Integrating evolutionary and regulatory information with multispecies approach implicates genes and pathways in obsessive-compulsive disorder

Hyun Ji Noh et al.

Obsessive-compulsive disorder is a severe psychiatric disorder linked to abnormalities in glutamate signaling and the cortico-striatal circuit. We sequenced coding and regulatory elements for 608 genes potentially involved in obsessive-compulsive disorder in human, dog, and mouse. Using a new method that prioritizes likely functional variants, we compared 592 cases to 560 controls and found four strongly associated genes, validated in a larger cohort. NRXN1 and HTR2A are enriched for coding variants altering postsynaptic protein-binding domains. CTTNBP2 (synapse maintenance) and REEP3 (vesicle trafficking) are enriched for regulatory variants, of which at least six (35%) alter transcription factor-DNA binding in neuroblastoma cells. NRXN1 achieves genome-wide significance ($p = 6.37 \times 10^{-11}$) when we include 33,370 population-matched controls. Our findings suggest synaptic adhesion as a key component in compulsive behaviors, and show that targeted sequencing plus functional annotation can identify potentially causative variants, even when genomic data are limited.
Obsessive-compulsive disorder (OCD) is a highly heritable (h² = 0.27–0.65)1, debilitating neuropsychiatric disorder characterized by intrusive thoughts and time-consuming repetitive behaviors. Over 80 million people worldwide are estimated to suffer from OCD, and most do not find relief with available therapeutics1, underscoring the urgency to better understand the underlying biology. Genome-wide association studies (GWAS) implicate glutamate signaling and synaptic proteins2,3, but specific genes and variants have not been validated. Isolating and characterizing such genes are important for understanding the biology and developing treatments for this devastating disease.

In mouse, genetically engineered lines have causally implicated the cortico-striatal neural pathway in compulsive behavior. Mice with a deletion of Sapap3 exhibit self-mutilating compulsive grooming and dysfunctional cortico-striatal synaptic transmission, with abnormally high activity of medium spiny neurons (MSNs) in the striatum. Resulting compulsive grooming is ameliorated by selective serotonin reuptake inhibitor (SSRI), a first-line medication for OCD4. Similarly, chronic optogenetic stimulation of the cortico-striatal pathway in normal mice leads to compulsive grooming accompanied by sustained increases in MSN activity5. Thus, excessive striatal activity, likely due to diminished inhibitory drive in MSN microcircuitry, is a key component of compulsive grooming. The brain region disrupted in this mouse model is also implicated by imaging studies in human OCD6.

Pet dogs are a natural model for OCD amenable to genome-wide mapping due to their unique population structure7. Canine compulsive disorder (canine CD) closely parallels OCD, with equivalent clinical metrics, including compulsive extensions of normal behaviors, typical onset at early social maturity, roughly a 50% rate of response to SSRIs, high heritability, and polygenic architecture8. Through GWAS and targeted sequencing in dog breeds with exceptionally high rates of canine CD, we associated genes involved in synaptic function and adhesion with CD, including neural cadherin (CDH2), catenin alpha2 (CTNNA2), ataxin-1 (ATXN1), and plasma glutamate carboxypeptidase (PGCP)5,9.

Human genetic studies of related disorders, such as autism spectrum disorders (ASD), suggest additional genes. Both ASD and OCD are characterized by repetitive behaviors, and high comorbidity suggests a shared genetic basis6. Genome-wide studies searching for de novo and inherited risk variants have confidently associated hundreds of genes with ASD; this set may be enriched for genes involved in OCD10.

Focusing on genes implicated by model organisms and related disorders could find variants underlying OCD risk, even with smaller sample sizes. Researchers, particularly in psychiatric genetics, are wary of “candidate gene” approaches, which often failed to replicate11. Closer examination of past studies suggests this approach is powerful and reliable when the set of genes tested is large, and the association is driven by rare variation11. A study testing 2000 candidate genes for association with diabetic retinopathy identified 25 genes, at least 11 of which achieved genome-wide significance in a GWAS of type 2 diabetes, a related disorder12, 13. A targeted-sequencing study of ASD, with 78 genes, identified four genes with recurrent, rare deleterious mutations; these four genes are also implicated by whole-exome sequencing studies14. Candidate gene studies also replicated associations to rare variants in APP, PSEN1, and PSEN2 for Alzheimer’s disease15, PCSK9 for low-density lipoprotein–cholesterol level16, and copy-number variants for autism and schizophrenia10.

Detecting associations driven by rare variants requires sequencing data, which captures nearly all variants. Although whole-genome sequencing studies of complex diseases are still prohibitively expensive, it is feasible to target a subset of the genome. Sequencing also facilitates identification of causal variants, accelerating discovery of new therapeutic avenues17, 18. For example, finding functional, rare variants in PCSK9 led to new therapies for hypercholesterolemia19. One approach is to target predominantly coding regions (whole-exome sequencing). Although successful in finding causal variants for rare diseases20, this approach misses the majority of disease-associated variants predicted to be regulatory21. A targeted-sequencing approach that captures both the regulatory and coding variation of a large set of candidate genes offers many advantages of whole-genome sequencing, and is feasible when cohort size and resources are limited.

Here we report a new strategy that overcomes limitations of less comprehensive candidate gene studies and exome-only approaches, and identifies functional variants associated with increased risk of OCD. We start by compiling a large set of 608 genes (~3% of human genes) using studies of compulsive behavior in dogs and mice, and studies of ASD and OCD in humans. By focusing on this subset of genes, targeting both coding and regulatory regions, and applying a new statistical method that incorporates regulatory and evolutionary information, we identify four associated genes, including NRXN1, the first genome-wide-significant association reported for OCD.

Results

Targeted-sequencing design. We compiled a list of 608 genes using three strategies (65 were implicated more than once) (Supplementary Table 1 and Supplementary Methods):

1. 263 “model-organism genes”, including 56 genes associated in canine CD GWAS and 222 genes implicated in murine-compulsive grooming.
2. 196 “ASD genes” from SFARI database (https://gene.sfari.org/) as of 2009.
3. 216 “human candidate genes” from small-scale OCD candidate gene studies (56 genes), family-based linkage studies of OCD (91 genes), and by other neuropsychiatric disorders (69 genes).

We targeted coding regions and 82,723 evolutionarily constrained elements in and around these genes, totaling 13.2 Mb (58 bp–16 kb size range, median size 237 bp), 34% noncoding22.

Variant detection. We sequenced 592 European ancestry DSM-IV OCD cases and 560 ancestry-matched controls using pooled sequencing, with 16 samples per bar-coded pool (37 “case” pools; 35 “control” pools). Overall, 95% of target regions were sequenced at ≥30× read depth per pool (median 112×; ~7× per individual; Supplementary Fig. 1), sufficient to identify variants occurring in just one individual, assuming 0.5–1% per base machine error rate.

We called 124,541 single nucleotide polymorphisms (SNPs) using Szyzygy (84,216)17 and SNVer (81,829)23. For primary analyses, we focused on 41,504 “high-confidence” SNPs detected by both, with highly correlated allele frequencies (AF) (Pearson’s ρ = 0.999, p < 2.2 × 10–15; Supplementary Fig. 2). We see no significant difference between case and control pools, indicating no bias in variant detection.

Variant annotation. We used three annotations shown to be enriched for disease-associated variation to identify likely functional variants in our targeted regions: coding, evolutionary conserved, and/or DNase hypersensitivity site (DHS)24–27. We annotated 67% (27,626) of high-confidence variants, with 16% coding (49% of those were non-synonymous), 36% DHS,
and 80% evolutionary conserved or divergent (Fig. 1a). We measured evolutionary constraint using mammalian GERP++ scores; scores > 2 were "conserved" and scores < -2 were "divergent".

**Gene-based burden analysis.** To identify genes with a significant load of non-reference alleles in OCD cases, relative to controls, we developed PolyStrat, a one-sided gene-based burden test that controls for gene length (Supplementary Fig. 3a) and incorporates variant annotation. We used four variant categories: (i) all (Overall), (ii) coding (Exon), (iii) regulatory (variants in DHS), and (iv) rare (1000 Genomes Project AF < 0.01). Each category is further stratified by evolutionary status: (i) all detected variants; (ii) slow-evolving conserved (Cons); (iii) fast-evolving divergent (Div); and (iv) evolutionary (Evo). "Evo" is the subset of "all" variants annotated as either "conserved" or "divergent". In total,
variants (Supplementary Fig. 4).

is the only gene with excess divergent (potentially fast REEP3 evolving) variants. No genes had a significance of association possibly inflated by linkage between markers.
study of the disease of interest is not available, targeting genes implicated in a model organism may be as effective as targeting genes implicated in a comorbid, phenotypically similar human disorder.

Fig. 3 Validated top candidate variants disrupt functional elements active in brain. a Frequencies of the top candidate variants show that most are rare (AF < 0.01) in our cohort, illustrating the value of sequencing rather than array-based genotyping for detecting candidate variants. b–g Allele frequencies from pooled sequencing of individuals in the original cohort (cohort 1) are validated by genotyping in controls; and for allele-frequency differences between cases and controls. For the vast majority of variants, the pooled-sequencing allele frequencies are also highly correlated with frequencies observed in 33,370 ExAC individuals in both cases (b, gray dots) and controls (c, gray dots). Genotyping an independent cohort (cohort 2) of cases and controls reveals genotyping allele frequencies from cohort 1 are correlated in e cases and f controls; and for g allele-frequency differences between cases and controls. Correlation test was performed using Fisher’s Z transform. h The top candidate variants in CTNNBP2 and REEP3, the two genes enriched for regulatory variants, disrupt DNase hypersensitivity sites (DHS), enhancers, and transcription factor-binding sites (TFBS) annotated as functional in brain tissues and cell lines in ENCODE and Roadmap Epigenomics. All 17 variants disrupt elements active in either the dorsolateral prefrontal cortex and/or substantia nigra, which are among the brain regions involved in the CSTC circuit implicated in OCD, illustrated with black arrows. Image adapted from Creative Commons original by Patrick J. Lynch and C. Carl Jaffe, MD. j The CSTC circuit requires a balance between a direct, GABAergic signaling pathway and an indirect pathway that involves both GABAergic and glutamate signaling. In OCD patients, an imbalance favoring the direct over the indirect pathway disrupts the normal functioning of the CSTC circuit.
The five genes most strongly implicated in canine CD and murine-compulsive grooming (CDH2, CTNNA2, ATXN1, POC1, and SapaP3) have significantly lower p-values than the other 603 sequenced genes (Wilcoxon unpaired, one-sided p = 2.6 × 10^{-3}). The difference becomes more significant when only rare variants are tested (Wilcoxon unpaired, one-sided p = 3.2 × 10^{-5}) (Fig. 1c). This is consistent with the hypothesis that severe disease-causing variants, rare in humans due to negative selection, may persist at higher frequencies in model organisms where selection is relaxed.

Applying the burden test across multiple genes with shared biological functions, we identified gene sets with high-variant burdens: “cell death” (uncorrected p = 3 × 10^{-4}, corrected p < 0.03) and “GO:0031334 positive regulation of protein complex assembly” (uncorrected p = 7 × 10^{-4}, corrected p < 0.06). Overlaying the burden test results onto the GO network topology highlights functional themes linking the enriched gene sets: regulation of protein complex assembly and cytoskeleton organization; neuronal migration; action potential; and cytoplasmic vesicle (Supplementary Fig. 6).

Validation of candidate variants by genotyping. We genotyped the top 67 candidate functional variants from the five significant genes, including 42 rare SNPs (AF < 0.01), in the pooled-sequencing cohort (Fig. 3a). This yielded, after QC, individual genotypes for 63 SNPs in 571 cases and 555 controls (98% of the largest effect variants in the combined cohort (1298 OCD cases and 1660 controls) alter enhancer elements in these two genes: 61% also had a higher frequency of the non-reference allele in cases in the second cohort (Wilcoxon paired one-sided test for 63 SNPs, p = 0.005). More specifically, of 54 SNPs that had a higher frequency of the non-reference allele in cases in the first cohort, 61% also had a higher frequency of the non-reference allele in cases in the second cohort. The 33 SNPs that failed to validate in either of the two cohorts had smaller allele-frequency differences in the first cohort (one-sided unpaired t-test p = 0.02).

In summary, the allele-frequency analysis described above identified four genes: NRXN1, HTR2A, CTNNBP2, and REEP3. LIPH is excluded because its association is likely slightly inflated by LD and the genotyping in the second cohort did not reproduce as clearly. To validate the associations, we employed distinct strategies depending on whether the association was driven by coding (NRXN1 and HTR2A) or regulatory variation (CTNNBP2 and REEP3).

**Table 2 Candidate regulatory variants**

| Chrpos         | Ref    | Alt   | rsID   | Transcription factor         | EMSA  | OR  | GERP |
|----------------|--------|-------|--------|-------------------------------|-------|-----|------|
| CTTNB2          |        |       |        |                               |       |     |      |
| chr7:117356081 | T      | G     | None   | CTCF (GB, NB), RAD21 (NB, ESC)| c     |     |     |
| chr7:117390966 | T      | Del   | None   | CTCF (NB, CB), RAD21 (ESC)   | d     |     |     |
| chr7:117417559 | A      | G     | rs75322384 |                              | b     |     |     |
| chr7:117421141 | C      | A     | None   |                               | b     |     |     |
| chr7:117431202 | C      | A     | None   |                               | b     |     |     |
| chr7:117431704 | C      | T     | None   | RAD21 (ESC)                   | b     |     |     |
| chr7:117431879 | G      | A     | None   |                               | b     |     |     |
| chr7:117450810 | C      | rs34868515 |      | SP1, YY1, EP300, JUND, TCF12, HDAC2, NANOG, BCL11A, TEAD4 (all ESC) | c     |     |     |
| chr7:117456904 | C      | T     | rs12706157 |                              | c     |     |     |
| chr7:117457141 | G      | C     | rs13248222 |                              | b     |     |     |
| chr7:117468056 | C      | T     | rs2067080 | EP300 (NB), FOXP2 (NB), JUND (ESC) | b     |     |     |
| chr7:117468334 | T      | C     | rs2111209 | EP300 (NB), FOXP2 (NB), JUND (ESC) | b     |     |     |
| REEP3           |        |       |        |                               |       |     |      |
| chr10:65307923 | A      | G     | rs78109635 | GATA2 (NB)                    | c     |     |     |
| chr10:65332906 | T      | C     | rs76646063 | GATA3, GATA2, EP300 (all NB) | c     |     |     |
| chr10:65387644 | C      | G     | rs531840 |                               | b     |     |     |
| chr10:65387722 | C      | Del   | None   | SIN3A (NB), POLR2A (NB), REST (NB), USF1 (ESC), EP300 (NB) | b     |     |     |
| chr10:65388750 | G      | A     | None   |                               | b     |     |     |

We identified twelve candidate regulatory SNPs in CTTNB2, including: seven intronic SNPs with DHS signals in neural stem cells (SK-N-MC) or neuroblasts (SK-N-SH, BE2-C, SH-SYSY, SK-N-SH-RA); four of which also overlap TF-binding sites in the brain-derived cell lines; two intronic SNPs near the top DHS variants and potentially altering the same regulatory elements; and three coding SNPs that lie within or near regulatory marks (Supplementary Fig. 7b). We also identified five candidate regulatory SNPs in or near REEP3, including: one intrinsic SNP (chr10:65307923) that alters a DHS active in neural stem cells and GATA2, GATA3, and EP300 binding sites active in neuroblasts; one intrinsic SNP (chr10:65332906) that alters a DHS active in neural stem cells and GATA2, GATA3, and EP300 binding sites active in neuroblasts; three noncoding SNPs (chr10:65387644, chr10:65387722, and chr10:65388750) that cluster ~3 kb upstream in a DHS active in multiple brain-related cells, including neuroblasts, and are seen only in OCD patients in our pooled sequencing (Supplementary Fig. 7c).

1. Coding, EMSA, electrophoretic mobility shift assay
2. No change
3. Strong TF-DNA binding change
4. Weak change; Gb, globloblast; NB, neuroblasts; ESC, embryonic stem cell; CB, cerebellum; “Transcription factor” column shows the TF bindings to the regions and brain/developmental cell types where the signals are found. OR (odds ratio) column reports data in the combined set, unless noted with
5. Indicating data from sequencing

For CTNNBP2 and REEP3, regulatory variants give a far stronger burden signal than does testing for either coding variants or all variants (Fig. 1b). Furthermore, the three largest effect variants in the combined cohort (1298 OCD cases and 1660 controls) alter enhancer elements in these two genes.
Using ENCODE and Roadmap Epigenomics data, we identified 17 candidate regulatory variants in **a** REEP3 and **b** CTTNB2 that disrupt chromatin structure in brain-related cell types\(^{26,29}\). All 17 likely to alter transcription factor-binding sites (TFBS) and/or binding changes are shown in Supplementary Fig. 8. Of the six extracellular LNS (laminin, nectin, sex-hormone-binding globulin) domains in the longer isoform NRXN1a, five with known protein structure are shown and labeled L2-L6. NRXN1a-2 includes four of these domains (L2-L5; dashed ellipse). **d** The isoform-based test comparing OCD cases to ExAC finds NRXN1 as the top-scoring gene with genome-wide significance. The ExAC analysis score is defined as the ratio of \( \chi^2 \) statistics \( \left( \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \right) \), where \( n = \) total number of isoforms, \( O_i = \) number of non-reference alleles observed in isoform \( i \), \( E_i = \) number of non-reference alleles expected from ExAC in isoform \( i \) between OCD vs. ExAC comparison and control vs. ExAC comparison.

We functionally tested 17 candidate regulatory SNPs in **a** CTTNB2 and REEP3, likely to alter transcription factor-binding sites (TFBS) and/or disrupt chromatin structure in brain-related cell types\(^{26,29}\). All 17 alter enhancers or transcription associated loci active in either the substantia nigra (SN), which relays signals from the striatum to the thalamus, and/or the dorsolateral prefrontal cortex (DL-PFC), which sends signals from the cortex to the striatum/thalamus (Fig. 3b, i). Both regions act in the CSTC pathway implicated by neurophysiological and genetic studies in OCD (Fig. 3j)\(^{30}\).

We functionally tested 17 candidate regulatory SNPs in **a** CTTNB2 and **b** REEP3 (Table 2; Supplementary Fig. 7b). We introduced each into a human neuroblastoma cell line (SK-N-BE(2)) and assessed transcription factor binding using electrophoretic mobility shift assays (EMSA). Both DHS SNPs in **a** CTTNB2, three of seven DHS SNPs in **b** REEP3, and one non-DHS variant in **b** CTTNB2 clearly alter specific DNA-protein binding (Fig. 4a, b). We see weak evidence of differential binding for one upstream DHS SNP in **a** REEP3, two DHS SNPs in **b** CTTNB2, and one non-DHS SNP in **b** CTTNB2 (Supplementary Fig. 8).

The high rate of functional validation by EMSA demonstrates that screening using both regulatory and evolutionary information is remarkably effective in identifying strong candidate OCD-risk variants. In total, eight of 12 tested DHS SNPs (67%) show evidence of altered protein binding, despite testing a single cell line at a single time point under standard-binding conditions (Table 2). This includes two SNPs with high ORs in the full genotyping data sets that strongly disrupt specific DNA-protein binding (chr10:65332906 with OR = 3.7; chr7:117417559 with OR = 2.2). Two of five non-DHS SNPs (40%) also show altered binding, illustrating that DHS mark alone is a powerful but imperfect predictor of regulatory function. Both of these SNPs alter highly constrained elements (SiPhy score 8.7), whereas only one of the three non-DHS SNPs is constrained. Although this is a small data set, our results suggest that incorporating both DHS and conservation may identify functional regulatory variants with greater specificity, an observation consistent with previously published research\(^{31}\).
Validation of coding variants using ExAC. In contrast to the regulatory-variant burden found in CTTNBP2 and REEP3, NRXN1 and HTR2A showed significant PolyStrat signals when only coding variants are considered. Of 12 candidate coding SNPs in NRXN1, seven are missense (Table 3). Four of these are SNPs private to OCD cases, and the other three are rare (AF in controls 0.0009–0.0036). All seven change amino acids in laminin G or EGF-like domains important for postsynaptic binding, potentially affecting the involvement of NRXN1 in synapse formation and maintenance (Fig. 4c). Of the three candidate coding SNPs in HTR2A, two (one missense and one synonymous) are in the last coding exon, and one (missense) is the cytoplasmic domain with a PDZ-binding motif, potentially affecting binding affinity or specificity.

We sought to improve our statistical power by combining our pooled-sequencing data with publicly available ExAC data. Using only our data, the associations of CTTNBP2, REEP3, NRXN1, and HTR2A with OCD are experiment-wise significant, but do not reach the genome-wide significance threshold \( p < 2.5 \times 10^{-6} \) (~20,000 human genes), with the strongest association to NRXN1, at \( p = 5.1 \times 10^{-5} \) (cohort 1 and 2; Fisher’s combined \( p \)). For the two genes with a burden of coding variants (NRXN1 and HTR2A), we used ExAC to assess variant burden in OCD cases compared with 33,370 non-Finnish Europeans. Such a comparison was not possible for CTTNBP2 and REEP3, for which associated variants are predominantly noncoding and thus not assayed in ExAC.

To assess the significance of the variant enrichment in each gene, we used an isoform-based test that incorporates a within-gene comparison to assess significance, effectively controlling for inflation due to the extremely large size of the ExAC cohort (Supplementary Methods). Of 542 genes with more than one isoform, we saw no significant difference between our control data and ExAC for over 90% (493 genes had corrected \( p > 0.05 \)). Focusing on the subset of 66 genes with nominally significant PolyStrat scores, NRXN1 had the largest difference between cases and ExAC (\( \chi^2 = 82.3, \text{df} = 16 \), uncorrected \( p = 6.37 \times 10^{-11} \), corrected \( p = 1.27 \times 10^{-6} \)) and no difference between controls and ExAC (\( \chi^2 = 10.5, \text{df} = 16, \text{uncorrected } p = 0.84 \) (Fig. 4d). No previous findings in OCD genetics have reached this level of significance despite >100 candidate gene studies, a dozen linkage studies, and two GWAS. HTR2A, while enriched for coding variants, had only two SNPs in cases, providing insufficient information for the isoform test.

The significant association of NRXN1 reflects an exceptional burden of variants in one of its 17 Ensembl isoforms. NRXN1a-2, which contains all 12 candidate coding SNPs, had the largest deviation between observed and expected variant counts, with a residual at least 1.4× higher than any other isoform (NRXN1a-2 = 22.3, NRXN1-001 = 16.3; median = 5.15). After adjusting for the residuals from the “null” control data and ExAC comparison, the NRXN1a-2 residual is still 1.3× higher (OCD residual/control residual NRXN1a-2 = 5.34, NRXN1-014 = 4.04).

### Discussion

By analyzing sequencing data for 608 OCD candidate genes, then prioritizing variants according to functional and conservation annotations, we identified four genes with a reproducible variant burden in OCD cases. Two genes, NRXN1 and HTR2A (Table 3), have a burden of coding variants, and the other two, CTTNBP2 and REEP3 (Table 2), have a burden of conserved regulatory variants. Notably, all four act in neural pathways linked to OCD, including serotonergic and glutamatergic signaling, synaptic connectivity, and the CSTC circuit, offering new insight into the biological basis of compulsive behavior (Fig. 5).

We used three independent approaches to validate our findings: (1) For the top candidate SNPs, allele-frequency differences from sequencing data were confirmed by genotyping of both the original cohort (Fig. 3d) and a larger, independent cohort (Fig. 3g). (2) For the two genes with a burden of coding variants (NRXN1 and HTR2A), comparison of our data to 33,370 population-matched controls from ExAC revealed genome-
Data 1)2-3. In addition, 60% of our variants have allele frequencies <0.01, and would be missed even through imputation with 1000 Genomes and UK10K27.

Our new analytical method, PolyStrat, analyzes targeted-sequencing data capturing all variants, and leverages public evolutionary and regulatory data to increase power. PolyStrat first filters out variants that are less likely to be functional, then performs gene-burden tests. In contrast to gene-based approaches focusing on ultra-rare, protein-damaging variants, PolyStrat considers variants of diverse frequencies, gaining power to identify genes with excess variants in cases.

PolyStrat is particularly advantageous when applied to studies with smaller cohorts. By testing for association at the gene level, it requires statistical correction only for the ~20,000 genes in the genome. It increases power further by using targeted-sequencing data to capture nearly all variation, including variants with higher allele frequencies and/or larger effect sizes, in regions that are coding or evolutionarily constrained, and enriching for causal variants by removing ~33% of variants unlikely to be functional18. PolyStrat tests ~82 times more functional variants than PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/), which focuses on protein-damaging variants (27,626 vs. 335 in our data).

Our PolyStrat results are consistent with expectations from simulations, which suggest that 200–700 cases should yield 90% power to detect associated genes with allele frequencies and effect sizes similar to our four genes39. Specifically, we would achieve 90% power to detect associations to NRXN1 (combined AF = 0.022; OR = 2.4) with ~600 cases, to HTR2A (combined AF = 0.03; OR = 1.56) with ~700 cases; to REEP3 (combined AF = 0.04; OR = 2.11) with ~200 cases, and to rare (AF <0.01) variants in CTTNBP2 (combined AF = 0.003; OR = 4.7) with ~500 cases.

Previous research on the four genes identified by PolyStrat revealed that all are expressed in the striatum, a brain region linked to OCD (http://human.brain-map.org/). All four genes are involved in pathways relevant to brain function, and harbor variants that could alter OCD risk (Table 4).

NRXN1 encodes the synapse cell-adhesion protein neurexin 1α, a component of cortico-striatal neural pathway40,41 implicated in ASD and other psychiatic diseases42, and functionally related to genes associated with OCD (CDH9/CDH10)8,9 and canine CD8 (CDH2) (Fig. 5). NRXN1 isoforms are implicated in distinct neuropsychiatric disorders. The non-synonymous variants in the NRXN1a-2 isoform (Fig. 4c) may alter synaptic function by disrupting cellular localization or interactions with binding partners, including neurexophilins. The five synonymous candidate variants in likely regulatory elements may affect protein folding by disrupting post-transcriptional regulation, seen in other neuropsychiatric disorders44.

The synaptic plasticity gene REEP3, also implicated in ASD45, encodes a protein that shapes tubular endoplasmic reticulum membranes found in highly polarized cells, including neurons46. The two EMSA-validated REEP3 variants change regulatory elements active in the cortico-striatal neural pathway (Fig. 3h) and bound by multiple TFs (Table 2) including GATA2, which may be required to actuate inhibitory GABAergic neurons47. Thus, variants disrupting GATA2 binding could change the balance between excitatory and inhibitory neurons in the CSTC circuit (Fig. 3j)30.

CTTNBP2 regulates post synaptic excitatory synapse formation. All four EMSA - confirmed variants in CTTNBP2 alter epigenetic marks active in the key structures of the cortico-striatal neural pathway48 (Table 2; Fig. 3h), potentially affecting the expression of this critical gene. CTTNBP2 proteins interact with both proteins encoded by STRN (striatin), which approached experiment-wide significance in this study (uncorrected p = 0.0016, corrected p <0.1; Fig. 1b) and the canine CD gene CDH2 (Fig. 5).

**Fig. 5** All four top candidate genes function at the synapse and interact with proteins implicated in OCD by previous studies in the three species. Human, dog, and mouse are marked with superscripts h, d, and m, respectively. Genes identified in this study are shown in red. Solid lines indicate direct interactions, and dashed lines indicate indirect interactions.
**Table 4 Summary of top genes**

| Gene product/brain relevance                                                                 | NRXN1                                                                 | REEP3                                                                 | CTTNBP2                                                                 | HTR2A                                                                 |
|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------|----------------------------------------------------------------------|-------------------------------------------------------------------------|-------------------------------------------------------------------------|
| Synaptic cell-adhesion protein/synapse functioning and development in the cortical-striatal pathway | Synaptic cell-adhesion protein/synapse functioning and development in the cortical-striatal pathway | Synaptic cell-adhesion protein/synapse functioning and development in the cortical-striatal pathway | Synaptic cell-adhesion protein/synapse functioning and development in the cortical-striatal pathway | Synaptic cell-adhesion protein/synapse functioning and development in the cortical-striatal pathway |
| ASD46                                                                                          | ASD46                                                                 | ASD46                                                             | ASD46                                                                   | ASD46                                                                   |
| Neurodevelopmental disorders incl. ASD42, 64                                                   | Neurodevelopmental disorders incl. ASD42, 64                            | Neurodevelopmental disorders incl. ASD42, 64                       | Neurodevelopmental disorders incl. ASD42, 64                            | Neurodevelopmental disorders incl. ASD42, 64                            |
| (1) Model (mouse) organism gene; (2) ASD gene                                                 | (2) ASD gene                                                           | (1) Model (mouse) organism gene; (2) ASD gene                     | (1) Model (mouse) organism gene; (2) ASD gene                         | (1) Model (mouse) organism gene; (2) ASD gene                         |
| Cytoskeleton organization                                                                     | Cytoskeleton organization                                              | Cytoskeleton organization                                          | Cytoskeleton organization                                              | Cytoskeleton organization                                              |
| By comparison to ExAC—genome-wide-significant association                                     | By comparison to ExAC—genome-wide-significant association              | By comparison to ExAC—genome-wide-significant association         | By comparison to ExAC—genome-wide-significant association              | By comparison to ExAC—genome-wide-significant association              |
| Inaccurate cellular localization of NRXN1, or altered binding competition to its partner, modifying synaptic adhesion | Inaccurate cellular localization of NRXN1, or altered binding competition to its partner, modifying synaptic adhesion | Inaccurate cellular localization of NRXN1, or altered binding competition to its partner, modifying synaptic adhesion | Inaccurate cellular localization of NRXN1, or altered binding competition to its partner, modifying synaptic adhesion | Inaccurate cellular localization of NRXN1, or altered binding competition to its partner, modifying synaptic adhesion |
| By EMSA—disrupt regulatory elements bound by various TF, including GATA2                     | By EMSA—disrupt regulatory elements bound by various TF, including GATA2 | By EMSA—disrupt regulatory elements bound by various TF, including GATA2 | By EMSA—disrupt regulatory elements bound by various TF, including GATA2 | By EMSA—disrupt regulatory elements bound by various TF, including GATA2 |
| REEP3 expression in GABA neurons inhibited by variants that reduce GATA2 binding47, leading to excitatory/inhibitory imbalance in CSTC circuit30 | REEP3 expression in GABA neurons inhibited by variants that reduce GATA2 binding47, leading to excitatory/inhibitory imbalance in CSTC circuit30 | REEP3 expression in GABA neurons inhibited by variants that reduce GATA2 binding47, leading to excitatory/inhibitory imbalance in CSTC circuit30 | REEP3 expression in GABA neurons inhibited by variants that reduce GATA2 binding47, leading to excitatory/inhibitory imbalance in CSTC circuit30 | REEP3 expression in GABA neurons inhibited by variants that reduce GATA2 binding47, leading to excitatory/inhibitory imbalance in CSTC circuit30 |
| By EMSA—alter epigenetic marks active in the cortico-striatal pathway                          | By EMSA—alter epigenetic marks active in the cortico-striatal pathway | By EMSA—alter epigenetic marks active in the cortico-striatal pathway | By EMSA—alter epigenetic marks active in the cortico-striatal pathway | By EMSA—alter epigenetic marks active in the cortico-striatal pathway |
| Modified binding affinity of HTR2A, changing the activation of downstream calcium signaling in neurons12 | Modified binding affinity of HTR2A, changing the activation of downstream calcium signaling in neurons12 | Modified binding affinity of HTR2A, changing the activation of downstream calcium signaling in neurons12 | Modified binding affinity of HTR2A, changing the activation of downstream calcium signaling in neurons12 | Modified binding affinity of HTR2A, changing the activation of downstream calcium signaling in neurons12 |

*For explanation of each category, see “Results” section*

**HTR2A** encodes a G-protein-coupled serotonin receptor expressed throughout the central nervous system, including in the prefrontal cortex, and has been implicated in ASD and OCD45. A related serotonin-receptor cluster (HTR3C/HTR3D/HTR3E) is associated with severe canine CD49, (Fig. 5). The three coding variants found in HTR2A may alter its binding affinity (Table 3)32, and one of the three, a rare missense variant (rs6314; AF = 0.004 in 1000G CEU population) is perfectly linked (D’ = 1; http://raggr.usc.edu) to a common variant (rs6314) associated with response to SSRIs50.

Taken together, our top four associated genes and our pathway analysis implicate three classes of neuronal functions in OCD, as described below.

First, synaptic cell-adhesion molecules help establish and maintain contact between the presynaptic and postsynaptic membrane, and are critical for synaptic development and neural plasticity. NRXN1 encodes a cell-adhesion molecule predominantly expressed in the brain, and CTTNBP2 regulates cortactin, another such molecule, echoing earlier findings linking cell-adhesion genes to compulsive disorders in dogs (CDH9/CDH10, CTNNA2, ATXN1, PGC1, and Sapap3) were significantly more enriched for rare variants in humans than the other 603 genes targeted, although they did not individually achieve significance (Fig. 1c). We propose that the enrichment of rare variants in humans reflects natural selective forces limiting the prevalence of severe disease-causing variants. Such forces are less powerful in selectively bred animal populations. Because risk variants identified through animal models are anticipated to be rare in humans, replication will require either family-based studies, or cohorts of magnitude not currently available.

We also find that the ratio of coding to regulatory variants is positively correlated with a gene’s developmental importance. Although single-gene p-values from PolyStrat tests are positively correlated across variant categories, as is expected given overlaps between different variant categories (Fig. 3a; Supplementary Fig. 9), this pattern breaks down for our four significantly associated genes. NRXN1 and HTR2A, which have burdens of coding variants, score poorly on regulatory-variant tests; CTTNBP2 and REEP3, which have burdens of regulatory variants, score poorly in coding-variant tests (Fig. 1b). This is consistent with the ExAC study showing that genes critical to viability or development do not tolerate major coding changes33. In that study, the authors infer that CTTNBP2 and REEP3 would be intolerant of homozygous loss of function variants (pRec = 0.99999015 and pRec = 0.953842585, respectively), whereas HTR2A (pRec = 0.225555783) and, most notably, NRXN1 (pRec = 0.99999015 and pRec = 0.953842585, respectively).
The remaining steps followed the LightShift Chemiluminescent EMSA Kit protocol (Thermo Scientific).

Statistical analysis. For gene-association/pathway-association, we used the sum of the differences of non-reference allele rates between cases and controls per gene as test statistic, and calculated the probability of observing a test statistic by chance from 10,000 permutations. Multiple testing was empirically corrected using "minP" procedure. See Supplementary Methods for details.

Code availability. The code used in this study was obtained from R package Rplinkseq and PLINK1.9.

Data availability. All data presented in this study are accessible at: https://data.broadinstitute.org/OCD_NatureCommunications2017.

Received: 25 January 2017 Accepted: 1 August 2017 Published online: 17 October 2017

References
1. Pauls, D. L. The genetics of obsessive-compulsive disorder: a review. Dialogues Clin. Neurosci. 12, 149–163 (2010).
2. Stewart, S. E. et al. Genome-wide association study of obsessive-compulsive disorder. Mol. Psychiatry 18, 788–798 (2013).
3. Mattheisen, M. et al. Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. Mol. Psychiatry 20, 337–344 (2014).
4. Welch, J. M. et al. Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice. Nature 448, 894–900 (2007).
5. Ahmari, S. E. et al. Repeated cortico-striatal stimulation generates persistent OCD-like behavior. Science 340, 1234–1239 (2013).
6. Ting, J. T. & Feng, G. Neurobiology of obsessive-compulsive disorder: insights into neural circuitry dysfunction through mouse genetics. Curr. Opin. Neurol. 21, 842–848 (2011).
7. Karlsson, E. K. & Lindblad-Toh, K. Leader of the pack: gene mapping in dogs and other model organisms. Nat. Rev. Genet. 9, 713–723 (2008).
8. Tang, R. et al. Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. Genome Biol. 15, R25 (2014).
9. Dodman, N. H. et al. A canine chromosome 7 locus confers compulsive disorder susceptibility. Mol. Psychiatry 15, 8–10 (2010).
10. Sullivan, P. F., Daly, M. J. & O’Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. 13, 537–551 (2012).
11. Farrell, M. S. et al. Evaluating historical candidate genes for schizophrenia. Mol. Psychiatry 20, 555–562 (2015).
12. Sobrin, L. et al. Candidate gene association study for diabetic retinopathy in persons with type 2 diabetes: the Candidate gene Association Resource (CARE). Invest. Ophthalmol. Vis. Sci. 52, 7593–7602 (2011).
13. Fuchberger, C. et al. The genetic architecture of type 2 diabetes. Nature 536, 41–47 (2016).
14. D’Gama, A. M. et al. Targeted DNA sequencing from autism spectrum disorder brains implicates multiple genetic mechanisms. Neuron 88, 910–917 (2015).
15. Bertram, L. & Tanzi, R. E. Thirty years of Alzheimer’s disease genetics: the implications of systematic meta-analyses. Nat. Rev. Neurosci. 9, 768–778 (2008).
16. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. 354, 1264–1272 (2006).
17. Rivas, M. A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43, 1066–1073 (2011).
18. Gutierrez-Achury, J. et al. Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. Nat. Genet. 47, 577–578 (2015).
19. Roth, E. M., McKenney, J. M., Hanotin, C., Asset, G. & Stein, E. A. Aorvastatin with or without an antibody to PCSK9 in primary hypercholesterolemia. N. Engl. J. Med. 367, 1891–1900 (2012).
20. Warren, A. et al. Exome sequencing: current and future perspectives. G3 5, 1543–1550 (2015).
21. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
22. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482 (2011).
23. 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 Res. 39, e132 (2011).
24. Pruitt, K. D. et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42, D756–D763 (2014).

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00831-x ARTICLE

Methods
Study design. We designed and carried out the study in two phases. In the first, discovery phase, we performed targeted sequencing of 592 individuals with DSM-IV OCD and 560 controls of European ancestry, and tested association for OCD at single variant-level, gene-level, and pathway-level. In the second, validation phase, we employed three distinct analyses. (1) We genotyped both the original cohort and a second, independent cohort containing 1834 DNA samples (729 DSM-IV OCD cases and 1105 controls) of European ancestry, including a total of 2986 individuals (1321 OCD cases and 1665 controls) to confirm the observed allele frequencies in the discovery phase. (2) We compared our sequencing data with 33,370 population-matched controls from ExAC to confirm the gene-based burden of coding variants as well as allele frequencies. (3) We performed EMSA to test whether our candidate variants have regulatory function. Uses of biospecimens in this study were reviewed and approved by the Office of Research Subject Protection, or the Partners HealthCare Human Research Committee. Informed consent was obtained from all subjects included in our study.

Targeted regions. We targeted 82,723 evolutionarily constrained regions in and around 608 genes, which included all regions within 1 kb of the start and end of each of 608 targeted genes with SiPhy evolutionarily constraint score ≥21. For the intergenic regions upstream and downstream of each gene, we found associations driven by both coding and regulatory variants, highlighting new potential therapeutic targets. Our method holds promise for elucidating the biological basis of complex disease, and for extending the power of precision medicine to previously excluded populations.

Pooled sequencing and variant annotation. Groups of 16 individuals were pooled together into 37 case pools and 35 control pools and bar-coded. Targeted-genomic regions were captured using a custom NimbleGen hybrid capture array and sequenced on Illumina GAII or Illumina HiSeq000. Sequencing reads were aligned and processed by Picard analysis pipeline (http://broadinstitute.github.io/picard). Variants and AFs were called using Szygy21 and SNVer23. We used ANNOVAR23 to annotate variants for RefSeq genes (hg19), GERP scores, ENCODE DHS cluster, and 1000 G data.

Genotyping. SNP genotyping was performed using the Sequenom MassARRAY iPLEX platform. The resulting data were analyzed using PLINK1.9 (www.cog-genomics.org/plink2).

EMSA. For each allelic of the tested variants, pairs of 5'-biotinylated oligonucleotides were obtained from IDT Inc. (Supplementary Data 4). Equal volumes of forward and reverse oligonucleotides (1 pmol/µl) were mixed and heated at 95°C for 5 min and then cooled to room temperature. Annealed probes were incubated at room temperature for 30 min with SK-N-BE(2) nuclear extract (Active Motif). The remaining steps followed the LightShift Chemiluminescent EMSA Kit protocol (Thermo Scientific).
25. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e128 (2010).

26. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).

27. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PloS Comput. Biol.* 6, e1001025 (2010).

28. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).

29. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 human reference epigenomes. *Nature* 518, 317–330 (2015).

30. Pauls, D. L., Abramovitch, A., Rauch, S. L. & Geller, D. A. Obsessive-compulsive disorder: an integrative genetic and neurobiological perspective. *Nat. Rev. Neurosci.* 15, 410–424 (2014).

31. Kheradpour, P. et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* 23, 800–811 (2013).

32. Becamel, C. et al. The serotonin 5-HT2A and 5-HT2C receptors interact with specific sets of PDZ proteins. *J. Biol. Chem.* 279, 20257–20266 (2004).

33. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016).

34. Schneider, J. W. Caveats for using statistical significance tests in research assessments. *J. Informetr.* 7, 50–62 (2013).

35. Taylor, S. Multiple mechanisms of obsessive-compulsive disorder: a comprehensive meta-analysis of genetic association studies. *Mol. Psychiatry* 18, 799–805 (2013).

36. Park, J.-H. et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl Acad. Sci. USA* 108, 18026–18031 (2011).

37. Jiang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* 6, 8111 (2015).

38. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 93, 805–814 (2013).

39. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 93, 805–814 (2013).

40. Liu, J. et al. LRTM2 interacts with Neurexin1 and regulates excitatory synapse formation. *Neuron* 64, 799–806 (2009).

41. Surmeier, D. J., Ding, J., Day, M., Wang, Z. & Shen, W. D1 and D2 dopamine receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends Neurosci.* 30, 228–235 (2007).

42. Sudhof, T. C. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 455, 903–911 (2008).

43. Rajasekaran, A. et al. Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet.* 18, 988–996 (2009).

44. Takata, A., Ionita-Laza, I., Gogos, J. A., Xu, B. & Karayiorgou, M. De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia. *Neuron* 92, 940–947 (2016).

45. Castermans, D. et al. Identification and characterization of the TRIP6 and REEP3 genes on chromosome 10q21.3 as novel candidate genes for autism. *Eur. J. Hum. Genet.* 15, 422–431 (2007).

46. Blackstone, C., O’Kane, C. J. & Reid, E. Hereditary spastic paraplegias: molecular mechanisms and clinical implications. *J. Neurol. Sci.* 317, 18–28 (2012).

47. Chen, Y.-K., Chen, C. Y.-H., Lee, E. & Hsu, T.-C. Cortactin-binding protein 2 modulates the mobility of cortactin and regulates dendritic spine formation and maintenance. *J. Neurosci.* 32, 1043–1055 (2012).

48. Lambe, E. K., Fillman, S. G., Webster, M. J. & Shannon Weickert, C. Serotonin receptor expression in human prefrontal cortex: balancing excitation and inhibition across postnatal development. *PLoS ONE* 6, e22799 (2011).

49. Jenkins, A. K. et al. Neurexin 1 (NRXN1) splice isoform expression during midbrain post-mitotic selector gene for midbrain and the motor pathway. *Cell Mol. Life Sci.* 73, 598–616 (2016).

50. Chen, C. Y.-H., Hsu, T.-C., Hsu, T.-C. and the motor pathway. *Cell Mol. Life Sci.* 73, 598–616 (2016).

51. Shmelkov, S. V. et al. Slitrk5 deletion leads to obsessive-compulsive-like behaviors in mice. *Nat. Neurosci.* 19, 779–789 (2011).

52. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 44, D7–D19 (2016).

53. Chen, Y.-K., Y.-K., C. & Y.-P., H. cortactin-binding protein 2 modulates the mobility of cortactin and regulates dendritic spine formation and maintenance. *J. Neurosci.* 32, 1043–1055 (2012).

54. Kala, K. et al. Gata2 is a tissue-specific main intercellular adhesion strengthening in fibroblasts. *J. Cell Sci.* 117, S117–S121 (2004).

55. Chen, Y.-K., Chen, C. Y.-H., Lee, E. & Hsu, T.-C. CTNTBP2, but not CTNTBP2NL, regulates dendritic spineogenesis and synaptic distribution of the striatum-PP2A complex. *Mol. Biol. Cell* 23, 4383–4392 (2012).

Acknowledgements
We thank the participating individuals for their support, Eric S. Lander, Steven E. Hyman, Jessica Alföldi, and Kaatlin Samocha for valuable input; Leslie Gaffney for help with illustrations; Jeremiah M. Scharf for sample collection and discussions; and Broad Genomics Platform for sample processing, sequencing, and genotyping. H.J.N. is supported by the AKC Health Foundation and Swedish Research Council, C.R. by the Swedish Research Council (K2013-61P-22168), K.L.-T. by the Swedish Medical Research Council and European Research Council, and E.K.K. by NIH NIMH (1R21MH109938-01). A Broad Institute SPARC grant supported part of this work.

Author contributions
K.L.-T., E.K.K., G.F., H.J.N., and R.T. conceived and designed the experiments. H.J.N., R.T., E.K.K., J.F., O.D., R.S., D.H., D.P.G., K.L.-T. analyzed the data. H.J.N., D.P.G., E.K.K., and K.L.-T. wrote the paper. R.T. and H.J.N. performed sequence capture. R.S. performed EMSA. M.W., H.-G., S.R., C.A.M., T.E.A., S.A.R., M.A.J., J.A.K., E.G., G.H., D.C.C., E.A., S.W., P.D.P., G.F., C.H., M.T.P., and C.N.P. diagnosed/colllected samples. J.L. and M.K. and G.V.G. coordinated/prepared samples and data generation.

Additional information
Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00831-x.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Hyun Ji Noh\(^1\), Ruqi Tang\(^1,2,3\), Jason Flannick\(^1\), Colm O’Dushlaine\(^1\), Ross Swofford\(^1\), Daniel Howrigan\(^1\), Diane P. Genereux\(^1\), Jeremy Johnson\(^1\), Gerard van Grootheest\(^4\), Edna Grünblatt\(^5,6,7\), Erik Andersson\(^8\),
Diana R. Djurfeldt\(^8,9\), Paresh D. Patel\(^10\), Michele Koltookian\(^1\), Christina M. Hultman\(^11\), Michele T. Pato\(^12\),
Carlos N. Pato\(^12\), Steven A. Rasmussen\(^13\), Michael A. Jenike\(^14\), Gregory L. Hanna\(^10\), S. Evelyn Stewart\(^15\),
James A. Knowles\(^12\), Stephan Ruhrmann\(^16\), Hans-Jörgen Grabe\(^17\), Michael Wagner\(^18,19\), Christian Rück\(^8,9\),
Carol A. Mathews\(^20\), Susanne Walitza\(^5,6,7\), Daniëlle C. Cath\(^21\), Guoping Feng\(^2,22\),
Elinor K. Karlsson\(^2,23\) & Kerstin Lindblad-Toh\(^1,24\)

\(^1\)Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA. \(^2\)Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA. \(^3\)Renji Hospital, School of Medicine, Shanghai Jiao Tong University, 145 Shandong Middle Road, Huangpu Qu, Shanghai 200001, China. \(^4\)GGZ inGeest and Department of Psychiatry, VU University Medical Center, De Boelelaan 1117, 1081 HV, Amsterdam, The Netherlands. \(^5\)Department of Child & Adolescent Psychiatry and Psychotherapy, University Hospital of Psychiatry Zurich, University of Zurich, Neumünsterallee 9, Zurich 8032, Switzerland.
\(^6\)Neuroscience Center Zurich, University of Zurich & ETH Zurich, Winterthurer Strasse 190, Zurich 8057, Switzerland. \(^7\)Zurich Center for Integrative Human Physiology, University of Zurich, Winterthurer Strasse 190, Zurich 8057, Switzerland. \(^8\)Department of Clinical Neuroscience, Centre for Psychiatry Research, Karolinska Institutet, Stockholm 17177, Sweden. \(^9\)Department of Clinical Neuroscience, Centre for Psychiatry Research, Karolinska Institutet, Stockholm 17177, Sweden. \(^10\)Department of Psychiatry, University of Michigan, 4250 Plymouth Road, Ann Arbor, MI 48109, USA.
\(^11\)Department of Medical Epidemiology & Biostatistics, Karolinska Institutet, Stockholm 17177, Sweden. \(^12\)Department of Psychiatry & Behavioral Sciences, USC, 2250 Alcazar Street, Los Angeles, CA 90033, USA. \(^13\)Department of Psychiatry & Human Behavior, Brown Medical School, 345 Blackstone Boulevard, Box G-BH, Providence, RI 02906, USA. \(^14\)Department of Psychiatry, Harvard Medical School, 401 Park Drive, Boston, MA 02215, USA. \(^15\)BC Mental Health & Addictions Research Institute, UBC, 2255 Wesbrook Mall, Vancouver, BC, Canada V6T 2A1. \(^16\)Department of Psychiatry & Psychotherapy, University of Cologne, Kerpener Street 62, Cologne 50937, Germany. \(^17\)Department of Psychiatry & Psychotherapy, University of Medicine Greifswald, Fleischmannstrasse 8, Greifswald 17475, Germany. \(^18\)Department of Psychiatry & Psychotherapy, University of Bonn, Regina-Pacis-Weg 3, Bonn 53113, Germany. \(^19\)German Center for Neurodegenerative Diseases, Sigmund-Freud-Strasse 27, Bonn 53127, Germany. \(^20\)Department of Psychiatry & Genetics Institute, University of Florida, 1149 Newell Drive, Gainesville, FL 32610, USA. \(^21\)Department of Clinical & Health Psychology, Utrecht University, Heidelberglaan 1, Utrecht, CS 3584, The Netherlands. \(^22\)Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA. \(^23\)Department of Bioinformatics & Integrative Biology and Program in Molecular Medicine, University of Massachusetts Medical School, 368 Plantation Street, Sherman Center, Worcester, MA 01605, USA. \(^24\)Science for Life Laboratory, IMBIM, Uppsala University, Uppsala 75236, Sweden. Elinor K. Karlsson and Kerstin Lindblad-Toh contributed equally to this work.