An oligogenic risk-model for Gilles de la Tourette syndrome based on whole-genome sequencing data

Malgorzata Borczyk*[1], Jakub P Fichna*[2], Marcin Piechota[1], Sławomir Gołda[3], Michal Korostyński[1], Piotr Janik[4], Cezary Żekanowski[2]
*These authors contributed equally
1 - Laboratory of Pharmacogenomics, Department of Molecular Neuropharmacology, Maj Institute of Pharmacology, Polish Academy of Sciences, Krakow, Poland
2 - Laboratory of Neurogenetics, Mossakowski Medical Research Centre, Polish Academy of Sciences, Warsaw, Poland
3 - Department of Molecular Neuropharmacology, Maj Institute of Pharmacology, Polish Academy of Sciences, Krakow, Poland
4 - Department of Neurology, Warsaw Medical University, Warsaw, Poland
corresponding author: Malgorzata Borczyk, gosborcz@if-pan.krakow.pl

Abstract

Gilles de la Tourette syndrome (GTS) is a neurodevelopmental disorder from the spectrum of tic disorders (TDs). GTS and other TDs have a substantial genetic component with the heritability estimated at between 60 and 80%. Here we propose an oligogenic risk model of GTS and other TDs using whole-genome sequencing (WGS) data from a group of Polish GTS patients and their families (n=185). The model is based on the overrepresentation of putatively pathogenic coding and non-coding genetic variants in genes selected from a set of 86 genes previously suggested to be associated with GTS. Based on the variant overrepresentation (SKAT test results) between unrelated GTS patients and controls based on gnomAD database allele frequencies five genes (*HDC, CHADL, MAOA, NAA11,* and *PCDH10*) were selected for the risk model. Putatively pathogenic variants (n = 98) with the median allele frequency of ~0.04 in and near these genes were used to build an additive classifier which was then validated on the GTS patients and their families. This risk model successfully assigned individuals from 22 families to either healthy or GTS groups (AUC-ROC = 0.6, $p < 0.00001$). These results were additionally validated using the GTS GWAS data from the Psychiatric Genomic Consortium. To investigate the GTS genetics further we identified 32 genes from the list of 86 genes as candidate genes in 14 multiplex families, including *NEGR1* and *NRXN* with variants overrepresented in multiple families. WGS data allowed the construction of an oligogenic risk model of GTS based on possibly pathogenic variants likely contributing to the risk of GTS and TDs. The model includes putatively deleterious rare and non-coding variants in and near GTS candidate genes that may cooperatively contribute to GTS etiology and provides a novel approach to the analysis of clinical WGS data.

Introduction

Gilles de la Tourette syndrome (GTS) is a neurodevelopmental disorder characterized by multiple motor and at least one vocal and/or phonic tic which persists for longer than 12 months. The clinical phenotype of GTS belongs to the spectrum of tic disorders (TDs), a broad diagnostic category that includes chronic motor and chronic phonic tic disorder, provisional tic disorder, transient tic disorder, and unspecified tic disorder (Selvini et al., 2019). In about 90% of cases, GTS is accompanied by comorbid psychiatric disorders, including obsessive-compulsive disorder (OCD), attention-deficit and hyperactivity disorder (ADHD), autism spectrum disorder (ASD), affective disorders, anxiety disorders, impulse control disorders, and personality disorders (Robertson, 2000). ADHD is associated with ca. 46% of GTS cases, and OCD with ca. 43% of cases (Hirschtritt et al., 2015). GTS and other TDs typically begin in childhood (average age at onset 4-6 years) or, less commonly, in

adolescence. Most juvenile cases improve during adolescence and the prevalence of GTS in adulthood is about 20 times lower than in childhood (Bloch & Leckman, 2009; Knight et al., 2012). Males are more commonly affected, with a male to female ratio of 3-4 to 1 (Freeman et al., 2000). The GTS prevalence in the general pediatric population ranges from 0.3% to 1% (Scharf et al. 2015). Other, non-GTS TDs, including isolated tics, are more common than GTS and affect up to 5% of the general population, however, this estimation varies depending on the methodology used (Robertson, 2000).

GTS and other TDs have a substantial genetic component, but the development of the clinical phenotype is complex and may be influenced by environmental, prenatal and perinatal factors, hormonal disturbances, as well as psychosocial stressors (Gloor & Walitza, 2016; Yu et al., 2019). The estimated heritability of GTS and other TDs based on twin, family, and population studies falls between 60 and 80% (Davis et al., 2013; Mataix-Cols et al., 2015). Accordingly, candidate gene approach, linkage studies, and structural variant studies have shown that the genetic basis of GTS is heterogeneous and at least several genes have been implicated in the etiology of the disease (Pagliaroli et al., 2016; Qi et al., 2019). These genes are related to various processes and neurochemical pathways including dopaminergic, serotonergic, and histaminergic signaling (*DRD2, DRD4, SERT, HDC, DAT1, 5-HTTLPR*), as well as synapse development, remodeling and functioning (*SLITRK1, NLGN4, NRXN1*), differentiation of axons, cell adhesion (*CNTNAP2*), and mitochondrial activity (*IMMP2L*) (Table 2) (Georgitsi et al., 2016; Qi et al., 2019). Still, the exact involvement of these genes in GTS etiology remains unknown, and the Human Phenotype Ontology database lists only two genes *HDC* and *SLITRK*, as associated with GTS (Köhler et al., 2021). Recently, epigenetic mechanisms involving DNA methylation, histone acetylation, and gene regulation by non-coding RNAs have been proposed to mediate the impact of environmental factors on the genetic background of GTS (Qi et al., 2019).

It is now evident that no single gene is responsible for a large fraction of GTS cases, albeit rare variants with large effects have been considered causative in single GTS families (Castellan Baldan et al., 2014). Genome-wide association studies (GWAS) of common neuropsychiatric disorders show that GTS is genetically correlated with OCD, ADHD, and major depressive disorder (MDD), diseases known to have an overlapping and highly polygenic background (Abdulkadir et al., 2018; Lee et al., 2019). However, the GWAS-based polygenic risk scores of GTS and other TDs calculated to date explain only about 0.5% of the variance in tic presence (Abdulkadir et al., 2019; Yu et al., 2015). It is possible that common variants with low impact identified by GWAS contribute to the disease in a polygenic manner, while rare deleterious variants are confined to a small fraction of GTS families.

We propose that a substantial part of GTS cases could be explained by an oligogenic model which assumes a compound effect of multiple low- and medium-impact variants with varied population allele frequencies. This hypothesis is supported by exomic data showing that *de novo* damaging variants in approximately 400 genes contribute to GTS risk in 12% of clinical cases (Willsey et al., 2017). GWAS results suggest that as much as 21% of GTS heritability can be explained by genotypes with a MAF (minor allele frequency) between 0.001 and 0.05 (Davis et al., 2013). Recently, a whole-exome sequencing study indicated a role of the rare variant burden in 13 families with TD history (Cao et al., 2021). Still, the role of rare, and particularly non-coding variants, particularly with MAF < 0.001 remains largely unexplored for GTS and other TDs.

To verify the above conjecture we investigated whether an oligogenic additive model could be used to distinguish healthy individuals from GTS and other TDs subjects. Whole-genome sequencing (WGS) was used to analyze known and identify novel variants in genes previously indicated with varying levels of evidence, to be associated with GTS and other TDs (Table 2). The search window encompassed also 20,000 bp of both flanks of the gene to identify variants located both in the gene itself and in distant regulatory regions as well. The first phase of analysis was conducted to select candidate genes that best differentiate unrelated GTS patients (discovery group) from the general population. Variants in five such genes were then used to build a model assigning individuals from GTS risk families to a group with clinical symptoms (GTS and other TDs) or to a healthy group, with the AUC-ROC (area under the receiver operating characteristic curve) metric of 0.6.

This is, to our knowledge, the first statistical model using WGS data, including non-coding and rare variants, to predict GTS/TD risk. The model includes 98 variants putatively associated with GTS and other TDs risks, including 75 non-coding variants. The obtained results were additionally validated with data from a large GTS GWAS study from the Psychiatric Genomic Consortium (Sullivan et al., 2018) Moreover, based on data from 14 sequenced multiplex families we propose other candidate genes from the preselected list with variants likely contributing to family-specific GTS risk. Overall, our results provide novel insights into the genetic architecture of GTS and other TDs and present a novel approach to WGS-based analysis of complex genetic disorders.

Methods

Ethics statement

The study was approved by the Ethics Committee of the Medical University of Warsaw (KB/2/2007, KB/53/A/2010, KB/63/A/2018) and has therefore been

performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All participants or their legal representatives gave written consent prior to inclusion in the study.

Patient selection and diagnostic criteria

All patients were recruited from a single Outpatient Clinic and were personally reviewed and evaluated by the same clinician well-experienced in tic disorders (PJ). The patients had been referred to the Clinic by a general neurologist and psychiatrist due to problematic diagnosis or tics refractory to treatment or sought medical advice on their own because of troublesome tics. The study was designed as a one-time registration study, as patients were registered in the database only once, and no additional clinical data obtained in follow-up visits were included in the analysis. In case of a positive family history DNA was collected from affected relatives and healthy members of the proband's family.

The patients were evaluated for GTS and other TDs according to the Diagnostic and Statistical Manual of Mental Disorders criteria valid at the time of evaluation (DSM-IV-TR, DSM-5). All patients were systematically interviewed with the aid of a semi-structured interview consisting of demographic and clinical data. This schedule was based on the TIC (Tourette syndrome International database Consortium) Data Entry Form developed by Freeman et al. (2000), in which study the investigator (PJ) participated and subsequently used this form in clinical practice.

The prevalence of the most common comorbid disorders encountered in GTS was evaluated on the basis of the above semi-structured clinical interview. The disorders listed in this interview included: ADHD, OCD, depression, anxiety disorder (different forms: phobias, panic disorders, generalized anxiety disorder, and separation anxiety disorder), Oppositional Defiant Disorder, Conduct Disorder, and Autism Spectrum Disorder. The list of obsessions and compulsions included in the Yale-Brown Obsessive Compulsive Scale (Y-BOCS) was used to establish the clinical spectrum of OCD. All patients were questioned thoroughly about all the symptoms included in the DSM as the diagnostic criteria of comorbid disorders mentioned above. Diagnoses of mental disorders from psychiatric clinics established before our evaluation was accepted and included in the analysis. The diagnosis of Autism Spectrum Disorder was made in specialized clinics. Children and adolescents with more complex psychopathology were assessed with the M.I.N.I. International Neuropsychiatric Interview for Children and Adolescents. Patients with severe psychiatric comorbidities were referred to a psychiatrist to confirm the diagnosis.

Tic severity was measured using the Yale Global Tic Severity Scale (YGTSS) on the day of a patient's evaluation (Leckman et al., 1989). An additional diagnosis of severe GTS in adult patients was also assessed. In contrast to children and

adolescents, for whom most of the information was provided by their parents, adults reported the symptoms themselves. Tics were qualified as severe if at any time they disrupted normal daily activities (e.g. led to repeating grades at school, job loss, or physical injuries), caused a significant deterioration of life quality, or required pharmacological therapy. Tics were also qualified as severe if the Yale Global Tic Severity Scale-Total Tic Score (YGTSS-TTS) was ≥ 35 points (range: 0-50) at the time of clinical evaluation (Leckman et al., 1998).

Overall, each patient was assigned to one of the following groups: GTS, non-GTS TDs, or healthy controls. In the GTS group, an additional subgroup of severe tics was distinguished according to the criteria described above. In the non-GTS TDs, a subgroup of non-chronic tic disorders (transient tics) was also distinguished and included patients with tic disorders that lasted less than 12 months.

Whole-genome sequencing

Genomic DNA of 186 patients (one sample was excluded during quality control, see details below) was extracted from peripheral blood leukocytes using a standard salting-out procedure (Miller et al. 1988) or from saliva collected with the Oragene DNA Self Collection Kit and using the Prep IT L2P Purification Kit (DNA Genotek Inc., Ottawa, Ontario, Canada). Whole-genome sequencing (WGS) was performed by Novogene (Beijing, China) according to the following protocol: sequencing libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, the UK) following the manufacturer's recommendations. Genomic DNA was randomly fragmented to 350 bp on average with a Bioruptor and DNA fragments were size-selected with sample purification beads. The selected fragments were then end-polished, A-tailed, and ligated with a full-length adapter. The fragments were then filtered with the beads again. Finally, the libraries were analyzed for size distribution on an Agilent 2100 Bioanalyzer, quantified using real-time PCR, and paired-end sequenced on an Illumina high-throughput HiSeq X Ten sequencer.

WGS data preprocessing

Fastq files were processed with the Intelliseq Germline Pipeline (https://gitlab.com/intelliseq/workflows) built with Cromwell (https://cromwell.readthedocs.io/en/stable/). The fastq files were assessed for quality with FastQC. The files were then aligned to the Broad Institute Hg38 Human Reference Genome with GATK 4.0.3. Duplicate reads were removed with Picard and base quality Phred scores were recalibrated using the GATK covariance recalibration. Variants were called with the GATK HaplotypeCaller to give genomic

variant calling files (gvcf) The gvcf files were then subjected to joint genotyping with GATK via GenomicsDB to generate a joint vcf file.

WGS data filtering and annotation

All analyses described below were performed with Hail (0.2.30, https://hail.is/). Briefly, the vcf file with all alternative allele calls in the study group was filtered to exclude repeated and low-quality sequences (UCSC RepeatMasker track). Multiallelic variants were split. Only loci with more than 90% gnomAD (v3) samples with a DP > 1 were kept for further analysis. Variants were annotated with gnomAD (v3; https://gnomad.broadinstitute.org/) (Karczewski et al., 2020), combined annotation dependent depletion (CADD) scores (Rentzsch et al., 2019), and human phenotype ontology (HPO) phenotypes (Köhler et al., 2021). A preliminary principal component analysis (PCA) of 0.01% of the genotypes revealed one sample (sequenced n = 186) that deviated by more than six standard deviations (SDs) on the first PC and this sample was excluded from the analysis, thus 185 samples were included in the study. Detailed information about file preparation, annotation, filtering and also a full analysis code are available at the GitHub repository (https://github.com/ippas/imdik-zekanowski-gts).

GnomAD controls

Non-Finnish European population minor allele frequency from gnomAD (v3; https://gnomad.broadinstitute.org/) was used to simulate presumably healthy controls termed *gnomAD controls* (Karczewski et al., 2020). The sex distribution of the simulated samples reflected the sex distribution in the study group. The Y chromosome was excluded from the analysis and simulated males were assigned homozygous genotypes in non-PAR X regions. Each locus was drawn independently (n = 40 controls). Generated vcf was then merged with the study group vcf. PCA of 0.01% of all genotypes in this merged vcf was used to determine the difference in the genetic background between the sequenced study group and the controls.

Gene set selection

Gene variants potentially associated with GTS in our patients vs the general non-Finnish European population were analyzed for 86 genes previously implicated in GTS (Table 2). Sequence kernel association test (SKAT) was performed in the logistic mode to investigate variant overrepresentation aggregated per gene (Wu et al., 2011) in GTS patients outside families (discovery group, n = 40) with 40 sex-matched gnomAD controls. CADD scores were used as weights for the SKAT test. Regions of 20,000 bp flanking the coding sequence were considered associated with the gene. The SKAT test accounts for the contribution of both common and rare

variants and SNP-SNP interactions. The first nine principal components (PCs) were used as covariates in the SKAT test.

Classifier

Between two and five top genes from the SKAT test were taken for oligogenic risk models of GTS. A total of 16 classifiers were tested. These differed in the number of genes: 2, 3, 4, or 5 and the threshold of CADD scores included: 5, 10, 15 or 20. For each gene and sample, the number of variants in the gene itself and in the flanking regions (+/- 20,000 bp) that passed above the chosen CADD threshold was summed. The model was run on the test group comprising individuals from 22 families. The AUC-ROC of the classifiers was calculated based on the sliding of the threshold of the difference between the calculated number of variants in the sample and in the gnomAD controls. To estimate type I error we used a Monte Carlo feature selection approach. In this approach features to build models were selected randomly in each iteration and AUC was calculated. The AUC for the original model was evaluated using the distribution of AUC values of random models. Random gene sets were selected from a list of human protein-coding genes that are assigned to at least one Gene Ontology term to be comparable with the selected gene set. The procedure for the random genes included: 1) determining the difference in the number of CADD > 5/10/15 or 20 variants within the gene itself and in the 5' and 3' flanks of 2/3/4 or 5 genes between the gnomAD controls and the discovery group, 2) running of the classifier on the test group. The procedure was repeated for 30 distinct gene sets giving 480 different classifiers of which two classifiers failed to reach 100% of false or true positives and were excluded from further calculations. The 478 classifiers remaining were used for the evaluation of the *p* value of the selected model.

Detailed variant investigation and GWAS validation

All 98 variants included in the risk model were further characterized regarding their position in relation to the gene, predicted molecular consequences, and predicted pathogenicity. Additionally, GTS GWAS data were obtained from the Psychiatric Genomics Consortium (Yu et al., 2019) (https://www.med.unc.edu/pgc/download-results/). The SNP coordinates were translated from Hg37 to Hg38 coordinates with the Hail liftover function. SNPs within the classifier genes and in 20,000 bp flanks of the genes were investigated in terms of their *p* value from GWAS. Random gene sets were used to determine whether these *p* values showed a distribution different from that for the whole genome (Yu et al., 2019).

Candidate gene investigation in multiplex families

To identify family-specific GTS/TD candidate genes, 14 multiplex families with five or more members, at least two individuals with GTS or other TDs and at least two healthy individuals were investigated. Variants were filtered according to the following criteria: location: within 20,000 bp from one of the investigated 86 genes (Table 2), MAF in the non-Finnish European population (Karczewski et al., 2020): < 0.05; CADD scores: > 10; AF in healthy controls: <0.2; AF in GTS and other TDs individuals: > 0.7. Such variants were identified in 9 of the 14 families.

Results

Description of the study group

The study group comprised GTS patients and their families from the Polish population. The family members were diagnosed towards GTS and other TDs. Each subject was assigned to one of the following sub-groups: healthy, GTS, and non-GTS TDs (transient tics, chronic motor or vocal tics, isolated tics, unclassified tics). Out of 185 subjects in the study group, 40 were unrelated individuals, including 38 adults diagnosed with GTS with severe tics. The remaining subjects (n=145) came from 22 families with 3 to 14 members each (median 6 members). Overall, 58 out of 185 subjects included in the study were diagnosed with comorbidities, including 11 with ASD and 26 with OCD. Table 1 presents the distribution of basic and additional diagnoses.

Table 1. General characteristics of study group. GTS - Gilles de la Tourette syndrome, non-GTS TD - tic disorder other than GTS, ASD - Autism Spectrum Disorder, OCD - Obsessive-Compulsive Disorder, ADHD - Attention-Deficit/Hyperactivity Disorder

diagnosis	number of patients	sex	additional details	comorbidities
discovery group (unrelated patients)				
GTS	**40**	6 female, 34 male	includes 38 adults with severe tics	3 ASD, 19 OCD, 22 Depression, 16 ADHD, 24 Anxiety Disorder
test group (families)				
GTS	**47**	11 female, 36 male	includes 3 adults with severe tics	6 ASD, 3 OCD, 2 Depression, 4 ADHD, 8 Anxiety Disorder
non-GTS TD	**42**	15 female 27 male	includes 17 with transient tics	1 ASD, 4 OCD
healthy controls without tics	**56**	35 female 21 male		1 ASD, 1 Depression

Comparison between study group sand gnomAD controls

To test whether the genetic background of the gnomAD controls is consistent with that for the Polish study group and to assess the homogeneity we conducted a principal component analysis (PCA) of a subset (~10,000) of genotypes for the study group (n = 186) and the gnomAD controls (n = 40) (Figure 1, Supplementary Figure S1). Standard deviations of the first two PCs were 0.12 and 0.11 and, after initial exclusion of one sample, no samples deviated by more than 6 SDs and the largest families dominated the largest variance. The gnomAD controls formed a uniform cluster only slightly deviating from the study group, which confirms a lack of substantial differences in the genetic background between the study group and non-Finnish Europeans (median of the first two PCs for controls [0.01; -0.01], for the study group [-0.03; -0.02]).

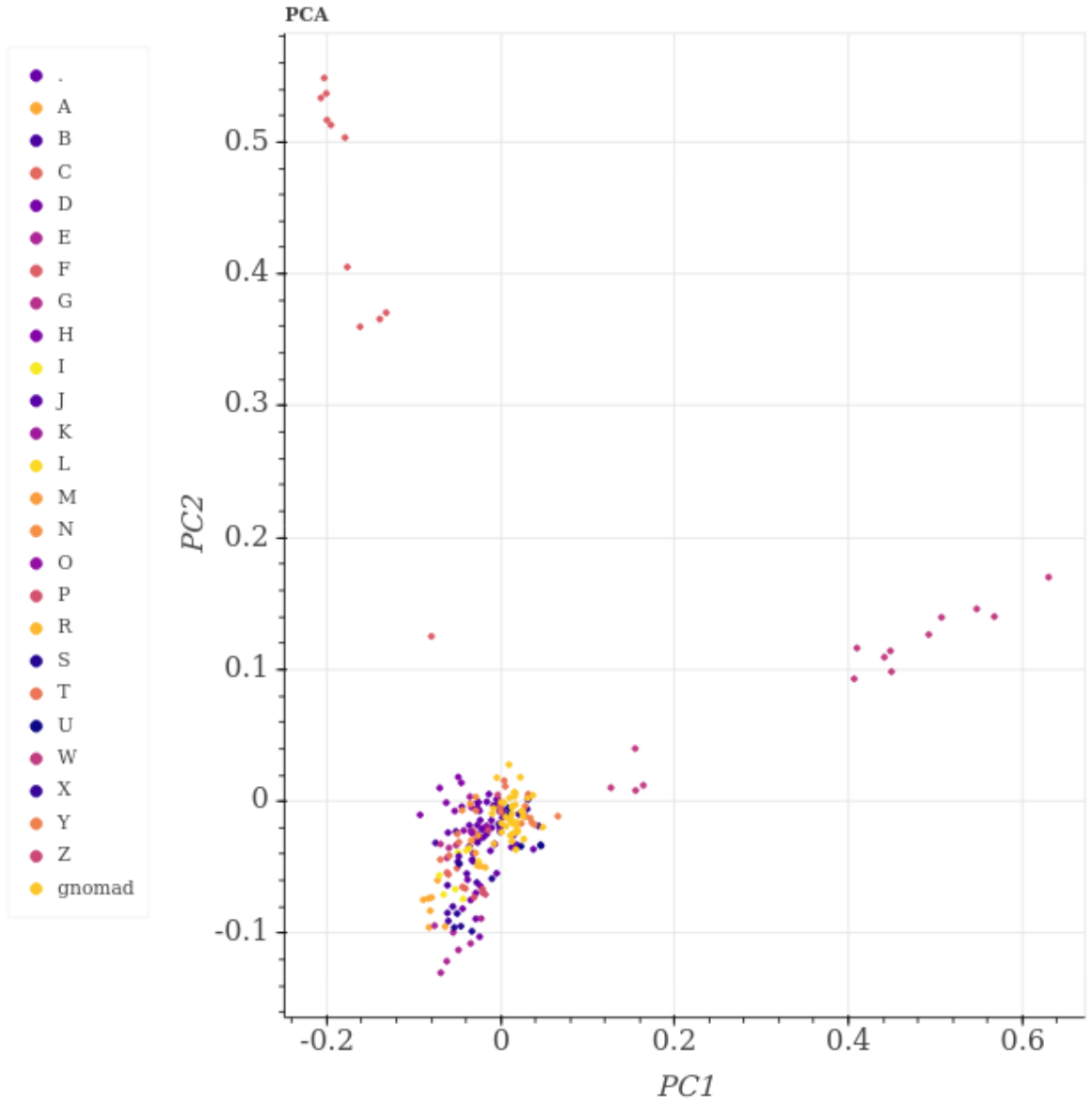


Figure 1. Principal component analysis of study group and gnomAD controls. A - Z study group families. – study group individuals outside families, gnomAD - controls. PCA was calculated based on ~7,000 genotypes (0.01%; after preliminary filtering). Further PCA plots up to the 10th PC are presented as supplementary figure S1.

Selection of genes for the risk model

In the first step, we attempted to identify the best candidate genes to construct an oligogenic model of GTS. Towards this end, we performed a sequence kernel association test (SKAT), a variant burden test, on variants within or in the vicinity of a preselected list of genes previously proposed to be linked with GTS. Combined annotation-dependent depletion (CADD) scores predicting the pathogenicity of both coding and non-coding variants were used as weights for the SKAT. The higher the

predicted pathogenicity of the variant the larger was its weight. CADD leverages a range of variant annotations (including SIFT and PolyPhen) and, importantly, provides scores for non-coding variants as well (Rentzsch et al., 2019). The list of genes (n=86) was selected from diverse genetic studies on GTS (association studies, linkage analysis, GWAS, WES/WGS) offering a varied quality of evidence (Table 2).

Sequencing of the study group revealed a total of 130 432 variants in and around these genes, including 6483 with CADD scores > 10 and 434 with CADD scores > 25. Of those, 718 variants were in exons, 77100 in introns, 1525 in UTRs, the rest were located in intergenic regions. Summary statistics of the allele frequencies in each genotyped locus including 20,000-bp 5' and 3' flanks is shown in Supplementary Material (Supplementary Table S1).

Table 2. **List of genes potentially related to GTS used for oligogenic models of GTS.** Gene symbols in bold font indicate genes that were included in the final developed model.

Genes	Method	reference
DRD2, DRD4, DAT1, DBH, ***MAOA****, HTR1A, HTR2C, HTR2A, SERT (SLC6A4), 5-HTTLPR, TPH2, EAAT1*	literature review of candidate gene studies	Georgitsi et al. (2016)
SLITRK1, SLITRK2, SLITRK3, SLITRK4, SLITRK5, SLITRK6, IMMP2L, CNTNAP2, NLGN4	literature review of chromosomal aberration studies	
NRXN1, AADAC, CTNNA3, FSCB, KCNE1, KCNE2, RCAN1, COL8A1, CHD8, SCUBE1	literature review of CNV research	
HDC	literature review of linkage studies	
COL27A1, NTN4	literature review of GWAS studies	
DCC, RBFOX1, SLC30A9, DCAF4L1,	meta-GWAS of multiple psychiatric disorders;	Lee et al. (2019)

SORCS3, KCNQ5, KCNQ-IT1, APOPT1, C14orf2, ***NAA11****, NEGR1,* ***CHADL****, SOX5*	genes near SNPs showing>0.85 probability of association with GTS	
RICTOR	WES of 9 trios	Eriguchi et al. (2017)
WWC1, CELSR3, OPA1, NIPBL, FN1, FBN2	combined WES of 802 trios	Wang et al. (2018); Willsey et al. (2017)
PNKD	linkage analysis in a multiplex family	Sun et al. (2018)
OPRK1, ***PCDH10****, NTSR2*	candidate gene study of rare variants in opioid-related genes	Depienne et al. (2019)
CDH26, CADM2, OPCML, CDH9, NCAM2, CD47, CDH5, CADM4, C1QBP, CTTN, LSAMP, PKP4, PCDH1, CNTNAP2, MBP, GABBR2, GABBR2, GRIK4, NCR1, FLT3, IL12A, HDAC9, CD180, CDH26, NCAM2, NTM, ROBO2	analysis of large GWAS based on preselected gene lists	Tsetsos et al. (2021)

In order to best leverage the study group composition, GnomAD controls (n = 40) and unrelated individuals diagnosed with GTS (discovery group, n = 40) were compared in this part of the analysis. The qq-plot of the relationship between the expected distribution of *p values* against the observed one showed some deviation from the expected distribution of *p values* (genomic inflation factor lambda = 1.39) (Figure 2). This is not unexpected as all of the selected genes could play a role in GTS etiology. Out of the 86 genes included in the SKAT test, none remained significant after Bonferroni correction for multiple comparisons (top gene *p* value: 0.0051 * 86 genes = 0.44, Table 3). However, this test still allowed us to identify top candidate genes in which the variant burden could differentiate between the GTS patients and controls.

Table 3. Selection of genes with possible overrepresentation of rare, damaging variants in unrelated GTS patients. Results of the SKAT test (unrelated GTS vs GnomAD controls). Top 20 genes are sorted according to ascending *p* value.

gene	number of variants	q stat	*p* value
CHADL	98	995	0.0051
HDC	120	511	0.0052
MAOA	97	960	0.014
NAA11	174	785	0.024
PCDH10	243	905	0.029
FSCB	68	317	0.031
OPRK1	199	427	0.034
RICTOR	269	695	0.043
SLC30A9	240	752	0.047
PKP4	702	354	0.048
SLITRK2	81	335	0.098
IL12A	94	441	0.12
PNKD	223	868	0.13
GABBR2	949	232	0.16
CDH26	197	166	0.18
SLITRK4	84	617	0.19
DCAF4L1	87	134	0.21
DRD2	224	616	0.22
CD47	180	431	0.23
IMMP2L	1319	289	0.26

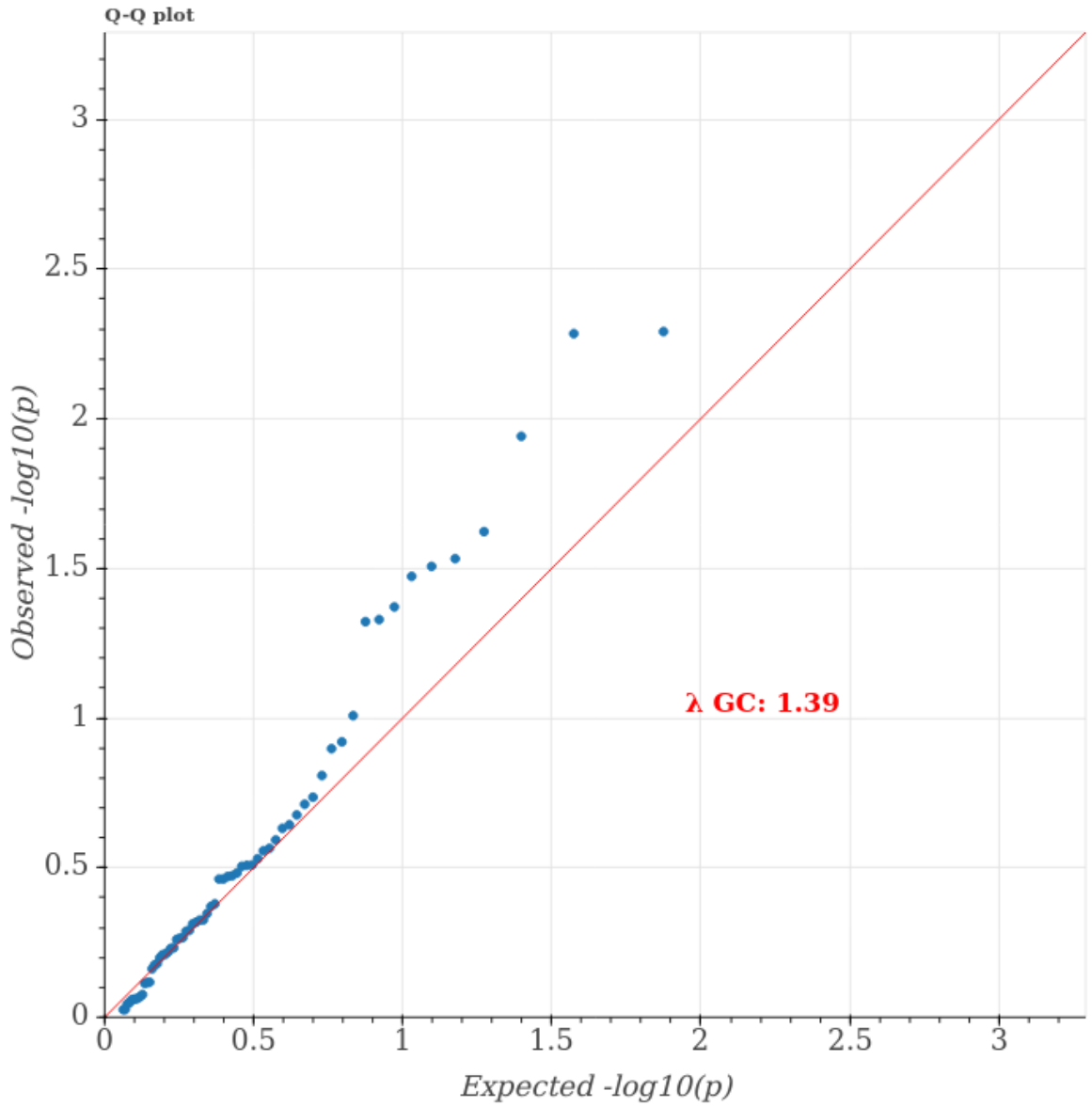


Figure 2. Comparison of observed and expected *p* values for associations between rare variants in 86 candidate genes and GTS phenotype. Quantile-quantile plot of expected and observed *p* values from SKAT test results of tested 86 GTS-candidate genes. x-axis - expected *p* values expressed as $-\log_{10}$, y-axis - observed *p* values expressed as $-\log_{10}$. Inflation factor lambda (1.39) is indicated in red.

Creation and evaluation of GTS risk model

Next, classifiers based on the SKAT results and tested on individuals from sequenced families were created. As CADD scores were used for the SKAT test, the same scores were used to create the classifiers. We applied a range of CADD thresholds as cut-off values for variant inclusion (5,10,15 or 20) and included

between two and five from among the top 20 genes (Figure 3A). These cut-off criteria were selected in order for the tested models to contain an optimal number of variants (within the range of tens to hundreds) for the model to be both oligogenic and interpretable. The difference in the mean number of non-reference calls between the discovery GTS group and gnomAD controls for each gene was summed into a score that assigned each sample to either the GTS or control group. Individuals with a non-GTS TD were considered correctly assigned if they were assigned to the GTS group. The best classifier obtained was that with the CADD cut-off of 10 and top five genes (*CHADL*, *HDC*, *MAOA*, *NAA11, PCDH10*). Table 4 shows the percentage of correct assignments of patients and their family members for two selected points of the classifier, A and B in Figure 3A. To gauge the specificity and sensitivity of the model, Area Under the Curve of the Receiver Operator Characteristics (AUC ROC) was calculated. The result obtained for all family members was ~0.60 (0.597). Since these individuals were not used to create the classifier and came from families and therefore were more genetically related than a random sample, one may conclude that the classifier used was fairly efficient incorrectly predicting their GTS status. To test the specificity of this classifier further we compared it with 478 control classifiers based on random sets of 2 to 5 protein-coding genes in a Monte Carlo approach (Figure 3B). The 99th percentile of the AUC of these random classifiers was ~0.57 (0.569) and the 99.999th ~0.59 (0.589). Thus the best classifier performed better than 0.99999 of the random ones and can be assigned a p value < 0.00001. Additional validation was performed against classifiers based on two other preselected gene lists: a list of all human protein-coding genes and a list of genes preferentially expressed in the human brain (Supplementary Figure S2). None of these classifiers reached an AUC > ~0.6.

Table 4. Assignment of family members to GTS or non-GTS groups. Two points of the classifier presented in Fig. 3A were used to assign individual family members to either group. Point A is the most optimal point. This table provides % of correctly classified individuals. Patients with tics were considered correctly assigned if they were assigned to the GTS group.

Family members	correct assignment at point A (%)	correct assignment at point B (%)
healthy (n = 56)	55%	64%
GTS (n = 47)	70%	59%
non-GTS TDs (n = 42)	60% (52% for transient tics)	47% (35% for transient tics)

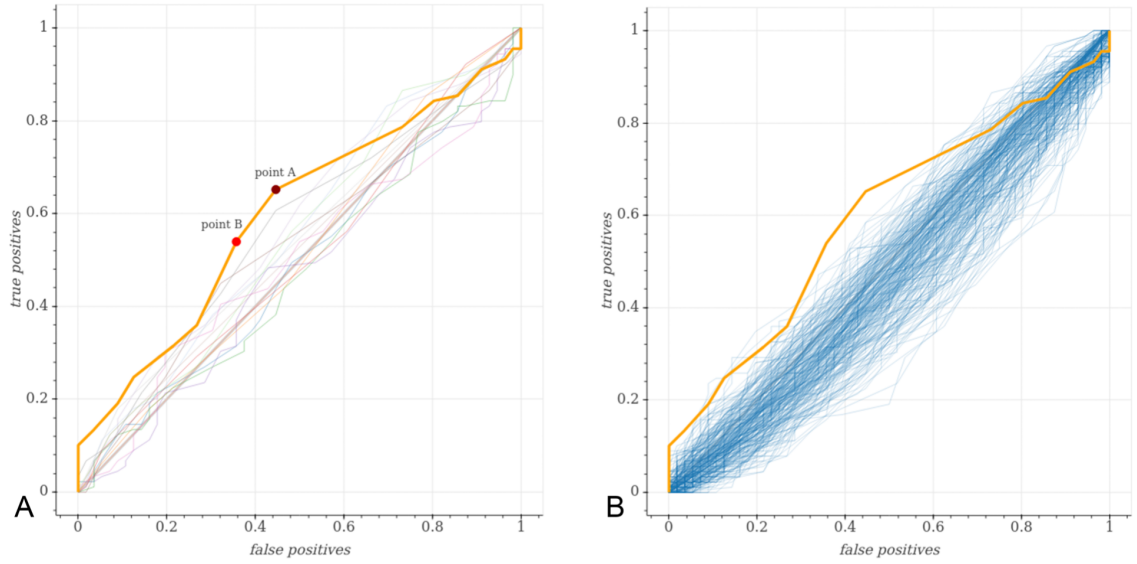


Figure 3. Comparison of the performance of tested classifiers of GTS risk. Graphs show ROCs of the chosen GTS risk model (orange line) and other tested classifiers. **A)** Comparison of diverse classifiers based on top genes considered with different CADD thresholds (5,10,15,20) and a different number of genes (2 - 5). The thick orange line represents the best classifier taken as a GTS risk model, with CADD threshold of 10 and five genes. Points A and B have the highest percentage of correct assignments of which details are provided in Table 4. **B)** Best classifier from A compared with 478 control classifiers (blue lines) based on sets of random genes. x-axes: fraction of false-positive results; y-axes: fraction of true positive results.

Analysis of gene variants included in the GTS risk model

The best classifier, with AUC-ROC ~0.6, was based on 98 variants (CADD score >10), differentially distributed across five genes: *CHADL* (12 variants), *HDC* (15 variants), *MAOA* (3 variants), *NAA11* (23 variants), and *PCDH10* (45 variants). Out of these variants, 11 were in the coding sequences (six missense and five same-sense), 48 in introns, 12 in 3' or 5' UTRs, and 27 further in the flanks (Figure 4A). Characterization of the variants revealed that their median CADD score was 13.7 (Figure 4B). The median allele frequency of these variants in the non-Finnish European population according to the gnomAD database (Karczewski et al., 2020) was 0.034 and the largest bin (30 variants) in the distribution histogram comprised variants of frequencies between 0 and 0.01 (Figure 4C). Nine variants had allele frequencies below 0.0001. A detailed description of all variants is provided in Supplementary Table S2.

At the optimal point, the classifier assigned a sample to the GTS and other TDs group when it had 21 or more putatively pathogenic variants amongst the tested genes. Figure 4D shows such an assignment for two families, the largest family in the dataset (family W, with 14 members) and a family with 8 members (family Y). For family W the false positive rate was 33% (2/6) and the true positive rate 75% (6/8), for family Y 50% and 83%, respectively. In both families individuals with GTS and other TDs have, on average, more putatively pathogenic variants from the list included in the best classifier (98 variants). In family Y this overrepresentation concerned *NAA1* (an average of 5.5 variants in healthy individuals vs 6.25 in GTS and other TDs), *MAOA* (0.5 vs 0.625), and *PCDH10* (7.16 vs 9.875). All the overrepresented variants were non-coding. In the second family, the variant overrepresentation was only found in *CHADL* variants (4 vs 5.67). A detailed analysis of variant distribution among family members is presented in Supplementary Table S2.

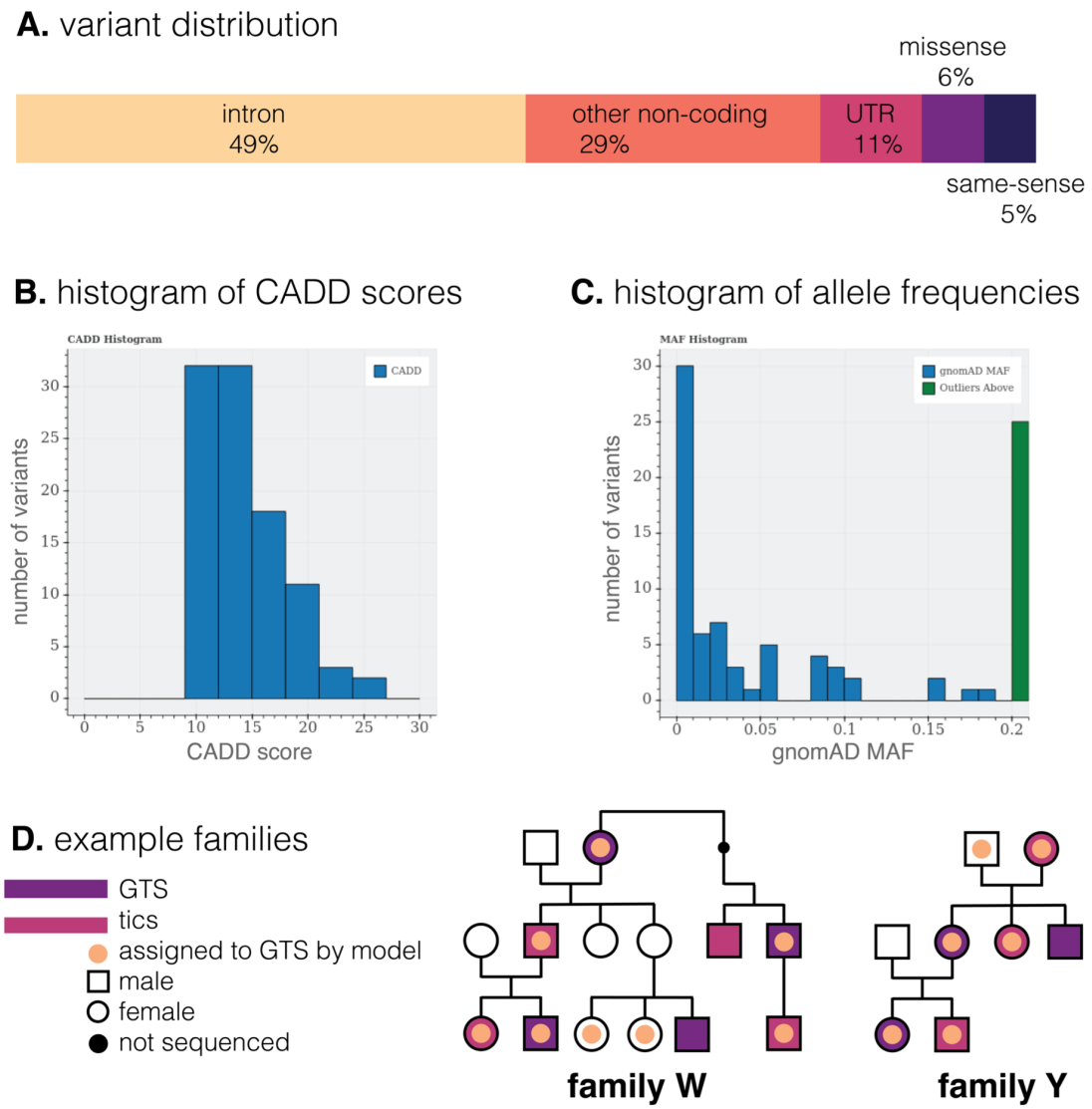


Figure 4. Characteristics of genetic variants included in the GTS risk model and example assignments. A. Localization of variants in genes and flanking regions; B. Histogram of CADD scores of 98 variants included in the risk model; The CADD score is scaled non-linearly in a Phred scale, where a score greater or equal to 10 indicates that a given variant is predicted to be within the top 10% of most deleterious variants substitutions possible in the human genome, whereas a score greater or equal 20 corresponds to the 1% of the most deleterious variants (Rentzsch et al., 2019) C. Distribution allele frequencies (MAF) of 98 chosen variants in non-Finnish Europeans in the gnomAD database; D. An example of classification of members of two families by the GTS risk model. Orange dots indicate individuals assigned to the GTS and other TDs group by the risk model;

Validation of GTS risk model with published GWAS data

To validate our GTS model we analyzed data from a much wider group including 4819 GTS patients and 9488 non-Finnish controls subjected to GWAS and reported in a meta-analysis by the Psychiatric Genomic Consortium (Yu et al., 2019). None of the SNPs reported in the GWAS corresponded to any of the variants included in our GTS model. Therefore, we gathered all SNPs located within the analysed windows for five genes from our model: *CHADL*, *HDC*, *MAOA*, *NAA11,* and *PCDH10*. Overall, 939 SNPs reported in GWAS were located within these regions. Although none of the SNPs reached a genome-wide significant *p* value of 5×10^{-8} for association with GTS, 9 of the SNPs had *p* values below 1×10^{-5}.

Next, we interrogated whether the *p* values of selected 939 SNPs were randomly distributed or did trend towards association with GTS phenotype. As a control, we used SNPs within 20,000 bp of randomly selected protein-coding genes collected into 100 sets of five random genes. For the random controls, the 99th percentile of *p* value for association with GTS phenotype was ~0.008 (-log10(*p* value) = 2.16). For the five genes from our GTS risk model, 0.11 (104 out of 939 SNPs) showed *p* values below this threshold (Figure 5). This suggests that although none of the SNPs reached a genome-wide significance threshold, numerous SNPs in or around the five genes from our model showed a higher-than-chance tendency for over-representation among GTS patients.

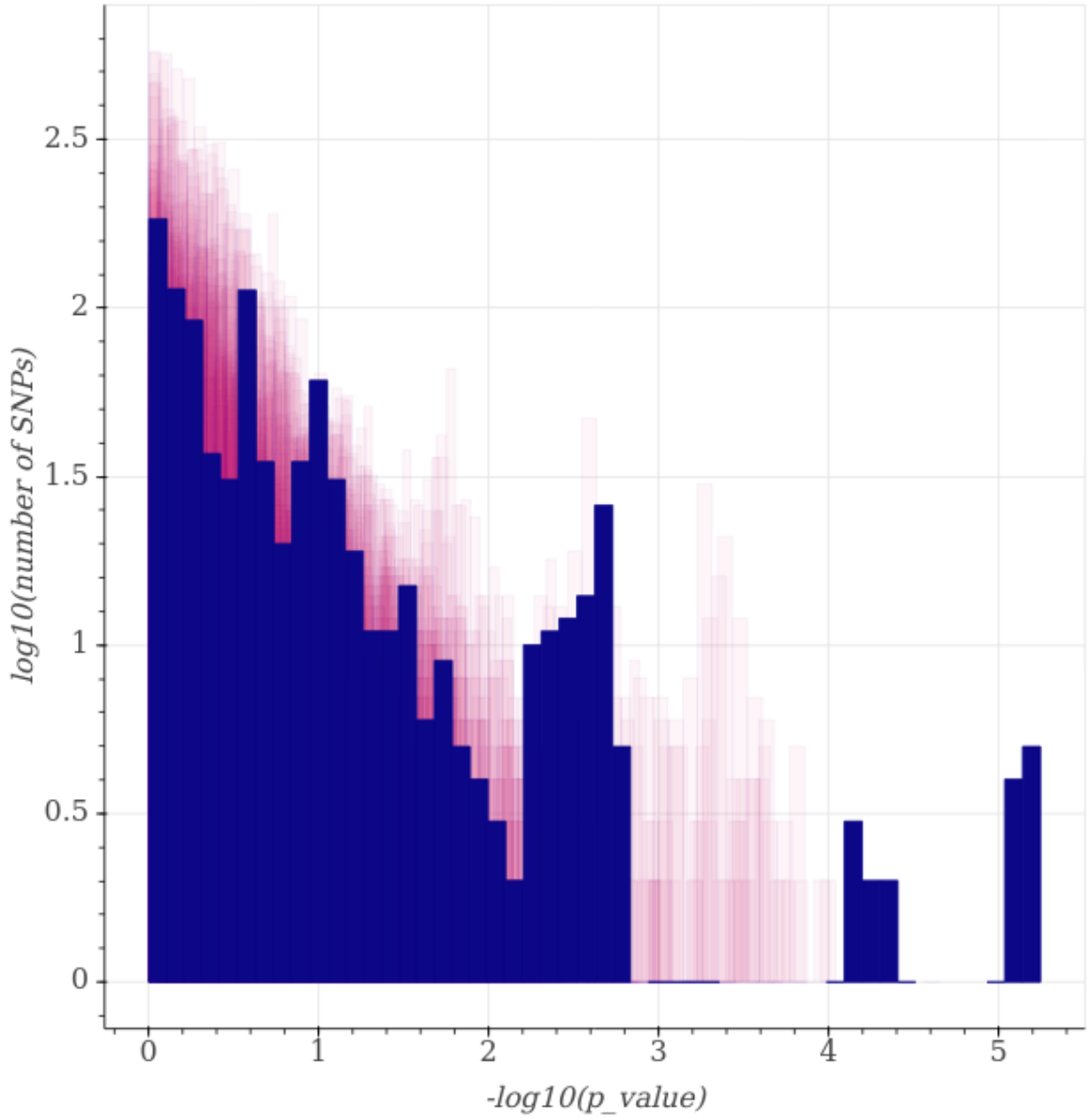


Figure 5. Association of GWAS SNPs in genes from the risk model with GTS phenotype. *P* values were obtained from the available GWAS data via the Psychiatric Genomic Consortium (Yu et al., 2019). Blue bars with solid fill: Histogram of *p* values of SNPs from GWAS detected within the regions included in the risk model (*CHADL*, *HDC*, *MAOA*, *NAA11,* and *PCDH10* with 20,000 flanks). Transparent violet bars: 100 overlaid control histograms of *p* values of SNPs in control sets of randomly selected 5 genes with 20,000 bp flanks. None of the SNPs in control sets reached a *p* value < 10^{-5}, while 9 of the SNPs in the regions included in the risk model passed this threshold. The height of each bar indicates the number of SNPs assigned a *p* value within each of the 50 bins. x-axis: binned *p values* (-log10); y-axis - number of SNPs (log10).

Family-specific candidate genes in large families

Our oligogenic classifier distinguished GTS-affected individuals with good specificity but undoubtedly other genes play a role in GTS risk and particular variants may be family-specific. Here, again using the preselected gene list and WGS data we sought family-specific candidate genes in 14 families with 5 or more members. All identified variants were first filtered according to their localization (within 20,000 bp of the genes of interest), MAF (<0.05), and deleteriousness (CADD > 10). A variant was considered overrepresented if it was present in > 70% of GTS and other TDs-afflicted subjects and < 20% of healthy family members. Ninety-two such variants in 32 genes were identified (Table 5 and Supplementary Tables S4): 74 intronic, seven in 3' UTR, four intergenic, one in a splice region, four missense, and two were other non-coding variants. Notably, 14 of those variants were found in more than one family, including variants in *NEGR1* and *NRXN1* present in four families.

Table 5. GTS-related genes with rare variants overrepresented among GTS patients from large families**.** All variants and their segregation are detailed in supplementary table S4.

family	number of family members	GTS-associated genes with overrepresented rare variants
A	7	*TPH2, SORCS3, NCAM2, HTR2A, IMMP2L, KCNE2, CELSR3, FBN2, NEGR1*
C	9	*FN1*
F	11	*PKP4*
I	6	*CNTNAP2, PDGFB, NRXN1*
J	5	*FN1, PCDH1, NRXN1, SOX5, PKP4, CTNNA3*
R	6	*NTM, OPCML, ROBO2, CDH5, GRIK4, NEGR1, LSAMP, NRXN1, CADM2, CTNNA3, CD47*
S	6	*WWC1, NTM, PKP4, SLITRK2, COL27A1, OPCML, DCC, LSAMP, SOX5, CTNNA3, HDAC9, NEGR1, NRXN1,*

		CADM2, CD47
T	9	*IMMP2L, WWC1*
Y	8	*CHADL, IMMP2L, SCUBE1, NEGR1, PDGFB*

Discussion

Here we proposed an oligogenic risk model of GTS and additionally provided data indicating a role of family-specific GTS-related variants. Studies published so far show that GTS has a polygenic background and that rare variants may have a high impact on GTS etiopathology (Lee et al., 2019; Wang et al., 2018; Willsey et al., 2017). We hypothesized that there is also a substantial contribution of non-coding, potentially regulatory, genomic variants linked to GTS-risk genes. Our study was performed on a group of 185 Polish subjects, including GTS patients and their healthy relatives, and data for non-Finnish Europeans from the gnomAD database as a population control (Karczewski et al., 2020). The presented oligogenic risk model was developed in two steps and is based on selected variants with a CADD score > 10 in and in the vicinity of *HDC, CHADL, MAOA, NAA11*, and *PCDH10* genes. The model was shown to effectively differentiate GTS and other TDs patients from healthy members of large families. Importantly, individuals diagnosed with chronic TDs were assigned to the GTS group with a similar frequency as GTS patients, whereas individuals with transient tics were more likely to be classified as healthy. This is in concert with the idea that psychiatric disorders form a continuum of phenotypes rather than being discrete units (Taylor et al., 2019), and that GTS and other tics constitute a spectrum of tic disorders with a common genetic background.

The proposed risk model, based on five selected genes, showed high sensitivity and specificity as compared to similar classifiers built on other gene sets ($p < 0.00001$), albeit it was not perfectly accurate. This limitation may be explained by other variants contributing to the disease risk and the existence of external factors that influence the development of tics. Furthermore, we tested it on families where the closely related healthy individuals were expected to (and did) carry some of the risk variants. One of the factors that may have influenced the model is the high genetic homogeneity of the Polish population (Jarczak et al., 2019; Soltyszewski et al., 2008). However, the overrepresentation of SNPs with low *p* values in selected genes in GWAS results from the Psychiatric Genomic Consortium (PGC) indicates that our model may perform similarly in other populations. Although the five genes included in the risk model have previously been claimed to be associated with GTS, for most of them no such link was confirmed by large-cohort studies. Our results further indicate

a role of *HDC*, *MAOA*, *PCDH10*, *CHADL*, and *NAA11* and identified novel, mostly rare and non-coding variants associated with GTS and other TDs.

Previously a nonsense mutation was shown to be associated with GTS in the *HDC* gene encoding L-histidine decarboxylase which synthesizes histamine. In addition, two intronic SNPs in this gene were found to be associated with GTS (Karagiannidis et al., 2013). We identified fifteen additional variants within the *HDC* locus or in its flanking regions (5 and 10 variants, respectively). The burden of these variants may collectively affect the histaminergic pathway and thereby contribute to GTS etiology.

The *MAOA* gene is implicated in multiple psychiatric and behavioral disorders as it encodes monoamine oxidase, a popular target of anti-depression pharmacotherapy (Shulman et al., 2013). The enzyme plays an important role in the inactivation of norepinephrine, dopamine, and serotonin. An association with GTS has been postulated for other genes involved in the dopaminergic and serotonergic pathways: dopamine transporter (*DAT1*), tryptophan hydroxylase-2 (*TPH2*), and dopamine receptors (*DRD1, DRD2, DRD4,* and *DRD5*) (Qi et al., 2019). To date, two studies have been published regarding tandem-repeat polymorphisms in the *MAOA* (Diaz-Anzaldua et al., 2004; Gade et al., 1998). Here we report three novel *MAOA* variants, including an intron one as contributing to GTS risk.

In *PCDH10,* two rare (MAF < 1%) variants that possibly lead to a loss of protein function have been detected in GTS patients (Depienne et al., 2019). Homozygous deletion of *PCDH10* has been associated with ASD, a common comorbidity of GTS (Morrow et al., 2008). Our study reports five coding and 40 non-coding variants in or near this gene encoding protocadherin associated with synapse formation (Mancini et al., 2020).

CHADL and *NAA11* were identified in a meta-analysis of GWAS results of multiple psychiatric disorders, including GTS, based on nearby SNPs (rs575265 for *CHADL* and rs1484144 for *NAA11*). The *CHADL* variant showed a high association with both GTS and OCD and the *NAA11* SNP was most strongly associated with ADHD, ASD, major depression, schizophrenia, bipolar disorder, and GTS (Lee et al., 2019). Another variant in the vicinity of *CHADL* (rs11090045) was also associated with neuroticism (Luciano et al., 2018). *NAA11* encodes the catalytic subunit of N-alpha-acetyltransferase complex responsible for alpha-acetylation of proteins. The function of *CHADL* remains to be fully elucidated and thus far it is only predicted to have a regulatory role in chondrocyte differentiation.

The majority of the variants included in the classifier are non-coding (75 variants), mostly intronic (48 variants) and UTR (ten in 3' UTR and two in 5' UTR). Importantly, all of these variants are predicted to be within the top 10% of the most damaging

substitutions in the human genome (their CADD scores calculated using a machine-learning algorithm were ≥10). Furthermore, most of them have low general-population allele frequencies (58 with MAF < 0.05 and 30 are very rare with MAF < 0.01). All the above features make our model fairly unique among GWAS-based studies in which the identified SNPs are usually more common (MAF > 0.05), their impact is unknown and in fact many represent associated haplotypes. Although the interpretation of the role of non-coding variants is challenging and interactions at a distance cannot be excluded, it seems plausible that the identified variants affect the expression of nearby genes, most likely those included in the classifier (Zhang & Lupski, 2015).

In addition to the general GTS risk model discussed above, we also identified possible family-specific genes contributing to GTS using sequencing data from multiplex families with multiple individuals with GTS and other TDs. With this approach we identified 32 genes, rare variants in or near which were over-represented in the GTS/TD patients. As before, most of these variants were non-coding. Notably, two genes, *NEGR1 and NRXN1,* had over-represented variants in four families. *NEGR1,* encoding a cell adhesion molecule, was reported in the same meta-analysis of GWAS which also reported *CHADL* and *NAA11* (Lee et al., 2019), whereas *NRXN1* was postulated to be associated with GTS in CNV studies (Huang et al., 2017).

Overall, the presented approach provides a promising path for further studies of the genomic basis of GTS. The obtained results support the concept that the additive effect of putatively deleterious variants in a small subset of key genes is a substantial risk factor of GTS. We have validated results currently available in the literature and identified a range of rare non-coding variants not previously associated with GTS that could contribute to its etiopathology. The ability of the classifier to distinguish GTS-affected from healthy individuals within families is of particular importance, as the availability of a burden test for affected families could be highly beneficial in genetic counseling. Although the clinical utility of the presented model is limited, it provides an insight into the variant burden associated with familial as well as sporadic GTS. A lack of unrelated controls from the same population is the main limitation of the study. We aimed to minimize this limitation by dividing the study group into separate sub-groups and providing multiple negative controls for the risk model which was additionally validated on independently published GWAS results. Further WGS studies of substantially larger groups and including an extended panel of genes should provide an even better tool for oligogenic GTS risk prediction. A similar approach could be used to predict the risk of other complex diseases.

Funding

This work was supported by the National Science Center, Poland (NCN) project UMO-2016/23/B/NZ2/03030. Computational resources for this research were supplied by PL-Grid Infrastructure.

Literature

Abdulkadir, M., Londono, D., Gordon, D., Fernandez, T. V., Brown, L. W., Cheon, K.-A., Coffey, B. J., Elzerman, L., Fremer, C., Fründt, O., Garcia-Delgar, B., Gilbert, D. L., Grice, D. E., Hedderly, T., Heyman, I., Hong, H. J., Huyser, C., Ibanez-Gomez, L., Jakubovski, E., … Dietrich, A. (2018). Investigation of previously implicated genetic variants in chronic tic disorders: A transmission disequilibrium test approach. *European Archives of Psychiatry and Clinical Neuroscience*, *268*(3), 301–316. https://doi.org/10.1007/s00406-017-0808-8

Abdulkadir, M., Mathews, C. A., Scharf, J. M., Yu, D., Tischfield, J. A., Heiman, G. A., Hoekstra, P. J., & Dietrich, A. (2019). Polygenic Risk Scores Derived From a Tourette Syndrome Genome-wide Association Study Predict Presence of Tics in the Avon Longitudinal Study of Parents and Children Cohort. *Biological Psychiatry*, *85*(4), 298–304. https://doi.org/10.1016/j.biopsych.2018.09.011

Bloch, M. H., & Leckman, J. F. (2009). Clinical course of Tourette syndrome. *Journal of Psychosomatic Research*, *67*(6), 497–501. https://doi.org/10.1016/j.jpsychores.2009.09.002

Cao, X., Zhang, Y., Abdulkadir, M., Deng, L., Fernandez, T. V., Garcia-Delgar, B., Hagstrøm, J., Hoekstra, P. J., King, R. A., & Koesterich, J. (2021). Whole-exome sequencing identifies genes associated with Tourette's disorder in multiplex families. *Molecular Psychiatry*, 1–15.

Castellan Baldan, L., Williams, K. A., Gallezot, J.-D., Pogorelov, V., Rapanelli, M., Crowley, M., Anderson, G. M., Loring, E., Gorczyca, R., Billingslea, E., Wasylink, S., Panza, K. E., Ercan-Sencicek, A. G., Krusong, K., Leventhal, B. L., Ohtsu, H., Bloch, M. H., Hughes, Z. A., Krystal, J. H., … Pittenger, C. (2014). Histidine Decarboxylase Deficiency Causes Tourette Syndrome: Parallel Findings in Humans and Mice. *Neuron*, *81*(1), 77–90. https://doi.org/10.1016/j.neuron.2013.10.052

Davis, L. K., Yu, D., Keenan, C. L., Gamazon, E. R., Konkashbaev, A. I., Derks, E. M., Neale, B. M., Yang, J., Lee, S. H., Evans, P., Barr, C. L., Bellodi, L., Benarroch, F., Berrio, G. B., Bienvenu, O. J., Bloch, M. H., Blom, R. M., Bruun, R. D., Budman, C. L., … Scharf, J. M. (2013). Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. *PLOS Genetics*, *9*(10), e1003864. https://doi.org/10.1371/journal.pgen.1003864

Depienne, C., Ciura, S., Trouillard, O., Bouteiller, D., Leitão, E., Nava, C., Keren, B., Marie, Y., Guegan, J., Forlani, S., Brice, A., Anheim, M., Agid, Y., Krack, P., Damier, P., Viallet, F., Houeto, J.-L., Durif, F., Vidailhet, M., … Hartmann, A. (2019). Association of Rare Genetic Variants in Opioid Receptors with Tourette Syndrome. *Tremor and Other Hyperkinetic Movements*, *9*, 10.7916/tohm.v0.693. https://doi.org/10.7916/tohm.v0.693

Diaz-Anzaldua, A., Joober, R., Riviere, J. B., Dion, Y., Lesperance, P., Richer, F., Chouinard, S., & Rouleau, G. A. (2004). Tourette syndrome and dopaminergic genes: A family-based association study in the French Canadian founder population. *Molecular Psychiatry*, *9*(3), 272–277.

Eriguchi, Y., Kuwabara, H., Inai, A., Kawakubo, Y., Nishimura, F., Kakiuchi, C., Tochigi, M., Ohashi, J., Aoki, N., Kato, K., Ishiura, H., Mitsui, J., Tsuji, S., Doi, K., Yoshimura, J., Morishita, S., Shimada, T., Furukawa, M., Umekage, T., … PhD1, Y. K. (2017). Identification of candidate genes involved in the etiology of sporadic Tourette syndrome by exome sequencing. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *174*(7), 712–723. https://doi.org/10.1002/ajmg.b.32559

Freeman, R. D., Fast, D. K., Burd, L., Kerbeshian, J., Robertson, M. M., & Sandor, P. (2000).

An international perspective on Tourette syndrome: Selected findings from 3500 individuals in 22 countries. *Developmental Medicine & Child Neurology*, *42*(7), 436–447. https://doi.org/10.1111/j.1469-8749.2000.tb00346.x

Gade, R., Muhleman, D., Blake, H., MacMurray, J., Johnson, P., Verde, R., Saucier, G., & Comings, D. E. (1998). Correlation of length of VNTR alleles at the X-linked MAOA gene and phenotypic effect in Tourette syndrome and drug abuse. *Molecular Psychiatry*, *3*(1), 50–60. https://doi.org/10.1038/sj.mp.4000326

Georgitsi, M., Willsey, A. J., Mathews, C. A., State, M., Scharf, J. M., & Paschou, P. (2016). The Genetic Etiology of Tourette Syndrome: Large-Scale Collaborative Efforts on the Precipice of Discovery. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00351

Gloor, F., & Walitza, S. (2016). Tic Disorders and Tourette Syndrome: Current Concepts of Etiology and Treatment in Children and Adolescents. *Neuropediatrics*, *47*. https://doi.org/10.1055/s-0035-1570492

Hirschtritt, M. E., Lee, P. C., Pauls, D. L., Dion, Y., Grados, M. A., Illmann, C., King, R. A., Sandor, P., McMahon, W. M., Lyon, G. J., Cath, D. C., Kurlan, R., Robertson, M. M., Osiecki, L., Scharf, J. M., Mathews, C. A., & for the Tourette Syndrome Association International Consortium for Genetics. (2015). Lifetime Prevalence, Age of Risk, and Genetic Relationships of Comorbid Psychiatric Disorders in Tourette Syndrome. *JAMA Psychiatry*, *72*(4), 325–333. https://doi.org/10.1001/jamapsychiatry.2014.2650

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., McGrath, L. M., Illmann, C., Sandor, P., Barr, C. L., Grados, M., Singer, H. S., Nöthen, M. M., Hebebrand, J., King, R. A., … Smit, J. (2017). Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. *Neuron*, *94*(6), 1101-1111.e7. https://doi.org/10.1016/j.neuron.2017.06.010

Jarczak, J., Grochowalski, Ł., Marciniak, B., Lach, J., Słomka, M., Sobalska-Kwapis, M., Lorkiewicz, W., Pułaski, Ł., & Strapagiel, D. (2019). Mitochondrial DNA variability of the Polish population. *European Journal of Human Genetics*, *27*(8), 1304–1314. https://doi.org/10.1038/s41431-019-0381-x

Karagiannidis, I., Dehning, S., Sandor, P., Tarnok, Z., Rizzo, R., Wolanczyk, T., Madruga-Garrido, M., Hebebrand, J., Nöthen, M. M., Lehmkuhl, G., Farkas, L., Nagy, P., Szymanska, U., Anastasiou, Z., Stathias, V., Androutsos, C., Tsironi, V., Koumoula, A., Barta, C., … Paschou, P. (2013). Support of the histaminergic hypothesis in Tourette syndrome: Association of the histamine decarboxylase gene in a large sample of families. *Journal of Medical Genetics*, *50*(11), 760–764. https://doi.org/10.1136/jmedgenet-2013-101637

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Knight, T., Steeves, T., Day, L., Lowerison, M., Jette, N., & Pringsheim, T. (2012). Prevalence of Tic Disorders: A Systematic Review and Meta-Analysis. *Pediatric Neurology*, *47*(2), 77–90. https://doi.org/10.1016/j.pediatrneurol.2012.05.002

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., & Brower, A. M. (2021). The human phenotype ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217.

Leckman, J. F., Riddle, M. A., Hardin, M. T., Ort, S. I., Swartz, K. L., Stevenson, J., & Cohen, D. J. (1989). The Yale Global Tic Severity Scale: Initial testing of a clinician-rated

scale of tic severity. *Journal of the American Academy of Child and Adolescent Psychiatry*, *28*(4), 566–573. https://doi.org/10.1097/00004583-198907000-00015

Leckman, J. F., Zhang, H., Vitale, A., Lahnin, F., Lynch, K., Bondi, C., Kim, Y.-S., & Peterson, B. S. (1998). Course of tic severity in Tourette syndrome: The first two decades. *Pediatrics*, *102*(1), 14–19.

Lee, P. H., Anttila, V., Won, H., Feng, Y.-C. A., Rosenthal, J., Zhu, Z., Tucker-Drob, E. M., Nivard, M. G., Grotzinger, A. D., Posthuma, D., Wang, M. M.-J., Yu, D., Stahl, E. A., Walters, R. K., Anney, R. J. L., Duncan, L. E., Ge, T., Adolfsson, R., Banaschewski, T., … Smoller, J. W. (2019). Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*, *179*(7), 1469-1482.e11. https://doi.org/10.1016/j.cell.2019.11.020

Luciano, M., Hagenaars, S. P., Davies, G., Hill, W. D., Clarke, T.-K., Shirali, M., Harris, S. E., Marioni, R. E., Liewald, D. C., Fawns-Ritchie, C., Adams, M. J., Howard, D. M., Lewis, C. M., Gale, C. R., McIntosh, A. M., & Deary, I. J. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics*, *50*(1), 6–11. https://doi.org/10.1038/s41588-017-0013-8

Mancini, M., Bassani, S., & Passafaro, M. (2020). Right Place at the Right Time: How Changes in Protocadherins Affect Synaptic Connections Contributing to the Etiology of Neurodevelopmental Disorders. *Cells*, *9*(12), 2711. https://doi.org/10.3390/cells9122711

Mataix-Cols, D., Isomura, K., Pérez-Vigil, A., Chang, Z., Rück, C., Larsson, K. J., Leckman, J. F., Serlachius, E., Larsson, H., & Lichtenstein, P. (2015). Familial Risks of Tourette Syndrome and Chronic Tic Disorders. A Population-Based Cohort Study. *JAMA Psychiatry*, *72*(8), 787–793. https://doi.org/10.1001/jamapsychiatry.2015.0627

Morrow, E. M., Yoo, S.-Y., Flavell, S. W., Kim, T.-K., Lin, Y., Hill, R. S., Mukaddes, N. M., Balkhy, S., Gascon, G., Hashmi, A., Al-Saad, S., Ware, J., Joseph, R. M., Greenblatt, R., Gleason, D., Ertelt, J. A., Apse, K. A., Bodell, A., Partlow, J. N., … Walsh, C. A. (2008). Identifying Autism Loci and Genes by Tracing Recent Shared Ancestry. *Science*, *321*(5886), 218–223. https://doi.org/10.1126/science.1157657

Pagliaroli, L., Vető, B., Arányi, T., & Barta, C. (2016). From genetics to epigenetics: New perspectives in tourette syndrome research. *Frontiers in Neuroscience*, *10*, 277.

Qi, Y., Zheng, Y., Li, Z., Liu, Z., & Xiong, L. (2019). Genetic studies of tic disorders and Tourette syndrome. *Psychiatric Disorders*, 547–571.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894.

Robertson, M. M. (2000). Tourette syndrome, associated conditions and the complexities of treatment. *Brain*, *123*(3), 425–462. https://doi.org/10.1093/brain/123.3.425

Selvini, C., Cavanna, S., & Cavanna, A. E. (2019). Gilles de la Tourette syndrome. In *Chromatin Signaling and Neurological Disorders* (pp. 331–345). Elsevier.

Shulman, K. I., Herrmann, N., & Walker, S. E. (2013). Current Place of Monoamine Oxidase Inhibitors in the Treatment of Depression. *CNS Drugs*, *27*(10), 789–797. https://doi.org/10.1007/s40263-013-0097-3

Soltyszewski, I., Plocienniczak, A., Fabricius, H. Å., Kornienko, I., Vodolazhsky, D., Parson, W., Hradil, R., Schmitter, H., Ivanov, P., & Kuzniar, P. (2008). Analysis of forensically used autosomal short tandem repeat markers in Polish and neighboring populations. *Forensic Science International: Genetics*, *2*(3), 205–211.

Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., Cichon, S., Edenberg, H. J., Faraone, S. V., Gelernter, J., Mathews, C. A., Nievergelt, C. M., Smoller, J. W., & O'Donovan, M. C. (2018). Psychiatric Genomics: An Update

and an Agenda. *The American Journal of Psychiatry*, *175*(1), 15–27. https://doi.org/10.1176/appi.ajp.2017.17030283

Sun, N., Nasello, C., Deng, L., Wang, N., Zhang, Y., Xu, Z., Song, Z., Kwan, K., King, R. A., Pang, Z. P., Xing, J., Heiman, G. A., & Tischfield, J. A. (2018). The PNKD gene is associated with Tourette Disorder or Tic disorder in a multiplex family. *Molecular Psychiatry*, *23*(6), 1487–1495. https://doi.org/10.1038/mp.2017.179

Taylor, M. J., Martin, J., Lu, Y., Brikell, I., Lundström, S., Larsson, H., & Lichtenstein, P. (2019). Association of Genetic Risk Factors for Psychiatric Disorders and Traits of These Disorders in a Swedish Population Twin Sample. *JAMA Psychiatry*, *76*(3), 280–289. https://doi.org/10.1001/jamapsychiatry.2018.3652

Tsetsos, F., Yu, D., Sul, J. H., Huang, A. Y., Illmann, C., Osiecki, L., Darrow, S. M., Hirschtritt, M. E., Greenberg, E., Muller-Vahl, K. R., Stuhrmann, M., Dion, Y., Rouleau, G. A., Aschauer, H., Stamenkovic, M., Schlögelhofer, M., Sandor, P., Barr, C. L., Grados, M. A., … Zinner, S. (2021). Synaptic processes and immune-related pathways implicated in Tourette syndrome. *Translational Psychiatry*, *11*(1), 56. https://doi.org/10.1038/s41398-020-01082-z

Wang, S., Mandell, J. D., Kumar, Y., Sun, N., Morris, M. T., Arbelaez, J., Nasello, C., Dong, S., Duhn, C., Zhao, X., Yang, Z., Padmanabhuni, S. S., Yu, D., King, R. A., Dietrich, A., Khalifa, N., Dahl, N., Huang, A. Y., Neale, B. M., … State, M. W. (2018). De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. *Cell Reports*, *24*(13), 3441-3454.e12. https://doi.org/10.1016/j.celrep.2018.08.082

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. (2017). De novo coding variants are strongly associated with Tourette disorder. *Neuron*, *94*(3), 486-499. e9.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

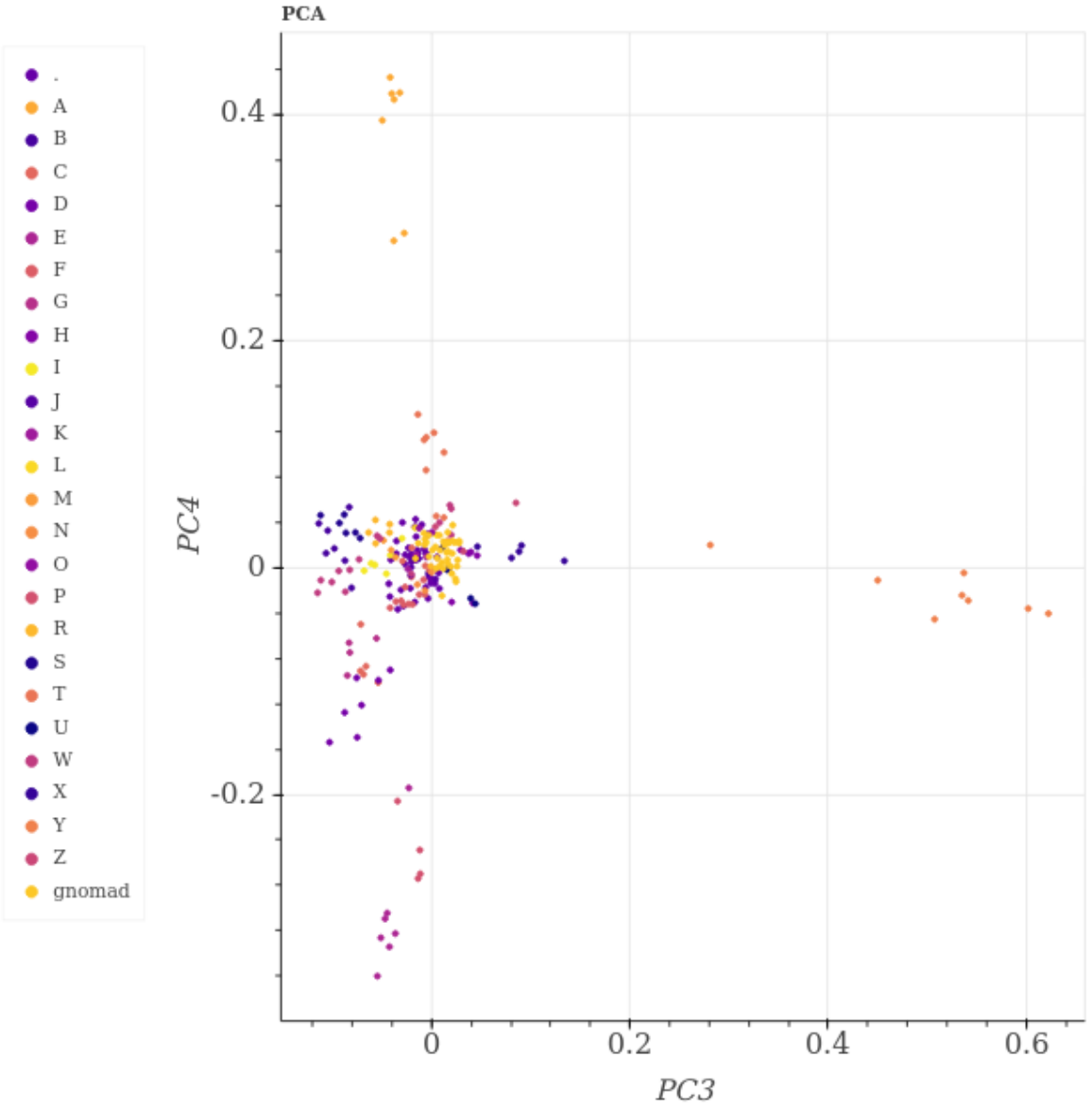
Yu, D., Mathews, C. A., Scharf, J. M., Neale, B. M., Davis, L. K., Gamazon, E. R., Derks, E. M., Evans, P., Edlund, C. K., Crane, J., Fagerness, J. A., Osiecki, L., Gallagher, P., Gerber, G., Haddad, S., Illmann, C., McGrath, L. M., Mayerfeld, C., Arepalli, S., … Pauls, D. L. (2015). Cross-disorder genome-wide analyses suggest a complex genetic relationship between Tourette's syndrome and OCD. *The American Journal of Psychiatry*, *172*(1), 82–93. https://doi.org/10.1176/appi.ajp.2014.13101306

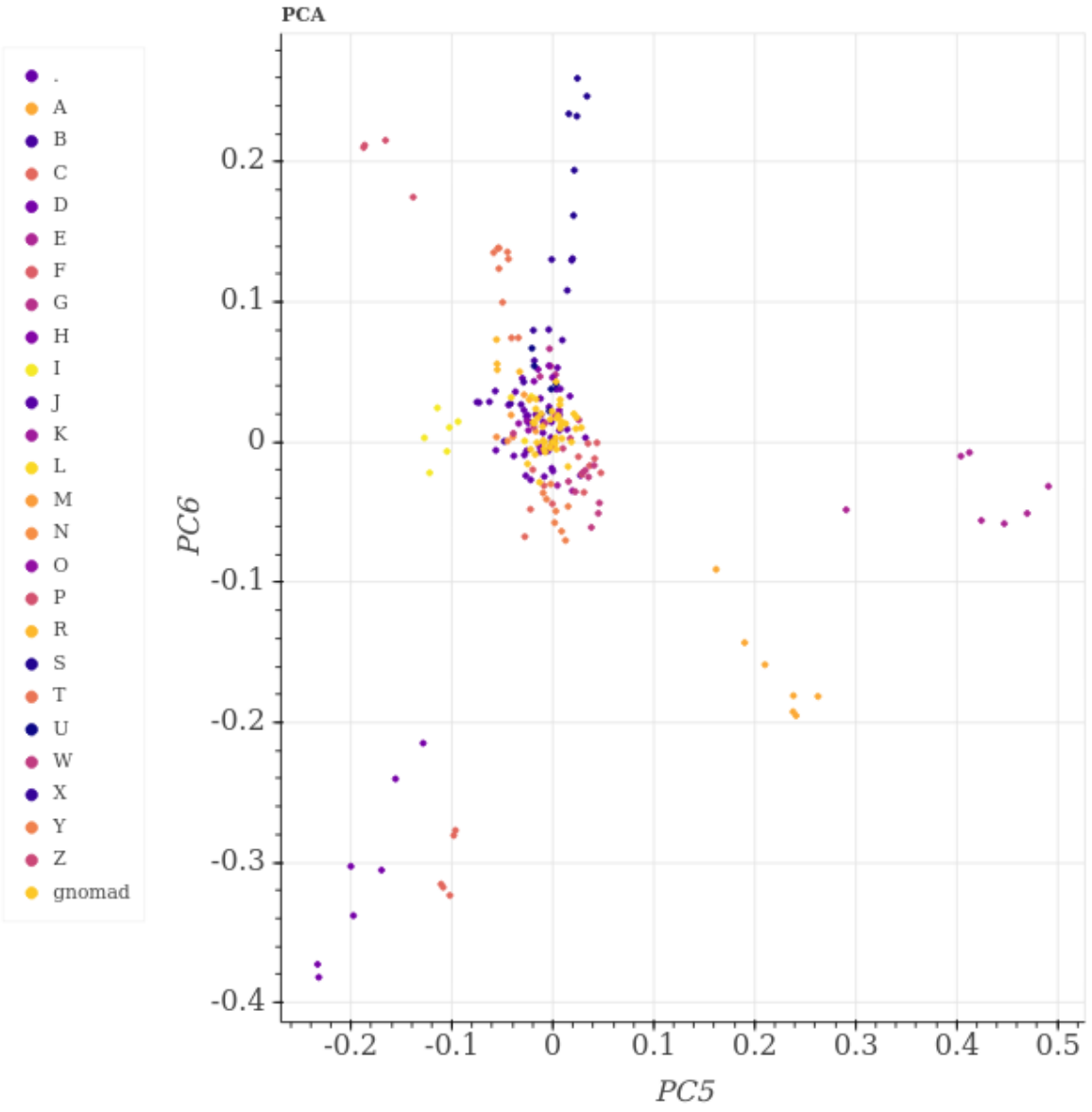
Yu, D., Sul, J. H., Tsetsos, F., Nawaz, M. S., Huang, A. Y., Zelaya, I., Illmann, C., Osiecki, L., Darrow, S. M., Hirschtritt, M. E., Greenberg, E., Muller-Vahl, K. R., Stuhrmann, M., Dion, Y., Rouleau, G., Aschauer, H., Stamenkovic, M., Schlögelhofer, M., Sandor, P., … Tourette Association of America International Consortium for Genetics, the Gilles de la Tourette GWAS Replication Initiative, the Tourette International Collaborative Genetics Study, and the Psychiatric Genomics Consortium Tourette Syndrome Working Group. (2019). Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies. *The American Journal of Psychiatry*, *176*(3), 217–227. https://doi.org/10.1176/appi.ajp.2018.18070857

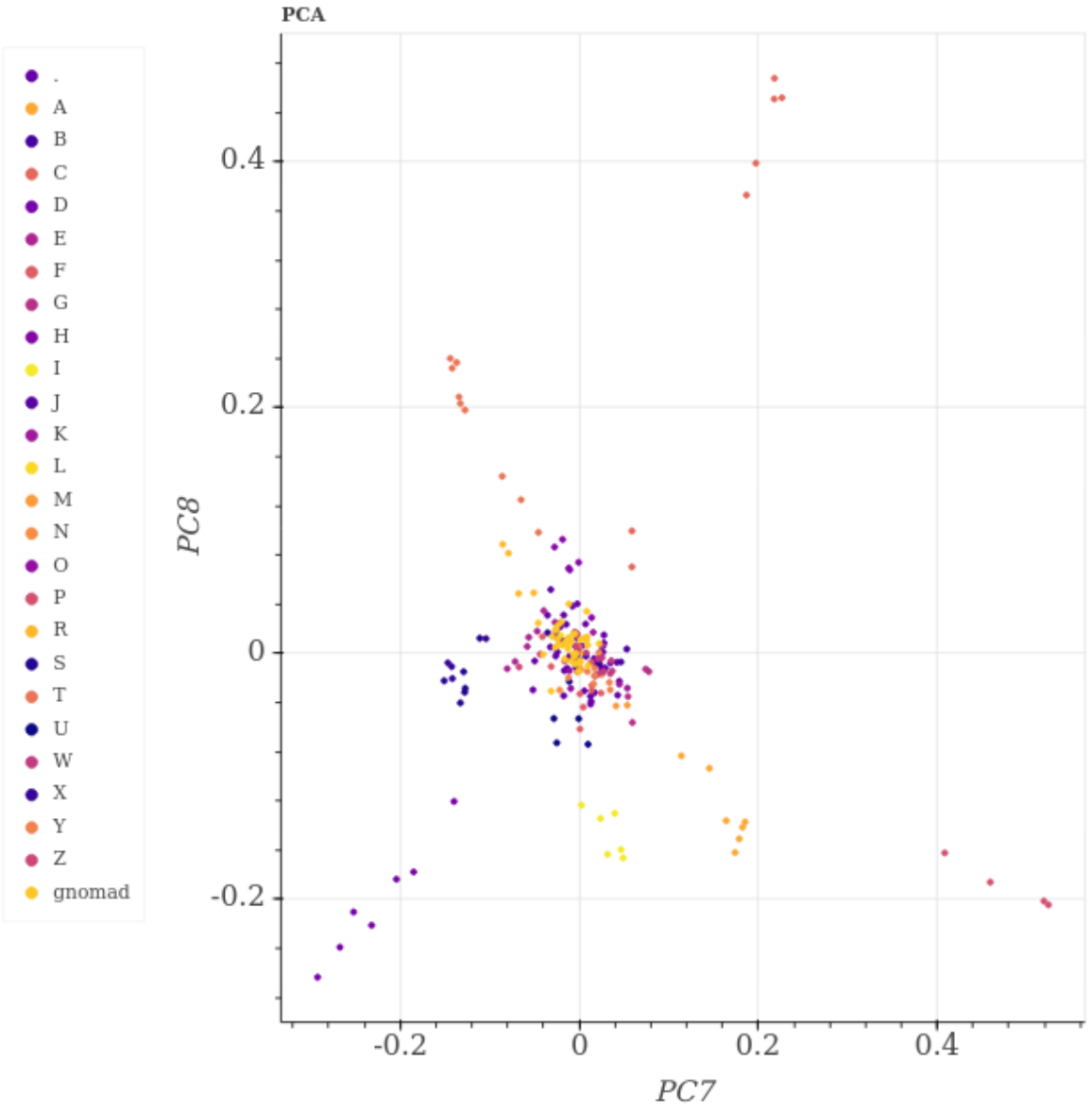
Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, *24*(R1), R102–R110. https://doi.org/10.1093/hmg/ddv259

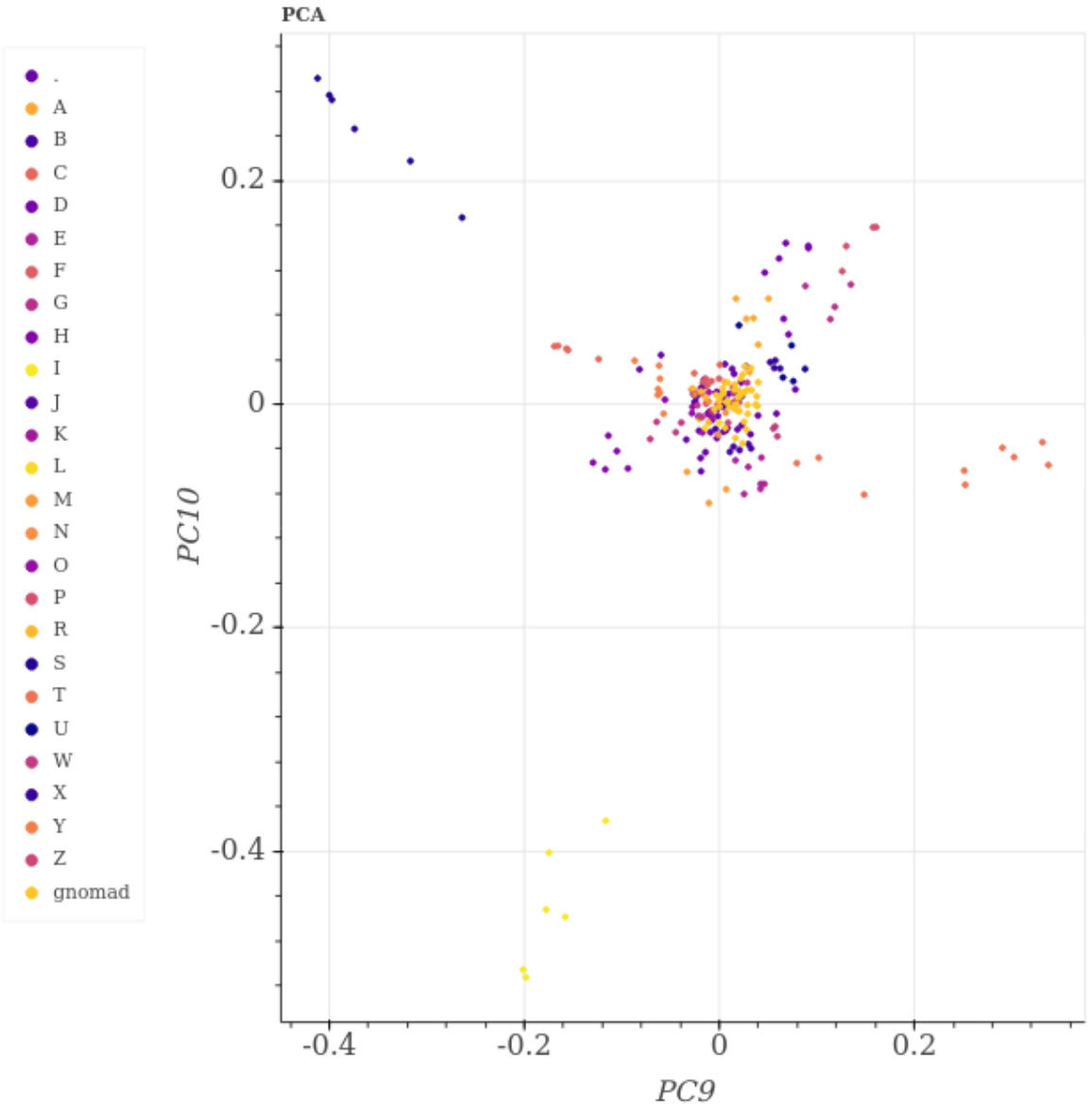
supplementary material : An oligogenic risk-model for Gilles de la Tourette syndrome based on whole-genome sequencing data

Malgorzata Borczyk*, Jakub P Fichna*, Marcin Piechota, Sławomir Gołda, Michal Korostyński, Piotr Janik, Cezary Żekanowski
*These authors contributed equally

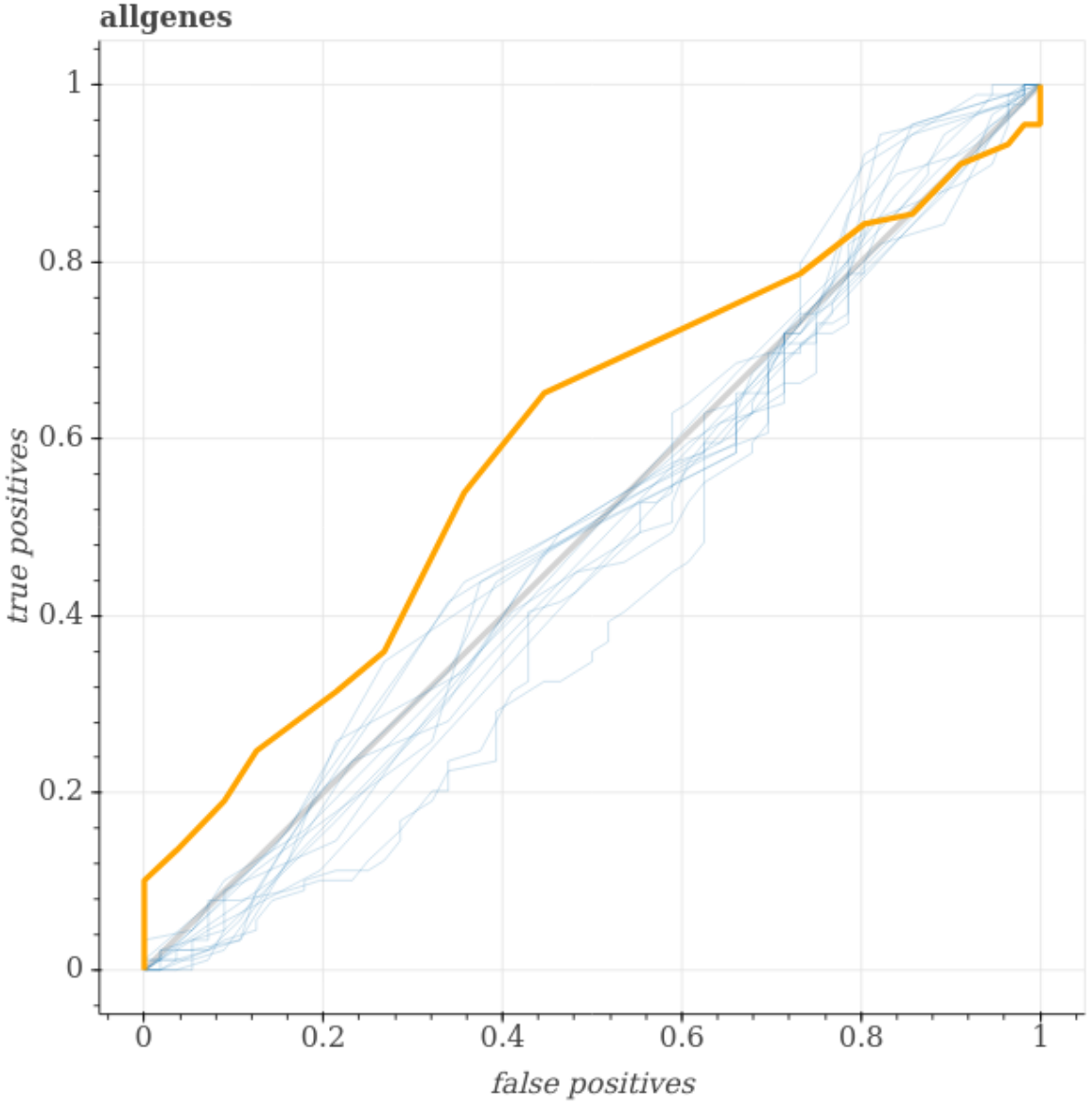


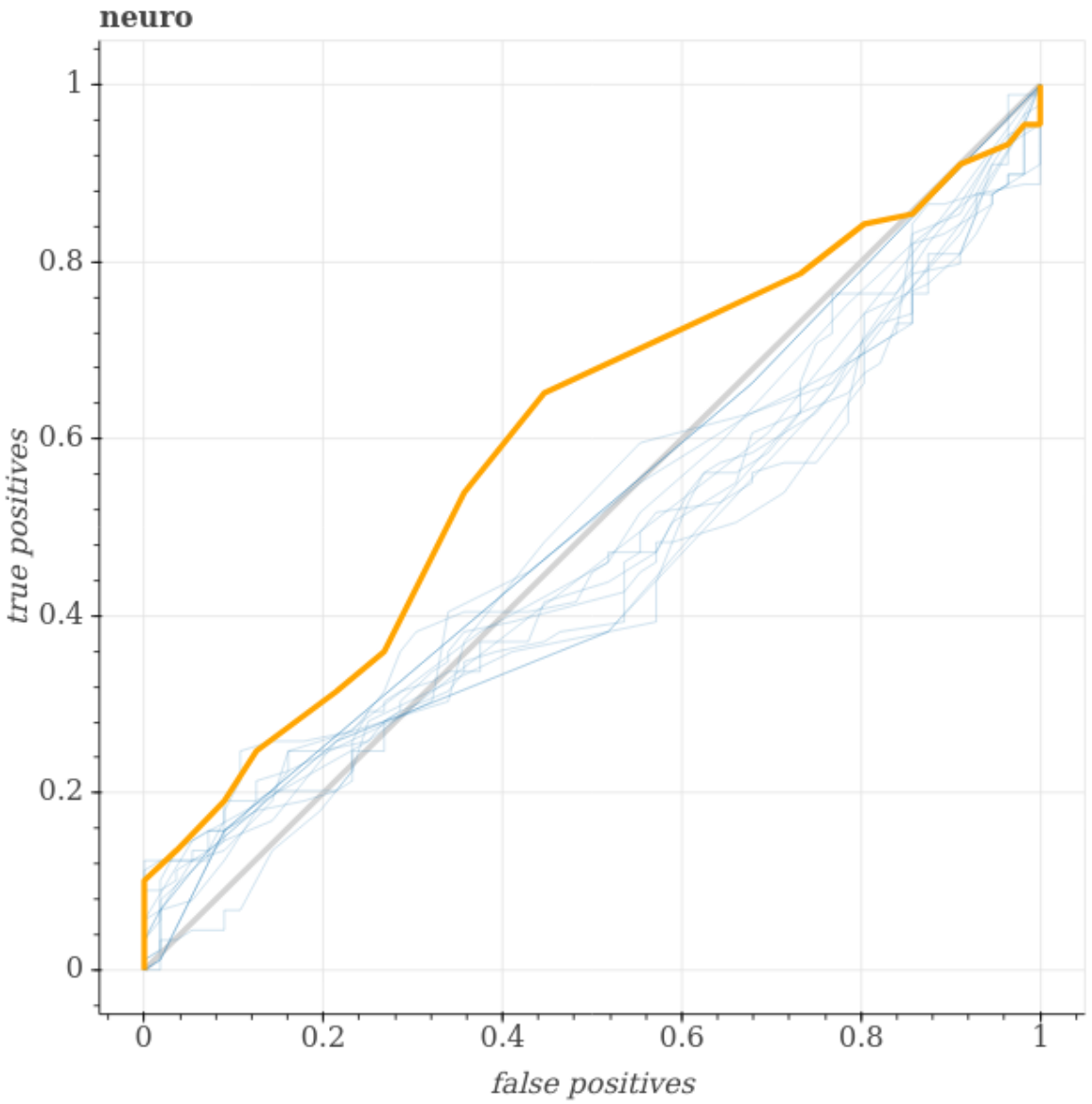






Supplementary Figure S1. Additional PCA plots (supplementary to Figure 1)





Supplementary Figure S2. Receiver-operator characteristics of classifiers built based on additional tested gene lists. The orange line represents the ROC of the chosen main classifier for comparison. allgenes - all human protein coding genes (n=23804); neuro - genes enriched in the brain according to the Human Protein atlas (https://www.proteinatlas.org/humanproteome/tissue) (n = 488);

Supplementary Table S1. Summary statistics of genotype frequencies in each of the genotyped groups (healthy controls, gnomAD simulated controls, familial cases (GTS and TDs) and sporadic cases (GTS) within/in the vicinity of 86 investigated candidate genes.

Supplementary Table S2. Details of variants included in the oligogenic classifier.

Supplementary Table S3. Segregation of variants included in the oligogenic classifier among sequenced members of families.

Supplementary Tables S4. Additional candidate variants identified in the multiplex families. The letter at the end of each file name represents the family code (same as provided in supplementary table T2).