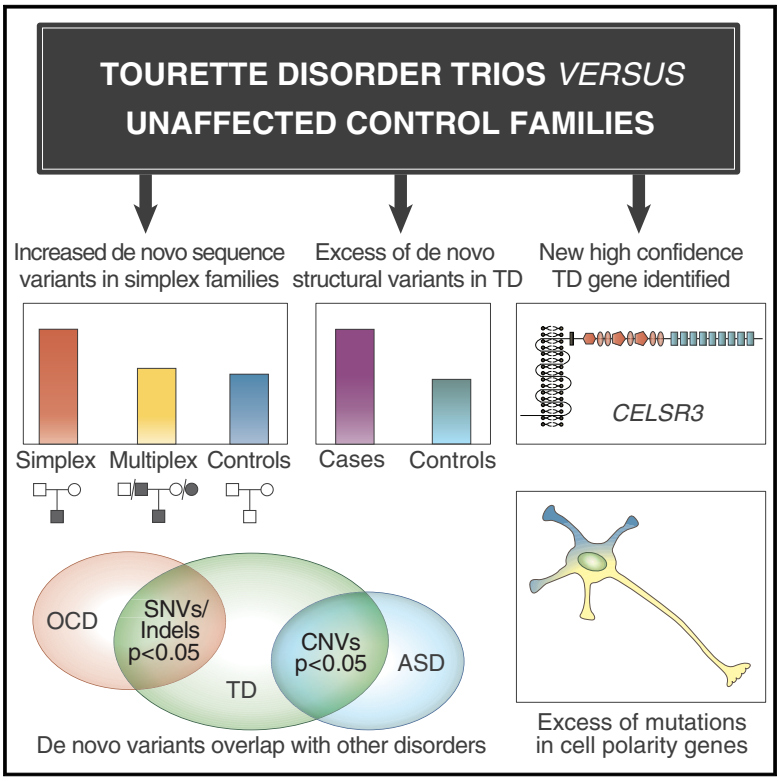
Article

Cell Reports

# *De Novo* Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis

## Graphical Abstract




## Highlights

- Recurrent *de novo* variants identify a new high-confidence TD risk gene: *CELSR3*
- Genes involved in cell polarity are more likely to be disrupted by *de novo* variants
- *De novo* sequence variants may carry more risk in simplex families, female probands
- *De novo* CNVs occur 2 to 3 times more often in TD probands than in matched controls

## Authors

Sheng Wang, Jeffrey D. Mandell, Yogesh Kumar, ..., Peristera Paschou, A. Jeremy Willsey, Matthew W. State

## Correspondence

ppaschou@purdue.edu (P.P.), jeremy.willsey@ucsf.edu (A.J.W.), matthew.state@ucsf.edu (M.W.S.)

## In Brief

Wang et al. expand their earlier exome-sequencing work in TD, adding 291 trios and conducting combined analyses suggesting *de novo* variants carry more risk in individuals with unaffected parents, establishing *de novo* structural variants as risk factors, identifying *CELSR3* as a risk gene, and implicating cell polarity in pathogenesis.

Wang et al., 2018, Cell Reports *24*, 3441–3454
September 25, 2018 
https://doi.org/10.1016/j.celrep.2018.08.082

CellPress

Cell Reports
Article



# *De Novo* Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis

Sheng Wang,[1,2,3,4] Jeffrey D. Mandell,[3,4] Yogesh Kumar,[5] Nawei Sun,[3,4] Montana T. Morris,[3,4] Juan Arbelaez,[3,4] Cara Nasello,[6] Shan Dong,[3] Clif Duhn,[3,4] Xin Zhao,[3,4,7] Zhiyu Yang,[5] Shanmukha S. Padmanabhuni,[5] Dongmei Yu,[8,9] Robert A. King,[10] Andrea Dietrich,[11] Najah Khalifa,[12,13] Niklas Dahl,[14] Alden Y. Huang,[15,16] Benjamin M. Neale,[8,9] Giovanni Coppola,[15,16] Carol A. Mathews,[17] Jeremiah M. Scharf,[8,9] Tourette International Collaborative Genetics Study (TIC Genetics), Tourette Syndrome Genetics Southern and Eastern Europe Initiative (TSGENESEE), Tourette Association of America International Consortium for Genetics (TAAICG), Thomas V. Fernandez,[10] Joseph D. Buxbaum,[18] Silvia De Rubeis,[18] Dorothy E. Grice,[18] Jinchuan Xing,[6] Gary A. Heiman,[6,20] Jay A. Tischfield,[6,20] Peristera Paschou,[5,20,*] A. Jeremy Willsey,[3,4,19,20,21,*] and Matthew W. State[3,19,20,*]

[1]College of Biological Sciences, China Agricultural University, Beijing, China
[2]National Institute of Biological Sciences, Beijing, China
[3]Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA
[4]Institute for Neurodegenerative Diseases, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA
[5]Department of Biological Sciences, Purdue University, West Lafayette, IN, USA
[6]Department of Genetics and the Human Genetics Institute of New Jersey, Rutgers, the State University of New Jersey, Piscataway, NJ, USA
[7]Department of Traditional Chinese Medicine, Xinhua Hospital Affiliated to Shanghai Jiatong University School of Medicine, Shanghai, China
[8]Center for Genomic Medicine, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[9]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[10]Yale Child Study Center and Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA
[11]Department of Child and Adolescent Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[12]Department of Neuroscience, Child and Adolescent Psychiatry Uppsala University, Uppsala, Sweden
[13]Centre for Research and Development, Region Gävleborg, Gävle, Sweden
[14]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
[15]Department of Neurology, University of California, Los Angeles, Los Angeles, CA, USA
[16]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, USA
[17]Department of Psychiatry, Genetics Institute, University of Florida, Gainesville, FL, USA
[18]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[19]Quantitative Biosciences Institute (QBI), University of California, San Francisco, San Francisco, CA, USA
[20]These authors contributed equally
[21]Lead Contact
*Correspondence: ppaschou@purdue.edu (P.P.), jeremy.willsey@ucsf.edu (A.J.W.), matthew.state@ucsf.edu (M.W.S.)
https://doi.org/10.1016/j.celrep.2018.08.082

## SUMMARY

We previously established the contribution of *de novo* damaging sequence variants to Tourette disorder (TD) through whole-exome sequencing of 511 trios. Here, we sequence an additional 291 TD trios and analyze the combined set of 802 trios. We observe an overrepresentation of *de novo* damaging variants in simplex, but not multiplex, families; we identify a high-confidence TD risk gene, *CELSR3* (*cadherin EGF LAG seven-pass G-type receptor 3*); we find that the genes mutated in TD patients are enriched for those related to cell polarity, suggesting a common pathway underlying pathobiology; and we confirm a statistically significant excess of *de novo* copy number variants in TD. Finally, we identify significant overlap of *de novo* sequence variants between TD and obsessive-compulsive disorder and *de novo* copy number variants between TD and autism spectrum disorder, consistent with shared genetic risk.

## INTRODUCTION

Tourette disorder (TD), an early onset neurodevelopmental disorder characterized by chronic motor and vocal tics, has a worldwide prevalence of approximately 0.3%–1% (CDC, 2009; Robertson, 2008; Scharf et al., 2015) and a pronounced sex bias with males much more likely to be affected (Freeman et al., 2000; Scharf et al., 2013). TD is highly comorbid with other psychiatric disorders, such as obsessive-compulsive disorder (OCD) and attention-deficit and hyperactivity disorder (ADHD) (Ghanizadeh and Mosallaei, 2009). Behavioral interventions have comparable effectiveness to medication for tic disorders, though both, unfortunately, have limited efficacy. Moreover, the most effective medications to suppress unwanted movements and vocalizations may lead

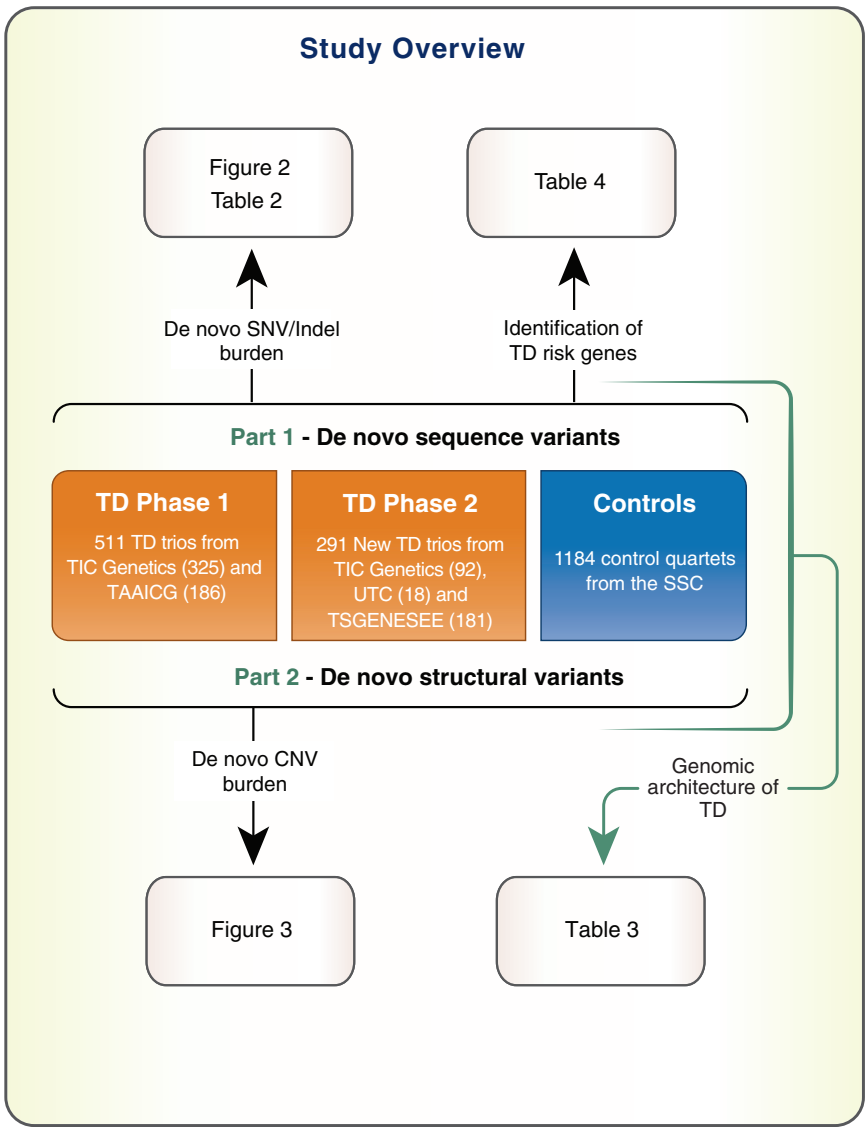



**Figure 1. Study Overview**

Our group previously generated and analyzed WES data from 511 TD trios, generated by the TIC Genetics (325 trios) and TAAICG (186 trios) consortia (Willsey et al., 2017). In this study, we expand the number of trios with WES data by 291 (92 from TIC Genetics, 18 from UTC, and 181 from TSGENESEE). We leverage recurrent *de novo* variants occurring within the same gene in unrelated individuals to identify a high-confidence gene, *CELSR3*. Next, we identify *de novo* CNVs from the WES data and significantly associate these variants with TD. Third, we replicate the association of *de novo* CNVs by analysis of microarray data from 399 partially overlapping TIC Genetics trios. Finally, based on the rate of *de novo* variants, we assess the genomic architecture of TD. CNVs, copy number variants; SSC, Simons Simplex Collection; TAAICG, Tourette Association of America International Consortium for Genetics; TD, Tourette disorder; TIC Genetics, Tourette International Collaborative Genetics consortium; TSGENESEE, Tourette Syndrome Genetics Southern and Eastern Europe Initiative; UTC, Uppsala Tourette Cohort.
See Figure S1 for an overview of quality control and sample filtering and Table S1 for sample metrics.

to long-term side effects, including chronic movement disorders (Quezada and Coffman, 2018). Development of a broader and more effective therapeutic armamentarium is currently profoundly limited by a lack of understanding of pathophysiology. However, given the significant role of genetic factors in TD (Huang et al., 2017; Pauls et al., 1981; Price et al., 1985; Willsey et al., 2017), the elucidation of genes and loci carrying large TD risks represents a promising path forward for clarifying the underlying biology. Indeed, in the past five years, advances in genomics technology, including microarray genotyping and whole-exome sequencing (WES), have resulted in an explosion of genetic data for neurodevelopmental disorders, including autism spectrum disorder (ASD), intellectual disability, epileptic encephalopathies, OCD, ADHD, and schizophrenia. With regard to early onset disorders in particular, it has become clear that the identification of recurrent *de novo* variants is a highly reliable and productive path forward for gene discovery, in the context of a demonstrated excess of these variants in cases versus controls (or versus expectation; Willsey et al., 2018).

Recently, our group reported the association of *de novo* damaging sequence variants (single-nucleotide variants [SNVs] and insertion or deletion variants [indels]) with TD risk (Willsey et al., 2017). We identified four TD risk genes, including one high-confidence TD (hcTD) risk gene (false discovery rate [FDR] < 0.1) and three probable TD (pTD) risk genes (FDR < 0.3). We also demonstrated that, similar to other early-onset neurodevelopmental disorders, the identification of recurrent *de novo* variants is a powerful strategy for gene discovery in TD. Our group and others have also demonstrated that rare copy number variants (CNVs) are associated with TD risk (Fernandez et al., 2012; Huang et al., 2017; McGrath et al., 2014; Nag et al., 2013; Sundaram et al., 2010). However, although suggestive evidence existed (Fernandez et al., 2012), *de novo* CNVs had not yet been firmly established as a risk factor.

In this study (Figure 1), we expand our earlier (phase 1) WES study by 291 additional trios (873 samples), increasing the total number of TD trios to 802 (2,406 samples). In the combined dataset, we identify a new high-confidence TD risk gene, *CELSR3*, as well as two probable risk genes (*OPA1* and *FBN2*). Analyses of the genes with *de novo* damaging variants implicate cell polarity in the pathogenesis of TD. We also conduct pilot analyses that suggest the yield of *de novo* sequence variants is increased in "apparently" simplex (neither of the parents had any reported history of a tic disorder) versus multiplex (at least one of the parents had a reported history of a tic disorder) TD families and in

**Table 1. Demographics and Sequencing Metrics by Cohort**

| Phase | Phase 1 | | | Phase 2 | | | Phases 1 and 2 |
|---|---|---|---|---|---|---|---|
| Cohort | TICGen | TAAICG (Broad) | TAAICG (UCLA) | TICGen | UTC | TSGENESEE | SSC Siblings |
| Samples (trios) sequenced | 325 | 149 | 37 | 92 | 18 | 181 | 1,184 |
| Samples (trios) passing QC for *de novo* sequence variant calling | 311 | 145 | 37 | 92 | 18 | 174 | 1,153 |
| Male:female (sex ratio) | 245:66 (3.71) | 116:29 (4.00) | 34:3 (11.33) | 73:19 (3.84) | 14:4 (3.50) | 144:30 (4.80) | 528:625 (0.84) |
| Paternal age[a] | 33.05 ± 0.63 | 33.35 ± 0.85 | 31.85 ± 1.90 | 33.98 ± 1.15 | NA | NA | 32.6 ± 0.33 |
| Maternal age[a] | 31.08 ± 0.57 | 31.64 ± 0.82 | 30.40 ± 1.46 | 31.16 ± 0.93 | NA | NA | 30.55 ± 0.29 |
| Simplex:multiplex[a,b] | 264:30 | 128:13 | 35:0 | 72:0 | 17:1 | 61:59 | NA (all simplex) |
| Comorbid:non-comorbid[a,c] | 216:86 | 101:39 | 26:10 | 64:22 | 0:18 | 84:64 | NA (all non-comorbid) |
| Exome array | Nimblegen EZ v2 | Agilent v1.1 | Nimblegen EZ v3 | IDT xGen | | | Nimblegen EZ v2 |
| Size of capture region (bp) | 44,001,748 | 32,760,120 | 63,564,965 | 33,337,769 | | | 44,001,748 |
| RefSeq hg19 coding region covered (bp) | 32,586,393 | 31,844,591 | 33,644,238 | 33,357,319 | | | 32,586,393 |
| RefSeq hg19 coding region covered (%) | 96.33 | 94.13 | 99.45 | 98.61 | | | 96.33 |
| Consensus region (bp)[d] | 19,343,430 | | | | | | |
| Coding region covered in consensus (%) | 59.36 | 60.74 | 57.49 | 57.99 | | | 59.36 |
| Mean consensus callable size (million bp)[e] | 18.97 ± 0.041 | 18.97 ± 0.059 | 18.32 ± 0.59 | 18.25 ± 0.20 | 18.87 ± 0.11 | 18.50 ± 0.0095 | 18.10 ± 0.064 |

Cohort characteristics as well as sequencing metrics are summarized per cohort and by phase. 95% confidence intervals are displayed as ±, where relevant. Agilent v1.1, Agilent SureSelect v1.1; IDT xGen, IDT xGen Exome Research Panel; Nimblegen EZ v2, Nimblegen EZ Exome v2; Nimblegen EZ v3, Nimblegen EZ Exome v3.

[a]Not all samples have data; we based calculations on those having records (e.g., we did not have parental age records for UTC and TSGENESEE cohorts).

[b]Simplex: parents unaffected with TD; multiplex: one or more parents have TD.

[c]Comorbid: probands comorbid with ADHD/OCD; non-comorbid: probands not comorbid with ADHD/OCD.

[d]We first calculated cumulative depth of coverages for each trio. For each cohort, we then generated a list of regions in which more than 50% of trios from that cohort have ≥20× joint coverage (i.e., each member of the trio has ≥20× depth at that position). We intersected these regions from each cohort to generate a list of consensus regions. To reduce any potential biases arising from differences in coverage, *de novo* burden analyses were restricted to these high-quality regions.

[e]We estimated the cumulative depth of coverage for each trio in the consensus regions and calculated the mean and 95% CI using one-sample t test in R. See STAR Methods for details.

female versus male probands. Additionally, we identify *de novo* CNVs in WES and complementary microarray data, and conclusively associate *de novo* CNVs with TD risk. We also revise our estimates on the contribution of *de novo* sequence and structural variants to TD risk: 9.7% of cases from TD simplex families carry a *de novo* damaging sequence variant and 1.5% carry a *de novo* structural variant likely mediating risk. Overall, this suggests that, in simplex families, approximately 10% of individuals meeting clinical diagnostic criteria for TD will carry a contributing *de novo* variant. Finally, we estimate that 483 genes contribute risk through disruption by *de novo* sequence variation.

## RESULTS

### *De Novo* Sequence Variants

To follow up our phase 1 study (Willsey et al., 2017), we conducted WES on 291 new "phase 2" TD trios (802 total trios across phase 1 and 2; Figure 1). We also analyzed 582 new phase 2 control trios from the Simons Simplex Collection (SSC) (1,184 total control trios across phase 1 and 2). After quality control, we trimmed to 777 TD trios and 1,153 SSC trios for *de novo* sequence variant calling (STAR Methods; Tables 1 and S1; Figure S1).

We leveraged GATK to conduct alignment, quality control, and variant calling (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). We conducted joint genotyping across the entire set of phase 1 and phase 2 TD trios, as well as the entire set of control trios, in order to reduce batch effects. We further modified our previous *de novo* calling pipeline (Willsey et al., 2017) to utilize the GATK genotype refinement workflow (STAR Methods; Table S2). We defined likely gene disrupting (LGD) variants as insertion of a premature stop codon, disruption of a canonical splice site, or a frameshift insertion or deletion, and probably damaging missense 3 (Mis 3) variants include

missense variants with a PolyPhen2 (HDIV) score $\geq$ 0.957 (Adzhubei et al., 2010, 2013). We refer to the set of LGD and Mis3 variants as "damaging".

We detected 309 *de novo* coding variants from phase 2 samples (1.09 variants per sample). Applying the new pipeline to the phase 1 samples, we detected a total of 466 *de novo* coding variants (0.94 variants per sample). The number of *de novo* variants per individual followed a Poisson distribution (Figure S2), and our new pipeline achieved a 95.9% validation rate across phase 1 and 2 TD samples. See STAR Methods for more details. We did not validate the *de novo* variants in control samples, and therefore, we conducted all burden analyses using all *de novo* variants identified in TD and control trios. However, for gene discovery, we considered validated *de novo* variants only. WES coverage varied across cohorts and phases because of the different capture arrays and sequencing protocols used (Table 1) and was positively correlated with the number of *de novo* variants observed per individual (STAR Methods). To account for these differences, we compared mutation rates, instead of the number of *de novo* variants observed per individual, to normalize for the number of bases with sufficient joint coverage for *de novo* calling (Willsey et al., 2017). To further reduce biases, we estimated mutation rates within a high-confidence region with high joint coverage across all cohorts (consensus region; Table 1; STAR Methods). We then compared the rate between TD probands and SSC siblings with a one-sided rate ratio test, as previously described (Willsey et al., 2017). We also confirmed that the overall rate of coding *de novo* sequence variants does not differ between phase 1 and phase 2 TD trios (rate ratio [RR] 1.03; p = 0.81; two-sided rate ratio test). See Table 2 for *de novo* rates by variant type and Table S3 for a detailed summary of all *de novo* variants called.

## *De Novo* Sequence Variants Contribute Strong Risk to Simplex TD

Our combined dataset consists of apparently simplex trios (the proband is the only individual with confirmed TD; 577 trios), multiplex trios (the proband and one or more parents have TD; 103 trios), and trios with insufficient phenotype data to make a determination (unknown; 97 trios). We did not consider affected status of other relatives, as this information was not consistently available across families. We first assessed whether *de novo* mutation rates vary by simplex versus multiplex trios. We observed a significant increase in simplex, but not multiplex, TD trios, particularly for LGD variants (simplex: RR 1.93, p = 0.0028; multiplex: RR 1.11, p = 0.50; Figure 2A; Table 2). Narrowing to mutation-intolerant genes (Kosmicki et al., 2017; Lek et al., 2016) further strengthens the statistical findings and increases the effect size in simplex families (e.g., for LGD variants; RR 3.61; p = 0.0023; Figure 2B; Table 2). For multiplex families, the effect size of LGD variants also increases, but the result remains non-significant (multiplex: RR 1.36; p = 0.55). Directly comparing the rate of *de novo* variants in simplex versus multiplex TD trios reveals significant differences for nonsynonymous variants in mutation-intolerant genes overall (RR 3.91; p = 0.023), as well as for missense variants in mutation-intolerant genes alone (RR 5.15; p = 0.047) and potentially for LGD variants too (RR 2.66; p = 0.28; Figure 2B). Together, these results suggest that *de novo* variants likely carry risk in multiplex TD but of lesser effect, although this remains to be confirmed with larger sample sizes. The *de novo* rate in unknown trios is similar to simplex trios, suggesting the unknown trios are largely composed of true simplex trios (Table 2; Figure S5). Therefore, although we excluded multiplex trios from *de novo* burden analyses, estimation of the total number of TD risk genes, and gene discovery, we included unknown trios in the estimation of the total number of TD risk genes and in gene discovery.

## Female Probands May Have More *De Novo* Sequence Variants

Given the strong male:female sex bias in TD, we next assessed whether sex of the proband influences *de novo* mutation rate in 577 simplex TD trios. We did not conduct analogous analyses in multiplex or unknown trios because of the small sample sizes available in this study. We first compared the rate of *de novo* variants in sex-matched TD probands and SSC controls. We observed an elevation in the rate of *de novo* LGD variants in female TD probands (RR 2.39; p = 0.018; female TD probands versus female SSC controls) as well as in male TD probands (RR 2.06; p = 0.015; male TD probands versus male SSC controls). A direct comparison of female and male TD probands does not reveal a statistically significant difference, though the result shows a trend toward enrichment in female probands (RR 1.57; p = 0.14; Figure S4A). Further narrowing to variants within mutation-intolerant genes increases the observed effect sizes (e.g., *de novo* LGD: female TD probands, RR 5.21, p = 0.027; male TD probands, RR 3.04, p = 0.04; Figure S4B). Again, however, a direct comparison of female versus male TD probands does not result in a statistically significant difference (e.g., *de novo* LGD: RR 1.45; p = 0.35; female versus male TD probands). We did not observe any difference between male and female SSC controls when comparing the overall rate of *de novo* coding variants (Figure S4A).

## *De Novo* Structural Variants

We detected *de novo* CNVs from the WES data from phase 1 and phase 2 TD samples with CoNIFER (Krumm et al., 2012; STAR Methods). This resulted in the identification of 27 *de novo* CNVs in the 789 TD trios passing CNV-specific quality control (0.034 per proband; 95% confidence interval [CI] 0.021–0.047; Figure S1; Table S5). In addition, we analyzed 1,136 SSC control quartets (mother, father, proband, and unaffected sibling). This provided the opportunity to compare the *de novo* CNV rate in TD probands versus SSC siblings as a negative control, as well as in SSC probands versus SSC siblings as a positive control. This also facilitated a comparison of the *de novo* CNV burden in ASD versus TD. Using identical methods, a total of 37 *de novo* CNVs were identified in 1,136 SSC probands (0.033 per proband; 95% CI 0.022–0.043) and 19 in the 1,136 SSC siblings (0.017 per sibling; 95% CI 0.0081–0.025). See Table S5 for details. We attempted qPCR-based confirmation of all *de novo* CNVs identified in TD probands (88.2% confirmation rate; STAR Methods; Table S3). We did not directly confirm *de novo* CNVs in the SSC quartets, but based on confirmations previously performed on a subset of these variants as reported in

**Table 2. *De Novo* Sequence Mutation Rates by Category**

| | Mutation Rate per Base Pair in RefSeq Coding Regions (× $10^{-8}$; ±95% CI) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | TD (n = 777) | | | | | | | | Controls (n = 1,153) | | |
| Cohort | Simplex (n = 577) | Multiplex (n = 103) | Unknown (n = 97) | Combined (n = 777) | Simplex Male (n = 461) | Simplex Female (n = 116) | Simplex Comorbid (n = 384) | Simplex Non-comorbid (n = 179) | Male (n = 528) | Female (n = 625) | Combined (n = 1,153) |
| Coding | 1.68 ± 0.17 | 1.58 ± 0.38 | 1.71 ± 0.44 | 1.67 ± 0.15 | 1.67 ± 0.20 | 1.69 ± 0.35 | 1.65 ± 0.21 | 1.72 ± 0.30 | 1.50 ± 0.17 | 1.55 ± 0.16 | 1.53 ± 0.12 |
| Syn | 0.42 ± 0.087 | 0.44 ± 0.23 | 0.41 ± 0.20 | 0.42 ± 0.075 | 0.42 ± 0.099 | 0.43 ± 0.18 | 0.43 ± 0.11 | 0.40 ± 0.16 | 0.39 ± 0.085 | 0.41 ± 0.085 | 0.40 ± 0.060 |
| Nonsyn | 1.25 ± 0.15 | 1.14 ± 0.35 | 1.30 ± 0.40 | 1.24 ± 0.13 | 1.25 ± 0.18 | 1.25 ± 0.32 | 1.22 ± 0.19 | 1.32 ± 0.27 | 1.10 ± 0.16 | 1.12 ± 0.14 | 1.11 ± 0.10 |
| Mis | 1.07 ± 0.14 | 1.03 ± 0.34 | 1.13 ± 0.38 | 1.08 ± 0.12 | 1.09 ± 0.16 | 1.01 ± 0.30 | 1.04 ± 0.18 | 1.13 ± 0.25 | 1.02 ± 0.15 | 1.02 ± 0.13 | 1.02 ± 0.098 |
| Mis3 | 0.60 ± 0.10 | 0.60 ± 0.23 | 0.69 ± 0.27 | 0.61 ± 0.090 | 0.62 ± 0.12 | 0.53 ± 0.20 | 0.62 ± 0.14 | 0.53 ± 0.17 | 0.53 ± 0.11 | 0.49 ± 0.093 | 0.51 ± 0.071 |
| LGD | 0.18 ± 0.057 | 0.10 ± 0.10 | 0.17 ± 0.13 | 0.17 ± 0.047 | 0.16 ± 0.062 | 0.25 ± 0.14 | 0.18 ± 0.072 | 0.19 ± 0.10 | 0.079 ± 0.039 | 0.11 ± 0.042 | 0.093 ± 0.029 |
| LGD + Mis3 | 0.78 ± 0.12 | 0.70 ± 0.24 | 0.86 ± 0.31 | 0.78 ± 0.10 | 0.78 ± 0.14 | 0.78 ± 0.22 | 0.80 ± 0.15 | 0.72 ± 0.19 | 0.61 ± 0.12 | 0.60 ± 0.11 | 0.60 ± 0.078 |
| LGD SNV | 0.073 ± 0.038 | 0.10 ± 0.10 | 0.082 ± 0.093 | 0.078 ± 0.033 | 0.057 ± 0.039 | 0.14 ± 0.11 | 0.082 ± 0.050 | 0.059 ± 0.058 | 0.027 ± 0.024 | 0.065 ± 0.033 | 0.048 ± 0.021 |
| LGD FS | 0.11 ± 0.042 | - | 0.085 ± 0.097 | 0.089 ± 0.034 | 0.10 ± 0.047 | 0.11 ± 0.098 | 0.096 ± 0.050 | 0.13 ± 0.085 | 0.051 ± 0.032 | 0.040 ± 0.026 | 0.045 ± 0.020 |
| In frame | 0.0045 ± 0.0088 | 0.026 ± 0.051 | - | 0.0067 ± 0.0093 | 0.0056 ± 0.011 | - | 0.0067 ± 0.013 | - | 0.020 ± 0.019 | 0.0042 ± 0.0082 | 0.011 ± 0.010 |
| Intolerant Mis | 0.13 ± 0.047 | 0.026 ± 0.051 | 0.19 ± 0.14 | 0.13 ± 0.040 | 0.12 ± 0.050 | 0.18 ± 0.12 | 0.12 ± 0.055 | 0.15 ± 0.090 | 0.077 ± 0.039 | 0.049 ± 0.029 | 0.062 ± 0.024 |
| Intolerant LGD | 0.069 ± 0.039 | 0.026 ± 0.051 | 0.055 ± 0.077 | 0.062 ± 0.031 | 0.064 ± 0.044 | 0.091 ± 0.089 | 0.070 ± 0.051 | 0.074 ± 0.065 | 0.020 ± 0.020 | 0.018 ± 0.017 | 0.019 ± 0.013 |
| Intolerant Nonsyn | 0.20 ± 0.063 | 0.051 ± 0.072 | 0.25 ± 0.16 | 0.19 ± 0.052 | 0.18 ± 0.068 | 0.27 ± 0.16 | 0.19 ± 0.076 | 0.22 ± 0.012 | 0.097 ± 0.043 | 0.067 ± 0.034 | 0.081 ± 0.027 |

We excluded any *de novo* variants located outside of the consensus regions and then calculated the mutation rate per base pair and 95% CI using t test in R. See also Figures S4 and S5. Comorbid, probands with TD and ADHD/OCD; damaging, LGD + Mis3; in frame, indel causing in-frame deletion or insertion (loss or gain of amino acids); intolerant LGD, *de novo* LGD variants occurring in genes with pLI greater than 0.9; intolerant Mis, *de novo* missense variants occurring in genes with missense Z score greater than 3.891; intolerant Nonsyn, intolerant Mis + intolerant LGD; LGD, likely gene disrupting (insertion of premature stop codon, disruption of canonical splice site, and insertion-deletion frameshift); LGD FS, insertion-deletion variant causing frameshift; LGD SNV, point mutation causing insertion of premature stop codon and disruption of canonical splice site; Mis, missense; Mis3, missense 3 (PolyPhen2 [HDIV] score $\geq$ 0.957; Adzhubei et al., 2010, 2013); multiplex, one or more parents have TD; non-comorbid, probands with TD only (without ADHD/OCD); Nonsyn, nonsynonymous; simplex, parents unaffected with TD; Syn, synonymous; unknown, phenotypic data unavailable for parents.

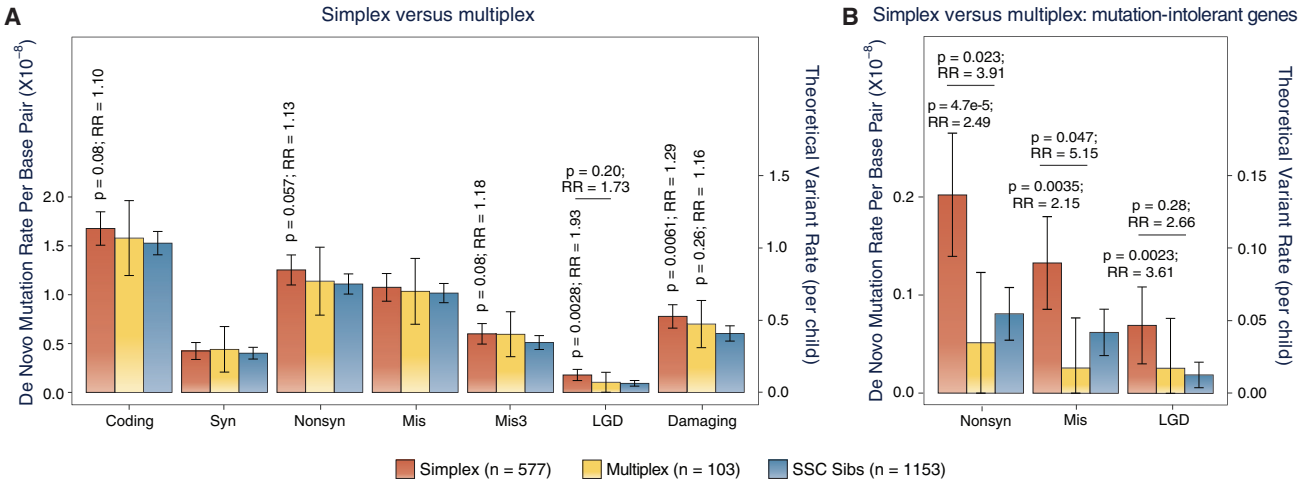



**Figure 2. Combined Burden Analysis Identifies Differences in *De Novo* Rate in Simplex versus Multiplex Families**

We defined a consensus region, consisting of a set of intervals with high-quality coverage across all samples. We then estimated the *de novo* mutation rates per base pair in this consensus region (STAR Methods). We converted the mutation rate per base pair to an expected rate per child (proband or control) by multiplying the mutation rate per base pair by the size of the total RefSeq hg19 "coding" region (33,828,798 bp).

(A) *De novo* variants are overrepresented in simplex TD trios only. LGD variants are significantly increased in simplex TD probands compared to SSC controls (RR 1.93; p = 0.0028; one-sided rate ratio test). Mis3 variants also trend toward enrichment (RR 1.18; p = 0.08). Therefore, *de novo* damaging variants as a group are overrepresented in simplex TD (RR 1.29; p = 0.0061). In contrast, *de novo* variants in any category are not significantly increased in multiplex TD families, though *de novo* damaging variants trend in that direction (RR 1.16; p = 0.26). Additionally, the rate of *de novo* LGD variants may be higher in simplex versus multiplex trios though the difference does not reach statistical significance (RR 1.73; p = 0.20).

(B) Restricting the analysis to *de novo* variants in mutation-intolerant genes (missense *Z* score $\geq$ 3.891 or pLI $\geq$ 0.9; Lek et al., 2016) reveals much larger effect sizes, particularly in simplex families. Comparing simplex to multiplex trios reveals significant differences for *de novo* nonsynonymous variants (RR 3.91; p = 0.023) and for *de novo* missense variants (RR 5.15; p = 0.047), but not for *de novo* LGD variants only (RR 2.66; p = 0.28; STAR Methods).

Damaging, LGD + Mis3; LGD, likely gene disrupting (insertion of premature stop codon, disruption of canonical splice site, and frameshift insertion-deletion variant); Mis, missense; Mis3, probably damaging missense variants (PolyPhen2 [HDIV] score $\geq$ 0.957; Adzhubei et al., 2010, 2013); Nonsyn, nonsynonymous; RR, rate ratio; Syn, synonymous. Error bars in (A) and (B) represent the 95% confidence interval (CI). When necessary, we truncated the lower bound of the CI to 0. See Figures S2, S4, and S5 and Table S3.

Sanders et al. (2015), we estimate a 97.7% confirmation rate. Therefore, as with *de novo* sequence variants, we based all burden analyses on all detected *de novo* CNVs, though we observed similar results when narrowing to confirmed *de novo* CNVs only (STAR Methods).

### *De Novo* CNVs Are Increased in TD

We normalized *de novo* CNV rate per individual per cohort based on the number of non-contiguous intervals captured on each array type to reduce potential bias arising from different capture arrays (STAR Methods; Figure S3A). We observed an increased rate of *de novo* CNVs in phase 1 TD samples (RR = 2.2; one-sided Wilcoxon rank-sum test; p = 0.004; Figure 3A), phase 2 TD samples (RR = 2.2; p = 0.024), and the combined dataset (RR = 2.2; p = 0.0025). De novo deletions (RR 2.13; p = 0.04) and duplications (RR 2.25; p = 0.015) are independently overrepresented in the combined TD dataset, suggesting both are risk factors (Table S5). As expected, we also observed an increased rate of *de novo* CNVs in SSC probands (RR = 1.9; p = 0.0026). We do not observe a significant difference between the ASD and TD samples (RR = 1.1; two-sided Wilcoxon rank-sum test; p = 0.83), suggesting that *de novo* CNVs occur at a similar rate in TD and ASD, although larger sample sizes will be needed to confirm this observation. We did not assess the *de novo* CNV rate in simplex versus multiplex families or in male versus female probands due to the limited number of *de novo* CNVs identified here and the corresponding lack of power.

### Association of *De Novo* CNVs Is Replicated in Microarray Genotyping Data

We used microarray genotyping data in an effort to replicate the association observed in the WES data. We obtained genotyping data generated from the Illumina HumanOmniExpressExome chip for 412 TD trios. We trimmed this number to 399 trios after quality control (Figure S3C and S3D). These 399 trios overlap with 279 of the 789 TD trios in the WES CNV analyses and with 35 of the 148 trios in Fernandez et al. (2012). We utilized 765 SSC quartets, previously genotyped with the Illumina HumanOmni chip, as controls (763 after quality control). To account for the different microarray platforms, we narrowed to high-quality SNPs present on both arrays (Figure S3B). We detected CNVs with PennCNV using an exome-specific Hidden Markov Model (HMM) file (Szatkiewicz et al., 2013). We identified 13 *de novo* CNVs in 399 TD samples (0.033 per proband; 95% CI 0.012–0.053; 81.8% validation rate), 28 in 763 SSC probands (0.037 per individual; 95% CI 0.021–0.052; 100% validation rate), and 9 in 763 SSC unaffected siblings (0.012 per individual; 95% CI 0.0041–0.020; 100% validation rate). Again, we observed an increased burden of *de novo* CNVs in TD samples versus SSC unaffected control siblings (Figure 3B; RR = 2.8; p = 0.024). *De*

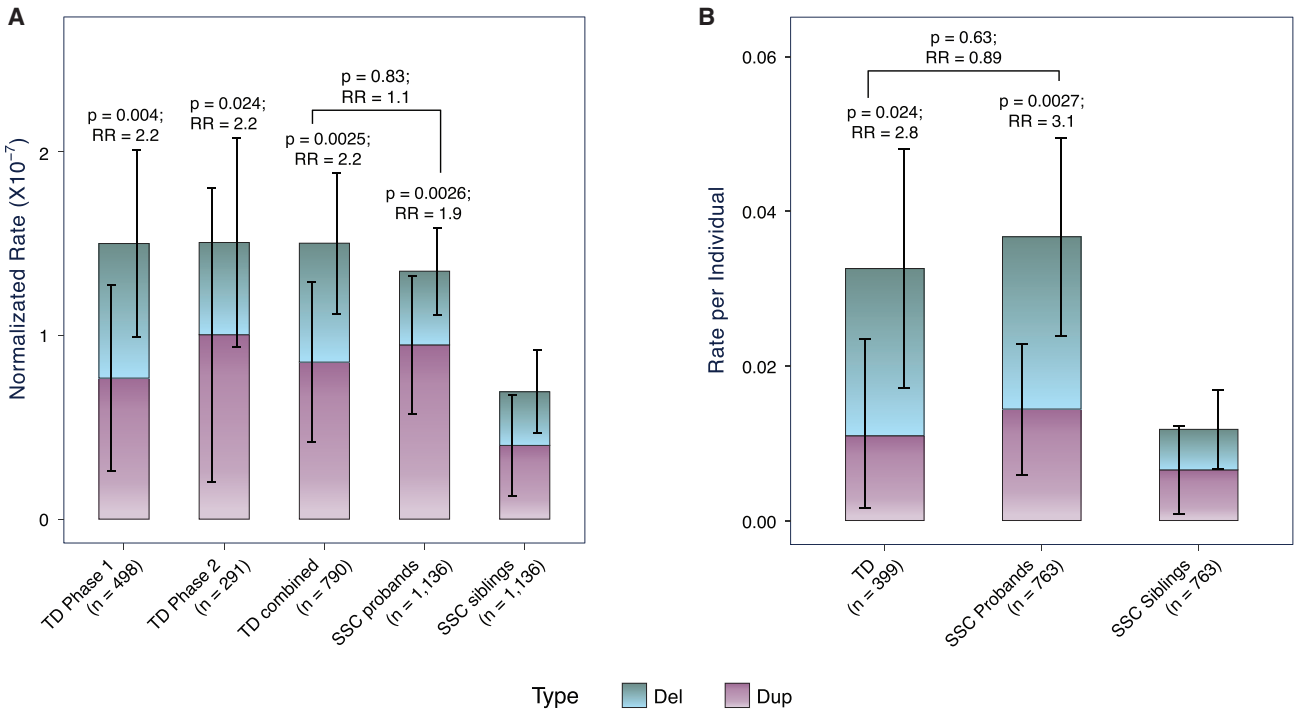



**Figure 3. *De Novo* CNV Burden Analysis**

We called *de novo* CNVs from WES data and array data with CoNIFER (Krumm et al., 2012) and PennCNV (Wang et al., 2007), respectively. We utilized different methods for normalization to make the results comparable across different samples sets.

For the WES data (A), we normalized the *de novo* CNV rate by the number of discontinuous capture array intervals in each cohort (Figure S3A).

For the microarray data (B), we restricted *de novo* CNV calling to a set of SNPs shared across all arrays and further removed any outlier SNPs based on the LRR (Figure S3B; see STAR Methods for details). We compared each group with SSC sibling controls using a Wilcoxon rank-sum test in R. We also used the SSC probands as positive controls to validate our *de novo* calling pipelines. We used all *de novo* calls (confirmed and unconfirmed) in the burden analysis.

Both the WES data (A) and array data (B) demonstrate that *de novo* CNVs are significantly increased in TD compared to SSC controls and that *de novo* CNVs occur at approximately the same rate in TD and in ASD. Error bars in (A) and (B) represent the 95% confidence interval (CI). When necessary, we truncated the lower bound of the CI to 0.

See also Tables S3, S4, and S5.

*novo* deletions are independently overrepresented in TD (RR 3.8; p = 0.02), but *de novo* duplications do not reach significance (RR 1.9; p = 0.15; Table S5). We also confirmed an increased rate of *de novo* CNVs in SSC probands (RR = 3.1; p = 0.0027). Direct comparison between TD probands and SSC probands again shows no difference (RR = 0.89; p = 0.63). We did not observe any recurrent *de novo* CNVs, even when combining across the WES and array data.

## Approximately 10% of Cases Have a *De Novo* Damaging Variant or CNV

We next explored the genomic architecture of simplex TD (Table 3). We restricted these analyses to the simplex trios with WES data that passed quality control for both *de novo* sequence variant and CNV analyses (577 TD trios; 1,134 SSC control trios). We predicted that 22.3% of *de novo* damaging sequence variants contribute TD risk (95% CI 4.7%–41.5%) and 46.3% of *de novo* CNVs carry risk (95% CI −8.5%–101.1%) in simplex families. Additionally, we estimated that 9.7% of TD cases in simplex families carry one or more *de novo* damaging sequence variants mediating risk (95% CI 5.2%–14.3%) and that 1.5% carry a *de novo* CNV mediating risk (95% CI 0.0%–3.0%). Overall, we estimated that approximately 10.5% of cases have a *de novo* damaging sequence variant and/or CNV mediating risk (95% CI 6.0%–15.2%).

## *De Novo* Variants in TD Probands Overlap with Those Identified in Other Disorders

We compared the list of genes with confirmed *de novo* damaging variants in TD probands with genes mutated in other disorders with established *de novo* contributions, including ASD (Sanders et al., 2015), epileptic encephalopathies (EuroEPINOMICS-RES Consortium et al., 2014), intellectual disability (Gilissen et al., 2014; Hamdan et al., 2014; de Ligt et al., 2012; Rauch et al., 2012), OCD (Cappi et al., 2017), schizophrenia (Fromer et al., 2014), developmental disorders in general (Deciphering Developmental Disorders Study, 2017), and congenital heart disease (Jin et al., 2017). There is a high degree of overlap between TD and OCD (44 of 315 genes with *de novo* damaging variants in TD overlap with 90 genes with *de novo* damaging variants in OCD; $p < 1 \times 10^{-4}$ by permutation test accounting for per gene mutability). However, a substantial proportion of TD

**Table 3. Contributions of *De Novo* Events to TD Risk**

| | Percent of Children Carrying ≥1 Variant[a] | | Theoretical Rate per Child (±95% CI)[b] | | % of Cases with a Variant Mediating Risk (±95% CI)[c] | % of Variants Carrying TD Risk (±95% CI)[d] |
|---|---|---|---|---|---|---|
| | TD Simplex (n = 577) | Control (n = 1,134) | TD Simplex (n = 577) | Control (n = 1,134) | | |
| LGD | 9.0% | 4.6% | 0.12 (0.082–0.16) | 0.061 (0.042–0.080) | 4.4% (1.8%–7.1%) | 49.5% (13.6%–83%) |
| Mis3 | 26.7% | 20.8% | 0.41 (0.33–0.48) | 0.34 (0.30–0.39) | 5.9% (1.6%–10.2%) | 15.3% (−5.9%–36.4%) |
| Damaging (LGD+Mis3) | 33.4% | 23.7% | 0.50 (0.42–0.57) | 0.39 (0.34–0.44) | 9.7% (5.2%–14.3%) | 22.3% (4.7%–41.5%) |
| Intolerant genes | 8.8% | 4.1% | 0.17 (0.12–0.22) | 0.076 (0.055–0.098) | 4.7% (2.2%–7.1%) | 56.0% (26.0%–86.0%) |
| *De novo* CNVs[e] | 2.9% | 1.4% | $1.29 \times 10^{-7}$ ($0.68 \times 10^{-7}$–$1.90 \times 10^{-7}$) | $0.69 \times 10^{-7}$ ($0.34 \times 10^{-7}$–$1.05 \times 10^{-7}$) | 1.5% (0.0%–3.0%) | 46.3% (−8.5%–101.1%) |
| Damaging + *de novo* CNVs | 35.5% | 25.0% | - | - | 10.5% (6.0%–15.2%) | - |
| Intolerant genes + *de novo* CNVs | 11.8% | 5.6% | - | - | 6.2% (3.3%–9.2%) | - |

To estimate the contribution of *de novo* events to TD risk, we assessed the simplex TD and SSC controls used in both analyses of *de novo* sequence variants and *de novo* CNVs (577 TD simplex trios and 1,134 SSC sibling control trios; Table S1).

[a]We calculated the percentage of children carrying *de novo* events as the total number of individuals carrying one or more *de novo* events/total number of individuals in the cohort; we denote the percentages of TD cases and SSC controls as p(TD) and p(Controls), respectively.

[b]We estimated the theoretical rate per child (proband or control) for sequence variants as described in Figure 2. We obtained the mean and 95% CI by t test in R.

[c]We estimated the percentage of cases with a variant mediating TD risk by p(TD) − p(Controls). We generated the 95% CI by bootstrapping.

[d]We estimated the percentage of variants carrying TD risk and the corresponding 95% CI by two-sample t test in R, using the theoretical rate per child as input.

[e]It is unclear how to estimate the theoretical *de novo* CNV rate per individual in WES data. We thus used the *de novo* CNV rate normalized by the number of continuous intervals captured to estimate the percentage of variants carrying TD risk (STAR Methods). To determine the percentage of cases with a *de novo* sequence variant or a *de novo* CNV mediating risk, we used the percent of children carrying ≥1 of any of these variants.

probands in our sample have comorbid OCD (361 of 777 overall). Nonetheless, narrowing to probands with TD only still results in significant enrichment (22 of 179 genes with *de novo* damaging variants in TD overlap with 90 genes with *de novo* damaging variants in OCD; $p < 1 \times 10^{-4}$), suggesting this is not driven by comorbid diagnoses. We do not observe significant overlap with other disorders, even before correction for multiple comparisons: intellectual disability (p = 1.00); schizophrenia (p = 0.95); epileptic encephalopathies (p = 0.81); congenital heart disease (p = 0.47); ASD (p = 0.14); and developmental disorders in general (p = 0.092), although the latter two show a trend toward enrichment and these analyses are likely underpowered.

We conducted a similar analysis for *de novo* CNVs identified in our TD cohort and *de novo* CNVs previously identified in the SSC. We restricted to the unique set of *de novo* CNVs called in TD probands across the WES and microarray data and compared them to published, validated CNVs from 2,591 SSC probands (Sanders et al., 2015). 9 of the 34 *de novo* CNVs detected in TD probands were also detected in SSC probands (p = 0.024 by permutation test), whereas only 1 was detected in SSC unaffected siblings (p = 0.27). Due to the relatively small samples sizes of studies investigating *de novo* CNVs in other disorders, we did not test the significance of overlap between *de novo* CNVs in TD and other conditions. However, we did observe *de novo* CNVs in TD cases that have also been detected in other disorders (Table S3), for example, CNVs in 15q13.2-13.3 have been observed in ASD (Sanders et al., 2015), schizophrenia (Georgieva et al., 2014; Malhotra et al., 2011), and epilepsy (Epilepsy Phenome/Genome Project Epi4K Consortium, 2015).

## Approximately 483 Genes Contribute Risk to TD

We next estimated the number of genes likely to contribute to TD risk when disrupted by a *de novo* damaging variant. We used a previously established maximum-likelihood estimation procedure (Homsy et al., 2015; Willsey et al., 2017) and excluded multiplex families in which *de novo* damaging variants might contribute low TD risk (Figures 2 and S5A). Our data fit best with a model of 483 TD risk genes (Figure S6), consistent with our previous estimate of 420 risk genes (Willsey et al., 2017). We are unable to estimate the number of loci vulnerable to *de novo* CNVs due to the absence of recurrent variants.

## Integrated Analysis Identifies Additional TD Risk Genes, Including a High-Confidence Gene *CELSR3*

We leveraged *de novo* damaging variants and the Transmission and De Novo Association (TADA) algorithm to estimate per-gene association with TD (De Rubeis et al., 2014; He et al., 2013; Sanders et al., 2015; Willsey et al., 2017). We did not observe overlap between genes with *de novo* sequence variants and genes affected by *de novo* CNVs, as has been observed in ASD (Sanders et al., 2015), and therefore, we did not include *de novo* CNVs in this analysis. We also did not include inherited variants, as we did not observe overrepresentation in our combined TD cohort (Figure S5B). We utilized a Poisson regression model to control for paternal age, sex, affected status (TD or unaffected),

**Table 4. TD Risk Genes Identified in this Study**

| Gene | LGD | Mis3 | p Value | q Value | q Value in Phase 1[a] | Risk Status in Phase 1[a] | Intolerant | pLI[b] | Missense *Z* Score[c] |
|---|---|---|---|---|---|---|---|---|---|
| *WWC1*[d] | 1 | 1 | $1.93 \times 10^{-5}$ | 0.069 | 0.096 | hcTD | no | 0.02 | 1.27 |
| *CELSR3*[d] | 0 | 3 | $2.23 \times 10^{-5}$ | 0.073 | 0.14 | pTD | yes (LGD and Mis) | 1.00 | 6.17 |
| *OPA1*[d] | 0 | 2 | $6.70 \times 10^{-5}$ | 0.11 | 0.72 | NA | yes (LGD) | 0.99 | 1.83 |
| *NIPBL*[d] | 0 | 2 | $1.13 \times 10^{-4}$ | 0.16 | 0.22 | pTD | yes (LGD and Mis) | 1.00 | 5.04 |
| *FN1*[d] | 0 | 2 | $1.22 \times 10^{-4}$ | 0.19 | 0.26 | pTD | no | 0.06 | 1.39 |
| *FBN2*[d] | 0 | 2 | $1.29 \times 10^{-4}$ | 0.22 | 0.98 | NA | yes (LGD) | 1.00 | 1.22 |

Six genes with recurrent *de novo* variants meet our thresholds for association: two of these are high-confidence TD (hcTD) risk genes (*CELSR3* and *WWC1*; FDR $\leq$ 0.1), and four of these are probable TD (pTD) risk genes (*OPA1*, *NIPBL*, *FN1*, and *FBN2*; FDR $\leq$ 0.3). Four of these six TD risk genes are considered intolerant to variation; determined based on PLI and missense *Z* score.We excluded genes with only one *de novo* variant from this table (3 pTD genes; see Table S7). See also Figure S6 and Table S6.

[a]Willsey et al., (2017).

[b]Probability of being loss-of-function (LoF) intolerant, from Exome Aggregation Consortium (ExAC). pLI $\geq$ 0.9 is considered intolerant.

[c]*Z* score for missense variants, from ExAC. Mis_z $\geq$ 3.891 is considered intolerant.

[d]We previously identified *WWC1* as an hcTD gene and *CELSR3*, *NIPBL*, and *FN1* as pTD genes (Willsey et al., 2017).

and number of callable bases within the consensus region (STAR Methods) when estimating the relative risk for *de novo* LGD and for *de novo* Mis3 variants. We included confirmed *de novo* damaging variants identified in all 674 non-multiplex trios (577 simplex trios and 97 unknown trios) passing quality control. We also integrated *de novo* damaging variants called and confirmed in Willsey et al. (2017), but not called under the new pipeline, which added 8 *de novo* damaging variants (Table S3). TADA identified 2 hcTD genes (FDR q value $\leq$ 0.1; $\geq$2 *de novo* variants) and 4 pTD genes (q $\leq$ 0.3; $\geq$2 *de novo* variants), including one new hcTD gene, *CELSR3* (*cadherin EGF LAG seven-pass G-type receptor 3*) and two new pTD genes (*OPA1* and *FBN2*; Table 4). Four of these six TD risk genes, including *CELSR3*, are intolerant to variation based on pLI and/or missense *Z* score. We identified three additional genes with q $\leq$ 0.3 but only one *de novo* damaging variant; we omitted these genes from Table 4, but they are included in Table S7.

Interestingly, we observed an additional *de novo* damaging variant in *CELSR3* within the 103 multiplex families. We also identified two additional inherited compound heterozygous damaging variants in *CELSR3* in two independent probands (each with one rare and one common inherited variant), which is highly unlikely by chance (p = 0.0069 by permutation test; STAR Methods; Table S6). We did not observe any compound heterozygous variants in the other 5 TD risk genes.

### The Top TD Risk Genes Highlight Cell Polarity

Both of the hcTD risk genes identified here (*WWC1* and *CELSR3*) encode proteins involved in cell polarity. Therefore, we assessed whether *de novo* damaging variants in TD affect other genes encoding cell polarity proteins. We obtained a list of genes related to cell polarity from the Gene Ontology database (Ashburner et al., 2000; The Gene Ontology Consortium, 2017) and annotated the *de novo* variant list (Table S3). 15 of the 292 *de novo* damaging variants in non-multiplex families impact genes related to cell polarity, representing a significant enrichment over the variants identified in the SSC control trios (7 of 350 *de novo* damaging variants; one-sided Fisher's exact test odds ratio [OR] 2.56; p = 0.030). We confirmed this result with permutation testing (13 of 315 unique genes with confirmed *de novo* damaging variants are related to cell polarity; p = 0.032). We observed additional variants in cell polarity genes in multiplex families (2 of 45 *de novo* damaging variants), and the combined set of variants from all 777 TD trios are also significantly enriched for variants affecting cell polarity genes (17 of 337 unique genes; one-sided Fisher's exact test OR 2.60, p = 0.024; permutation test, p = 0.014).

## DISCUSSION

We previously established the contribution of *de novo* damaging sequence variants to TD risk and identified one hcTD risk gene, *WWC1*, based on *de novo* LGD variants observed in two unrelated probands. Furthermore, we demonstrated that sequencing of larger cohorts coupled with the identification of recurrent *de novo* variants would be a productive and reliable method for gene discovery in TD (Willsey et al., 2017). In this study, we sequenced an additional 291 trios, bringing the total sample size to 802 trios. After quality control, we used 674 non-multiplex trios for gene discovery (577 simplex families and 97 unknown families). Given this sample size and our previously estimated trajectory of gene discovery (Willsey et al., 2017), we expected to identify 1.4 hcTD genes and 5.4 pTD genes. In actuality, this study implicated 2 hcTD genes and 7 pTD genes, which fits well with our previous prediction. Note that we did not present three of the pTD genes in the main text, as they only carried one *de novo* damaging variant (Table S7).

We observed a strong effect of plexity on *de novo* mutation rate, particularly with respect to *de novo* variants in mutation-intolerant genes (Figure 2B). Therefore, this suggests that the recruitment and sequencing of simplex families should be the highest priority, at least in studies examining *de novo* variants. Of course, it still remains to be determined whether *de novo* variants (particularly *de novo* LGD variants) carry risk in multiplex families, as the effects observed here trend toward significance (e.g., RR 1.16; p = 0.26 for *de novo* LGD variants) in an underpowered analysis (103 multiplex trios) and *de novo* variants appear to carry risk in multiplex families for other neurodevelopmental disorders (Leppa et al., 2016; Martin et al., 2017).

We also observed preliminary evidence for an increased rate of *de novo* damaging sequence variants in female TD probands

compared to male TD probands, as has been observed in ASD (De Rubeis et al., 2014; Iossifov et al., 2014; Sanders et al., 2015). Given the TD sex bias for affected males (male:female = 3:1–4:1), this suggests a potential female protective effect similar to that which has been postulated in ASD (De Rubeis et al., 2014; Dong et al., 2014; Gockley et al., 2015; Iossifov et al., 2014; Jacquemont et al., 2014; Levy et al., 2011; Sanders et al., 2011, 2015). Larger sample sizes are, of course, required to confirm this preliminary observation, and it should be noted that we observe a significant excess of *de novo* variants in both male and female TD probands independently when compared to sex-matched controls, indicating these variants carry risk for both sexes. We do not observe any differences in the overall rate of coding *de novo* variants by sex in the TD cohort (RR 1.02; two-sided rate ratio test p = 0.90) or the SSC cohort (RR 1.03; p = 0.72), suggesting no systematic differences in the rate or detection of *de novo* variants overall.

We observed a significant increase in the rate of rare *de novo* CNVs in TD. We confirmed this association using both WES and microarray genotyping data. Of note, many of the samples assessed are represented only in the WES data (511 trios) or only in the array data only (120 trios), and *de novo* CNV calling was conducted with independent methods. Taken together, then, these results strongly support the conclusion that *de novo* CNVs carry risk for TD. Although rare CNVs have already been definitively associated with TD risk (Huang et al., 2017; McGrath et al., 2014; Nag et al., 2013), *de novo* CNVs had not been definitively implicated, though previous results suggested association (Fernandez et al., 2012). The number of WES samples in this current study is more than five-fold larger than that in Fernandez et al. (2012; 789 versus 148 trios), and the array data are more than two-fold larger (399 versus 148 trios), suggesting the main difference in these studies was the greater power to identify this association, especially given the similar effect sizes across the studies (RR 2.2 in our WES data and RR 2.8 in our array data versus RR 2.4 in Fernandez et al., 2012). Our observation of an increased rate of *de novo* sequence variants in simplex TD suggests that a similar phenomenon may also occur with respect to *de novo* CNVs. However, we did not assess this question here due to a very small number of *de novo* CNVs identified in multiplex families (3 *de novo* CNVs in 103 multiplex trios).

We estimated that 4.4% of TD probands have a *de novo* LGD variant mediating risk and 5.9% have a *de novo* missense 3 variant mediating risk (Table 3). Although *de novo* missense variants in general are not yet significantly associated, we can similarly estimate that 5.0% of TD probands carry a *de novo* missense variant mediating risk. At first glance, these estimates appear much lower than estimates in ASD (e.g., 9% and 12% for *de novo* LGD and *de novo* missense, respectively; Iossifov et al., 2014). However, the ASD estimates are based on different methods. Indeed, by applying our methods to their data, we achieved highly similar estimates (5.4% of ASD probands have a *de novo* LGD variant contributing risk and 3.1% have a *de novo* missense variant contributing risk). We believe the higher estimates in Iossifov et al. (2014) are due to two major factors. First, they use a much larger set of regions for analysis (~83 mb compared to ~30 mb here), and we expect the ascertainment differential to increase proportionally to target size if the mutation rate per base pair is constant. Second, their method counts multiple *de novo*s per individual (rate in probands minus rate in controls), whereas here, we are counting a maximum of one *de novo* per individual (percentage with $\geq$1 *de novo*). We previously observed similar rate ratios between TD probands versus SSC controls and ASD probands versus SSC controls (Willsey et al., 2017), further suggesting similar architecture.

Likewise, we did not observe a difference in the rate of *de novo* CNVs in TD probands compared to ASD probands (Figure 3). This suggests that the rate of *de novo* CNVs is not different in TD and ASD and that published data showing a higher proportion of *de novo* CNVs in ASD (e.g., 4.1% of individuals have a *de novo* CNV mediating ASD risk in Sanders et al., 2015 versus 1.5% reported here in TD) is likely due to the genome-wide coverage in those studies versus exome-wide coverage only here (i.e., whole-exome sequencing data and HumanOmniExpressExome-8-v1 genotyping data).

We observed significant overlap between TD and OCD for *de novo* damaging sequence variants, even when restricting to TD probands without comorbid OCD. We also observed significant overlap across *de novo* CNVs identified in TD and in ASD, consistent with previous results (Fernandez et al., 2012), and a suggestion of overlap of *de novo* sequence variants between TD and ASD (uncorrected p = 0.14). This suggests that TD and OCD as well as TD and ASD may share a subset of genetic risk loci, but this hypothesis warrants follow-up with larger sample sizes. By the same token, the lack of overlap between TD and other psychiatric disorders is inconclusive and may simply reflect underpowered analyses, and therefore, it will be important to revisit these analyses as data accumulate in these and other disorders not yet characterized. For example, enrichment of ultra-rare variants in ADHD (Satterstrom et al., 2018) suggests that *de novo* variants will carry risk in this condition. Coupled with the high degree of TD and ADHD comorbidity, this indicates that there may be strong overlap at the level of *de novo* variants as observed with OCD here.

We identified a total of six likely TD risk genes, including two hcTD genes, *CELSR3* (new; promoted from pTD status in phase 1) and *WWC1*, and four pTD genes, *OPA1* (new), *NIPBL*, *FN1*, and *FBN2* (new). Notably, both of the two hcTD genes encode proteins that are related to cell polarity, defined broadly in the Gene Ontology database as anisotropic intracellular organization or cell growth patterns (Ashburner et al., 2000; The Gene Ontology Consortium, 2017). Additionally, we observed general enrichment for cell polarity annotation among the genes carrying *de novo* damaging variants, including four mutation-intolerant genes (*SPRY2*, *MARK2*, *PSMC1*, and *UBC*; Table S3). Furthermore, recent rare CNV analyses have definitively implicated *NRXN1* deletions and *CNTN6* duplications with TD risk (Fernandez et al., 2012; Huang et al., 2017; Sundaram et al., 2010), and other studies have highlighted *CNTN4* and *CNTNAP2* (Fernandez et al., 2012; Verkerk et al., 2003). All of the proteins encoded by these genes have putative roles in cell polarity or axon pathfinding and/or organization (Bel et al., 2009; Fernandez et al., 2004; Kamei et al., 1998; Ushkaryov et al., 1992), suggesting that perturbation of cell polarity may contribute to TD. We do not observe convergence in other pathways, including histaminergic neurotransmission, as has been previously identified

(Fernandez et al., 2012). Given clear evidence that we are in the early phases of gene discovery in TD, it is very likely that further studies will clarify these results and generate additinonal testable hypotheses regarding the underlying neurobiology of TD. Generating more TD genomic risk data should also better address the extent to which TD-associated *de novo* variants overlap with CNVs and genes implicated in other neurodevelopmental disorders. As these data accumulate, functional genetics will be critical to translate findings into an actionable understanding of pathobiological mechanisms.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample collection
- METHOD DETAILS
  - Whole exome sequencing
  - Quality Control
  - Variant detection
  - Microarray genotyping
  - De novo variant validation
  - Burden analysis
  - Gene discovery
  - Systems biological analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY
  - Data
  - Software

## SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.08.082.

## CONSORTIA

The members of the Tourette International Collaborative Genetics (TIC Genetics) consortium are Mohamed Abdulkadir, Juan Arbelaez, Benjamin Bodmer, Yana Bromberg, Lawrence W. Brown, Keun-Ah Cheon, Barbara J. Coffey, Li Deng, Andrea Dietrich, Shan Dong, Clif Duhn, Lonneke Elzerman, Thomas V. Fernandez, Carolin Fremer, Blanca Garcia-Delgar, Donald L. Gilbert, Dorothy E. Grice, Julie Hagstrøm, Tammy Hedderly, Gary A. Heiman, Isobel Heyman, Pieter J. Hoekstra, Hyun Ju Hong, Chaim Huyser, Eun-Joo Kim, Young Key Kim, Young-Shin Kim, Robert A. King, Yun-Joo Koh, Sodahm Kook, Samuel Kuperman, Bennett L. Leventhal, Andrea G. Ludolph, Marcos Madruga-Garrido, Jeffrey D. Mandell, Athanasios Maras, Pablo Mir, Astrid Morer, Montana T. Morris, Kirsten Müller-Vahl, Alexander Münchau, Tara L. Murphy, Cara Nasello, Kerstin J. Plessen, Hannah Poisner, Veit Roessner, Stephan J. Sanders, Eun-Young Shin, Dong-Ho Song, Jungeun Song, Matthew W. State, Nawei Sun, Joshua K. Thackray, Jay A. Tischfield, Jennifer Tübing, Frank Visscher, Sina Wanderer, Sheng Wang, A. Jeremy Willsey, Martin Woods, Jinchuan Xing, Yeting Zhang, Xin Zhao, and Samuel H. Zinner.

The members of the Tourette Syndrome Genetics Southern and Eastern Europe Initiative (TSGENESEE) are Christos Androutsos, Csaba Barta, Luca Farkas, Jakub Fichna, Marianthi Georgitsi, Piotr Janik, Iordanis Karagiannidis, Anastasia Koumoula, Peter Nagy, Peristera Paschou, Joanna Puchala, Renata Rizzo, Natalia Szejko, Urszula Szymanska, Zsanett Tarnok, Vaia Tsironi, Tomasz Wolanczyk, and Cezary Zekanowski.

The members of the Tourette Association of America International Consortium for Genetics (TAAICG) are Cathy L. Barr, James R. Batterson, Cheston Berlin, Ruth D. Bruun, Cathy L. Budman, Danielle C. Cath, Sylvain Chouinard, Giovanni Coppola, Nancy J. Cox, Sabrina Darrow, Lea K. Davis, Yves Dion, Nelson B. Freimer, Marco A. Grados, Matthew E. Hirschtritt, Alden Y. Huang, Cornelia Illmann, Robert A. King, Roger Kurlan, James F. Leckman, Gholson J. Lyon, Irene A. Malaty, Carol A. Mathews, William M. MacMahon, Benjamin M. Neale, Michael S. Okun, Lisa Osiecki, David L. Pauls, Danielle Posthuma, Vasily Ramensky, Mary M. Robertson, Guy A. Rouleau, Paul Sandor, Jeremiah M. Scharf, Harvey S. Singer, Jan Smit, Jae-Hoon Sul, and Dongmei Yu.

## ACKNOWLEDGMENTS

We wish to thank the families who have participated in and contributed to this study. We also thank the NIMH Repository and Genomics Resource (U24MH068457 to J.A.T.) at RUCDR Infinite Biologics for transforming cell lines and providing DNA samples, Liping Wei at Peking University for her support in this project, and Sarah Pyle for graphic design. This study was supported by grants from the National Institute of Mental Health (R01MH092290 to Lawrence W. Brown, R01MH092291 to Samuel Kuperman, R01MH092292 to Barbara J. Coffey, R01MH092293 to G.A.H. and J.A.T., R01MH092513 to Samuel H. Zinner, R01MH092516 to D.E.G., R01MH092520 to Donald L. Gilbert, R01MH092289 to M.W.S., and K08MH099424 to T.V.F.), from the Human Genetics Institute of New Jersey (to G.A.H. and J.A.T.), and the New Jersey Center for Tourette Syndrome and Associated Disorders (to G.A.H. and J.A.T.). We are also grateful to the NJCTS for facilitating the inception and organization of the TIC Genetics study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This study was also supported by the Weill Institute for Neurosciences (Startup Funding to A.J.W.) and the Overlook International Foundation (to M.W.S. and A.J.W.).

The Yale Center for Mendelian Genomics (NIH M#UM1HG006504-05) is funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute. The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

This work was additionally supported by grants from Spain (to Pablo Mir): the Instituto de Salud Carlos III (PI10/01674 and PI13/01461); the Consejería de Economía, Innovación, Ciencia y Empresa de la Junta de Andalucía (CVI-02526 and CTS-7685); the Consejería de Salud y Bienestar Social de la Junta de Andalucía (PI-0741/2010, PI-0437-2012, and PI-0471-2013); the Sociedad Andaluza de Neurología; the Fundación Alicia Koplowitz; the Fundación Mutua Madrileña; and the Jaques and Gloria Gossweiler Foundation, (to Astrid Morer): Alicia Koplowitz Foundation; grants from Germany (to Astrid Morer): Deutsche Forschungsgemeinschaft (DFG) (MU 1692/3-1and MU 1692/4-1 and project C5 of the SFB 936); and grants from Sweden: the Swedish Research Council 2015-02424 (to N.D.). This research was also supported in part by an Informatics Starter Grant from the PhRMA Foundation (to Yana Bromberg), the Mindich Child Health and Developmental Institute at the Icahn School of Medicine at Mount Sinai (to D.E.G.), the Seaver Foundation (to D.E.G.), and the Stanley Center for Psychiatric Research (to D.E.G.). All research at Great Ormond Street Hospital NHS Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, and E. Wijsman). We also appreciate obtaining access to whole-exome sequencing, microarray genotyping, and phenotype data on SFARI Base. Approved researchers can obtain the SSC population dataset

described in this study by applying at https://base.sfari.org. Finally, we thank all of the individuals involved in recruitment and assessment of the subjects reported in this study: Denmark: Nikoline Frost and Heidi B. Biernat (Copenhagen); Germany: Yvonne Friedrich (Dresden), Daniela Ihlenburg-Schwarz (Hannover), and Jenny Schmalfeld (Lübeck); Spain: Fátima Carrillo, Marta Correa, Pilar Gómez-Garre, and Laura Vargas (Sevilla); the Netherlands: Vivian op de Beek (Amsterdam); Jolanda Blom, Rudi Bruggemans, and MariAnne Overdijk (Barendrecht); and Marieke Messchendorp, Thaïra Openneer, and Anne Marie Stolte (Groningen); UK: Anup Kharod (London GOSH); USA: Sarah Jacobson (Cincinnati), Angie Cookman (Iowa City), Laura Ibanez-Gomez and Zoey Shaw (Mount Sinai/NKI), Shannon Granillo and J.D. Sandhu (Seattle Children's), and Yanran Wang (Rutgers); and to all who may not have been mentioned.

## AUTHOR CONTRIBUTIONS

Conceptualization, S.W., T.V.F., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Methodology, S.W., R.A.K., T.V.F., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Software, S.W., J.D.M., and A.J.W.; Validation, S.W., N.S., M.T.M., J.A., C.D., D.Y., A.Y.H., G.C., J.M.S., T.V.F., J.X., and A.J.W.; Formal Analysis, S.W., J.D.M., S.D., X.Z., and A.J.W.; Investigation, S.W., J.D.M., J.M.S., T.V.F., J.X., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Resources, R.A.K., A.D., N.K., N.D., B.M.N., G.C., C.A.M., J.M.S., TIC Genetics, TSGENESEE, TAAICG, T.V.F., J.D.B., S.D.R., D.E.G., J.X., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Data Curation, S.W., J.D.M., D.Y., A.D., N.K., N.D., C.A.M., J.M.S., T.V.F., J.D.B., S.D.R., D.E.G., J.X., G.A.H., J.A.T., P.P., and A.J.W.; Writing – Original Draft, S.W., J.D.M., A.J.W., and M.W.S.; Writing – Review & Editing, S.W., J.D.M., Y.K., N.S., M.T.M., J.A., C.N., S.D., C.D., X.Z., Z.Y., S.S.P., D.Y., R.A.K., A.D., N.K., N.D., A.Y.H., B.M.N., G.C., C.A.M., J.M.S., TIC Genetics, TSGENESEE, TAAICG, T.V.F., J.D.B., S.D.R., D.E.G., J.X., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Visualization, S.W. and A.J.W.; Supervision, R.A.K., T.V.F., J.D.B., S.D.R., D.E.G., J.X., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Project Administration, B.M.N., G.C., C.A.M., J.M.S., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.; Funding Acquisition, N.K., N.D., B.M.N., G.C., C.A.M., J.M.S., T.V.F., J.X., G.A.H., J.A.T., P.P., A.J.W., and M.W.S.

## DECLARATION OF INTERESTS

Donald L. Gilbert has received salary/travel/honoraria from the Tourette Association of America, the Child Neurology Society, U.S. National Vaccine Injury Compensation Program, Ecopipam Pharmaceuticals, EryDel Pharmaceuticals, Elsevier, and Wolters Kluwer. A.J.W. is a paid consultant for Daiichi Sankyo. M.W.S. is a consultant to BlackThorn and ArRett Pharmaceuticals.

Received: May 3, 2018
Revised: July 13, 2018
Accepted: August 27, 2018
Published: September 25, 2018; corrected online: December 14, 2018

## REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *Chapter 7*, Unit7.20.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bel, C., Oguievetskaia, K., Pitaval, C., Goutebroze, L., and Faivre-Sarrailh, C. (2009). Axonal targeting of Caspr2 in hippocampal neurons via selective somatodendritic endocytosis. J. Cell Sci. *122*, 3403–3413.

Cappi, C., Oliphant, M.E., Peter, Z., Zai, G., Sullivan, C.A.W., Gupta, A.R., Hoffman, E.J., Virdee, M., Jeremy Willsey, A., Shavitt, R.G., et al. (2017). De novo damaging coding mutations are strongly associated with obsessive-compulsive disorder and overlap with autism. bioRxiv. https://doi.org/10.1101/127712.

Centers for Disease Control and Prevention (CDC) (2009). Prevalence of diagnosed Tourette syndrome in persons aged 6-17 years - United States, 2007. MMWR Morb. Mortal. Wkly. Rep. *58*, 581–585.

de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al.; DDD Study; Homozygosity Mapping Collaborative for Autism; UK10K Consortium (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dietrich, A., Fernandez, T.V., King, R.A., State, M.W., Tischfield, J.A., Hoekstra, P.J., and Heiman, G.A.; TIC Genetics Collaborative Group (2015). The Tourette International Collaborative Genetics (TIC Genetics) study, finding the genes causing Tourette syndrome: objectives and methods. Eur. Child Adolesc. Psychiatry *24*, 141–151.

Dong, S., Walker, M.F., Carriero, N.J., DiCola, M., Willsey, A.J., Ye, A.Y., Waqar, Z., Gonzalez, L.E., Overton, J.D., Frahm, S., et al. (2014). De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. Cell Rep. *9*, 16–23.

Epilepsy Phenome/Genome Project Epi4K Consortium (2015). Copy number variant analysis from exome data in 349 patients with epileptic encephalopathy. Ann. Neurol *78*, 323–328.

EuroEPINOMICS-RES Consortium; Epilepsy Phenome/Genome Project; Epi4K Consortium (2014). De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. Am. J. Hum. Genet. *95*, 360–370.

Fernandez, T., Morgan, T., Davis, N., Klin, A., Morris, A., Farhi, A., Lifton, R.P., and State, M.W. (2004). Disruption of contactin 4 (CNTN4) results in developmental delay and other features of 3p deletion syndrome. Am. J. Hum. Genet. *74*, 1286–1293.

Fernandez, T.V., Sanders, S.J., Yurkiewicz, I.R., Ercan-Sencicek, A.G., Kim, Y.-S., Fishman, D.O., Raubeson, M.J., Song, Y., Yasuno, K., Ho, W.S.C., et al. (2012). Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. Biol. Psychiatry *71*, 392–402.

Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192–195.

Freeman, R.D., Fast, D.K., Burd, L., Kerbeshian, J., Robertson, M.M., and Sandor, P. (2000). An international perspective on Tourette syndrome: selected findings from 3,500 individuals in 22 countries. Dev. Med. Child Neurol. *42*, 436–447.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature *506*, 179–184.

Georgieva, L., Rees, E., Moran, J.L., Chambert, K.D., Milanova, V., Craddock, N., Purcell, S., Sklar, P., McCarroll, S., Holmans, P., et al. (2014). De novo CNVs in bipolar affective disorder and schizophrenia. Hum. Mol. Genet. *23*, 6677–6683.

Ghanizadeh, A., and Mosallaei, S. (2009). Psychiatric disorders and behavioral problems in children and adolescents with Tourette syndrome. Brain Dev. *31*, 15–19.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.

Gockley, J., Willsey, A.J., Dong, S., Dougherty, J.D., Constantino, J.N., and Sanders, S.J. (2015). The female protective effect in autism spectrum disorder is not mediated by a single genetic locus. Mol. Autism *6*, 25.

Guo, Y., He, J., Zhao, S., Wu, H., Zhong, X., Sheng, Q., Samuels, D.C., Shyr, Y., and Long, J. (2014). Illumina human exome genotyping array clustering and quality control. Nat. Protoc. *9*, 2643–2662.

Hamdan, F.F., Srour, M., Capo-Chichi, J.-M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D., Diallo, O., et al. (2014). De novo mutations in moderate or severe intellectual disability. PLoS Genet. *10*, e1004772.

He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. *9*, e1003671.

Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H., Gorham, J., et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science *350*, 1262–1266.

Huang, A.Y., Yu, D., Davis, L.K., Sul, J.H., Tsetsos, F., Ramensky, V., Zelaya, I., Ramos, E.M., Osiecki, L., Chen, J.A., et al. (2017). Rare copy number variants in NRXN1 and CNTN6 increase risk for Tourette syndrome. Neuron *94*, 1101–1111.e7.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Jacquemont, S., Coe, B.P., Hersch, M., Duyzend, M.H., Krumm, N., Bergmann, S., Beckmann, J.S., Rosenfeld, J.A., and Eichler, E.E. (2014). A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. Am. J. Hum. Genet. *94*, 415–425.

Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat. Genet. *49*, 1593–1601.

Kamei, Y., Tsutsumi, O., Taketani, Y., and Watanabe, K. (1998). cDNA cloning and chromosomal localization of neural adhesion molecule NB-3 in human. J. Neurosci. Res. *51*, 275–283.

Karagiannidis, I., Rizzo, R., Tarnok, Z., Wolanczyk, T., Hebebrand, J., Nöthen, M.M., Lehmkuhl, G., Farkas, L., Nagy, P., Barta, C., et al.; TSGeneSEE (2012). Replication of association between a SLITRK1 haplotype and Tourette Syndrome in a large sample of families. Mol. Psychiatry *17*, 665–668.

Kosmicki, J.A., Samocha, K.E., Howrigan, D.P., Sanders, S.J., Slowikowski, K., Lek, M., Karczewski, K.J., Cutler, D.J., Devlin, B., Roeder, K., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat. Genet. *49*, 504–510.

Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., and Eichler, E.E.; NHLBI Exome Sequencing Project (2012). Copy number variation detection and genotyping from exome sequence data. Genome Res. *22*, 1525–1532.

Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.-X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nat. Genet. *47*, 582–588.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Leppa, V.M., Kravitz, S.N., Martin, C.L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, S.J., Lowe, J.K., Cantor, R.M., and Geschwind, D.H. (2016). Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. Am. J. Hum. Genet. *99*, 540–554.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., et al. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, 886–897.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.

Malhotra, D., McCarthy, S., Michaelson, J.J., Vacic, V., Burdick, K.E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, E.S., et al. (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron *72*, 951–963.

Martin, H.C., Jones, W.D., Stephenson, J., Handsaker, J., Gallone, G., McRae, J.F., Prigmore, E., Short, P., Niemi, M., Kaplanis, J., et al. (2017). Quantifying the contribution of recessive coding variation to developmental disorders. bioRxiv. https://doi.org/10.1101/201533.

McGrath, L.M., Yu, D., Marshall, C., Davis, L.K., Thiruvahindrapuram, B., Li, B., Cappi, C., Gerber, G., Wolf, A., Schroeder, F.A., et al. (2014). Copy number variation in obsessive-compulsive disorder and tourette syndrome: a cross-disorder study. J. Am. Acad. Child Adolesc. Psychiatry *53*, 910–919.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Nag, A., Bochukova, E.G., Kremeyer, B., Campbell, D.D., Muller, H., Valencia-Duarte, A.V., Cardona, J., Rivas, I.C., Mesa, S.C., Cuartas, M., et al.; Tourette Syndrome Association International Consortium for Genetics (2013). CNV analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. PLoS ONE *8*, e59061.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

Pauls, D.L., Cohen, D.J., Heimbuch, R., Detlor, J., and Kidd, K.K. (1981). Familial pattern and transmission of Gilles de la Tourette syndrome and multiple tics. Arch. Gen. Psychiatry *38*, 1091–1093.

Price, R.A., Kidd, K.K., Cohen, D.J., Pauls, D.L., and Leckman, J.F. (1985). A twin study of Tourette syndrome. Arch. Gen. Psychiatry *42*, 815–820.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Quezada, J., and Coffman, K.A. (2018). Current approaches and new developments in the pharmacological management of Tourette syndrome. CNS Drugs *32*, 33–45.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.

Robertson, M.M. (2008). The prevalence and epidemiology of Gilles de la Tourette syndrome. Part 1: the epidemiological and prevalence studies. J. Psychosom. Res. *65*, 461–472.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al.; Autism Sequencing Consortium (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron *87*, 1215–1233.

Satterstrom, F.K., Walters, R.K., Singh, T., Wigdor, E.M., Lescai, F., Demontis, D., Kosmicki, J.A., Grove, J., Stevens, C., Bybjerg-Grauholm, J., et al. (2018). ASD and ADHD have a similar burden of rare protein-truncating variants. bioRxiv. https://doi.org/10.1101/277707.

Scharf, J.M., Yu, D., Mathews, C.A., Neale, B.M., Stewart, S.E., Fagerness, J.A., Evans, P., Gamazon, E., Edlund, C.K., Service, S.K., et al.; North American Brain Expression Consortium; UK Human Brain Expression Database (2013). Genome-wide association study of Tourette's syndrome. Mol. Psychiatry *18*, 721–728.

Scharf, J.M., Miller, L.L., Gauvin, C.A., Alabiso, J., Mathews, C.A., and Ben-Shlomo, Y. (2015). Population prevalence of Tourette syndrome: a systematic review and meta-analysis. Mov. Disord. *30*, 221–228.

Sundaram, S.K., Huq, A.M., Wilson, B.J., and Chugani, H.T. (2010). Tourette syndrome is associated with recurrent exonic copy number variants. Neurology *74*, 1583–1590.

Szatkiewicz, J.P., Neale, B.M., O'Dushlaine, C., Fromer, M., Goldstein, J.I., Moran, J.L., Chambert, K., Kähler, A., Magnusson, P.K.E., Hultman, C.M., et al. (2013). Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. Mol. Psychiatry *18*, 1178–1184.

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. *45* (*D1*), D331–D338.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Ushkaryov, Y.A., Petrenko, A.G., Geppert, M., and Südhof, T.C. (1992). Neurexins: synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin. Science *257*, 50–56.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–11.10.33.

Verkerk, A.J., Mathews, C.A., Joosse, M., Eussen, B.H., Heutink, P., and Oostra, B.A.; Tourette Syndrome Association International Consortium for Genetics (2003). CNTNAP2 is disrupted in a family with Gilles de la Tourette syndrome and obsessive compulsive disorder. Genomics *82*, 1–9.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Willsey, A.J., Fernandez, T.V., Yu, D., King, R.A., Dietrich, A., Xing, J., Sanders, S.J., Mandell, J.D., Huang, A.Y., Richer, P., et al. (2017). De novo coding variants are strongly associated with Tourette disorder. Neuron *94*, 486–499.e9.

Willsey, A.J., Morris, M.T., Wang, S., Willsey, H.R., Sun, N., Teerikorpi, N., Baum, T.B., Cagney, G., Bender, K.J., Desai, T.A., et al. (2018). The Psychiatric Cell Map Initiative: a convergent systems biological approach to illuminating key molecular pathways in neuropsychiatric disorders. Cell *174*, 505–520.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Biological Samples | | |
| TIC Genetics trios (n = 417) | Tourette International Collaborative Genetics Study | https://tic-genetics.org/ |
| TAAICG trios (n = 186) | Tourette Association of America International Consortium for Genetics | https://www.findtsgene.org/tique |
| TSGENESEE trios (n = 181) | Tourette Syndrome Genetics Southern and Eastern Europe Initiative | http://tsgenesee.mbg.duth.gr/index.html |
| UTC trios (n = 18) | Upsala Tourette Cohort | N/A |
| Deposited Data | | |
| Whole exome sequencing data from TIC Genetics trios in Phase 1 project (n = 325) | Willsey et al., 2017 | BioProject: PRJNA384374 |
| Whole exome sequencing data from TSAICG trios in Phase 1 project (n = 186) | Willsey et al., 2017 | BioProject: PRJNA384389 |
| Whole exome sequencing data from TIC Genetics trios in Phase 2 project (n = 92) | This paper | BioProject: PRJNA384374 |
| Genotyping data from TIC Genetics trios (n = 412) | This paper | BioProject: PRJNA384374 |
| Whole exome sequencing data from SSC control quartets (n = 1,184) | Iossifov et al., 2014 | NDAR: DOI:10.15154/1149697 |
| Genotyping data from SSC control quartets (n = 765) | Sanders et al., 2011, 2015 | http://www.sfari.org/resources/sfari-base |
| Software and Algorithms | | |
| Genome Analysis Toolkit (GATK) | DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013 | https://software.broadinstitute.org/gatk/best-practices/ |
| BWA-mem | Li and Durbin, 2009 | http://bio-bwa.sourceforge.net/ |
| Picard Tools | Broad Institute. | https://broadinstitute.github.io/picard/ |
| Annovar | Wang et al., 2010 | http://annovar.openbioinformatics.org/en/latest/ |
| PLINK/SEQ | Fromer et al., 2014 | https://atgu.mgh.harvard.edu/plinkseq/ |
| Primer Design | This paper. | https://primerdesign.willseylab.com/ |
| Python script code for data processing & analysis | This paper. | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |
| R code for data analysis | This paper. | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |
| TADA | He et al., 2013 | http://www.compgen.pitt.edu/TADA/TADA_homepage.htm |
| GenomeStudio 2.0 | Illumina | https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html |
| Other | | |
| 1000 Genomes GRCh37 hg19 genome build | N/A | http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz |
| RefSeq hg19 gene annotation | N/A | http://genome.ucsc.edu/cgi-bin/hgTables?command=start |
| Intervals file for NimbleGen SeqCap EZ Exome v2 | Roche NimbleGen, Madison, WI, USA | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |
| Intervals file for NimbleGen SeqCap EZ Exome v3 | Agilent Technologies, Santa Clara, USA | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Intervals file for Agilent SureSelect v1.1 | Roche NimbleGen, Madison, WI, USA | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |
| Intervals file for IDT xGen | Integrated DNA Technologies, Inc., Skokie, Illinois, USA | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |
| Coding regions only from RefSeq hg19 gene annotation | This paper. | https://bitbucket.org/willseylab/tourette_phase2/src/master/ |

## CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for data should be directed to and will be fulfilled by the Lead Contact, Jeremy Willsey (jeremy.willsey@ucsf.edu).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Sample collection

#### *TD Trios*

We utilized 511 TD trios (affected child and both parents) characterized in our previous "Phase 1" study (Willsey et al., 2017). Ascertainment of these samples has also been described previously (Dietrich et al., 2015). In this study ("Phase 2"), we sequenced 291 additional TD trio samples from three independent collaborative groups: the Tourette International Collaborative Genetics group (TIC Genetics; 92 new trios), the European Multicenter Tics in Children Studies (TSGENESEE; 181 new trios) and the Uppsala Tourette Cohort (UTC, 18 new trios). We ascertained the TIC Genetics trios as previously described (Dietrich et al., 2015). We also ascertained the TSGENESEE trios as previously described (Karagiannidis et al., 2012). All adult participants and parents of children provided written informed consent along with written or oral assent of their participating child. The Institutional Review Board (IRB) of each participating site approved the study.

The UTC was collected under a study in Sweden called, "Mapping of Hereditary Factors in Neuropsychiatric Conditions, Focusing on Tourette Syndrome." Individuals with a TD diagnosis were asked to participate and signed informed consent documents that described the nature of the study. Inclusion criteria for patients were individuals meeting the DSM-IV criteria for TD. All patients were ascertained by a specialist in child psychiatry or child neurology. After a 60-90 minute assessment, blood samples were processed and DNA stored in biobank of the Academic Hospital. All adult participants and parents of children provided written informed consent along with written or oral assent of their participating child. The regional ethical committee of Uppsala approved the study (equivalent to IRB).

Phenotypic data available for each cohort is described in Table 1, including sex, parental age (where available, we were not able to obtain parental age for TSGENESEE and UTC samples), comorbid OCD and/or ADHD in probands, and the history of tic disorders in the first degree relatives for most data. Among the 789 TD trios that passed quality control (see 'Quality control'), we defined 582 trios (73.8%) as 'apparently simplex', which means neither of the parents had any reported history of a tic disorder; 103 trios (13.1%) as 'multiplex', which means at least one of the parents had a reported history of a tic disorder, and 104 trios (13.2%) as 'unknown', which means that we were not able to assign status based on incomplete parental phenotypic data.

#### *Control Samples*

We obtained a total of 1,184 quartets from the Simons Simplex Collection (SSC) (Fischbach and Lord, 2010). These quartets consist of an ASD proband, an unaffected sibling, and both unaffected parents. 602 of these quartets were used as controls in our Phase 1 study (Willsey et al., 2017) and 582 are new. 1,174 quartets passed quality control (see 'Quality control').

## METHOD DETAILS

### Whole exome sequencing

#### *Exome capture and sequencing*

We derived DNA samples for the Phase 2 trios (873 total samples from 291 trios) from a combination of whole blood (858 samples), lymphoblastoid cell lines (13 samples, 8 parental samples and 5 child samples), and saliva (2 samples, 1 parental sample and 1 child sample). We did not observe an excess of *de novo* variants in any of the non-blood samples (excess defined as $\geq$ 10 *de novo*s). We utilized the IDT xGen kit (https://www.idtdna.com/pages/products/next-generation-sequencing/hybridization-capture/lockdown-panels/xgen-exome-research-panel) to capture the exome and then performed whole exome sequencing (WES) with the Illumina Hiseq 4000 platform to 100 base pair long paired end reads. For the 511 trios previously characterized, DNA was derived from whole blood, exome capture was performed with three different capture arrays–Nimblegen EZ v2, Agilent v1.1, and Nimblegen EZ v3–and the exome was sequenced with the Illumina Hiseq 2500 platform (Willsey et al., 2017).

### *Control data for whole exome sequencing*

We obtained a total of 1,184 quartets from the SSC, which were previously captured on the Nimblegen EZ v2 array and sequenced with the Illumina Hiseq 2000 platform (Iossifov et al., 2012, 2014; Krumm et al., 2015; O'Roak et al., 2012; Sanders et al., 2012, 2015). All the WES control data were generated from blood-derived DNA. We summarized metadata and sequencing metrics from all TD trios and SSC control trios in Tables 1 and S1.

### *Variant calling pipeline summary*

We used GATK best practices to process all raw whole exome sequencing data across both Phase 1 and 2 (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). We aligned sequencing reads in FASTQ format to the GRCh37 build of the human reference genome with BWA-mem (Li and Durbin, 2009). For consistency, we reverted sequencing alignment data (BAM files) from SSC control families to FASTQ format and then processed identically. We sorted, indexed, and marked duplicate reads in the alignment files (BAM format) with Picard Tools, and then locally realigned reads containing indels with GATK's IndelRealigner tool. Next, we used GATK to perform Base Quality Score Recalibration using the training data recommended by GATK. We used the recalibrated alignment data produced by this step in all downstream analyses, including quality control. We produced variant calls in gVCF format for all samples with GATK HaplotypeCaller. Finally, we produced a list of joint recalibrated variant calls for the entire sample collection by running GATK GenotypeGVCFs followed by GATK's SNP and indel Variant Quality Score Recalibration steps.

### *Variant calling pipeline details*

Below are descriptions of specific software tools used to perform data processing, along with the runtime options that can be used to reproduce our work. Any arguments or options not specified either retain their default values or are system-specific. For example, file paths, memory allocation, and multithreading options are not included.

0) Sequencing data was acquired from SSC in BAM format. To transform this data into a fastq format so that alignments could be re-generated, and variants called, using the same methods as with the TD cohorts, we took the following steps:

- Randomly re-order alignments in BAM files with the Samtools bamshuf function
- Using Picard's RevertSam function, remove alignment information and restore original quality scores to reads. Options used: SORT_ORDER = unsorted, RESTORE_ORIGINAL_QUALITIES = true, VALIDATION_STRINGENCY = LENIENT
- Using Picard's SamToFastq function, convert BAM files to paired-end fastq format. Option used: VALIDATION_STRINGENCY = LENIENT

1) BWA
Tool: BWA 0.7.12
Runtime options:
mem –R [sample-specific header] [GRCh37 reference fasta file]

2) SAM sorting
Tool: Picard 2.1.1
Runtime options:
SortSam SO = coordinate
Note: In this version of Picard, specifying an output filename ending in ''.bam'' automatically compresses alignments into BAM format, which we did.

3) Duplicate marking and BAM index creation
Tool: Picard 2.1.1
Runtime options:
CREATE_INDEX = TRUE

4) Indel realignment
Tool: GATK 3.5
Training file, available online in the GATK Resource Bundle:
Mills_and_1000G_gold_standard.indels.b37.v*cf.* (aka ''Golden Indels'')
Runtime options (RealignerTargetCreator step):
- T RealignerTargetCreator
- –intervals [exome-capture-array-specific interval file]
- –interval_padding 100
- R [GRCh37 fasta reference]
- known [Golden Indels file]
- –filter_mismatching_base_and_quals

Runtime options (IndelRealigner step):
- T IndelRealigner
- R [GRCh37 fasta reference]
- –target-intervals [interval list created by RealignerTargetCreater)
- –filter_mismatching_base_and_quals

5) Base Quality Score Recalibration
Tool: GATK 3.5
Training files, available online in the GATK Resource Bundle:
- Mills_and_1000G_gold_standard.indels.b37.v*cf.* (aka ''Golden Indels'')
- dbsnp_138.b37.v*cf.*
Runtime options (BaseRecalibrator step):
- T BaseRecalibrator
- –intervals [exome-capture-array-specific interval file]
- –interval_padding 100
- R [GRCh37 fasta reference]
- knownSites Mills_and_1000G_gold_standard.indels.b37.v*cf.*
- knownSites dbsnp_138.b37.v*cf.*
Runtime options (PrintReads step):
- T PrintReads
- R [GRCh37 fasta reference]
- BQSR [recal data table from BaseRecalibrator step]
6) Variant calling
Tool: GATK 3.5
Runtime options:
- T HaplotypeCaller
- R [GRCh37 fasta reference]
- ERC GVCF
- variant_index_type LINEAR
- variant_index_parameter 128000
- –read_filter BadCigar
- an StrandOddsRatio -an AlleleBalanceBySample
- an DepthPerSampleHC -an MappingQualityZeroBySample
- an StrandBiasBySample -an GenotypeSummaries
7) Joint genotyping

First, gVCFs were combined into batches of around 50 samples each using GATK's CombineGVCFs tool. In order to speed up data processing, we created 13 separate combined GVCFs for each batch of samples for the following subsets of chromosomes: 1, 2, 3/21, 4/22, 5/19, 6/Y, 7/20/MT, 18/X, 8/17, 9/16, 10/15, 11/14, 12/13. (These subsets were chosen to have similar combined sizes to increase the efficiency of parallel processing.) For each of the 13 chromosome subsets, GATK's GenotypeGVCFs was run using all associated combined gVCF files as inputs. Then, GATK's CatVariants was used to combine the 13 separate joint VCF output files into one comprehensive joint VCF file before continuing to the final variant quality score recalibration steps.

Tool: GATK 3.5
Runtime options, CombineGVCFs step:
- T CombineGVCFs
- R [GRCh37 fasta reference]
- I [ ~50 gVCF samples]
- –intervals [run-specific chromosome subset (see above) ]
Runtime options, GenotypeGVCFs step:
- T GenotypeGVCFs
- R [GRCh37 fasta reference]
- I [all combined gVCF files corresponding to the current run's chromosome subset]
- –pedigree [Plink-style pedigree file including all samples/families]
- an InbreedingCoeff -an StrandOddsRatio -an BaseQualityRankSumTest
- an ChromosomeCounts -an Coverage -an FisherStrand
- an MappingQualityRankSumTest -an MappingQualityZero -an QualByDepth
- an RMSMappingQuality -an ReadPosRankSumTest -an VariantType
- an DepthPerAlleleBySample -an AlleleBalanceBySample
- an MappingQualityZeroBySample -an StrandBiasBySample
- an DepthPerSampleHC -an GenotypeSummaries
CatVariants invokation:
java -cp [GenomeAnalysisTK.jar] org.broadinstitute.gatk.tools.CatVariants
- R [GRCh37 fasta reference]
- V [.v*cf.*gz file] x13 (chromosome-specific joint VCFs)

- assumeSorted
- o [output file]

8) SNP variant quality score recalibration
Tool: GATK 3.5
Recalibration training files, available online in the GATK Resource Bundle:

1. hapmap_3.3.b37.v*cf.*
2. 1000G_omni2.5.b37.v*cf.*
3. 1000G_phase1.snps.high_confidence.b37.v*cf.*
4. dbsnp_138.b37.v*cf.*

Runtime options (VariantRecalibrator step):
- T VariantRecalibrator
- mode SNP
- R [GRCh37 reference fasta]
- resource:hapmap,known = false,training = true,truth = true,prior = 15.0 [recal training file 1]
- resource:omni,known = false,training = true,truth = true,prior = 12.0 [recal training file 2]
- resource:1000G,known = false,training = true,truth = false,prior = 10.0 [recal training file 3]
- resource:dbsnp,known = true,training = false,truth = false,prior = 2.0 [recal training file 4]
- an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR

Runtime options (ApplyRecalibration step):
- T ApplyRecalibration
- mode SNP
- R [GRCh37 reference fasta]
- tranchesFile [tranches output file from VariantRecalibrator]
- recalFile [recal output file from VariantRecalibrator]
- –ts_filter_level 99.5

9) Indel Variant Quality Score Recalibration
Tool used: GATK, version 3.5
Recalibration training files, available online in the GATK Resource Bundle:
- Mills_and_1000G_gold_standard.indels.b37.v*cf.* (aka "Golden Indels")
- dbsnp_138.b37.v*cf.*

Runtime options (VariantRecalibrator step):
- T VariantRecalibrator
- R [GRCh37 reference fasta]
- mode INDEL
- –maxGaussians 4
- resource:mills,known = false,training = true,truth = true,prior = 12.0 [golden indels file]
- resource:dbsnp,known = true,training = false,truth = false,prior = 2.0 [dbSNP 138 file]
- an QD -an FS -an SOR -an ReadPosRankSum -an MQRankSum

Runtime options (ApplyRecalibration step):
- T ApplyRecalibration
- mode INDEL
- R [GRCh37 reference fasta]
- tranchesFile [tranches output file from VariantRecalibrator]
- recalFile [recal output file from VariantRecalibrator]
- –ts_filter_level 90.0

## Quality Control

### *Pedigree Check*

We verified sample pedigree information by running PLINK (Purcell et al., 2007) on SNP-site genotype calls derived from our WES data. More specifically, we confirmed familial relationships and sex with an in-house 'family check' script (see "Quality control"). This script also checks for higher than expected relatedness across independent trios.

Among the TD data, 789 of the 802 trios remained after we removed 11 trios with unexpected familial relationships (e.g., proband not related to parents) and 2 trios that were sequenced in both Phase 1 and Phase 2 project (We only used the Phase 2 data for these two samples). We manually fixed annotation errors where possible (e.g., wrong sex indicated). We considered these 789 trios only in downstream analyses. We removed 10 of the 1,184 SSC control quartets due to unexpected familial relationships, leaving 1,174 quartets for downstream analysis.

### *Whole Exome Sequence Data Quality*

We used Picard Tools to obtain quality metrics related to target capture, sequencing, and alignment, and we ran GATK's DepthOfCoverage tool to measure coverage across the exome at base-pair resolution. We then identified the sites within each trio that had at least 20X coverage across all trio members (≥20X joint coverage). We performed principal components analysis (PCA) on these metrics to identify outliers. We treated any samples more than 3 standard deviations (SD) from the mean in any of the first four principal components as outliers and removed them from subsequent analyses.

When considering *de novo* sequence variants, the above QC removed 0 TD trios and 18 SSC sibling trios, leaving 789 TD trios and 1,156 SSC sibling trios. We further removed any samples with > 10 *de novo* sequence variants, leaving 777 TD trios (12 trios removed) and 1,153 SSC sibling trios (3 trios removed) for *de novo* sequence variant analysis.

Our methods differed slightly for *de novo* CNV analysis: we used SSC probands as positive controls to assess our pipeline. Therefore, we did quality control on complete quartets, which removed additional families due to probands failing quality control resulting in removal of the entire quartet. In total we removed 38 such quartets, leaving 1,136 SSC quartets for downstream analysis. We did not remove any samples with excess *de novo* sequence variants. Therefore, we included 789 TD trios and 1,136 SSC quartets in *de novo* CNV analyses (Table S1).

## Variant detection

### *De novo sequence variant detection*

We optimized our *de novo* variant calling pipeline by integrating the GATK genotype refinement workflow (GRW) (https://software.broadinstitute.org/gatk/documentation/article.php?id=4727). We re-estimated the genotype likelihood for each individual at each position by utilizing SNP information from 1000 Genomes project as well as pedigree information. We then marked variants with genotype quality (GQ) ≥ 20 and allele count (AC) < max(4, 0.1% samples) (i.e., variant is present in a max of 4 or 0.1% of samples) as putative *de novo* variants. After these standard GRW steps, we further applied several empirical error filters to remove false positives: (1) homozygous in father and mother with allele balance (AB) < 0.05; (2) heterozygous in child with AB between 0.3-0.7; (3) depth in all trio samples DP ≥ 20; (4) mapping quality: MQ ≥ 30; (5) allele frequency in cohort AF < 0.1%; (6) GQ ≥ 90 in child sample; (7) *de novo* mutation count ≤ 10 (See "Determine cutoff for de novo mutation per child"). Finally, we visualized all the *de novo* indels by IGV (Thorvaldsdóttir et al., 2013) to remove false positives. We considered the resulting set of *de novo* variants as 'high confidence' *de novo* variants.

To validate the new *de novo* calling pipeline, we compared the new *de novo* calls to those from the old pipeline (published in Willsey et al., (2017). For comparability, we ran both pipelines on the VCF file from the Phase 1 study, which did not undergo joint-genotyping with Phase 2 samples. We only used these *de novo* calls for pipeline optimization (the *de novo*s presented in the main text were derived from the VCF generated by joint genotyping across Phase 1 and 2 samples, and the SSC samples). The Phase 2 pipeline has increased sensitivity for *de novo* calling (520 total variants in the Phase 1 data, including 83 new *de novo* coding variants, and missing 17 previously called *de novo* coding variants; Table S2).

All results presented in the main text are derived from the VCF generated from joint genotyping across all the TD and SSC samples. This decreases the number of detected *de novo* coding variants (from 520 to the 466 total reported in the main text for Phase 1 samples), likely because rare variant detection may be penalized by joint calling across a large number of samples. Because of this, we included the confirmed *de novo* damaging variants from the Phase 1 study (Willsey et al., 2017), that were missed here, into gene discovery with TADA in order to increase yield. We did not use these variants for burden analyses.

### *Determining cutoff for* de novo *mutation per child*

The distribution of *de novo* mutations across individuals should theoretically follow a Poisson distribution. To determine the the cutoff for the *de novo* calling, we determined how well our observations fit a Poisson distribution under different cutoffs (Figure S2). To normalize the different capture regions in different sample sets, we only used mutations in consensus regions (See "Estimation of mutation rate per base pair" for definition). First, we called all the mutations with the error filters 1-6 in "De novo variant detection" and summarized the *de novo* counts for each individual. Then we fixed the cutoff from 1 to 20 mutations (filter #7). With each cutoff, we generated a *de novo* mutation list. We then summarized the *de novo* mutation counts in consensus regions and used the mutation rate per individual (# of *de novo* mutations / # of passed individuals) as the lambda of theoretical Poisson distribution. We then utilized lambda to generate a list of values based on Poisson distribution by *npos* function in R. The goodness-of-fit was performed by chisq.test in R to obtain the p value. We used the maximum cutoff that is not significantly different with theoretical distribution ($p > 0.05$) to increase the sensitivity of *de novo* calling.

### *Inherited sequence variants detection*

We annotated GRW-processed VCF files with ANNOVAR and then detected inherited variants in coding regions with an in-house script. As in the original TADA study (He et al., 2013), we considered three categories of inherited variants based on the genotypes of the trios: alternative homozygous (*0/1* x *0/1* - > *1/1*), transmitted heterozygous (*0/1* x *0/1*(*0/0*) - > *0/1*), and non-transmitted (non-transmitted: *0/1* x *0/1*(*0/0*) - > *0/0*). We then utilized informative genotypes to identify "paternal" and "maternal" transmitted mutations. We defined rare mutations as population allele frequency less than 0.1% in Exome Aggregation Consortium (ExAC) version 0.3 which contains ~65,000 whole exome allele frequency data (Lek et al., 2016). Relevant filters from *de novo* variant calling were applied in the inherited variant calling including:

1) heterozygous AB: 0.3-0.7; homozygous AB: < 0.05;
2) depth in all the trio samples: DP $\geq$ 20.
3) genotype quality (GQ) $\geq$ 20

#### *De novo CNV detection from WES data*

We identified *de novo* CNVs from whole exome sequencing data using CoNIFER and the standard workflow (Krumm et al., 2012). Briefly, we defined each continuous capture region as a 'probe', then calculated the RPKM (reads per thousand bases per million reads sequenced) for all the samples and transformed to standardized z-score (ZRPKM). We then implemented singular value decomposition (SVD) and made final calls based on SVD-ZRPKM values to help correct for biases arising from data generation. Post-CoNIFER, we merged neighbor CNVs if the distance between two adjacent CNVs was less than half of the larger one. Finally, we generated a list of high confidence *de novo* CNVs by implementing additional filtering criteria on raw *de novo* CNVs, including:

1) not detected in both parents (i.e., without any overlap).
2) less than 50% overlap with common CNVs (MacDonald et al., 2014).
3) less than 50% overlap with telomeric, centromeric and immunoglobulin regions.
4) covering more than 12 probes.
5) Manual visualization blinded to affected status.

### Microarray genotyping

#### *TD data*

We genotyped 412 trio samples with the HumanOmniExpressExome-8-v1 platform. We utilized GenomeStudio to generate high quality final reports according to a previously published protocol (Guo et al., 2014). Specifically, after loading and automatic clustering of the raw intensity data in GenomeStudio, we excluded all samples with less than 98% call rate. We then re-clustered the remaining samples and manually checked and adjusted the following terms:

1) excluded the abnormal SNPs in chrX, chrY, chrXY and chrMT;
2) adjusted clusters with low GenTrain scores (which means less than 0.7).

As a result, 3 samples with less than 98% call rate and all the SNPs with less than 95% call frequency were excluded. After these steps, we exported final reports as one file per sample.

#### *Control data*

We obtained final report level data (i.e., post Genome-Studio data) for 765 previously genotyped quartets from the SSC (Sanders et al., 2011, 2015) as control data. In each family, there are two unaffected parents, one affected proband and one unaffected sibling. Because the microarray platforms used in TD samples and SSC controls were different (Table S4), we further trimmed the SNPs that only existed in the dataset we used in TD samples to make the results comparable. Then we calculated the standard deviation (SD) of the LRR ratio for each remained SNP in all the passed samples. All the SNPs with SD of LRR > 0.5 were removed. This process resulted in 686,180 SNPs, which we used for quality control and CNV detection (Figure S3B).

#### *Quality control for genotyping data*

The quality controls of genotyping data consisted of three steps (Figures S3C and S3D): (1) We calculated the mean of LRR in chromosome X and Y for each individual and removed any individuals with abnormal sex karyotype; (2) we estimated the contamination of these samples by calculation of heterozygous ratio (heterozygous to total SNPs) and duplicate sites ratio (SNPs with BAF of 0.25-0.4 or 0.6-0.75 to total SNPs) in each individual and excluded any outliers defined by 2 standard deviations from the mean; and (3) we checked the pedigrees in each cohort with an in-house script based on PLINK and removed any failing families. We excluded 11 families from the TD dataset and 2 families from the SSC dataset, leaving 399 TD trios and 763 SSC quartets for *de novo* CNV calling. Among the 399 TD trios, 279 trios overlap with the WES data. Additionally, 35 of these 399 TD trios were studied previously on a different microarray platform (Fernandez et al., 2012).

#### *De novo CNV detection from genotyping data*

We detected *de novo* CNVs based in the families passing QC using PennCNV with an exome-specific Hidden Markov Model file (Szatkiewicz et al., 2013). After merging neighbor CNVs with PennCNV default settings, the output was further filtered by a series of criteria:

1) not detected in both parents (i.e., without any overlap);
2) less than 50% overlap with common CNVs (MacDonald et al., 2014);
3) less than 50% overlap with telomeric, centromeric and immunoglobulin regions;
4) covered more than 10 SNPs.
5) remove the samples with a) waviness factor > 0.055; b) SD of LRRs > 0.3; c) detected CNVs number > 10.

All the outputs were checked by visualization blind to affected status to obtain the final *de novo* CNV list.

## De novo variant validation

### *De novo sequence variants validation*

We attempted to validate all 309 Phase 2 *de novo* coding variants (including those that disrupt canonical splicing sites) through PCR and Sanger sequencing. We designed the PCR primers for these sites with a primer3-based web tool developed by our lab (https://primerdesign.willseylab.com). We generated an amplicon for each variant site using PCR from blood-derived DNA, if available (13 samples from lymphoblastoid cell line DNA, and 2 samples from saliva DNA). Due to failure in primer design, PCR reaction, and/or Sanger sequencing, we were unable to validate of 55 of 309 sites. For the remaining 254 sites, we validated 243 of them as true *de novo* mutations (95.7%), including 95% for SNVs (232/243) and 100% for indels (11/11).

For the 511 Phase 1 TD trios, our new methods captured the majority of *de novo* coding variants identified in our Phase 1 analysis (379 of 419, 40 variants missed). Within these variants, 286 of the 293 with validation data from Willsey et al., 2017 confirmed as true *de novo* variants (97.6%). We also identified 87 additional *de novo* variants. These differences are likely due to joint calling across a larger number of samples as well as the GRW workflow described above. Within these 87 new *de novo* variants in Phase 1 samples, we confirmed 31/37 (83.8%; 50 variants were not validated because of sample accessibility or difficulty in primer design). Hence, we estimate our Phase 1 confirmation rate as 96.1% (317 of 330 *de novo* variants). Overall, we therefore achieved a confirmation rate of 95.9% (560 of 584) across Phases 1 and 2.

### *De novo CNV validation*

We attempted to validate all detected *de novo* CNVs with qPCR. We aimed to design three primers for each candidate and to ensure primers did not overlap common SNPs and were not within repeat regions (https://genome.ucsc.edu/cgi-bin/hgTables). We checked fidelity for each pair of primers *in silico*. We used *TERT* and *ZNF423* as controls to calculate the copy number (Sanders et al., 2011). We were able to generate primers for 17 *de novo* CNV candidates in WES data (17/27) and 15 of them were validated as true *de novo* CNVs (15/17, 88.2%). With respect to the microarray data, we conducted validation in 11 of 13 *de novo* CNV candidates and 9 were confirmed (9/11, 81.8%). We did not explicitly validate *de novo* CNVs identified in SSC controls. However, based on published confirmation results for *de novo* CNVs identified in the SSC quartets (Sanders et al., 2012, 2015), we were able to assess the expected confirmation rate of 44 of the 56 *de novo* CNVs identified in this study. Overall, 43 of these 44 *de novo* CNVs (97.7%) were previously confirmed in SSC samples in WES data, including 31/32 *de novo* CNVs in SSC probands and 12/12 *de novo* CNVs in SSC siblings. We conducted a similar analysis for *de novo* CNVs detected from the SSC microarray genotyping data. Aside from 1 *de novo* CNV not attempted previously, all were confirmed as true *de novo*s (27/27 *de novo* CNVs in SSC probands and 9/9 *de novo* CNVs in SSC siblings). Together, these comparisons suggest high specificity in *de novo* CNV calling. Given the good performance of our *de novo* CNV calling pipeline on SSC WES data (97.7% confirmed) or microarray data (100% confirmed), it is unclear why we have a lower confirmation rate in the TD samples (WES data, 88.2%; microarray data, 81.8%), though this is perhaps due to different validation methods, different sequencing and genotyping platforms, and/or criteria used in our study and in Sanders et al., 2015.

## Burden analysis

### *Estimation of mutation rate per base pair*

Since the capture platforms varied across different cohorts, we defined high confidence ''consensus regions'' to minimize bias when comparing the mutation rate in cases versus controls. We obtained the consensus callable regions by conducting the following steps:

1) By Family: within RefSeq hg19 coding regions, we produced a list of regions that have $\geq$ 20X coverage in all members of the trio (to match the minimum joint coverage required in *de novo* variant calling).
2) By Cohort: we filtered to regions from (1) covered in at least 50% of the trios
3) Across all cohorts (TD and SSC): we intersected the lists from (2) to generate the consensus regions.

These steps resulted in a set of consensus regions spanning 19,343,430 bp. We restricted comparisons of *de novo* and transmitted mutation rates to these regions. More specifically, to estimate the mutation rate rate per base pair, we considered *de novo* mutations occurring in consensus regions only. We then calculated mutation rates per individual as *number of de novo mutations in consensus region*/*number of base pairs with* $\geq$ *20X joint coverage within the consensus regions*. We then further divided the mutation rate by two to account for the diploid genome. We obtained the mean as well as the 95% confidence interval (CI) of the *de novo* mutation rate for each sample set by using *t.test* in R. Finally, we estimated the theoretical *de novo* mutation number by multiplying the rate per base pair by the total size of the RefSeq hg19 coding region (33,828,798 bp).

### *De novo sequence variant burden analysis*

We compared the *de novo* mutation rate in cases versus controls by one-sided Poisson test in R (Willsey et al., 2017):

$$poisson.test(x, T, alternative = ''greater''),$$

where x is a vector of length two, containing the *de novo* mutation counts in cases and *de novo* mutation counts in controls (number of events). T is also a vector of length two, containing the sum of the number of base pairs with $\geq$ 20X joint coverage in the consensus regions across all TD trios and SSC control trios, respectively (number of opportunities). We also obtained the estimated rate ratio and also 95% CI from this function. We truncated the lower bound of the 95% CI to 0 if negative.

### De novo sequence variants burden analysis on mutation-intolerant genes

We also narrowed to mutation-intolerant genes and conducted burden analyses. We identified missense intolerant genes based on missense Z $\geq$ 3.891 and LGD intolerant genes based on pLI $\geq$ 0.9 (Lek et al., 2016). Again, we restricted these analyses to *de novo* variants in the consensus regions only. For each class of variants, we narrowed to the corresponding filtered list of intolerant genes to calculate the mutation rate (e.g., missense variants in missense intolerant genes, LGD variants in LGD intolerant genes). We then combined these two lists of genes (missense plus LGD variants) to calculate the nonsynonymous mutation rate in mutation-intolerant genes. As in the overall analyses, the t.test function in R estimated mean and 95% CI for mutation rates, and we compared sample sets with a one-sided rate ratio test.

### De novo CNV burden analysis

Within the WES data, we observed that the number of *de novo* CNVs is positively associated with the number of probes (i.e., discontinuous capture regions) on the respective capture arrays (Figure S3A). To address this issue, we normalized our *de novo* CNVs rate per individual by dividing by the number of probes on each capture platform. We estimated the mean and 95% CI of the normalized *de novo* CNV rate by t.test in R. Comparison was carried out by Wilcoxon rank-sum test (WRST) in R: wilcoxon.test(x, y, alternative = "greater"), where x, y are vectors containing normalized *de novo* CNV rates for cases and controls respectively. We did not use a one-sided rate ratio test as in the *de novo* sequence variant burden analyses because we could not determine if *de novo* CNV occurence follows a Poisson distribution. We then calculated the rate ratio (RR) of *de novo* CNVs in cases versus controls as:

$$\mathrm{RR} = \frac{(\textit{total number of de novo CNVs in cases}) * (\textit{total number of callable probes in controls})}{(\textit{total number of de novo CNVs in controls}) * (\textit{total number of callable probes in cases})}$$

For the microarray genotyping data, we used the *de novo* CNV rate directly for the burden analysis because we had already trimmed to a common set of high quality SNP sites prior to *de novo* CNV calling. We also estimated the mean and 95% CI of the normalized *de novo* CNV rate by t.test in R. We used the WRST for comparison: wilcoxon.test(x, y, alternative = "greater"), where x, y are vectors containing *de novo* CNV rates per individual for cases and controls respectively. We calculated the RR of *de novo* CNVs in probands versus siblings as:

$$\mathrm{RR} = \frac{(\textit{total number of de novo CNVs in cases}) * (\textit{total number of controls})}{(\textit{total number of de novo CNVs in controls}) * (\textit{total number of cases})}$$

Given that we have partial confirmation results and the confirmation rate is lower in TD cases, we conducted the same analysis only using confirmed *de novo* CNVs for the burden analysis. In the WES data we confirmed 15 *de novo* CNVs from TD trios and 12 *de novo* CNVs from SSC sibling trios. As a result, we estimated RR 1.93, p = 0.040 in TD probands versus SSC controls. Within the small sample size of confirmed *de novo* CNVs in the microarray data, we obtained RR 1.91, p = 0.13 using the same method.

### De novo CNV burden analysis using exact binomial test

In addition to the WRST method, we utilized the binomial exact test to to confirm the increased rate of *de novo* CNVs in TD probands. Specifically, we detected *de novo* CNVs from 789 TD trios and 26 of them carry *de novo* CNVs. In comparison, 16 of 1,136 SSC siblings carry *de novo* CNVs. Thus, we carried out the binomial exact test as below:

$$\textit{binom.test}(26, 789, p = 16/1136, \textit{alternative} = \textit{"greater"}),$$

This generated the p value as $8.37 \times 10^{-5}$ which is consistent with the result from WRST that *de novo* CNVs are significantly increased in TD versus SSC siblings in WES data. Using the same method, we estimated the p value equals to 0.0087 for microarray data, which further indicated the increased *de novo* CNV rate in TD versus SSC siblings.

We also checked with the exact binomial test in R if the increased rate of *de novo* CNVs could be still observed based on the validated *de novo* CNVs only. We estimated the p value as 0.016 for WES data and 0.15 for microarray data. Therefore, even restricted to confirmed *de novo* CNVs, our results generally suggested the *de novo* CNV rate is increased in TD probands compared with SSC siblings.

### Genomic architecture of TD risk factors

We estimated the percentage of TD probands with *de novo* events (sequence variants and/or CNVs) mediating risk, as well as the percentage of *de novo* events carrying TD risk based on the passing simplex trios with WES data (i.e., 577 trios TD trios, 1,134 SSC control trios). This allowed us to assess both types of variation in these individuals, in the same dataset.

To estimate the the percentage of TD probands with *de novo* events we counted individuals with one or more *de novo* events as one and marked the remaining individuals as 0. We thus calculated the percentage of individuals with *de novo* events in cases and controls as pCases and pControls respectively. The percentage of cases with *de novo* events mediating risk was calculated as: $pCases - pControls$, the 95% confidence interval were estimated by bootstrapping for 1000 replicates. We calculated these values for sequence variants alone, CNVs alone, and for any *de novo* event.

We determined the percentage of *de novo* events carrying TD risk for sequence variants and CNVs separately. For sequence variants, we calculated the theoretical rate per base pair as before for each individual in the consensus regions. And then the theoretical rate per child was obtained by multiplying the entire refseq coding size (33,828,798 bp). The difference and 95% CI between cases and controls was estimated by two samples t.test in R. We divided difference by the theoretical rate in cases to obtained the percentage of *de novo* events carrying TD risk. We generated the 95% CI for this using the upper and lower of the difference in the

same formula. For *de novo* CNVs, we cannot obtain the theoretical rate per person as did for sequence mutations. Thus we treated the normalized mutation rate as theoretical rate per child and used the same strategy as before to generate the percentage of *de novo* CNVs carrying TD risk as well as the 95% CI.

## Gene discovery

### *Estimation of Total TD risk Genes by MLE*

The detection of recurrent damaging mutations enabled us to estimate the number of TD risk genes with a previously established maximum likelihood estimation (MLE) procedure (Homsy et al., 2015; Willsey et al., 2017). We assessed the percentage of damaging mutations carrying TD risk (E) and did 50,000 permutations for every possible number of risk genes from 1 to 2500. For each permutation, we randomly generated 292 de *de novo* variants (the total number of damaging mutations identified in all the TD samples; see below). We also selected a certain number of risk genes according weighted by their respective mutation probability, which accounts for gene size and GC content (He et al., 2013). We then randomly assigned a percentage of the 292 variants to risk genes, based on (E), and the remaining percent as non-risk. In each iteration, we combined the risk genes and non-risk genes and checked whether the recurrent mutation count was consistent with what we observed in our study. We estimated E using all *de novo* damaging mutations identified in TD non-multiplex probands (probands from simplex [577 samples] and unknown [97 samples] trios, 674 total probands) and SSC controls (1,153 samples) in consensus regions to reduce the bias that plexity and different capture platform introduced. We did not use use the mutations from multiplex families due to their unclear risk (Figures 2 and S5). We calculated E as:

$$E = \frac{M1 - M2}{M1} = \frac{(199/674) - (258/1153)}{(199/674)} = 0.2421256.$$

Among the 292 *de novo* damaging mutations detected in non-multiplex families (we removed the 10 variants failing confirmation, thus remained 282 variants), we observed 6 genes with two recurrent variants and 2 genes with three recurrent variants. With these observations, we determined the MLE of TD risk genes to be 483 genes based on the frequency of occurrences versus possible TD risk gene number (Figure S6).

### *Identification of TD risk genes with TADA-denovo model*

We did not observe an increase in rare transmitted mutations in simplex TD trios compared with SSC control trios (Figure S5B). Therefore, we used the TADA *de novo* only model and all *de novo* damaging variants identified in the non-multiplex families to identify TD risk genes. We updated the following parameters from Phase 1:

1) Fraction of causal genes ($\pi$)

we used the new estimated risk gene number (483),

$$\pi = \frac{483\ risk\ genes}{17726\ refseq\ hg19\ genes} = 0.02724811$$

2) Fold-enrichment ($\lambda$) for Mis3 and LGD

Instead of using synonymous mutations as controls, we used poisson regression to control the effects of paternal age, sex bias, and consensus callable size. To reduce bias from different exome capture kits, we only used the results from Phase 1 Yale non-multiplex samples (281 trios) and Phase 1 and 2 SSC controls (1,153 trios) to estimate the fold-enrichment values since they were captured by the same exome capture platform. Additionally, as in our other analyses we restricted the regression to the consensus regions to further reduce batch effects. The formula for the regressions were:

$$Number\ of\ Mis3\ mutations \sim paternalAge + sex + affect\ status + offset(log\ 10(consensus\ callable\ size))$$

and

$$Number\ of\ LGD\ mutations \sim paternalAge + sex + affect\ status + offset(log\ 10(consensus\ callable\ size))$$

We estimated the $\lambda$ for Mis3 and LGD as 1.383366 and 2.492502 respectively.

3) Relative risk ($\gamma$) for Mis3 and LGD

$$\text{For Mis3: } \gamma = 1 + \frac{\lambda - 1}{\pi} = 1 + \frac{1.383366 - 1}{0.04129527} = 10.28354$$

$$\text{For LGD : } \gamma = 1 + \frac{\lambda - 1}{\pi} = 1 + \frac{2.492502 - 1}{0.04129527} = 37.1422$$

With these parameters, we ran TADA-Denovo to estimate the p value and q value (false discovery rate, FDR) for each gene (1000 permutations). We considered genes with recurrent variants (> 1 *de novo* variant) and $q \leq 0.3$ as true TD risk genes (Table 4), but see Table S7 for an exome-wide summary of p and q values (3 genes have $q \leq 0.3$ but only one *de novo* variant).

### *Prediction of the Number of Risk Genes Identified by Cohort Size*

We took advantage of the estimated TD risk gene number above to predict the gene discovery yield while additional TD trios are whole-exome sequenced. As done previously (Willsey et al., 2017), we fixed the gene number at 483 and varied the cohort size. As in the MLE above, we randomly selected the risk genes, and then assigned a fraction of them to TD risk gene and the remaining as non-TD risk genes. We permuted 10,000 iterations for each cohort size and generated LGD and Mis3 variants separately based on their mutation rate (He et al., 2013; Sanders et al., 2012). We then combined the permuted variants and ran the TADA *de novo* only algorithm using the same sample parameters as above to assess the per gene q value. Then we counted the number of probable genes ($q < 0.3$) and high confidence genes ($q < 0.1$) for each cohort size and plotted the smoothed trend line using ggplot in R ("loess" function). We predicted the number of genes identified in a particular cohort size by regression model.

We estimated the fractions of LGD and Mis3 variants carrying TD risk ($E_{LGD}$ and $E_{Mis3}$). We did not exclude the variants that failed in confirmation due to the lack of confirmation data in SSC controls. We only used *de novo* damaging mutations in non-multiplex families and restricted the mutations to the consensus regions as in the MLE section. Specifically, the observed rate of *de novo* LGD variants, M1 = 44/674 for TD probands while M2 = 39/1153 for controls. Therefore, $E_{LGD}$ = (M1-M2)/M1 = 0.482. For Mis3 mutations, M1 = 155/674 and M2 = 219/1153, and therefore $E_{Mis3}$ = (M1-M2)/M1 = 0.174.

### *Permutation test for the occurrence of compound heterozygous mutations in hcTD genes*

We first detected compound heterozygous mutations within any of the genes with one or more *de novo* damaging variants in TD probands. We only considered compound heterozygous where at least one allele is rare in the population (AF < 0.1% based on ExAC v0.3) and both are Mis3 or LGD mutations. In total, we identified 189 mutations. Since the q value of *OPA1* is very close to the cutoff of hcTD genes ($q \leq 0.1$) while only concerning non-multiplex families in TADA, we treated *OPA1* as a potential hcTD genes in this permutation test. To estimate the probability of observing recurrent compound heterozygous mutations in any of the three genes (*CELSR3, WWC1* or *OPA1*), we thus sampled exactly 189 genes genome-wide with replacement and weighted by the probability of a damaging mutation (He et al., 2013; Sanders et al., 2012). We defined a "success" as *CELSR3, WWC1* or *OPA1* appears more than twice in the generated gene list. We permuted the process 10,000 times and calculated the p value as the total number of success in permutations.

## Systems biological analyses

### *Comparison of* de novo *damaging variants in TD and other disorders*

To assess the overlap of genes affected by *de novo* mutations between TD and other disorders, we used a one-sided permutation test to estimate the significance of overlap. We focused on the *de novo* damaging mutations from OCD (Cappi et al., 2017), ASD (Sanders et al., 2015), congenital heart disease (Jin et al., 2017), intellectual disability (Gilissen et al., 2014; Hamdan et al., 2014; de Ligt et al., 2012; Rauch et al., 2012), schizophrenia (Fromer et al., 2014), epileptic encephalopathies (EuroEPINOMICS-RES Consortium et al., 2014), and developmental disorders in general (Deciphering Developmental Disorders Study, 2017). For each disorder, we randomly selected the number of unique genes with one or more detected *de novo* damaging variants, weighted by damaging mutation probability (He et al., 2013; Sanders et al., 2012) and compared, at the gene level, the permuted list with our observations in TD. We defined a success as the amount of overlap derived from permutation greater than or equal to our observation. We permuted 10,000 iterations to estimate the p value for each disorder. Then we iterated the permutation test for each disorder and obtained the p values respectively. We removed *de novo* damaging mutations that failed in confirmation and treated genes with more than one mutations as one. Since a proportion of TD probands in our sample set comorbid with OCD, we therefore also did the same permutation test using the genes detected from TD probands without comorbid with OCD only.

### *Comparison of* de novo *CNVs in TD and ASD*

To assess the overlap of *de novo* CNVs detected from TD probands and ASD probands, we utilized the results from a previously published study characterizing the Simons Simplex Collection (Sanders et al., 2015). We removed *de novo* CNVs with more than 50% overlap between ASD probands and siblings which likely carry lower risk. We further removed the telomeric, centromeric and immunoglobulin regions as did in our *de novo* CNV calling workflow. These filters resulted in 290 and 57 *de novo* CNVs in ASD probands and siblings respectively. Then we intersected either *de novo* CNV list from ASD study with the *de novo* CNVs detected from either WES data or microarray data under the cutoff as 50% using bedtools. We combined CNVs detected in the same TD probands from either WES data or microarray data prior to the intersection. We observed 9 *de novo* CNVs overlapping between ASD probands and TD probands. In comparison, we observed only 1 *de novo* CNV shared across ASD siblings and TD probands.

We then estimated the significance of this observation by permutation test. We randomly picked a region list according to the length of the given CNVs and the chromosome location. We avoided the telomeric, centromeric and immunoglobulin regions in the permutation. We permuted 10,000 times for ASD probands and siblings. We defined a "success" as the intersection between the permuted list with *de novo* CNVs in TD was greater than or equal to the observations (i.e., 9 for TD versus ASD probands and 1 for TD versus ASD siblings). We estimated the final p value as the rate of success in 10,000 permutations.

### *Enrichment of* de novo *mutations in cell polarity*

To check whether the *de novo* mutations are enriched in cell polarity, we extracted all the genes related to cell polarity from Gene Ontology (http://www.geneontology.org/) and annotated the *de novo* variants in TD and SSC controls. We utilized three methods to assess the enrichment.

First, we compared *de novo* damaging variants affecting cell polarity genes in cases and controls. In total, we observed 337 *de novo* damaging mutations in TD probands and 350 *de novo* damaging mutations in SSC controls. Among these, 17/337 and 7/350 affected cell polarity genes, respectively. We did not consider confirmation status here as variants in the SSC samples did not undergo validation.

| | TD probands | SSC controls |
|---|---|---|
| # of *de novo* damaging mutations | 337 | 350 |
| # of hits in cell polarity | 17 | 7 |

We then compared using fisher's exact test in R as:

$$fisher.test(matrix(c((17, 337-17, 7, 350-7), ncol=2)),\ alternative="greater"),$$

The results in an estimated odds ratio (OR) = 2.60 and p value = 0.024.

Second, we used a permutation test at the gene-level to assess the enrichment. We removed all the *de novo*s that failed confirmation. In total, we observed 327 genes with one or more *de novo* damaging mutations, across Phase 1 and 2. 15 of these are cell polarity genes. For each of 10,000 permutations, we randomly selected 327 genes without replacement according to the damaging mutation probability of each gene (He et al., 2013; Sanders et al., 2012). We tabulated how many of these were cell polarity genes. We defined success as $\geq$ 15 cell polarity genes. We calculated the p value as the total number of successes in the 10,000 iterations. We estimated the p value as 0.014.

## QUANTIFICATION AND STATISTICAL ANALYSIS

We conducted all statistical analyses in Python (v $\geq$ 3.6) and R (v $\geq$ 3.31). We have made the scripts used in these analyses available on bitbucket at https://bitbucket.org/willseylab/tourette_phase2/src/master/. Where appropriate, we present data as mean ± the 95% confidence interval (CI). We estimate mean and 95% CI with the t.test function. We describe the value of n in the main text and/or in Tables 2 and 3, and n stands for number of samples (trios), number of base pairs, or number of variants as indicated. We conducted the primary burden analyses for sequence variants with a rate ratio test, using the poisson.test function in R, and comparing, across two cohorts, the number of *de novo* variants per the number of callable base pairs assessed. We did burden analyses for copy number variants using Wilcoxon rank-sum test using wilcox.test function in R. When comparing TD probands versus SSC controls, we utilized a one-sided test (alternative = "greater"), given the prior evidence for the role of *de novo* sequence/copy number variants in TD and other neurodevelopmental disorders. However, we compared rates between TD cohorts with a two-sided test because we did not expect these rates to differ. In secondary burden analyses, one-sided binomial exact tests (binom.test in R) and Fisher's exact tests (fisher.test in R), as well as a Poisson regression in R (*glm* with *family = poisson*, *link = "log"*) (see "Determining cutoff for de novo mutation per child") also assessed significance.

We did not correct p values for multiple comparisons because our primary hypotheses focused on *de novo* damaging variants followed by secondary characterization of individual variant classes, and because we previously implicated *de novo* variants in TD (Willsey et al., 2017). We considered a p value < 0.05 statistically significant and we list individual p values in the main text, Figures 2 and 3, and Tables 3 and 4.

As described above in the STAR Methods, we estimated p- and q-values for individual association with TD risk with the algorithm, TADA, which is described in detail in (He et al., 2013).

## DATA AND SOFTWARE AVAILABILITY

### Data

We have deposited aligned whole exome sequencing data (.bam files) in the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) under BioProject: PRJNA384374 (https://www.ncbi.nlm.nih.gov/bioproject/384374) (TIC Genetics data) and BioProject: PRJNA384389 (https://www.ncbi.nlm.nih.gov/bioproject/384389) (TAAICG data). We have also deposited the microarray genotyping data (final report files) from the TIC Genetics cohort under BioProject: PRJNA384374.

### Software

Perl, Python, and R code used to process these data and complete statistical analyses are available on bitbucket at https://bitbucket.org/willseylab/tourette_phase2/src/master/. Our in-house primer design software that generated primer sets for variant confirmations is located at https://primerdesign.willseylab.com/.