclear, several plausible scenarios have been suggested [ref. 37]. First, a correct tubulation may be required to distribute membranes during synaptic plasticity. Second, morphogenesis of the ER and other membrane compartment could be important in axonal transport to the distal location of the neuron. Third, membrane modeling defects could affect trafficking of receptors involved in signaling pathways. Lastly, dysregulated ER morphogens may directly affect the intracellular calcium release from the ER that is the downstream effect of the activation of several neuronal signals including HTR2A (FIG. 4). In REEP3, 2/3 candidate regulatory sites (two DHS and the 3'UTR variant cluster) overlapped with binding sites for GATA2, EP300, and POLR2A in brain-derived cells. While any of these transcription factors have the potential to be relevant to the OCD phenotype, GATA2 is of particular interest: without Gata2, all the precursors in the mouse embryonic midbrain fail to actuate GABAergic neuron-specific gene expression and instead switch to a glutamatergic phenotype [ref. 38]. Without wishing to be bound by theory, GATA2 may be a key transcription factor that regulates REEP3 expression, and the candidate mutations affecting its binding may dysregulate the REEP3 that shapes the tubular ER in axons, which is the internal storage for calcium whose release can be activated by HTR2A signaling.

Integrating the findings into the neurobiology of OCD

The findings described herein, integrated into what was previously known about OCD, emphasize specific protein groups and signaling pathways (FIG. 4). First and most notably, many lines of evidence now suggest roles for cell adhesion molecules in OCD. NRXN1 and CTTNBP2 identified in this study, CDH2 and CTNNA2 identified in dog OCD, DLGAP3 disrupted in OCD-like mouse, and DLGAP1, the top signal from the first major human OCD GWAS (IOCDF-GC) [ref. 39], are all cell adhesion molecules or their interacting partners and promote the synaptic connectivity and maintenance. This view is further supported by the top polygenic associations related to the regulation of protein complex assembly and cytoskeleton organization (FIG. 2 and Table 4).

Second, the candidate genes *HTR2A*, *REEP3* and *LIPH* corroborate a role for 5-HT signaling in OCD, with additional insights in connection with LPA signaling. The ER shaped by REEP proteins is the main storage for the internal calcium release by HTR2A activation. 5-HT signaling is significantly affected through lowered 5-HT turnover in the brains of the LPA1 receptor knockout mice [ref. 12]. In addition, LPA interferes with the signaling of an atypical antipsychotic agent risperidone, a HTR2A blocker on glial cells [ref. 12]. LPA can also induce calcium release as HTR2A signaling does. LPA-induced calcium responses were altered in cell lines originating from bipolar disorder patients [ref. 12]. Without wishing to be bound by theory, the calcium release from the ER that is activated by cross-talking HTR2A and LPA signaling may be relevant to OCD and an improper generation of tubular ER and its membrane in the axon by mutated REEP3 may affect their calcium-dependent downstream effect (FIG. 4).

Third, the findings herein also suggest a possible role for differentiation of glutamatergic-GABAergic systems in OCD. *PTPRD*, the top gene in the OCGAS GWAS, encodes a presynaptic protein that promotes the differentiation of glutamatergic synapses and regulate inhibitory GABAergic synapse development by interacting with a postsynaptic adhesion molecule, Slit and NTRK-like family member 3 (SLITRK3) [ref. 24] (FIG. 4). SLITRK3 is a protein in the same family as SLITRK5 whose knockout induces OCD-like behaviors in mice [ref. 40]. Neurexins induce GABAergic and glutamatergic postsynaptic differentiation via neuroligins [ref. 41] (FIG. 4). Without wishing to be bound by theory, the REEP3 candidate mutations likely disrupt the binding sites for GATA2 that is a key transcription factor that also selectively differentiate GABAergic neurons from glutamatergic

phenotype. Together, this suggests that inaccurate differentiation from the excitatory glutamatergic to inhibitory GABAergic neurons occur in OCD, proposing a plausible mechanism of how the imbalance between the excitatory signals and inhibitory signals in the CSTC circuit materializes in OCD, which is the central theory of the CSTC-OCD model [ref. 42].

Fourth, the polygenic burden analysis described herein suggests that programmed cell death and cell migration during brain development may be relevant to OCD. The significant polygenic associations of programmed cell death may appear surprising at first. However, the embryonic development including the neural tube closure and neuronal development are dependent on cell migration and specific sites and rates of programmed cell death [ref. 11]. Thus, subtle abnormalities in programmed cell death could impact the development of complex neurocircuits that requires delicate mapping of the correct cell types. In support of this speculation, telencephalic tangential migration was among the top polygenic associations in the study herein (FIG. 2). Telencephalic tangential migration refers to one of the two major neuronal migration events during the development of the cerebrum, which consists of the cerebral cortex and several subcortical structures such as the basal ganglia, from the embryonic telencephalon. Cells follow the tangential trajectories to migrate from the subcortical telencephalon, consisting primarily of the basal ganglia, to the specific destinations including the cortex, the striatum, and the thalamus [ref. 43, 44]. This suggests that inadequate telencephalic tangential migration probably results in the abnormal connectivity of the CSTC loop. In particular, it is known that a majority of cortical GABAergic neurons reach the cortex via the tangentially migrating stream, implicating a role for telencephalic tangential migration in populating inhibitory signals in the CSTC loop, which points back to the model of imbalanced excitatory-inhibitory signals in the CSTC loop of OCD brain.

Mutation types and embryonic development

It is intriguing that some OCD-associated genes have excessive coding candidate mutations and others have excessive regulatory candidate mutations. What causes these genes to be enriched with different types of candidate mutations? Without wishing to be bound by theory, it is hypothesized that genes with indispensable functions could only be affected by

mutations with subtle impact (e.g., regulatory changes), while genes with redundant or particular functions could withstand severe protein-damaging mutations in OCD. In fact, all of the OCD-associated genes with excessive regulatory mutations whose orthologues have been deleted in mice result in lethal knockouts. *CDH2*, a dog OCD-associated gene with excessive case-only regulatory variants, determines the left-right symmetry during development, and results in complete embryonic lethality during organogenesis when homozygously knocked out [ref. 45]. Homozygous deletion of *CTNNA2* in mice, another dog OCD-associated gene with regulatory variants, in mice also causes neonatal lethality, with abnormal cerebellum and neuron morphology [ref. 46]. The third dog OCD-associated gene with regulatory variants, *ATXN1*, also causes neonatal lethality or premature death with many developmental problems including abnormal nervous system phenotypes in several *ATXN1* disrupted mice with different allelic compositions [ref. 47]. There are no reports on phenotypes resulting from mouse knockouts of *CTTNBP2* or *REEP3* yet, but *CTTNBP2* is highly expressed in the mouse embryonic cerebral cortex and *REEP3* is expressed in the mouse myelinating oligodendrocytes at birth [ref. 48], suggesting developmental roles. Conversely, the only two OCD-associated genes with excessive coding mutations in this study do not yield lethality when disrupted in mice, despite the multiple mouse knockout studies with various allelic compositions. *NRXN1* null mice exhibit abnormal phenotypes including increased grooming, a phenotype similar to a form of canine OCD, acral lick dermatitis, and synaptic transmission ($Nrxn1^{tm1Sud}$), or no apparent abnormalities ($Nrxn1^{tm2Sud}$ and $Nrxn1^{tm4Sud}$). HTR2A null mice also do not result in lethality, while exhibiting abnormal phenotypes including anxiety-related response ($Htr2a^{tm1Grch}$, $Htr2a^{tm1Rhn}$, and $Htr2^{atm2Grch}$). Also, a deletion of the OCD-like mouse gene DLGAP3/ SAPAP3, where at least seven protein-damaging mutations have been identified in human OCD and trichotillomania [ref. 49], is not lethal.

Overall, without wishing to be bound by theory, genes with critical roles during embryonic development and later in the adult brain, appear to be mostly affected by regulatory mutations, while genes with synaptic roles but with limited developmental roles tend to be affected by more severely damaging protein-coding mutations in OCD. Non-developmental genes may also have regulatory mutations in addition to protein-coding mutations. Additionally, considering that the genes *NRXN1* and *HTR2A* that showed

excessive coding mutations directly constitute the transmembrane proteins at synapse and *DLGAP3/ SAPAP3* which showed rare missense mutations [ref. 49] induces other post-synaptic density proteins, while *CTTNBP2* and *REEP3* which show excessive regulatory mutations aid the synaptic maintenance and downstream effects, a protein's synaptic role might indicate how much damaging mutations it can harbor. For example, coding mutation proteins tend to play roles in synaptic formation, while regulatory mutation proteins tend to play roles in synaptic maintenance. However, functional characterization and identification of candidate mutations would be required for more OCD-relevant genes to support this argument.

Evaluation of genes from OCD animal models

Another interesting observation is that the five genes identified from animal models of OCD showed stronger association signals when considering rare variants only. This rare variant burden may be due to mutations in these genes substantially reducing the fitness of an individual, thus mutations are being purged from a population via purifying selection. In fact, four of the five genes likely possess vital developmental functions, whose mutations may be detrimental to an individual's survival. Three (*CDH2, ATXN1* and *CTNNA2*) genes are lethal when homozygously deleted in mice. While phenotype information for *PGCP/ CPQ* null mouse is not yet available, it is expressed in many mouse embryonic tissues from the blood and the liver to the eye and neurons [ref. 48], implicating its developmental role. While *SAPAP3/ DLGAP3* is not embryonic lethal in mice, the severe behavioral consequences by a single gene knockout [ref. 50] and its deep sequence conservation from humans to fruit flies [ref. 51] may imply that *SAPAP3/ DLGAP3* plays a central role in fundamental behaviors, thus its mutation might reduce an individual's fitness by affecting the behaviors before or during the reproductive age.

While the mutation of these genes was rare in humans, their disease association was relatively straightforward to find with animal models. This is probably due to the simpler genetic architectures in animal models compared to humans. The artificial mouse model for OCD involves only one gene, and the natural dog model, while it is multigenic, has presumably a simpler genetic architecture for OCD compared to humans due to the strong artificial selection during breed creation. Consequently, in these animal models, mutations

with large effect sizes, which would be otherwise rare in the natural populations, were retained or induced, producing OCD-like symptoms. Considering that the animal model genes do have relatively strong associations jointly, it may imply that the OCD genetic risks from these five genes represent the rare allele tail of the frequency spectrum for OCD risk alleles. Accordingly, a replication of association signals from these five genes in humans might be achieved only by very large sample sizes that would yield enough power to detect rare allele effects.

Towards the integration of functional evidence into sequencing studies

Here, an intelligent integration of evolutionary and regulatory information for coding and noncoding regions of the genome enabled confident associations of genes with a complex disorder, even with a small scale case-control sequencing. Separate analyses of coding and regulatory variants that were further stratified by evolutionary conservation status led to an identification of 4/5 associated genes in the study, demonstrating the power of variant stratification based on evolutionary and biochemical evidences. This approach also provided functional annotations for a large proportion of the candidate variants, generating experimentally tractable hypotheses for many candidate variants, some of which were validated here. Together, these results show that regulatory and coding variants are both critical in OCD, indicating that systematic surveys of functional noncoding regions as well as the exome are required for a complex trait like OCD. With major advances in evolutionary and functional genomics, it is now possible to design a sequencing study where both the whole exome and functional noncoding regions are targeted, covering 5-7% of the human genome, still providing a cost- and time-effective alternative to the whole genome sequencing. Therefore, this study suggests the benefit of integrating evolutionary and functional genomics information into sequencing studies for complex traits, from the initial sequence capture to the analysis and interpretation. Such a strategy will continue to improve the detection of disease-causing mutations, as the knowledge about the functional regions of the genome and their activities grow via accelerated advances in evolutionary and functional genomics.

EXPERIMENTAL PROCEDURES

Samples

The sample cohort consisted of 592 DSM-IV52 OCD cases and 560 matched controls that were sequenced, and additional 180 DSM-IV OCD cases and 180 matched controls that were genotyped for replication. All study subjects were self-reported white Caucasian with Northwestern European ancestry. Relevant institutional review boards approved this work at all participating locations.

Sequencing and variant detection

Groups of 16 individuals were pooled together to 37 case pools and 35 control pools and barcoded. Targeted genomic regions were captured using a NimbleGen hybrid capture custom array and sequenced by Illumina GAII or Illumina HiSeq2000. Sequencing reads were aligned by Picard analysis pipeline. Syzygy53 was used for variant discovery and allele frequency estimations in case and control pools. Validation of allele frequency estimation for high-confidence variants was done by comparison with the allele frequencies estimated by an independent software SNVer54. Validation or replication of selected variants in the sequencing samples and in an additional cohort used Sequenom.

Variant annotation

ANNOVAR55 was used to annotate variants (-buildver hg19) and to download relevant datasets when available in ANNOVAR database. Coding, synonymous, and nonsynonymous status were annotated based on RefSeq genes using ANNOVAR's gene-based function (--geneanno-dbtype refGene), evolutionary status based on GERP++ mammalian constraints using filter-based function (--filter -dbtype gerp++gt2 for sites with GERP>2), regulatory status based on ENCODE DNaseI hypersensitivity sites using region-based function (--regionanno -dbtype wgEncodeRegDnaseClustered), and the frequencies and novelty of variants based on the 1000 Genomes data (--filter -dbtype 1000g2012apr all). Sites with GERP scores<-2 in the targeted regions were downloaded in BED format from the UCSC Table Browser to annotate variants at 'evolutionarily divergent' sites (-- regionanno -dbtype bed).

Capture array design and origins of targeted genes

In order to compile the list of 608 genes sequenced, the following sources were analyzed which included genetic, animal model, and neurobiological studies relevant to OCD:

- 91 genes within the genome-wide OCD linkage signals
- 79 OCD/OCD-related disorder genetic candidates
- 55 genes whose canine orthologues within dog OCD GWAS-intervals [ref. 22, 23]
- 151 genes that encode post-synaptic density proteins that are highly expressed in the mouse striatum
- 68 genes that are differentially expressed between D1R- and D2R-expressing medium spiny neuron
- 80 genes whose orthologues have been disrupted in mouse models for OCD [ref. 56] or autism
- 89 genes implicated in human autism genetic studies
- 36 genes from 22q11.2 deletion syndrome, 16p11.2 duplication syndrome, and 15q11.2 deletion syndrome
- 17 genes that are microRNAs implicated in autism and their targets

Association analyses

Single-site association used t-test, comparing the estimated allele frequencies between 37 case and 35 control pools. For gene-based association, the basic burden test as described previously was used [ref. 58]. Briefly, the test statistic is sum of the standardized differences of non-reference allele rates between all cases and controls per gene. The test statistics were assessed by 10000 times permutation of case-control labels in one-sided manner, expecting a greater burden of non- reference alleles in cases than in controls. Polygenic association used the same burden test.

The null expectations were calculated using an empirical and/or a theoretical model(s). For the theoretical null, uniform distribution was assumed. For the empirical null, permuted test statistics were generated from 100 iterations [ref. 59] of burden tests. The

expected values were then compared with the observed values using quantile-quantile plot and the observed data was evaluated for deviation from the expectation globally.

Multiple testing procedure

The empirical 'minP' procedure was employed as used previously [ref. 58] to control for multiple testing in genic and polygenic burden tests. For genic burden tests, correction was done jointly for all 16 filters, 'Overall', 'Exon', 'DHS' and 'Rare' categories and their four sub-categories stratified by evolutionary status, and for all 608 sequenced genes. The permuted null was generated from the minimum empirical significance obtained from each permuted dataset as if the permuted dataset was the observed dataset, for all 16 filters and 10,000 times permutation. The observed values were then compared against the distribution of the permuted null across all tests and genes considered (FIG. 9). For polygenic burden test, all 989 tested GO sets were corrected for, using the same procedure and permutation number (FIG. 10).

Candidate variant analysis

Among all the variants identified within or near the genes of interest, a variant was selected as a candidate variant if it was more common in cases than in controls, and met any of the following 'stringent' criteria: i) $P_{single} < 0.05$, ii) present only in cases, effectively estimated allele count in controls < 1 (i.e. control AF < 8.93×10^{-4}), and iii) AFcase/AFcontrol > 2, or met at least two of the following 'relaxed' criteria: i) $P_{single} < 0.1$, ii) estimated allele count in controls < 2 (i.e. control AF < 1.79×10^{-3}), iii) AFcase/AFcontrol > 1.5, and iv) novel. A variant was 'Novel' if it was not described before, (i.e. did not have a dbSNP ID). A 'private' variant meant that the mutation was found in only one individual in the cohort.

For the protein sequence analysis with regards to the impact of a candidate variant, a protein sequence of interest containing a candidate variant and the location of corresponding amino acid residues were obtained by blastx on UniProtKB, then a protein domain search was run using InterPro sequence search. Amino acid modification information was extracted from UniProt.

GO network visualization and functional themes

DAVID 6.7 [ref. 60] was used to compute the enrichment of the targeted genes for all GO sets. 989 GO sets that showed weak enrichment (nominal $p<0.1$) were generously deemed to represent the search space, thus selected for testing polygenic association.

To study the background functions of the 989 GO sets, a treemapping method provided by REVIGO with default parameters was used (i.e. allowed similarity, Medium [0.7]; GO term weight, enrichment p-value; database with GO term size, whole UniProt; semantic similarity measure, SimRel; treemap option, abs log10 pvalue), which clustered GO sets based on similarity.

Network visualization of the 989 GO sets used CytoScape 3.1.0 [ref. 61] with Enrichment Map Plugin [ref. 4]. Nodes and edges were automatically placed based on the parameters recommended by the Enrichment Map manual, i.e. enrichment p-value cutoff for building enrichment map, 0.05; overlap metric, Jaccard coefficient; Jaccard coefficient cutoff for building maps, 0.25. The network was arranged by force directed layout weighted mode, using only the interactions that passed the threshold for the similarity coefficient.

Genotyping

For both the validation of allele frequency estimates in sequenced samples and replication of findings in additional samples, the Sequenom MassArray iPlex system was used to genotype selected subset of variants.

Gel shift assay

For each allele of the tested SNPs, pairs of 5'-biotinylated oligonucleotides were obtained from IDT Inc. (Coralville, IA, USA). Equal volumes of forward and reverse oligos were mixed and heated at 95°C for 5 minutes and then cooled to room temperature. Fifty femtomoles of annealed probes were incubated at room temperature for 30 minutes with 10 mg SK-N-BE(2) nuclear extract (Active Motif Carlsbad, CA, USA). The remaining steps followed the LightShift Chemiluminescent EMSA Kit protocol (Thermo Scientific).

REFERENCES

1. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences 106, 9362–9367 (2009).

2. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory dna. Science 337, 1190–1195 (2012).

3. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482 (2011).

4. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network- based method for gene-set enrichment visualization and interpretation. PloS one 5, e13984 (2010).

5. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. cell 120, 15–20 (2005).

6. NCBI. Nrxn1 neurexin 1 [homo sapiens (human)] gene id: 9378 (2014). URL http://www.ncbi.nlm.nih.gov/gene/9378.

7. Be´camel, C. et al. The serotonin 5-ht2a and 5-ht2c receptors interact with specific sets of pdz proteins. Journal of Biological Chemistry 279, 20257–20266 (2004).

8. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).

9. Shugart, Y. et al. Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q. Molecular psychiatry 11, 763–770 (2006).

10. NCBI. Liph lipase, member h [homo sapiens (human)] gene id: 200879 (2014). URL http://www.ncbi.nlm.nih.gov/gene/200879.

11. Fotopoulou, S. et al. Atx expression and lpa signalling are vital for the development of the nervous system. Developmental biology 339, 451–464 (2010).

12. Lin, M.-E., Herr, D. R. & Chun, J. Lysophosphatidic acid (lpa) receptors: signaling properties and disease relevance. Prostaglandins & other lipid mediators 91, 130–138 (2010).

13. PEDRAZA, C., CASTILLA-ORTEGA, E. & DE FONSECA, F. R. Role of lysophosphatidic acid (lpa) in behavioral processes: implications for psychiatric disorders. Lysophospholipid Receptors: Signaling and Biochemistry 451 (2013).

14. microRNA.org. microrna.org - targets and expression (2010). URL http://www.microrna.org/.

15. Betel, D., Wilson, M., Gabow, A., Marks, D. S. & Sander, C. The microrna. org resource: targets and expression. Nucleic acids research 36, D149–D153 (2008).

16. Li, X. & Jin, P. Roles of small regulatory rnas in determining neuronal identity. Nature Reviews Neuroscience 11, 329–338 (2010).

17. OMIM. 600565 neurexin i; nrxn1 (2014). URL http://www.omim.org/entry/600565.

18. Su¨dhof, T. C. Neuroligins and neurexins link synaptic function to cognitive disease. Nature 455, 903–911 (2008).

19. EMBL-EBI. Laminin g domain(ipr001791) (2014). URL http://www.ebi.ac.uk/interpro/entry/IPR001791.

20. EMBL-EBI. Epidermal growth factor-like domain (ipr000742) (2014). URL http://www.ebi.ac.uk/interpro/entry/IPR000742.

21. Rujescu, D. et al. Disruption of the neurexin 1 gene is associated with schizophrenia. Human molecular genetics 18, 988–996 (2009).

22. Tang, R. et al. Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. Genome Biology 15, R25 (2014).

23. Dodman, N. et al. A canine chromosome 7 locus confers compulsive disorder susceptibility. Molecular psychiatry 15, 8–10 (2010).

24. Mattheisen, M. et al. Genome-wide association study in obsessive-compulsive disorder: results from the ocgas. Molecular psychiatry (2014).

25. OMIM. 182135 5-hydroxytryptamine receptor 2a; htr2a (2014). URL http://omim.org/entry/182135.

26. Marek, G. J., Carpenter, L. L., McDougle, C. J. & Price, L. H. Synergistic action of 5-ht2a antagonists and selective serotonin reuptake inhibitors in neuropsychiatric disorders. Neuropsychopharmacology 28, 402–412 (2003).

27. Pittenger, C., Bloch, M. H. & Williams, K. Glutamate abnormalities in obsessive compulsive disorder: neurobiology, pathophysiology, and treatment. Pharmacology & therapeutics 132, 314–332 (2011).

28. UniProt. P28223 (5ht2a human) (2014). URLhttp://www.uniprot.org/uniprot/P28223.

29. Cheung, J. et al. Identification of the human cortactin-binding protein-2 gene from the autism candidate region at 7q31. Genomics 78, 7–11 (2001).

30. Chen, Y.-K. & Hsueh, Y.-P. Cortactin-binding protein 2 modulates the mobility of cortactin and regulates dendritic spine formation and maintenance. The Journal of Neuroscience 32, 1043–1055 (2012).

31. El Sayegh, T. Y. et al. Cortactin associates with n-cadherin adhesions and mediates intercellular adhesion strengthening in fibroblasts. Journal of cell science 117, 5117–5131 (2004).

32. Pemberton, H. et al. Separase, securin and rad21 in neural cell growth. Journal of cellular physiology 213, 45–53 (2007).

33. Heinrichs, A. Gene expression: Cohesin branches out. Nature Reviews Molecular Cell Biology 9, 268–269 (2008).

34. Wendt, K. S. et al. Cohesin mediates transcriptional insulation by ccctc-binding factor. Nature 451, 796–801 (2008).

35. Castermans, D. et al. Identification and characterization of the trip8 and reep3 genes on chromosome 10q21. 3 as novel candidate genes for autism. European Journal of Human Genetics 15, 422–431 (2007).

36. OMIM. 609348 receptor expression-enhancing protein 3;reep3 (2014). URL http://omim.org/entry/609348.

37. Blackstone, C., O'Kane, C. J. & Reid, E. Hereditary spastic paraplegias: membrane traffic and the motor pathway. Nature Reviews Neuroscience 12, 31–42 (2010).

38. Kala, K. et al. Gata2 is a tissue-specific post-mitotic selector gene for midbrain gabaergic neurons. Development 136, 253–262 (2009).

39. Stewart, S. et al. Genome-wide association study of obsessive-compulsive disorder. Molecular psychiatry 18, 788–798 (2012).

40. Shmelkov, S. V. et al. Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive- compulsive-like behaviors in mice. Nature medicine 16, 598–602 (2010).

41. Graf, E. R., Zhang, X., Jin, S.-X., Linhoff, M. W. & Craig, A. M. Neurexins induce differentiation of gaba and glutamate postsynaptic specializations via neuroligins. Cell 119, 1013–1026 (2004).

42. Pauls, D. L., Abramovitch, A., Rauch, S. L. & Geller, D. A. Obsessive-compulsive disorder: an integrative genetic and neurobiological perspective. Nature Reviews Neuroscience 15, 410– 424 (2014).

43. Mar´ın, O. & Rubenstein, J. L. A long, remarkable journey: tangential migration in the telencephalon. Nature Reviews Neuroscience 2, 780–790 (2001).

44. Lo´pez-Bendito, G. et al. Tangential neuronal migration controls axon guidance: a role for neuregulin-1 in thalamocortical axon navigation. Cell 125, 127–142 (2006).

45. Radice, G. L. et al. Developmental defects in mouse embryos lacking n-cadherin. Develop- mental biology 181, 64–78 (1997).

46. Togashi, H. et al. Cadherin regulates dendritic spine morphogenesis. Neuron 35, 77–89 (2002).

47. Watase, K. et al. A long cag repeat in the mouse sca1 locus replicates sca1 features and reveals the impact of protein solubility on selective neurodegeneration. Neuron 34, 905–919 (2002).

48. Edgar, R. et al. Lifemap discovery?: the embryonic development, stem cells, and regenerative medicine research portal. PloS one 8, e66629 (2013).

49. Zu¨chner, S. et al. Multiple rare sapap3 missense variants in trichotillomania and ocd. Molecular psychiatry 14, 6 (2009).

50. Welch, J. M. et al. Cortico-striatal synaptic defects and ocd-like behaviours in sapap3-mutant mice. Nature 448, 894–900 (2007).

51. NCBI. Dlgap3 discs, large (drosophila) homolog-associated protein 3 [homo sapiens (human)] gene id: 58512 (2014). URL http://www.ncbi.nlm.nih.gov/gene/58512.

52. Association, A. P., Association, A. P. et al. Diagnostic and statistical manual-text revision (DSM-IV-TRim, 2000) (American Psychiatric Association, 2000).

53. Rivas, M. A. et al. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics 43, 1066–1073 (2011).

54. Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. Snver: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic acids research 39, e132–e132 (2011).

55. Wang, K., Li, M. & Hakonarson, H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164–e164 (2010).

56. Burguie`re, E., Monteiro, P., Feng, G. & Graybiel, A. M. Optogenetic stimulation of lateral orbitofronto-striatal pathway suppresses compulsive behaviors. Science 340, 1243–1246 (2013).

57. Banerjee-Basu, S. & Packer, A. Sfari gene: an evolving database for the autism research community. Disease models & mechanisms 3, 133–135 (2010).

58. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature
(2014).

59. Howrigan, G. J. I. H. H. D. M. J. N. B. M., Daniel P. Re-calibrating test statistics distributions when testing rare genetic variants (World Congress of Psychiatric Genetics, Boston, MA, USA, 2013).

60. Da Wei Huang, B. T. S. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature protocols 4, 44–57 (2008).

61. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13, 2498–2504 (2003).

62. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100–1104 (2008).

63. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).

64. Gerstein, M. B. et al. Architecture of the human regulatory network derived from encode data. Nature 489, 91–100 (2012).

Without further elaboration, it is believed that one skilled in the art can, based on the above description, utilize the present invention to its fullest extent. The specific embodiments are, therefore, to be construed as merely illustrative, and not limitative of the remainder of the disclosure in any way whatsoever. All publications cited herein are incorporated by reference for the purposes or subject matter referenced herein.

The indefinite articles “a” and “an,” as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to mean “at least one.”

From the above description, one skilled in the art can easily ascertain the essential characteristics of the present invention, and without departing from the spirit and scope thereof, can make various changes and modifications of the invention to adapt it to various usages and conditions. Thus, other embodiments are also within the claims.

5

CLAIMS

What is claimed is:

1. A method, comprising:

(a) analyzing genomic DNA from a subject for the presence of a mutation within or near a gene selected from NRXN1, CTTNBP2, HTR2A, REEP3, or LIPH; and

(b) identifying a subject having the mutation as a subject at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder,

wherein if the gene is HTR2A, the mutation is within an exon or is a SNP provided in Table 2.

2. The method of claim 1, wherein the mutation is within 100 kb, upstream or downstream, of the chromosomal boundaries/co-ordinates provided in Table 1.

3. The method of claim 1 or 2, wherein the gene is selected from CTTNBP2 or REEP3.

4. The method of any one of claims A1 to A3, wherein the mutation is within an untranslated region (UTR), exon, or DNase1 hypersensitivity site of the gene.

5. The method of claim 1 or 2, wherein the gene is NRXN1 and the mutation is within the isoform AK093260.

6. The method of claim 1 or 2, wherein the gene is NRXN1 and the mutation is a SNP located at a chromosome location selected from chr2:51256161, chr2:50762143, chr2:51153020, chr2:50733992, chr2:51000979, chr2:50842279, chr2:51067620, chr2:50606585, chr2:50927403, chr2:50703012, chr2:50956849, chr2:50606521, chr2:50733958, chr2:50733841, chr2:50755127, chr2:50542571, chr2:50619213, chr2:50699305, chr2:50927347, chr2:50448284, chr2:50400991, chr2:50924511, chr2:50850245, chr2:50570754, chr2:50171755, chr2:50343508, chr2:50762346,

chr2:51148378, chr2:51244839, chr2:50607797, chr2:50941362, chr2:51236675, chr2:50187207, chr2:50848555, chr2:50198372, chr2:50981817, chr2:50693162, chr2:50570237, chr2:50683609, chr2:50849551, chr2:50735998, chr2:51246886, chr2:50200776, chr2:50323549, chr2:50675110, chr2:50922035, chr2:51057550, chr2:50386107, chr2:50386080, chr2:50847195, chr2:51252712, chr2:51245440, chr2:50354237, chr2:50719598, chr2:50952610, chr2:50792080, chr2:50542527, chr2:50750575, chr2:50155007, chr2:50779791, chr2:50733581, chr2:50400809, chr2:50201255, chr2:50178130, chr2:51146148, chr2:50575137, chr2:51148372, chr2:51171979, chr2:50779943, chr2:50848551, chr2:50165016, chr2:51149889, chr2:50774153, chr2:50389636, chr2:50434866, chr2:50724642, chr2:50981813, chr2:51085557, chr2:50463984, chr2:50724745, chr2:50981807, chr2:50598207, chr2:50675639, chr2:50653833, chr2:51145459, chr2:50542372, chr2:50952571, chr2:50548140, chr2:50765412, chr2:50850686, chr2:50934666, chr2:50682914, chr2:50709350, chr2:50979527, chr2:50386109, chr2:50542308, chr2:50607943, chr2:50735814, chr2:50981815, chr2:50155737, chr2:50683701, chr2:50842256, chr2:50148728, chr2:50952482, chr2:51153206, chr2:50560998, chr2:50996952, chr2:50458593, chr2:50924466, chr2:51005207, chr2:50602031, chr2:50178059, chr2:50850340, chr2:51016384, chr2:50175865, chr2:50571910, chr2:50570602, chr2:50548103, chr2:50518040, chr2:50236859, chr2:50464065, chr2:50598321, chr2:50282777, chr2:51245472, chr2:50735943, chr2:50927534, chr2:50941367, chr2:50952709, chr2:51067726, chr2:51079254, chr2:50277539, chr2:50424938, chr2:50765589, chr2:50699377, chr2:51149368, chr2:50723068, chr2:50723000, chr2:51245656, chr2:50571784, chr2:50148783, chr2:50598280, chr2:50850307, chr2:50850394, chr2:50563875, chr2:50614848, chr2:50531295, chr2:50877741, chr2:50733745, chr2:50919652, chr2:50570601, chr2:50981811, or chr2:51021463.

7. The method of claim 6, wherein the mutation is a SNP located at a chromosome location selected from chr2:50847195, chr2:50779791, chr2:50779943, chr2:50724642, chr2:50463984, chr2:50724745, chr2:50765412, chr2:50850686, chr2:50464065, chr2:50765589, chr2:50723068, or chr2:50733745.

8. The method of claim 1 or 2, wherein the gene is CTTNBP2 and the mutation is within or near a DNase1 hypersensitivity site or within an exon of CTTNBP2.

9. The method of claim 1 or 2, wherein the gene is CTTNBP2 and the mutation is a SNP located at a chromosome location selected from chr7:117430669, chr7:117358107, chr7:117431704, chr7:117396664, chr7:117374935, chr7:117391129, chr7:117368123, chr7:117446174, chr7:117456904, chr7:117356081, chr7:117427551, chr7:117354909, chr7:117452215, chr7:117431202, chr7:117358129, chr7:117359713, chr7:117457141, chr7:117450810, chr7:117431879, chr7:117386178, chr7:117385978, chr7:117468334, chr7:117396706, chr7:117501314, chr7:117390966, chr7:117354258, chr7:117352306, chr7:117351979, chr7:117431079, chr7:117417559, chr7:117427686, chr7:117421141, or chr7:117468056.

10. The method of claim 9, wherein the mutation is a SNP located at a chromosome location selected from chr7:117456904, chr7:117356081, chr7:117450810, chr7:117390966, chr7:117417559, chr7:117421141, or chr7:117468056.

11. The method of claim 1 or 2, wherein the gene is HTR2A and the mutation is within an exon of HTR2A.

12. The method of claim 11, wherein the mutation is within the last exon of HTR2A.

13. The method of claim 1 or 2, wherein the gene is HTR2A and the mutation is a SNP located at a chromosome location selected from chr13:47454997, chr13:47440198, chr13:47409048, chr13:47440301, chr13:47466592, chr13:47418543, chr13:4743474, chr13:47448370, chr13:47466622, chr13:47409701, chr13:47440209, chr13:47421746, chr13:47418629, chr13:47408946, chr13:47455071, or chr13:47469335.

14. The method of claim 13, wherein the mutation is a SNP located at a chromosome location selected from chr13:47409048, chr13:47466622, or chr13:47409701.

15. The method of claim 1 or 2, wherein the gene is REEP3 and the mutation is within or near a DNase1 hypersensitivity site of REEP3.

16. The method of claim 1 or 2, wherein the gene is REEP3 and the mutation is a SNP located at a chromosome location selected from chr10:65339450, chr10:65368263, chr10:65358911, chr10:65326034, chr10:65359513, chr10:65332906, chr10:65354650, chr10:65357754, chr10:65287863, chr10: 65387644, chr10: 65387722, chr10: 65388750, chr10: 65384621, or chr10:65307923.

17. The method of claim 16, wherein the mutation is a SNP located at a chromosome location selected from chr10:65332906, chr10: 65387644, chr10: 65387722, chr10: 65388750, chr10: 65384621, or chr10:65307923.

18. The method of claim 1 or 2, wherein the gene is LIPH and the mutation is within an untranslated region (UTR), intron, or exon of LIPH.

19. The method of claim 18, wherein the mutation is within the 3'UTR of LIPH or near a splice site of LIPH.

20. The method of claim 1 or 2, wherein the gene is LIPH and the mutation is a SNP located at a chromosome location selected from chr3:185241792, chr3:185229283, chr3:185226492, chr3:185229464, chr3:185226396, chr3:185225638, or chr3:185225644.

21. The method of claim 20, wherein the mutation is a SNP located at a chromosome location selected from chr3:185226492, chr3:185226396, chr3:185225638, or chr3:185225644.

22. The method of any one of claims 1 to 21, wherein the genomic DNA is obtained from a bodily fluid or tissue sample of the subject.

23. The method of any one of claims 1 to 22, wherein the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array.

24. The method of any one of claims 1 to 22, wherein the genomic DNA is analyzed using a bead array.

25. The method of any one of claims 1 to 22, wherein the genomic DNA is analyzed using a nucleic acid sequencing assay.

26. The method of any one of claims 1 to 25, wherein the subject is a human subject.

27. The method of any one of claims 1 to 26, wherein the method further comprises:

(c) administering a therapeutic agent to the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

28. The method of any one of claims 1 to 27, wherein the method further comprises:

(c) performing behavioral therapy on the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

29. The method of any one of the preceding claims, wherein the neuropsychiatric disorder is obsessive-compulsive disorder.

30. The method of any one of the preceding claims, wherein the mutation is at least two mutations.

31. The method of any one of the preceding claims, wherein the gene is at least two genes.

32. A method, comprising:

(a) analyzing genomic DNA from a subject for the presence of a SNP in Table 2; and

(b) identifying the subject having the SNP as a subject at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

33. The method of claim 32, wherein the SNP is a SNP located at a chromosome location selected from chr2:50847195, chr2:50779791, chr2:50779943, chr2:50724642, chr2:50463984, chr2:50724745, chr2:50765412, chr2:50850686, chr2:50464065, chr2:50765589, chr2:50723068, chr2:50733745, chr7:117456904, chr7:117356081, chr7:117450810, chr7:117390966, chr7:117417559, chr7:117421141, chr7:117468056, chr13:47409048, chr13:47466622, chr13:47409701, chr10:65332906, chr10: 65387644, chr10: 65387722, chr10: 65388750, chr10: 65384621, chr10:65307923, chr3:185226492, chr3:185226396, chr3:185225638, or chr3:185225644.

34. The method of claim 32 or 33, wherein the genomic DNA is obtained from a bodily fluid or tissue sample of the subject.

35. The method of any one of claims 32 to 34, wherein the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array.

36. The method of any one of claims 32 to 34, wherein the genomic DNA is analyzed using a bead array.

37. The method of any one of claims 32 to 34, wherein the genomic DNA is analyzed using a nucleic acid sequencing assay.

38. The method of any one of claims 32 to 37, wherein the method further comprises:

(c) administering a therapeutic agent to the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

39. The method of any one of claims 32 to 38, wherein the method further comprises:

(c) performing behavioral therapy on the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

40. The method of any one of claims 32 to 39, wherein the neuropsychiatric disorder is obsessive-compulsive disorder.

41. The method of any one of claims 32 to 40, wherein the mutation or SNP is two mutations or SNPs.

42. A method, comprising:

(a) analyzing genomic DNA from a subject for the presence of a first mutation within or near HTR2A and a second mutation within a gene selected from LIPH, NRXN1, CTTNBP2, or REEP3; and

(b) identifying a subject having the first and second mutation as a subject at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

43. The method of claim 42, wherein the first mutation and the second mutation are each independently within 100 kb, upstream or downstream, of the chromosomal boundaries/co-ordinates provided in Table 1.

44. The method of claim 42 or 43, wherein the mutation is within an untranslated region (UTR), exon, or DNase1 hypersensitivity site of the gene.

45. The method of any one of claims 42 to 44, wherein the first mutation is within an exon of HTR2A.

46. The method of claim 45, wherein the first mutation is within the last exon of HTR2A.

47. The method of any one of claims 42 to 46, wherein the gene is HTR2A and the second mutation is a SNP provided in Table 2.

48. The method of any one of claims 42 to 47, wherein the gene is LIPH and the second mutation is within an untranslated region (UTR), intron, or exon of LIPH.

49. The method of claim 48, wherein the second mutation is within the 3'UTR of LIPH or near a splice site of LIPH.

50. The method of any one of claims 42 to 47, wherein the gene is LIPH and the second mutation is a SNP provided in Table 2.

51. The method of any one of claims 42 to 47, wherein the gene is NRXN1 and the second mutation is within the isoform AK093260.

52. The method of any one of claims 42 to 47, wherein the gene is NRXN1 and the second mutation is a SNP provided in Table 2.

53. The method of any one of claims 42 to 47, wherein the gene is CTTNBP2 and the second mutation is within or near a DNase1 hypersensitivity site or within an exon of CTTNBP2.

54. The method of any one of claims 42 to 47, wherein the gene is CTTNBP2 and the second mutation is a SNP provided in Table 2.

55. The method of any one of claims 42 to 47, wherein the gene is REEP3 and the second mutation is within or near a DNase1 hypersensitivity site of REEP3.

56. The method of any one of claims 42 to 47, wherein the gene is REEP3 and the second mutation is a SNP provided in Table 2.

57. The method of any one of claims 42 to 56, wherein the genomic DNA is obtained from a bodily fluid or tissue sample of the subject.

58. The method of any one of claims 42 to 57, wherein the genomic DNA is analyzed using a single nucleotide polymorphism (SNP) array.

59. The method of any one of claims 42 to 57, wherein the genomic DNA is analyzed using a bead array.

60. The method of any one of claims 42 to 57, wherein the genomic DNA is analyzed using a nucleic acid sequencing assay.

61. The method of any one of claims 42 to 60, wherein the subject is a human subject.

62. The method of any one of claims 42 to 61, wherein the method further comprises:

(c) administering a therapeutic agent to the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

63. The method of any one of claims 42 to 62, wherein the method further comprises:

(c) performing behavioral therapy on the subject identified as at elevated risk of developing a neuropsychiatric disorder or as having a neuropsychiatric disorder.

64. The method of any one of claims 42 to 63, wherein the neuropsychiatric disorder is obsessive-compulsive disorder.

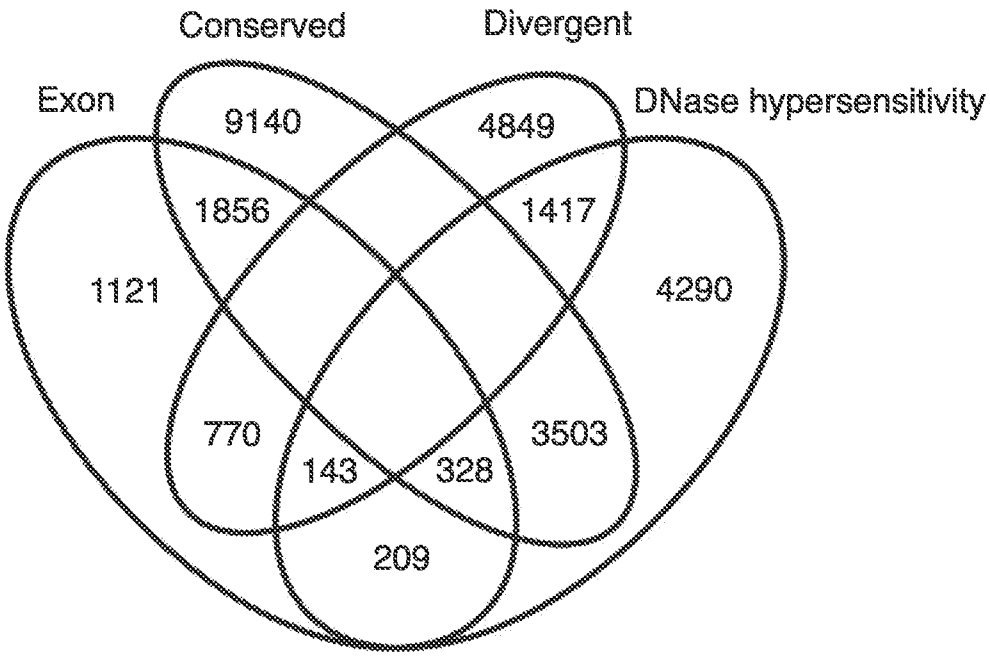

Conserved
Divergent
Exon
DNase hypersensitivity
9140
4849
1856
1417
1121
4290
770
3503
143
328
209

FIG. 1A

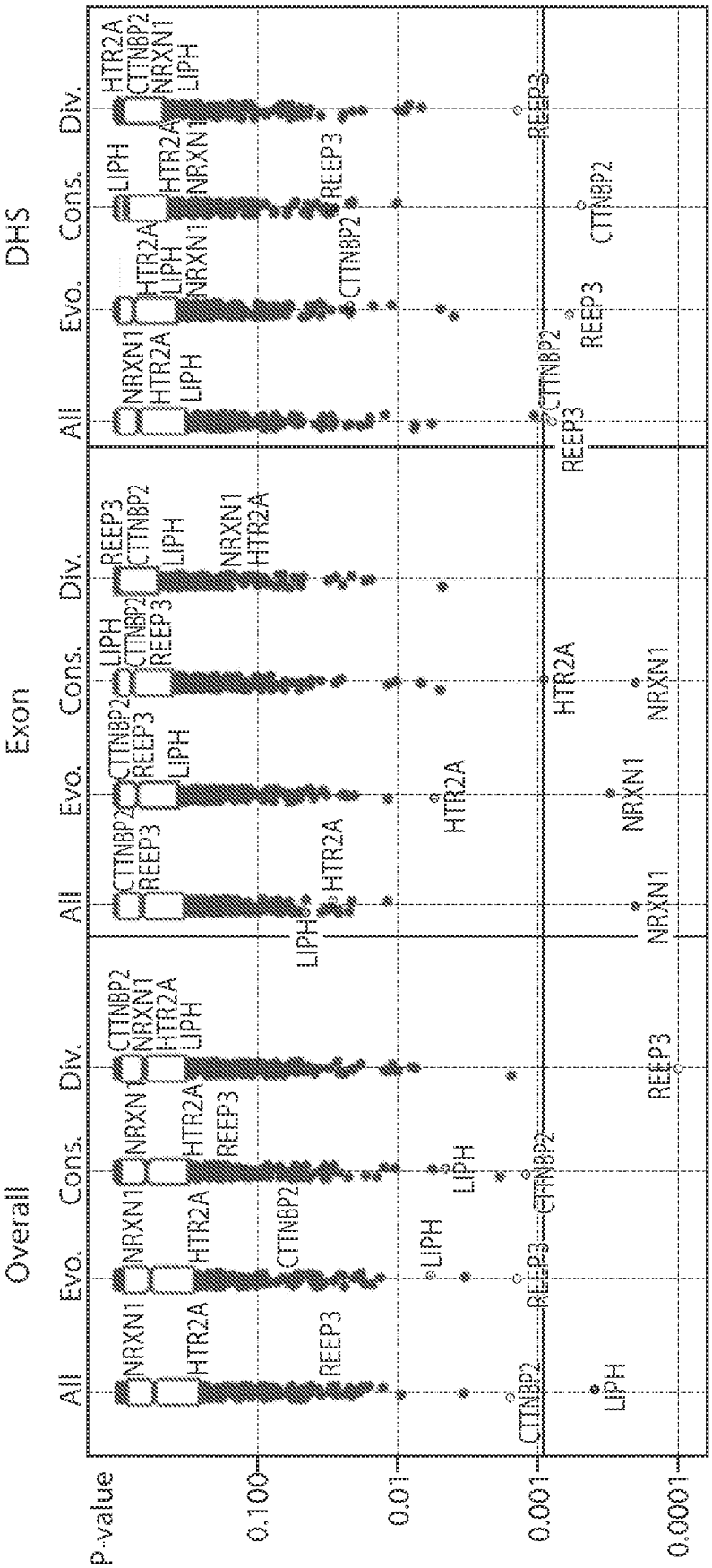

DHS
Exon
Overall
All
Evo.
Cons.
Div.
P-value
0.100
0.01
0.001
0.0001
NRXN1
HTR2A
LIPH
CTTNBP2
REEP3

FIG. 1B

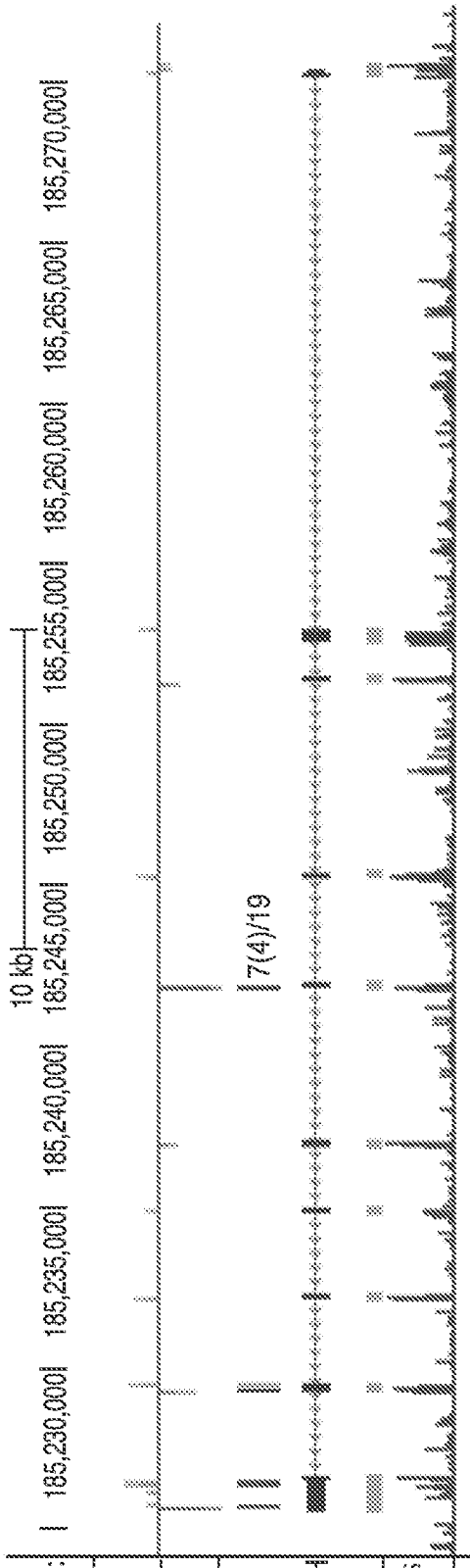

chr3:
185,230,000
185,235,000
185,240,000
10 kb
185,245,000
185,250,000
185,255,000
185,260,000
185,265,000
185,270,000
-Log10Psingle
3
0
-3
AF ca/co ratio
7(4)/19
Candidate variants
LIPH
1
Conserved elements
0

FIG. 2A

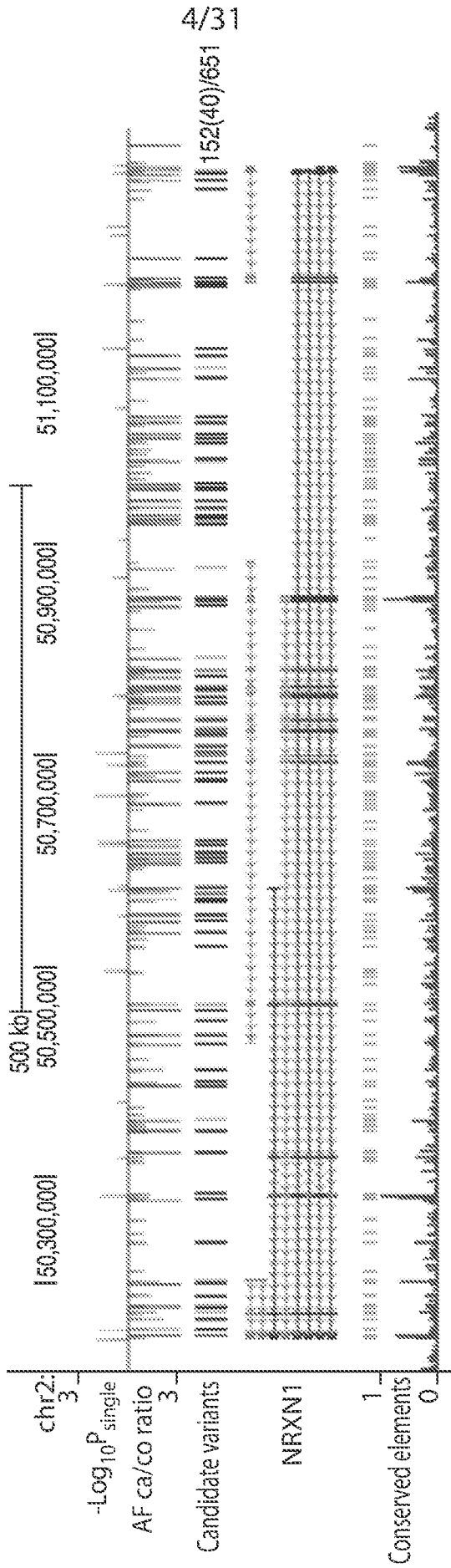

chr2:
150,300,000
500 kb
50,500,000
50,700,000
50,900,000
51,100,000
-Log10Psingle
AF ca/co ratio
3
0
Candidate variants
152(40)/651
NRXN1
Conserved elements
1
0

FIG. 2B

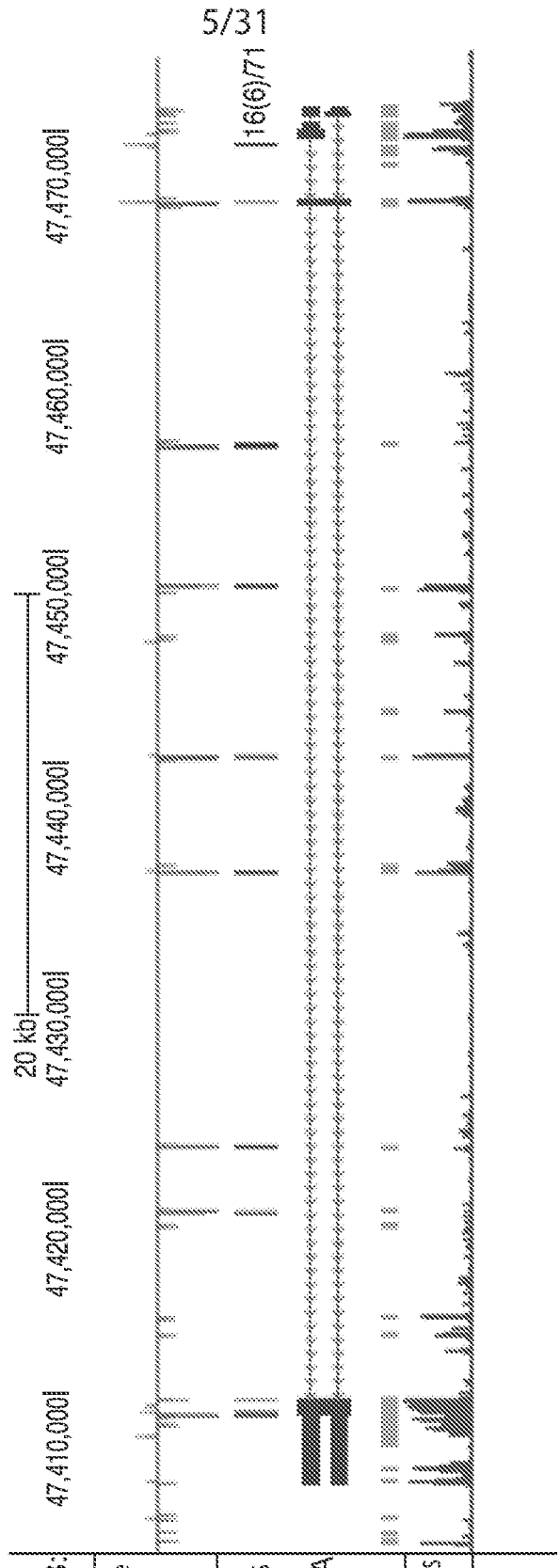

chr3:
47,410,000
47,420,000
20 kb
47,430,000
47,440,000
47,450,000
47,460,000
47,470,000
-Log10Psingle
3
AF ca/co ratio
3
5/31
16(6)/71
Candidate variants
HTR2A
1
Conserved elements
0

FIG. 2C

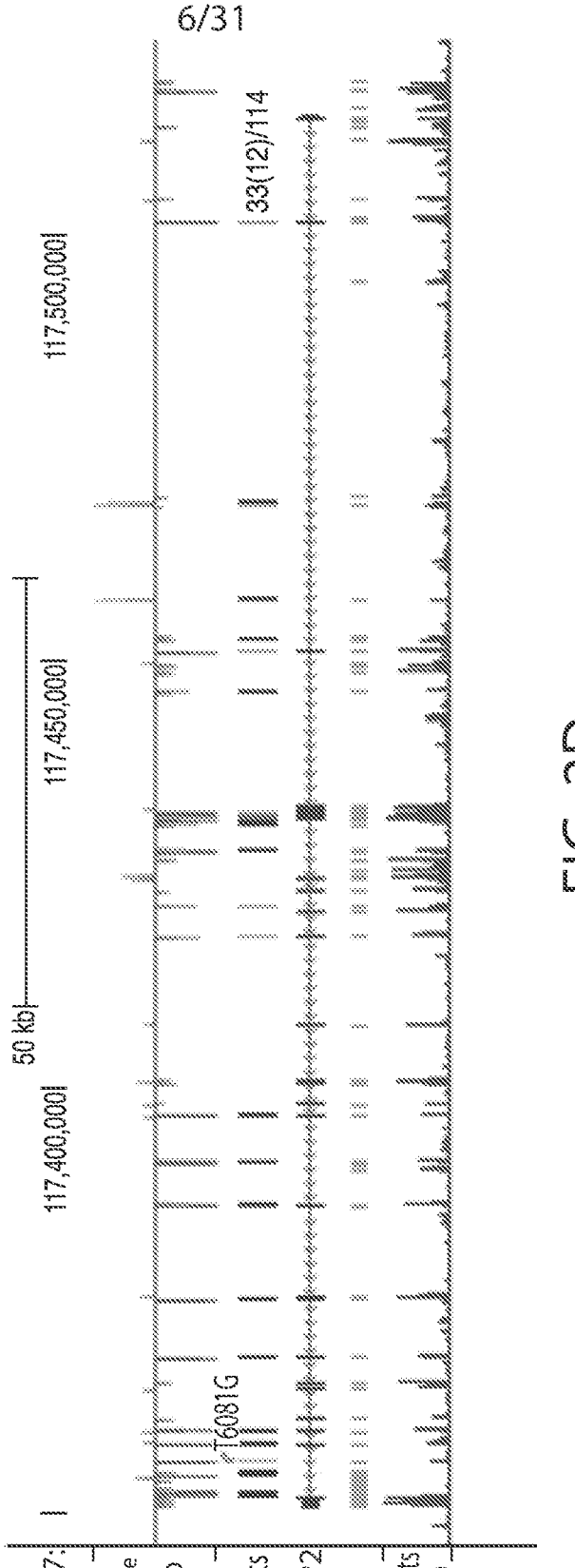

chr7:
117,400,000
50 kb
117,450,000
117,500,000
-Log10Psingle
AF ca/co ratio
T6081G
Candidate variants
33(12)/114
CTTNBP2
Conserved elements

FIG. 2D

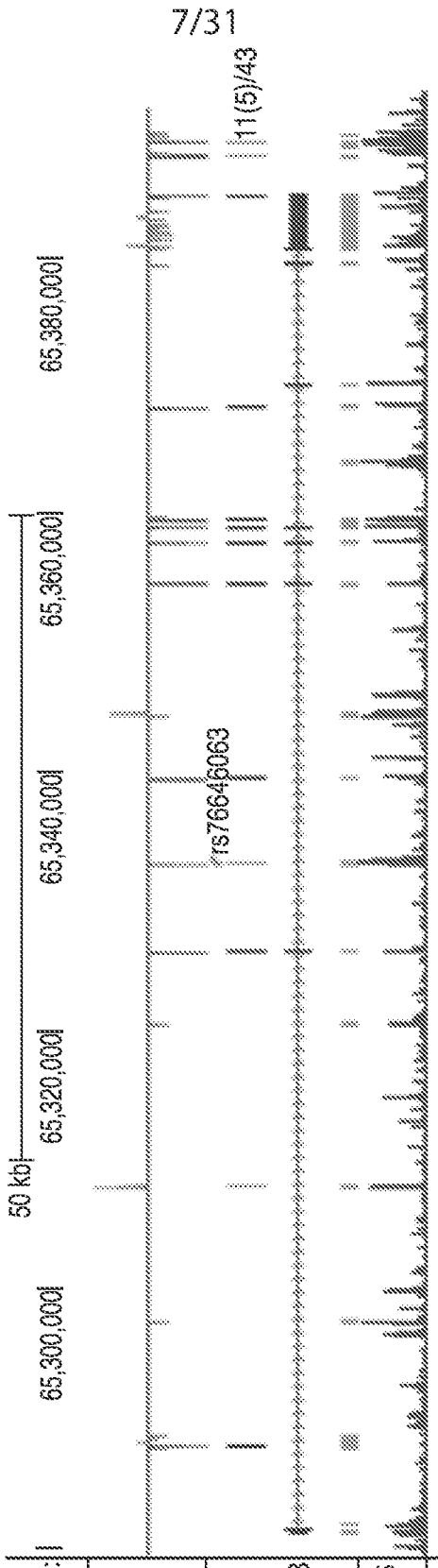

chr10:
65,300,000
65,320,000
65,340,000
65,360,000
65,380,000
50 kb
-Log10Psingle
3
AF ca/co ratio
3
Candidate variants
rs76646063
11(5)/43
REEP3
Conserved elements
1
0

FIG. 2E

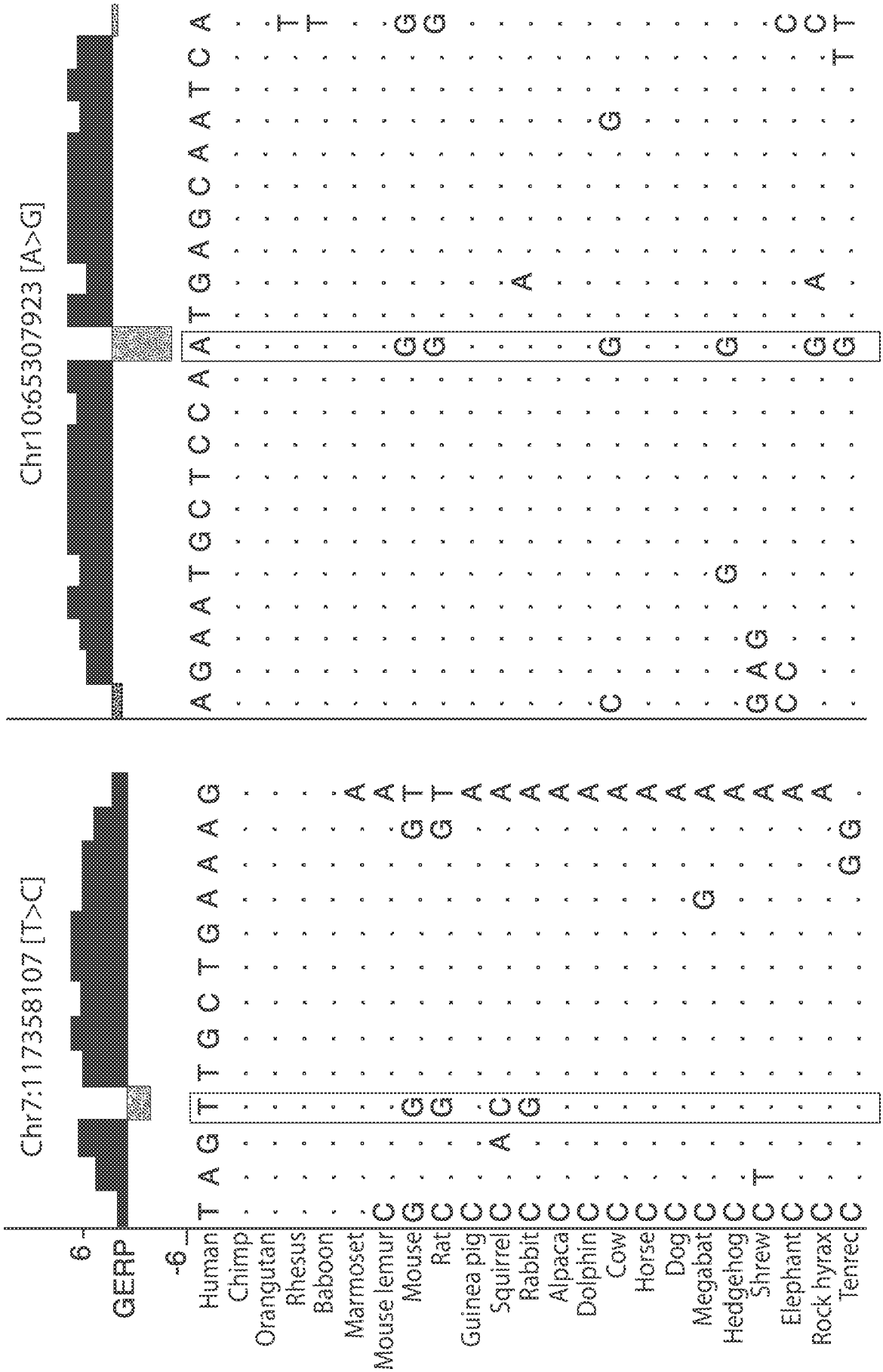

Chr7:117358107 [T>C]
Chr10:65307923 [A>G]
GERP
6
-6
Human
Chimp
Orangutan
Rhesus
Baboon
Marmoset
Mouse lemur
Mouse
Rat
Guinea pig
Squirrel
Rabbit
Alpaca
Dolphin
Cow
Horse
Dog
Megabat
Hedgehog
Shrew
Elephant
Rock hyrax
Tenrec

FIG. 3A

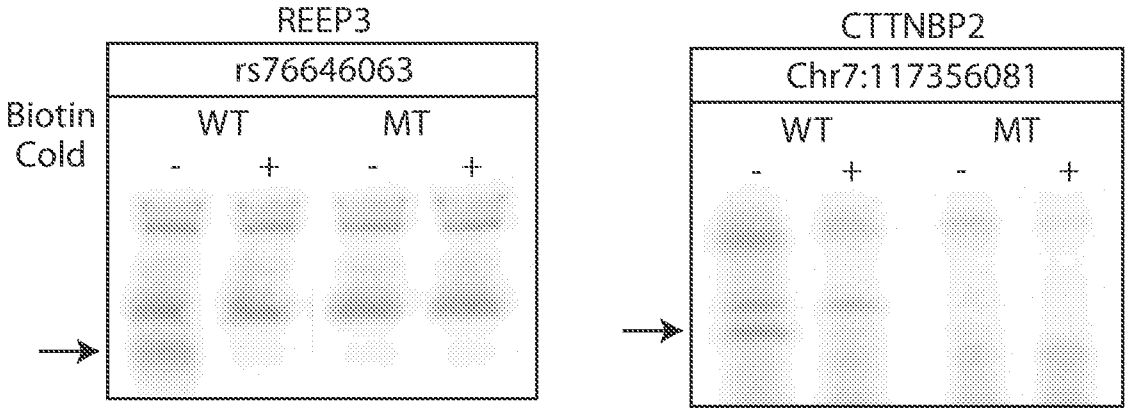

REEP3
rs76646063
Biotin
Cold
WT
MT
-
+
-
+
CTTNBP2
Chr7:117356081
WT
MT
-
+
-
+

FIG. 3B

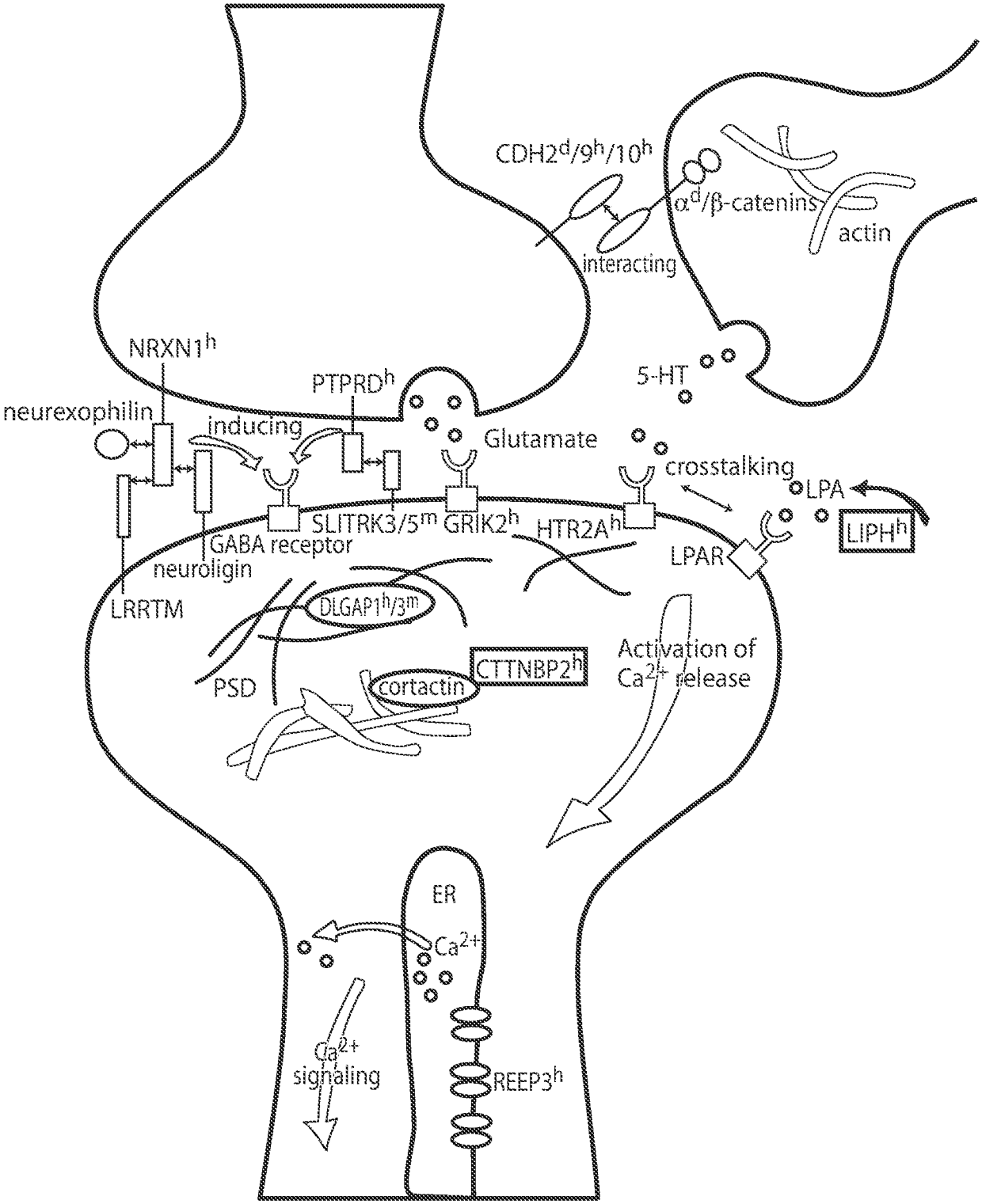

CDH2d/9h/10h
αd/β-catenins
actin
interacting
NRXN1h
PTPRDh
5-HT
neurexophilin
inducing
Glutamate
crosstalking
LPA
SLITRK3/5m
GRIK2h
HTR2Ah
LIPHh
GABA receptor
neuroligin
LPAR
LRRTM
DLGAP1h/3m
Activation of
Ca2+ release
CTTNBP2h
PSD
cortactin
ER
Ca2+
Ca2+
signaling
REEP3h

FIG. 4

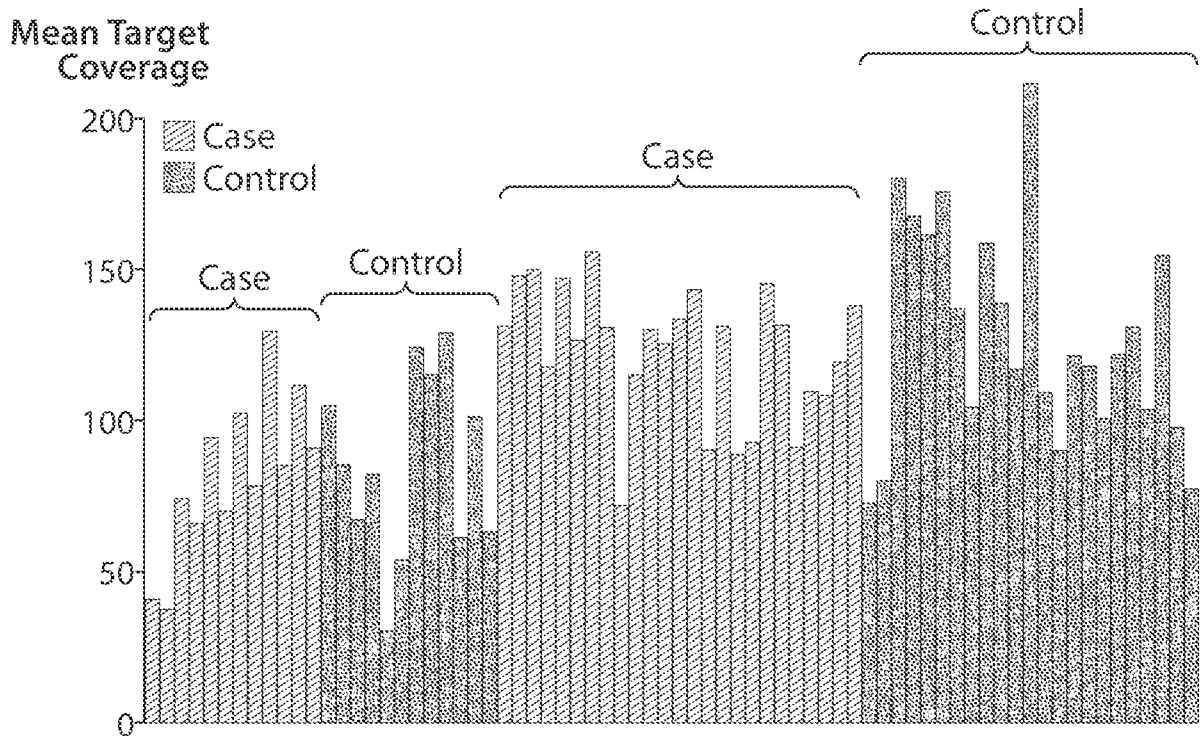

Mean Target Coverage
200
150
100
50
0
Case
Control
Case
Control
Case
Control

FIG. 5A

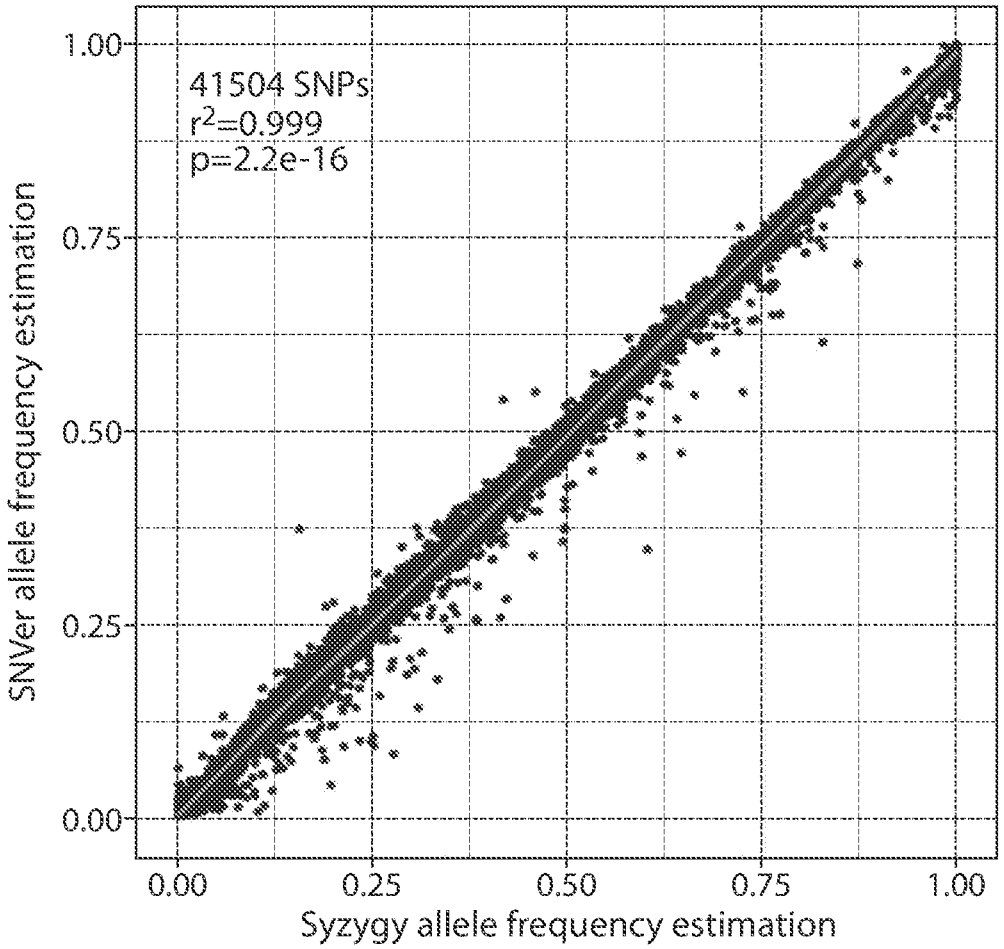

41504 SNPs
r2=0.999
p=2.2e-16
SNVer allele frequency estimation
Syzygy allele frequency estimation
0.00
0.25
0.50
0.75
1.00

FIG. 5B

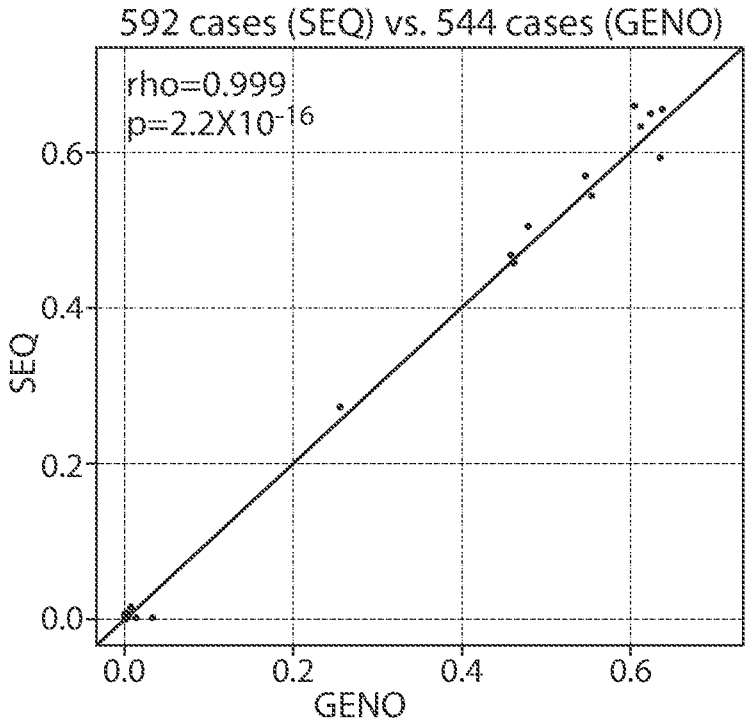

592 cases (SEQ) vs. 544 cases (GENO)
rho=0.999
p=2.2X10-16
SEQ
GENO
0.0
0.2
0.4
0.6

FIG. 6A

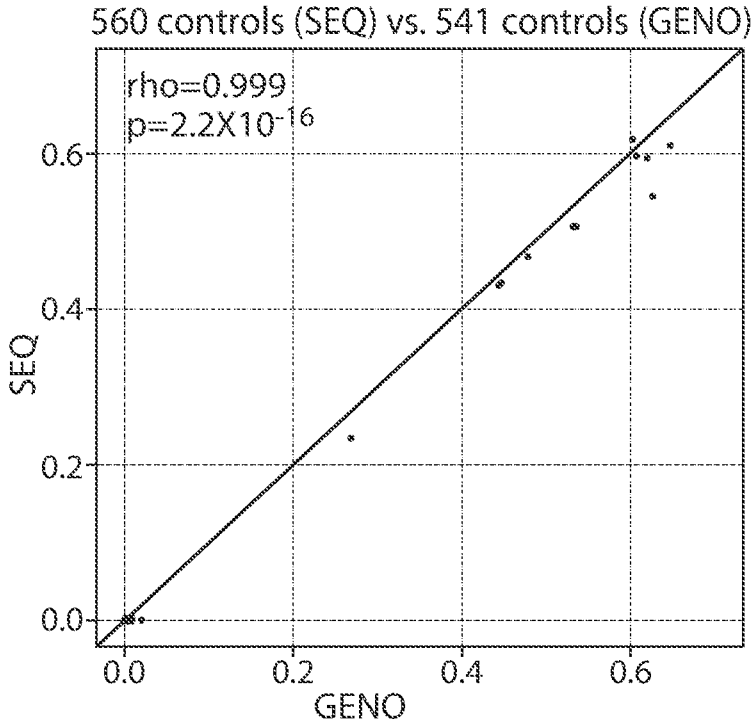

560 controls (SEQ) vs. 541 controls (GENO)
rho=0.999
p=2.2X10-16
SEQ
GENO
0.0
0.2
0.4
0.6

FIG. 6B

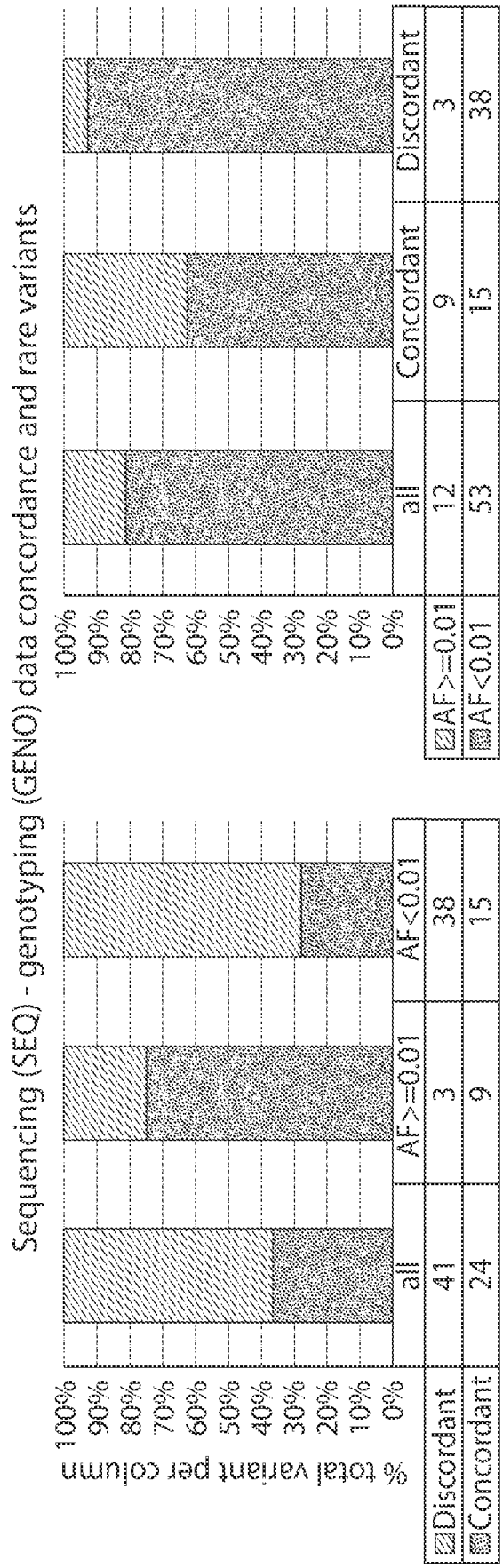
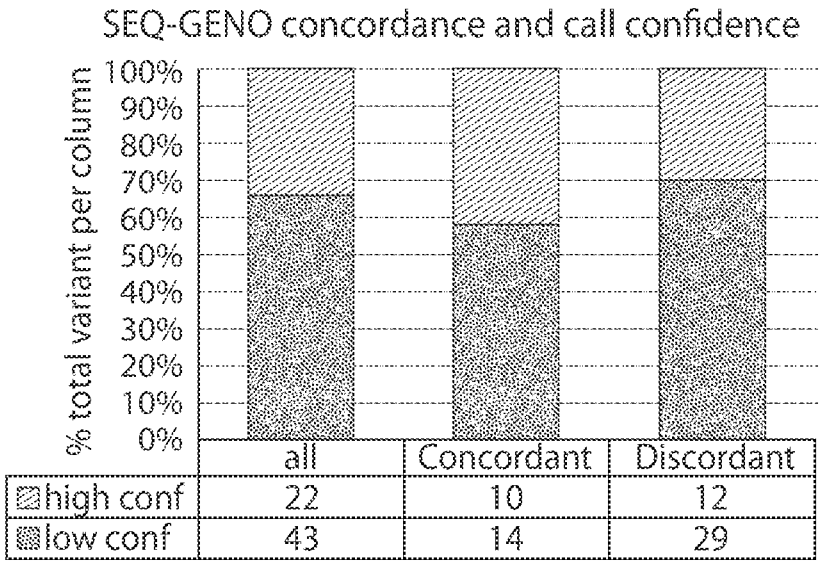
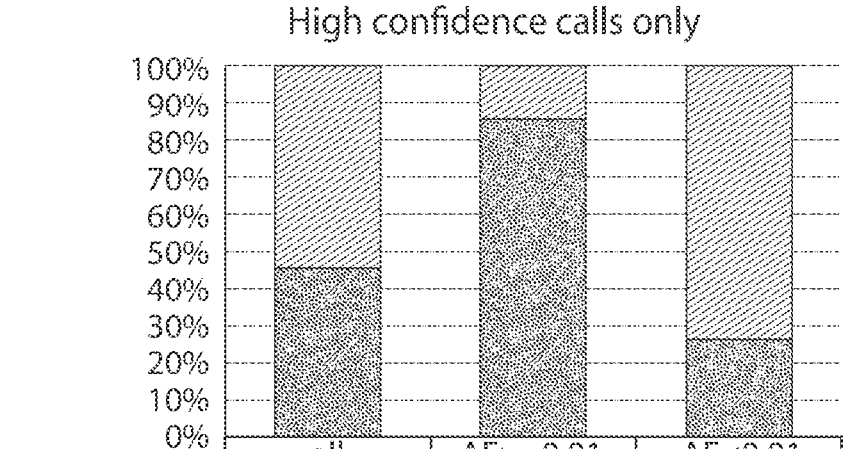

Sequencing (SEQ) - genotyping (GENO) data concordance and rare variants
% total variant per column
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Discordant
Concordant
all
AF>=0.01
AF<0.01
41
24
3
9
38
15
12
53
9
15
3
38

FIG. 7A



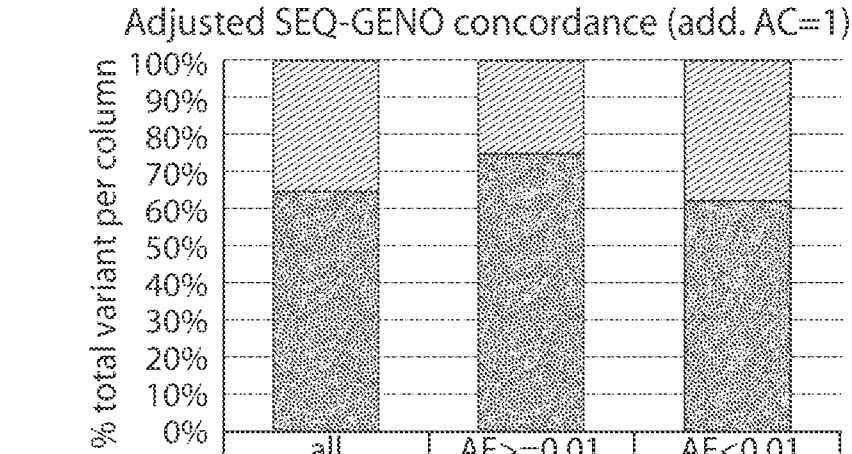
	all	Concordant	Discordant
high conf	22	10	12
low conf	43	14	29

FIG. 7B



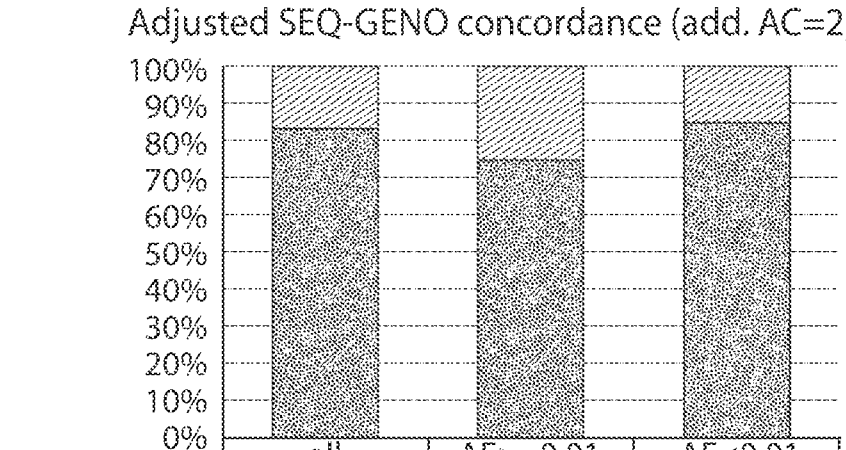
	all	AF>=0.01	AF<0.01
Discordant	12	1	11
Concordant	10	6	4

FIG. 7C



	all	AF>=0.01	AF<0.01
Discordant	23	3	20
Concordant	42	9	33

FIG. 7D



	all	AF>=0.01	AF<0.01
Discordant	11	3	8
Concordant	54	9	45

FIG. 7E

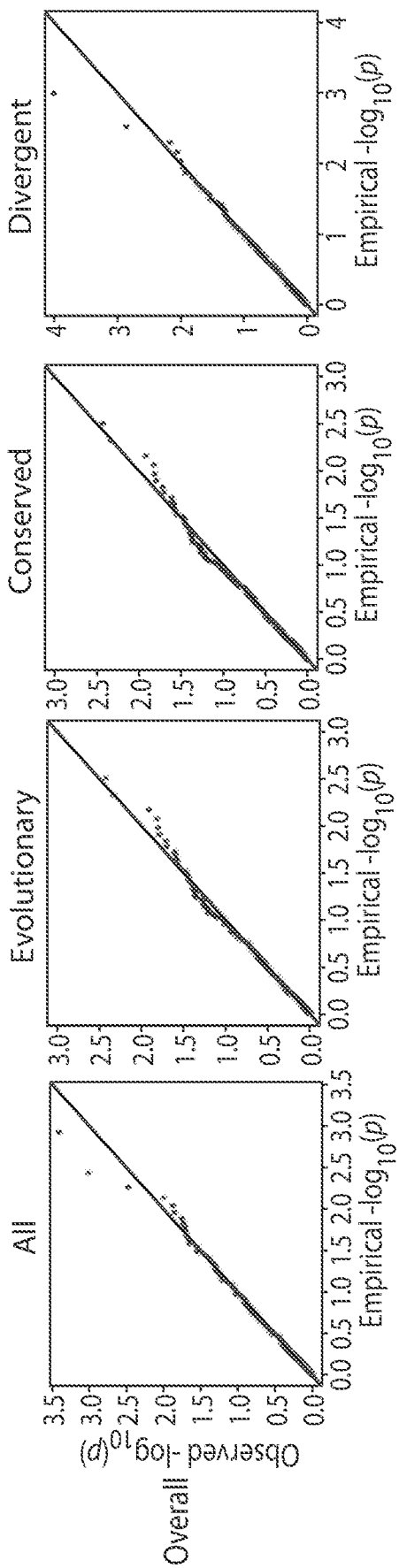

Overall
All
Evolutionary
Conserved
Divergent
Observed -log10(p)
Empirical -log10(p)

FIG. 8A

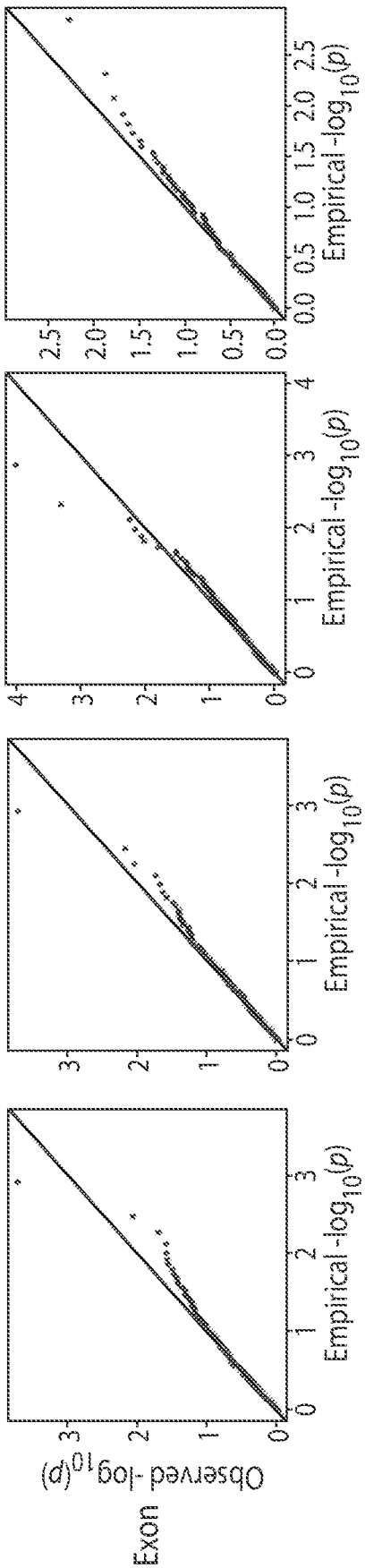

Exon
Observed -log10(p)
Empirical -log10(p)
Empirical -log10(p)
Empirical -log10(p)
Empirical -log10(p)

FIG. 8B

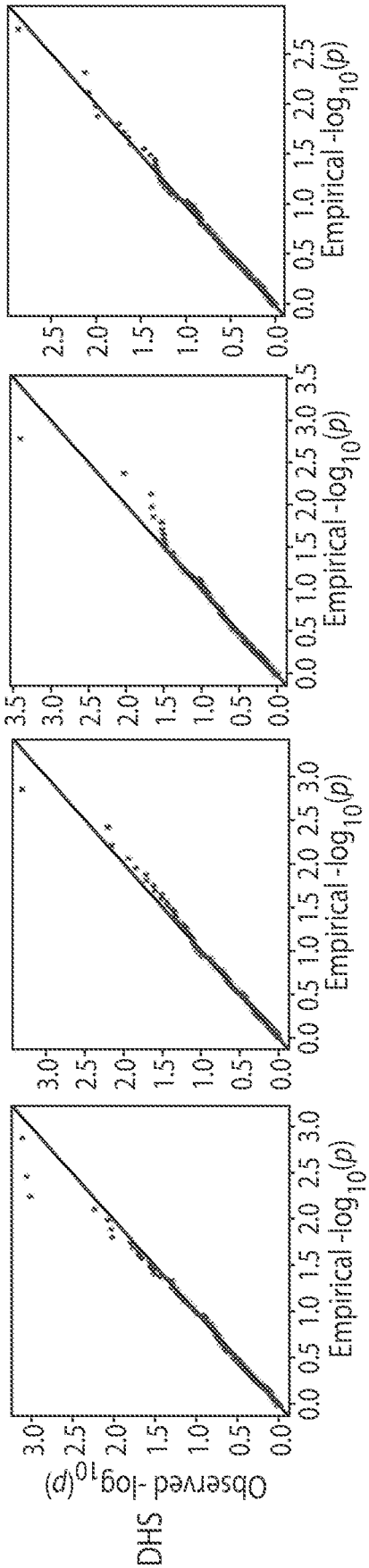


FIG. 8C

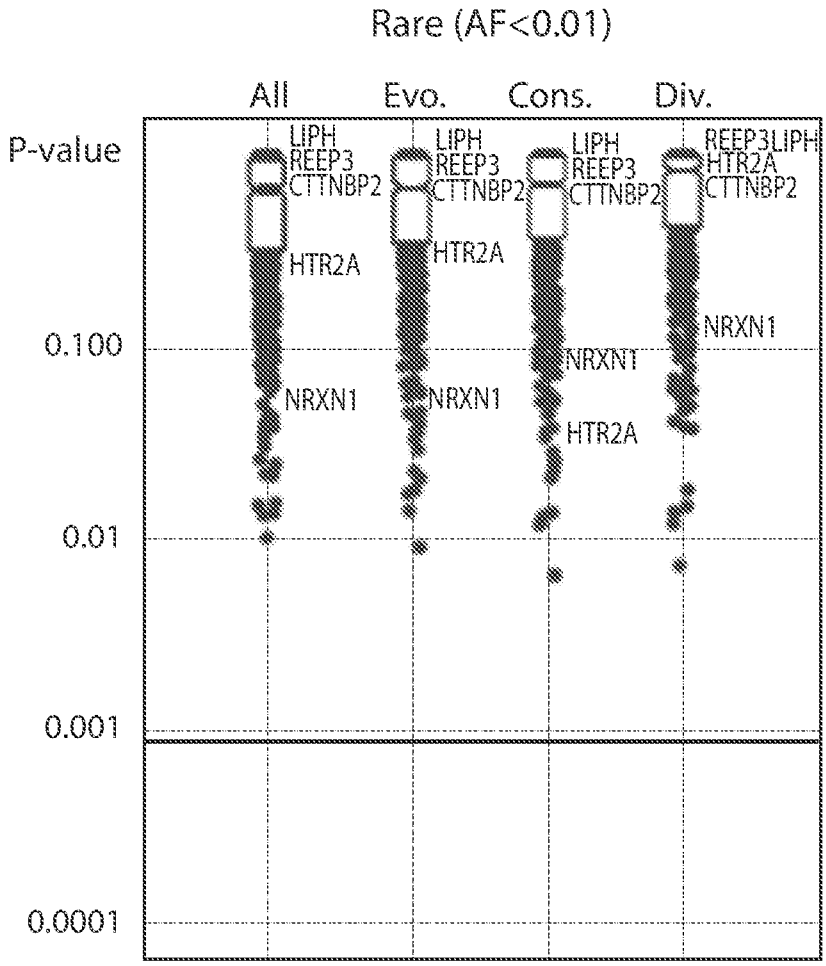

Rare (AF<0.01)
All
Evo.
Cons.
Div.
P-value
0.100
0.01
0.001
0.0001
LIPH
REEP3
CTTNBP2
HTR2A
NRXN1
REEP3LIPH
HTR2A
CTTNBP2

FIG. 8D

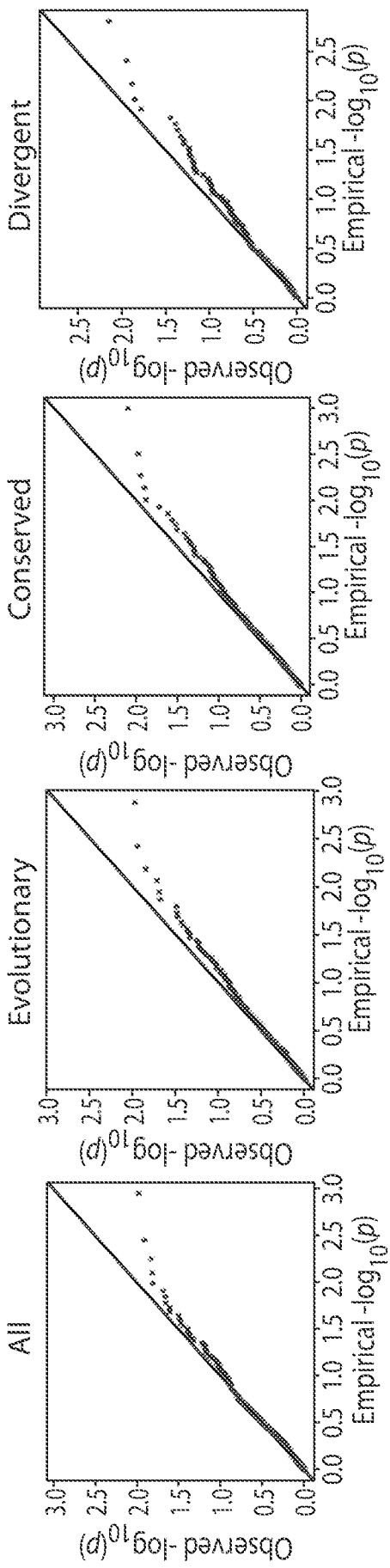

All
Evolutionary
Conserved
Divergent
Observed -log10(p)
Empirical -log10(p)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

FIG. 8E

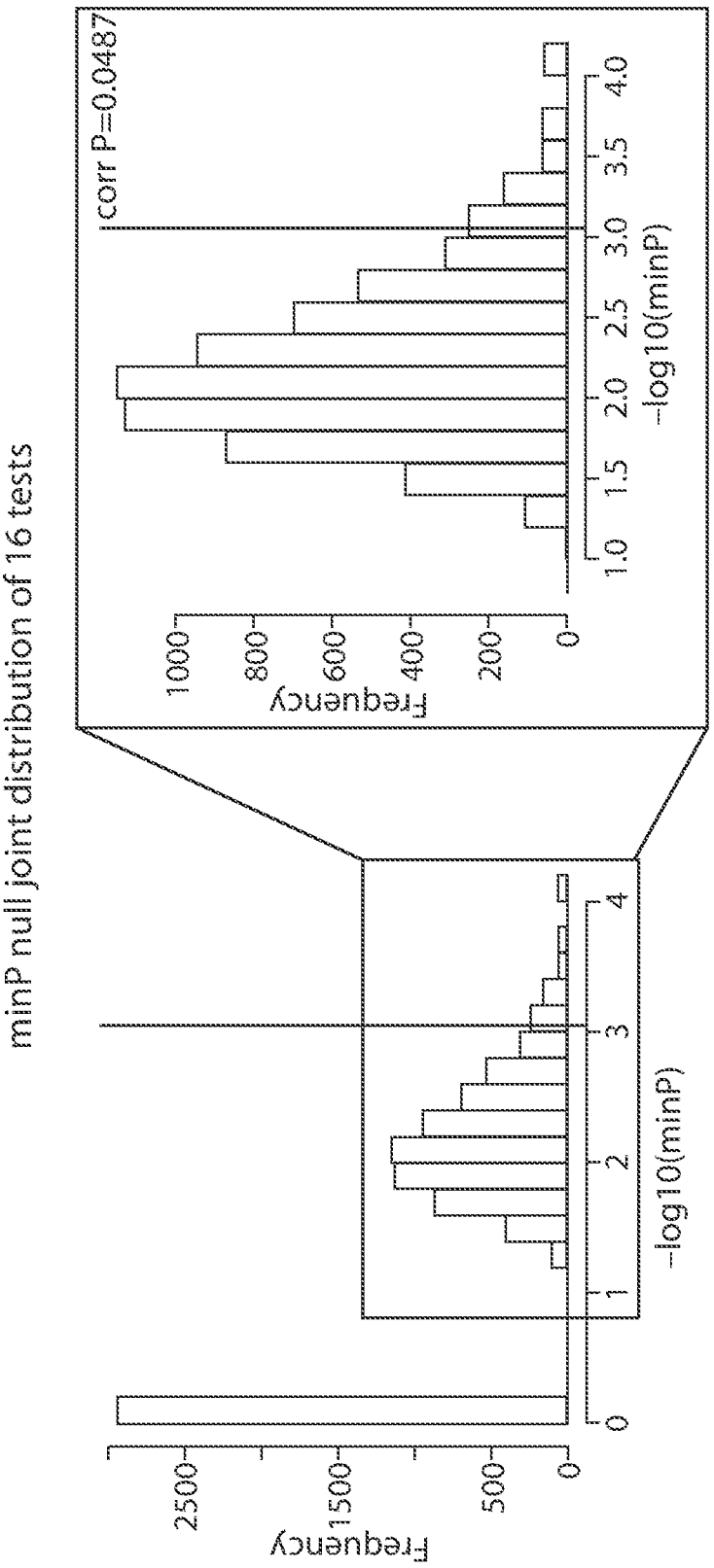

minP null joint distribution of 16 tests
Frequency
2500
1500
500
0
0
1
2
3
4
−log10(minP)
corr P=0.0487
Frequency
1000
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
−log10(minP)

FIG. 9

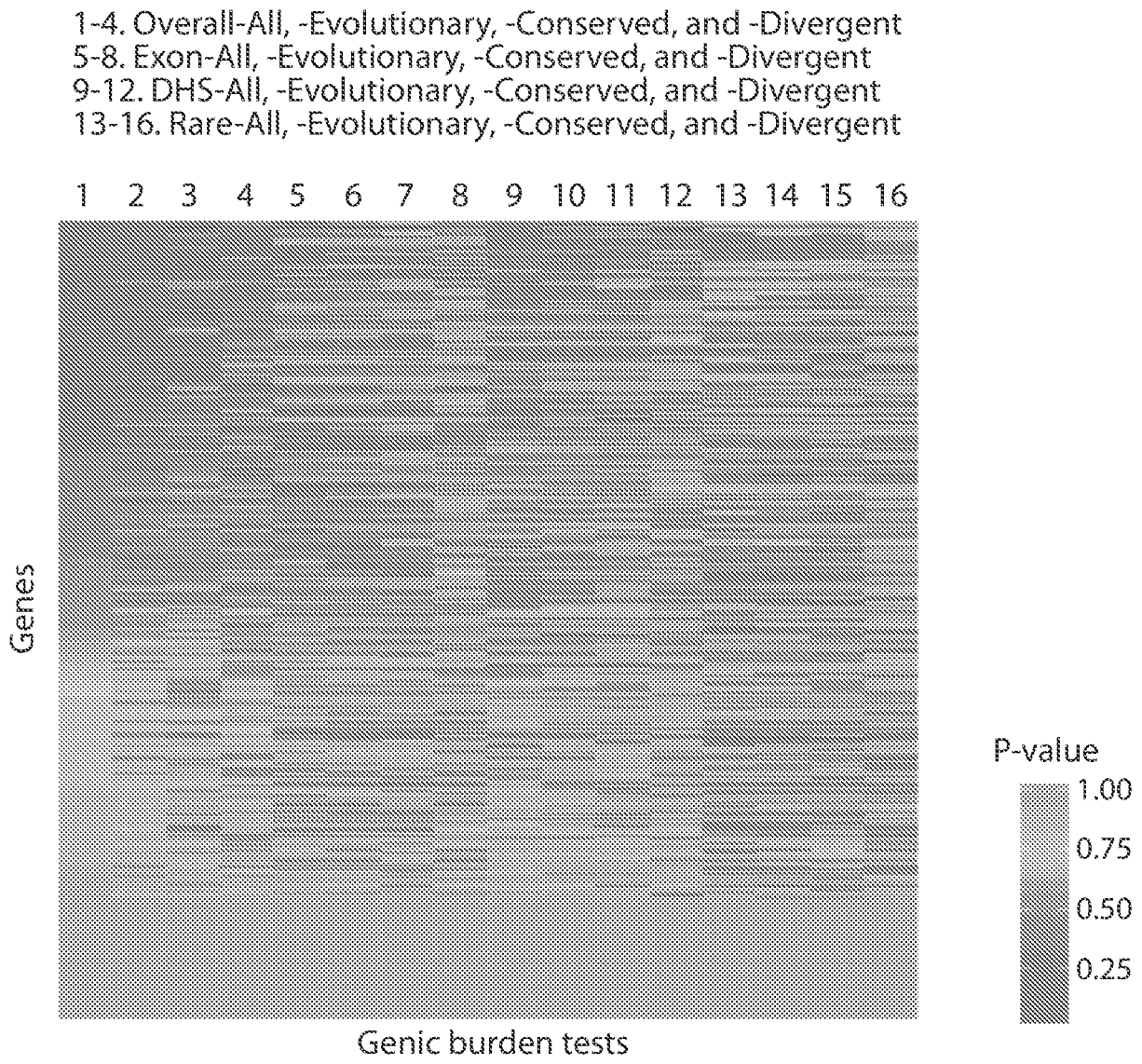

1-4. Overall-All, -Evolutionary, -Conserved, and -Divergent
5-8. Exon-All, -Evolutionary, -Conserved, and -Divergent
9-12. DHS-All, -Evolutionary, -Conserved, and -Divergent
13-16. Rare-All, -Evolutionary, -Conserved, and -Divergent
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Genes
Genic burden tests
P-value
1.00
0.75
0.50
0.25

FIG. 10

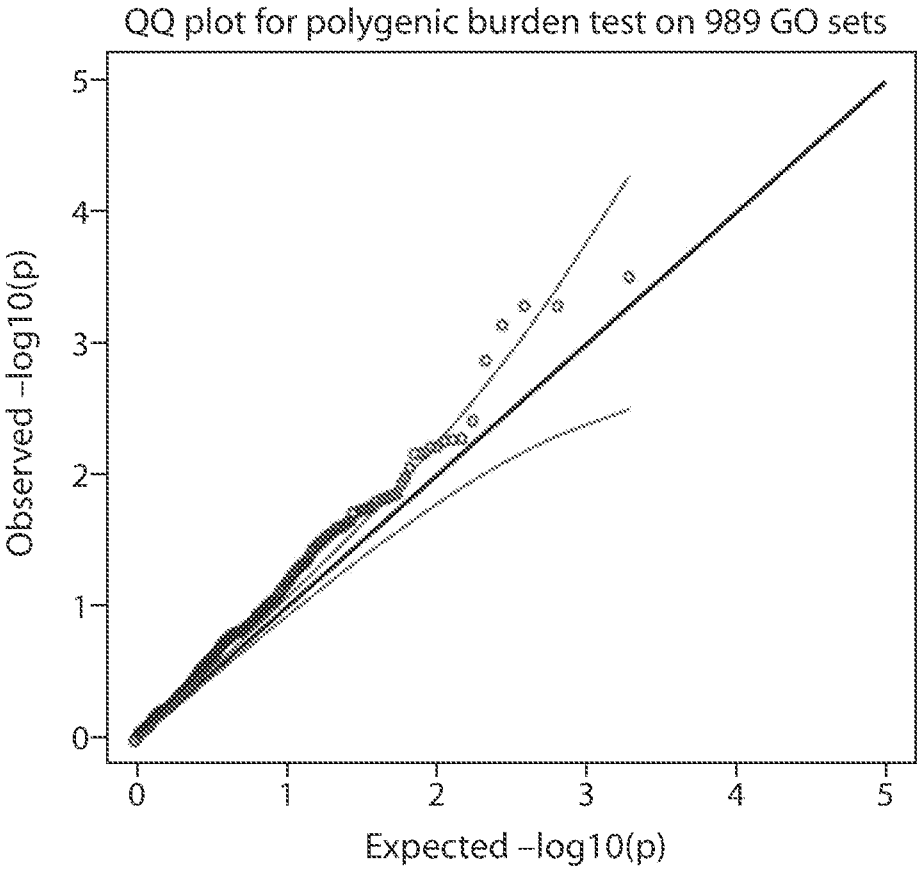


FIG. 11A

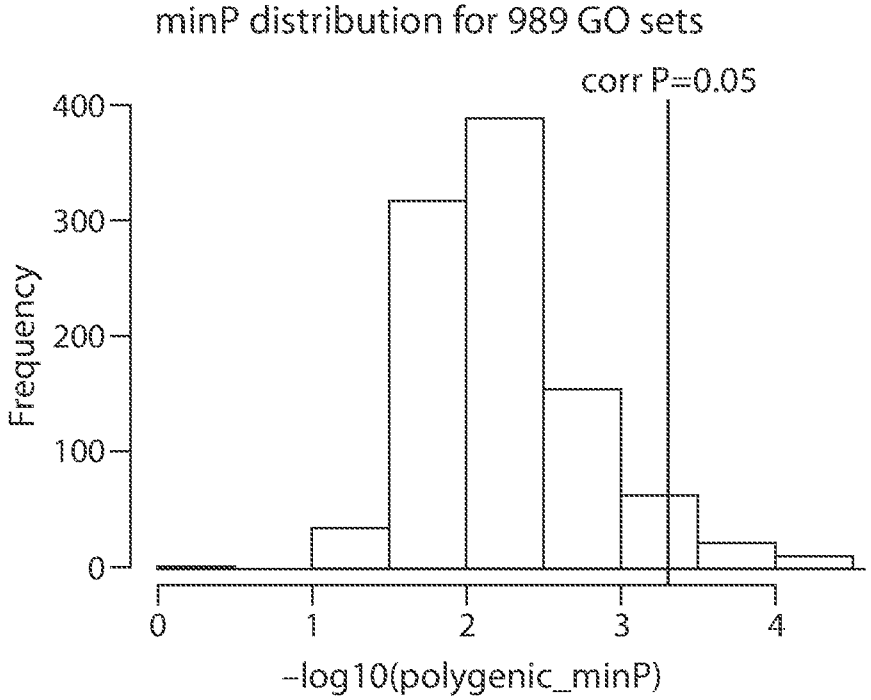


FIG. 11B

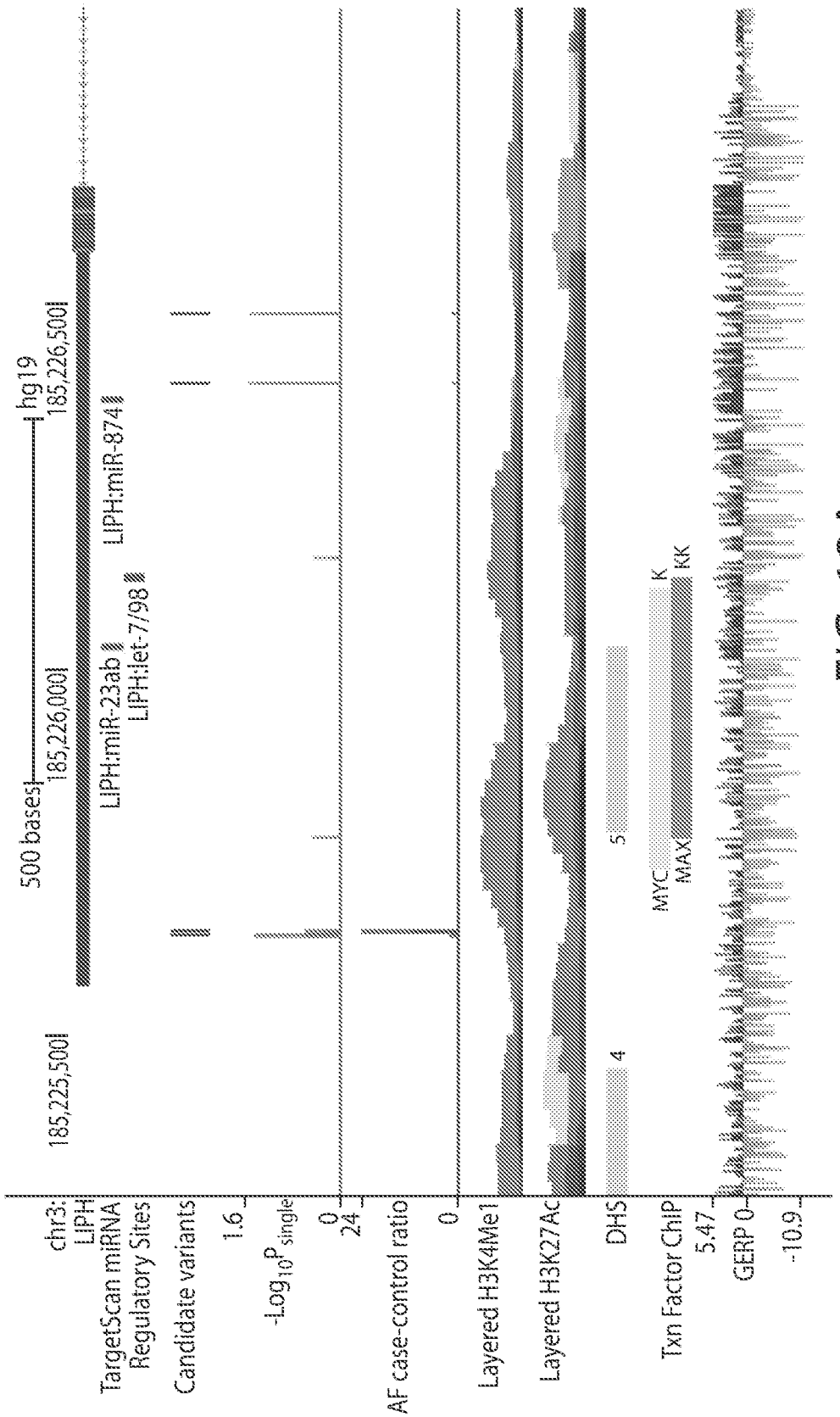

chr3: LIPH
185,225,500
500 bases
185,226,000
hg19
185,226,500
TargetScan miRNA
LIPH:miR-23ab
LIPH:miR-874
LIPH:let-7/98
Regulatory Sites
Candidate variants
1.6
-Log10Psingle
0
24
AF case-control ratio
0
Layered H3K4Me1
Layered H3K27Ac
DHS
4
5
Txn Factor ChIP
MYC
MAX
K
KK
5.47
GERP 0
-10.9

FIG. 12A

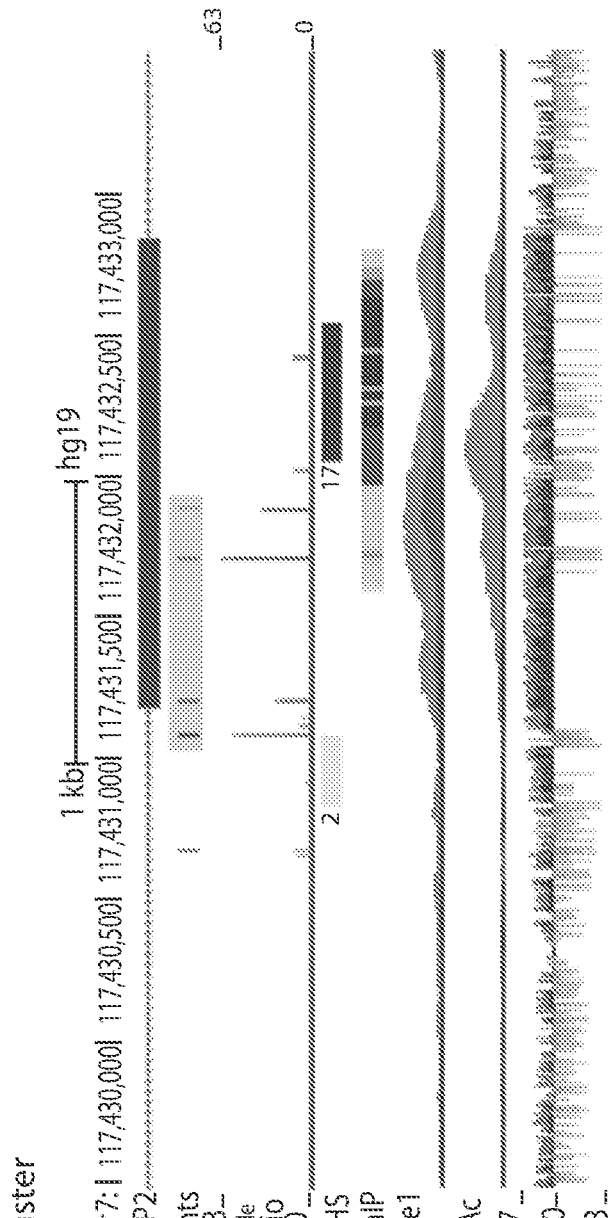

3 coding/1 nonDHS cluster
1 kb
hg19
chr7: 117,430,000 117,430,500 117,431,000 117,431,500 117,432,000 117,432,500 117,433,000
CTTNBP2
Candidate variants
-Log10Psingle
AF case-control ratio
DHS
Txn Factor ChIP
Layered H3K4Me1
Layered H3K27Ac
GERP

FIG. 12B

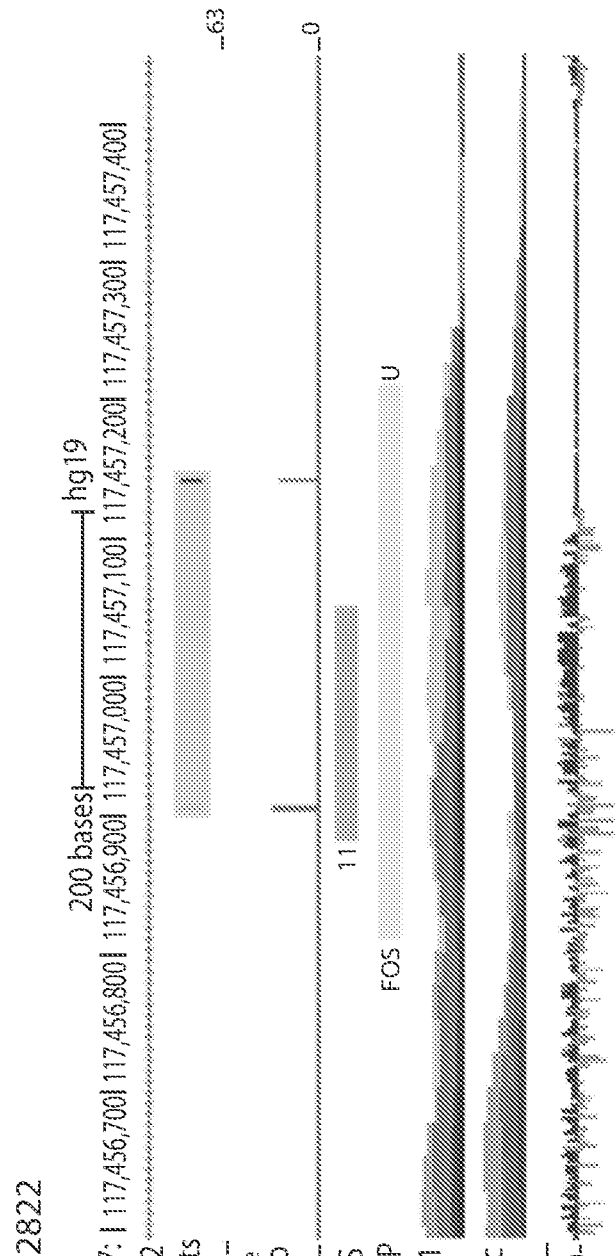


FIG. 12B (CONTINUED)

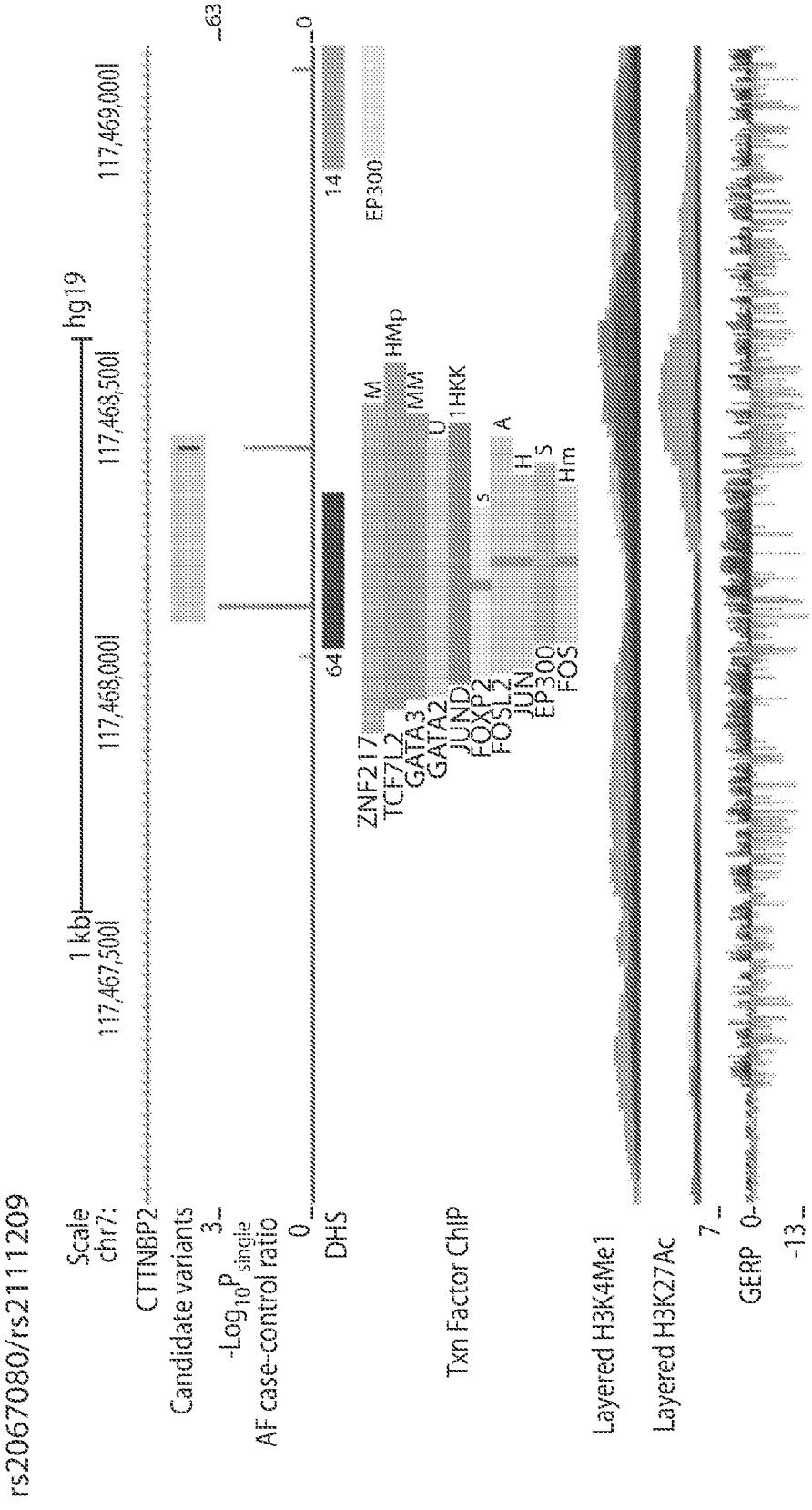


FIG. 12B (CONTINUED)

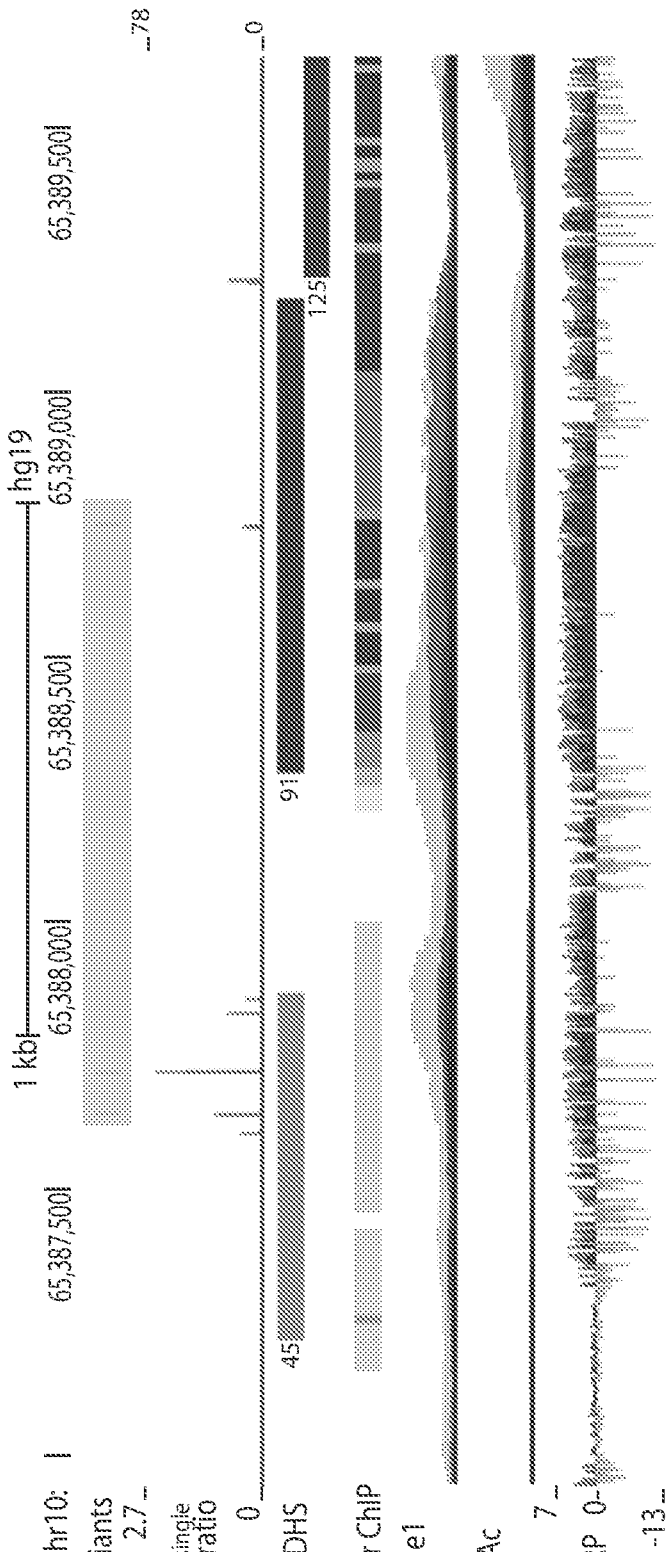

chr10:
65,387,500
1 kb
hg19
65,388,000
65,388,500
65,389,000
65,389,500
Candidate variants
2.7
-Log10Psingle
AF case-control ratio
0
78
0
DHS
45
91
125
Txn Factor ChIP
Layered H3K4Me1
Layered H3K27Ac
7
GERP 0
-13

FIG. 12C

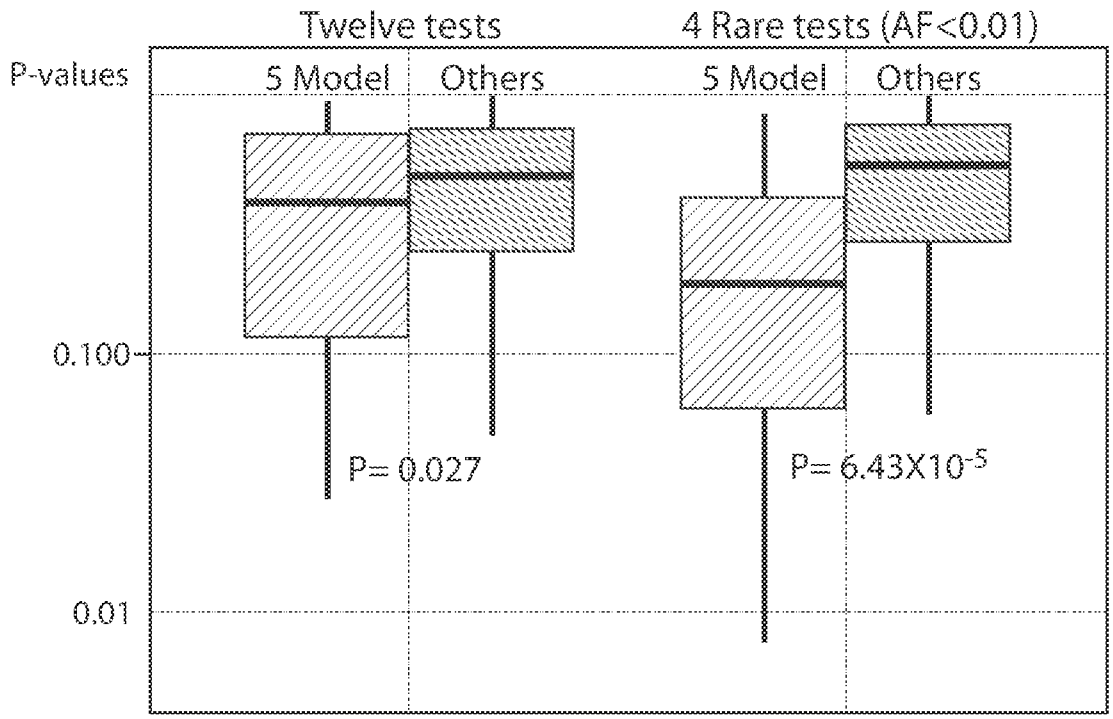

Twelve tests
4 Rare tests (AF<0.01)
P-values
5 Model
Others
5 Model
Others
0.100
0.01
P= 0.027
P= 6.43X10-5

FIG. 13

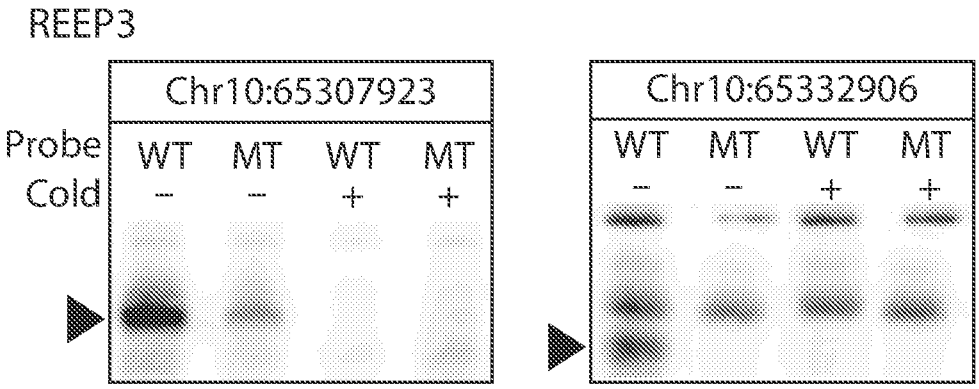


FIG. 14A

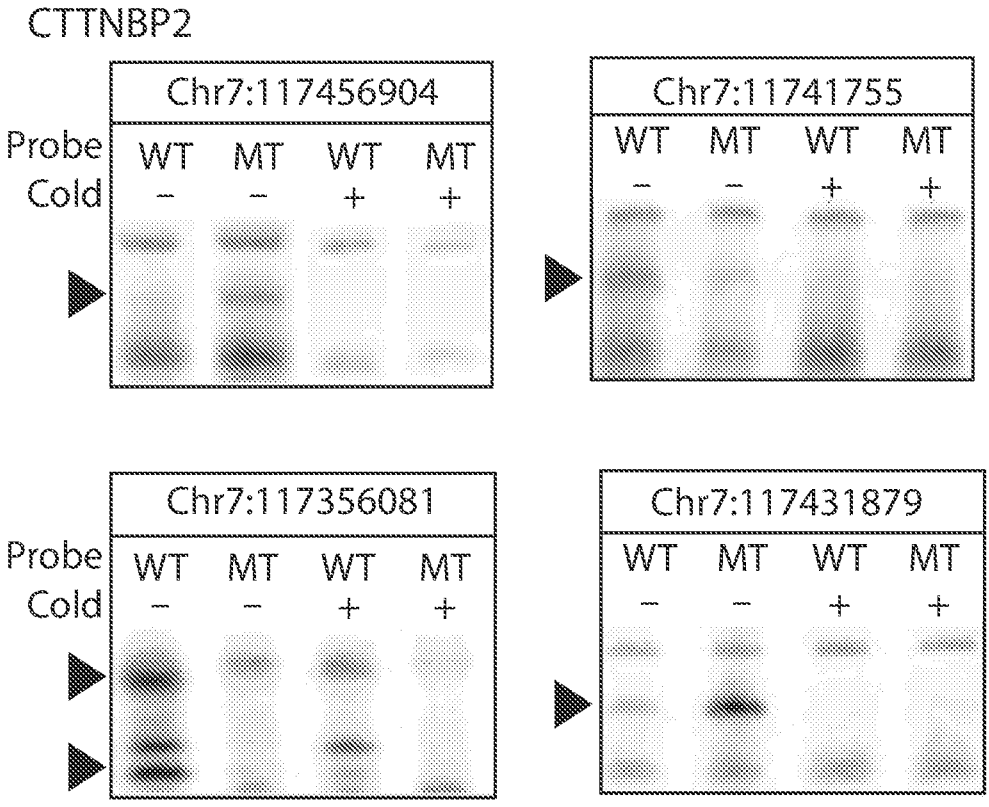


FIG. 14B