A comprehensive multiomics approach towards understanding obsessive-compulsive disorder

Laura Domènech Salgado

TESI DOCTORAL UPF / 2018

Thesis supervisors

Dra. Raquel Rabionet

Dra. Eulàlia Martí

Dr. Xavier Estivill

BIOINFORMATICS AND GENOMICS PROGRAMME

CENTRE FOR GENOMIC REGULATION

Universitat Pompeu Fabra Barcelona

A mis padres

«Cuando creíamos que teníamos todas las respuestas, de pronto, cambiaron todas las preguntas».

Mario Benedetti

ABSTRACT

Obsessive-compulsive disorder (OCD) is a clinically heterogeneous neuropsychiatric disorder that affects around 1-3% of the population. It is characterized by intrusive and unwanted thoughts, urges or images (called obsessions) and repetitive behaviours or mental acts (called compulsions), which are performed to partially relieve the anxiety or distress caused by the obsessions. Family and twin studies have consistently reported that OCD involves both environmental and polygenic risk factors. However, despite a number of genetic linkage, candidate genes and genome-wide association studies have been performed, very little progress has been made towards elucidating the genetic causes of OCD. In this project we have applied new omics approaches, including rare variant association studies (RVAS) and transcriptomics and metagenomics analyses, to focus on areas relatively underexplored in OCD, which could explain part of the missing heritability observed in this disorder. We have identified and replicated an enrichment of rare variants in *TMEM63A*, a gene that encodes for a calcium-permeable cation channel, through whole-exome sequencing, RVAS and targeted resequencing analyses. Moreover, we have observed an overrepresentation of genes enriched in rare variants in OCD cases related to calcium signalling, suggesting a potential role of calcium signalling dysfunction in the aetiology of OCD. Transcriptomic studies have identified differential expression of genes involved in neuronal development and function in OCD patients, such as *NRCAM*, which encodes for a neuronal cell adhesion molecule. Integration of our RVAS and transcriptomic results also uncover a possible role of semaphorins and axon guidance in OCD. Finally, metagenomics studies have confirmed the previously reported increase of the *Rikenellaceae* bacterial family in the gut microbiome as a potential biomarker of OCD and have shown a specific oro-pharyngeal dysbiotic signature in OCD patients, characterised by a significant higher *Actinobacteria*/*Fusobacteria* ratio compared to controls. In summary, our results support the high complexity of OCD and actively encourage further research in these areas through multiple omics approaches.

RESUM

El trastorn obsessiu compulsiu (TOC) és un trastorn neuropsiquiàtric clínicament heterogeni que afecta al voltant d'un 1-3% de la població. Es caracteritza per pensaments intrusius i no desitjats, ànsies o imatges (anomenades obsessions) i per comportaments o actes mentals repetitius (anomenats compulsions), que es realitzen per alleujar parcialment l'ansietat causada per les obsessions. Estudis familiars i de bessons han reportat consistentment que el TOC implica factors de risc ambientals i poligènics. No obstant, tot i que s'han realitzat molts estudis de lligament genètic, de gens candidats i d'associació del genoma complet, s'ha avançat molt poc a l'hora d'elucidar les causes genètiques del TOC. En aquest projecte hem aplicat nous enfocaments òmics, incloent estudis d'associació de variants rares (RVAS) i anàlisis de transcriptòmica i metagenòmica, per centrar-nos en àrees relativament poc explorades del TOC que podrien explicar part de l'*heretabilitat perduda* observada en aquest trastorn. Hem identificat i replicat un enriquiment de variants rares a *TMEM63A*, un gen que codifica un canal catiònic permeable per calci, a través d'anàlisis de seqüenciació de l'exoma complet, RVAS i reseqüenciació dirigida. A més, hem observat una sobrerepresentació de gens enriquits en variants rares en casos de TOC relacionats amb la senyalització de calci, suggerint un possible paper de la disfunció de la senyalització de calci a l'etiologia del TOC. Els estudis de transcriptòmica han identificat una expressió diferencial de gens involucrats en el desenvolupament i la funció neuronal, com *NRCAM,* que codifica per una molècula d'adhesió cel·lular neuronal. La integració dels resultats dels nostres estudis de RVAS i transcriptòmica també revelen un possible paper de les semaforines i del guiatge axonal al TOC. Finalment, els estudis de metagenòmica han confirmat l'increment prèviament reportat de la família bacterial *Rikenellaceae* en el microbioma intestinal com a possible biomarcador de TOC i han mostrat una signatura disbiòtica específica de l'orofaringe en els pacients de TOC, caracteritzada per una relació significativa més alta d'*Actinobacteris*/*Fusobacteris* en comparació als controls. En resum, els nostres resultats donen suport a l'alta complexitat del TOC i fomenten activament la recerca en aquestes àrees a través de la utilització de multiples òmiques.

INDEX

ABBREVIATONS

ABB: Allele Balance Bias
ACC: Anterior Cingulate Cortex
ACE: Abundance-based Coverage Estimator
Agilent 35: Agilent SureSelect Human All Exon 35Mb Kit
Agilent 50: Agilent SureSelect Human All Exon 50Mb Kit
BATI: Bayesian rare variant Association Test using Integrated Nested Laplace Approximation (INLA)
BWA-MEM: Burrows-Wheeler Alignment Maximal Exact Matches
CBT: Cognitive Behavioural Therapy
CGH: Comparative Genomic Hybridization
CNS: Central Nervous System
CNV: Copy Number Variant
CSTC: Cortico–Striato–Thalamo–Cortical
CY-BOCS: Children's Yale-Brown Obsessive-Compulsive Scale
dbSNP: SNP Database
DE: Differential Expression
DIC: Deviance Information Criteria
DSM: Diagnostic and Statistical Manual of Mental Disorders
EBV: Epstein-Barr virus
ENS: Enteric nervous system
EVS: Exome Variant Server
ExAC: Exome Aggregation Consortium
FC: Fold Change
FDR: False Discovery Rate
FS: Fisher strand
GATK: GenomeAnalysisTK
GO: Gene Ontology
GPe: Globus Pallidus externa
GPi: Globus Pallidus interna

GRAPE: Grape RNA-Seq Analysis Pipeline Environment

GWAS: Genome-Wide Association Study

IOCDF-GC: International OCD Foundation Genetics Collaborative

JSD: Jensen-Shannon Divergence

KBAC: Kernel-Based Adaptive Cluster

LDA: Linear Discriminant Analysis

LEfSe: Linear discriminant analysis Effect Size

LoF: Loss-of-Function

MAF: Minor Allele Frequency

MCC: Multi Case-Control

MFI: Median Fluorescence Intensity

MiST: Mixed effects Score Test

NCBI: National Center for Biotechnology Information

NGS: Next Generation Sequencing

NimbleGen v3: NimbleGen SeqCap EZ Library v3.0

OCD: Obsessive-Compulsive Disorder

ODC T0: OCD sample before treatment

OCD T3: OCD sample after, at least, three months of treatment

OCGAS: OCD Collaborative Genetics Association Study

OFC: Orbitofrontal Cortex

OR: Odds Ratio

PANDAS: Paediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal infections

PCA: Principal Component Analysis

PCoA: Principal Coordinate Analysis

PCR: Polymerase Chain Reaction

PPI: Protein-Protein Interaction

QC: Quality Control

RLE: Relative Log ration Expression

RNA-Seq: RNA-Sequencing

rRNA: ribosomal RNA

RUV: Remove Unwanted Variation

RVAS: Rare Variant Association Study

SCFA: Short Chain Fatty Acid

SKAT: Sequence Kernel Association Test

SKAT-O: SKAT Optimal test

SNP: Single Nucleotide Polymorphism

SNr: Substantia Nigra

SSRI: Selective Serotonin-Reuptake Inhibitor

STN: Subthalamic Nucleus

SV: Structural Variant

Ti/Tv: Transition to Transversion

VQRS: Variant Quality Score Recalibration

WES: Whole-Exome Sequencing

WGS: Whole-Genome Sequencing

Y-BOCS: Yale-Brown Obsessive Compulsive Scale

1000G: 1000 Genomes Project

^{1}H-MRS: Proton Magnetic Resonance Spectroscopy

INTRODUCTION

1. Obsessive-compulsive disorder

1.1. History

Obsessive-compulsive disorder (OCD) has probably existed since time immemorial. In fact, there are early descriptions about this disorder dated in the 7th century, when Saint John Climacus (570-649) reported the story of a monk who was overcome by intrusive, blasphemous thoughts[1].

In the European Renaissance, from 14th to 16th century, it was believed that people who suffered obsessions and compulsions were possessed by the Devil. Based on this, the treatment of individuals suffering OCD was done by the clergy, and involved the banishment of the "evil" from the "possessed" person through exorcism[1].

By the 1700s, physicians started treating OCD as a lunatic disorder and, until the 1850s, it was included in the old notion of “insanity”. Some years later, OCD became a separate disease: first, as a member of the old class of neuroses; then, as a variant of a new concept called psychosis; and finally, as a neurosis proper. But it was not until the late 1880s that OCD achieved full clinical and nosological definition[2].

1.2. Definition, symptomatology and diagnosis

OCD is a neuropsychiatric disorder characterized by intrusive and unwanted thoughts, urges or images (called obsessions) and repetitive behaviours or mental acts (called compulsions) which are performed to partially relieve the anxiety or distress caused by the obsessions (Figure 1)[3].

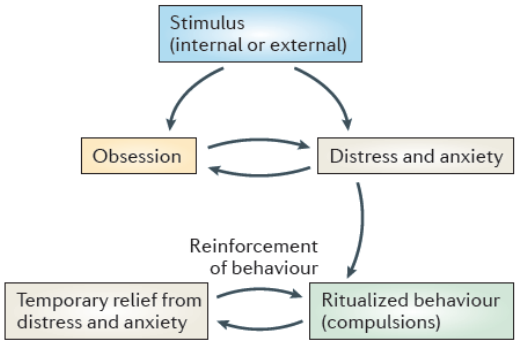


Figure 1. Theoretical basis of obsessive-compulsive behaviour. From Pauls *et al*., 2014[4]

The clinical expression of obsessions and compulsions varies greatly among the individuals with OCD and, for this reason, this disorder is catalogued as "clinically heterogeneous". Moreover, frequency and severity of OCD symptoms also vary across patients, ranging from moderate symptoms to constant and incapacitating intrusive thoughts or compulsions[3–6].

The symptomatology of OCD can be divided into different dimensions or subtypes. It has been suggested that each dimension may have distinct genetic or aetiological origins, as well as distinct neural circuitry[5,6]. Besides, individuals can have symptoms in more than one dimension[3]. A meta-analysis[7] of 21 factor analytic studies published between 1994 and 2008, that included 5,124 patients, reported that OCD symptoms could be divided into four factors, accounting for 79% of the variance: i) symmetry (with symmetry obsessions and repeating, ordering and counting compulsions); ii) forbidden thoughts (with aggressive, sexual, religious and somatic obsessions and checking compulsions); iii) cleaning (with contamination obsessions and cleaning compulsions); and iv) hoarding (with hoarding obsessions and compulsions) (Table 1). It is important to highlight, however, that in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), hoarding is classified as a distinct but OCD-related disorder[3].

Table 1. Factor structure of obsessive-compulsive disorder. From Bloch *et al*., 2008[7]

Factor (% variance)	Obsessions	Compulsions
Symmetry (26.7)	Symmetry	Repeating Ordering Counting
Forbidden thoughts (21.0)	Aggressive Sexual Religious Somatic	Checking
Cleaning (15.9)	Contamination	Cleaning
Hoarding (15.4)	Hoarding	Hoarding

The DSM provides clinicians with official definitions and criteria for diagnosing OCD and, although not all experts agree on the definitions and criteria set forth there, it is considered the "gold standard" by most mental health professionals. The diagnosis of OCD requires the presence of distressing and time-consuming obsessions and compulsions that interfere with the normal functioning of an individual, and neither obsessions nor compulsions can be attributed to the physiological effects of a substance. It is also important to exclude those cases whose obsessions and compulsions belong to another OCD-related disorder (e.g. generalized anxiety disorder, body dysmorphic disorder, trichotillomania, tic disorder, major depression disorder, etc.)[3].

Clinicians also use the Yale-Brown Obsessive Compulsive Scale (Y-BOCS) and its children's version, the Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS), to rate the severity of OCD symptoms[8,9]. These are clinician-rated scales that measure severity for obsessions and compulsions separately, and independently of the type of obsessions and compulsions.

1.3. Epidemiology

The age of onset for OCD ranges from very early childhood (before age 10) into adulthood, with a median age of onset of 19 years old[10]. Between 30 to 50% of individuals with OCD have childhood-onset OCD, a form that can lead to a lifetime disorder[3,11–13]. It is thought that there may be genetic and/or epigenetic factors that affect the age at which symptoms manifest in an individual[4]. Indeed,

some studies reported the possibility of classifying childhood-onset OCD as a distinct neurodevelopmental form of the disorder, as childhood-onset OCD is associated with a specific set of correlates that differ from findings reported in studies of adult OCD subjects[14–16].

According to the DSM-5, the worldwide 12-month prevalence of OCD is 1.1% - 1.8%, although some studies have reported a greater percentage[13,17,18]. Nevertheless, the prevalence of OCD may be underestimated, as many individuals with OCD are secretive about their symptoms and may lie to clinicians, denying having unusual behaviours or providing limited insight about them. Notably, there is a well-established gender imbalance in the prevalence of OCD: childhood-onset OCD is more prevalent in males than females, whereas this ratio is inverted in adulthood-onset OCD[3,19,20].

Although OCD is a common disorder, there are no clear environmental risk factors associated with the disease. Some reported environmental triggers consist on adverse perinatal events[21], inflammatory processes[22], and biologically or emotionally stressful and traumatic life events[23]. The most well documented environmental risk factor is a childhood streptococcal infection. Swedo *et al*.[24] reported a subgroup of paediatric patients with early and brusque onset of OCD and/or tic disorders after pharyngitis or upper respiratory distress caused by streptococcal infections. This subgroup is known by the acronym PANDAS (paediatric autoimmune neuropsychiatric disorders associated with streptococcal infections). The proposed theory is that an initial autoimmune reaction to a streptococcal infection produces antibodies that interfere with basal ganglia function, causing symptom exacerbations that can result in a broad range of neuropsychiatric symptoms.

1.4. Prognosis and impact

OCD has a significant impact on public health, especially if it is not treated, as it becomes a chronic disorder. It is associated with reduced quality of life and high levels of social and occupational impairment. This functional impairment is

associated, in turn, with symptom severity, time spent obsessing and doing compulsions, and avoidance of situations that can trigger obsessions or compulsions[3,25].

Moreover, individuals with OCD often show other comorbid disorders, having a lifetime diagnosis of anxiety disorder (76%), depressive or bipolar disorder (63%), or tic disorder (up to 30%). Comorbid obsessive-compulsive personality disorder is also common in individuals with OCD (23% - 32%)[3]. Furthermore, as a worrying fact, suicide attempts are reported in up to one-quarter of individuals with OCD[26].

To make matters worse, OCD does not have a good prognosis. Skoog *et al*.[27] performed a follow-up study of 122 OCD patients during 40 years and showed that only 48% of the OCD patients recovered. Besides, 48% of the patients had OCD for more than 30 years.

1.5. Neurobiology

Classical theoretical models suggest that OCD is underpinned by structural and functional abnormalities in orbito-fronto-striatal circuits. For this reason, since the 1980s, structural and functional imaging research has been used to elucidate the pathophysiology of OCD.

Over the years, investigators have used different techniques, such as region of interest (ROI), voxel-based morphometry (VBM) and diffusion tensor imaging (DTI) methods –in structural imaging studies- or positron emission tomography (PET), single photon emission computed tomography (SPECT), proton magnetic resonance spectroscopy (^{1}H-MRS) and functional magnetic resonance imaging (fMRI) methods –in functional imaging studies.

However, it is important to highlight that there are a lot of inconsistent findings in these neuroimaging studies, which could reflect confounding factors of the disorder, such as the comorbidity or the heterogeneity of OCD. Moreover, most

of the studies had small sample sizes and used heterogeneous samples that varied in severity, frequency and duration of the symptoms, gender, age at onset, and/or treatment conditions.

1.5.1. Structural imaging studies

Several studies have reported brain structural abnormalities in OCD patients. In 2009, Rotge *et al*.[28] published a meta-analysis of brain volume changes in OCD patients using magnetic resonance imaging (MRI) data from 14 case-control studies. OCD subjects showed a reduced volume of the left anterior cingulate cortex (ACC) and of the left and right orbitofrontal cortex (OFC), and an increased volume of the left and right thalamus. Moreover, they found that the severity of obsessive or compulsive symptoms correlated significantly with the thalamic volumes. This study suggested a structural alteration of the thalamo-cortical circuitry that may contribute to the pathophysiology of OCD. Other studies reported abnormalities in additional brain systems, including the parietal lobe (particularly the angular and supramarginal gyri) and the dorsolateral prefrontal cortex[29,30]. This supported the new idea that other brain regions and circuits could also be involved in the pathophysiology of OCD[29]. Of note, some structural imaging studies also reported that different dimensions of OCD may involve abnormalities in different neuronal systems[31,32].

1.5.2. Functional imaging studies

Functional imaging studies of OCD have mainly focused in the OFC and the striatum. In fact, the most replicated finding consists on an increased activation of the lateral and medial OFC[29,33], although hypermetabolic rates or hyperactivity have also been shown in the ACC and in the basal ganglia, which have been observed to decrease after treatment[30]. Recent functional studies suggested that other regions in the brain, such as dorsolateral and dorsoventral prefrontal cortices, and pre-frontal connections, could participate in the cognitive deficits observed in OCD subjects[34]. Some investigators also observed a general neuronal dysfunction in the brain of OCD patients[30]. Moreover, connectivity

studies have shown connectivity dysfunction between prefrontal and striatal regions[33,35]. As in structural imaging studies, some functional studies also reported that different dimensions of OCD may involve different neural correlates[6,36,37].

1.6. Neuropsychology

Given the structural and functional abnormalities found in some regions of the brain of individuals with OCD, it is logical to hypothesize that OCD patients would show neurocognitive impairment of tasks carried out by these regions. Following this hypothesis, several investigators have conducted a large number of neuropsychological studies of OCD. However, the results yielded are inconsistent.

In order to summarize the results obtained during the last years, Abramovitch *et al.*[38] performed a meta-analysis of 115 studies that involved over three thousand OCD patients and results from tests of 10 neuropsychological domains (distinct types of functions that the brain uses to execute behaviours). They found a reduced performance across all domains among OCD patients compared to healthy controls. Specifically, medium to large effect sizes were found for the memory domain; medium effect sizes were found for attention, executive functions and processing speed; and small effect sizes were found for working memory and visuospatial abilities.

It is thought that neurocognitive indices could be used as endophenotypes of OCD[39]. Some studies have suggested that different neuropsychological profiles could be associated with different dimensions of OCD[40] and that there is significant correlation between severity of OCD and neuropsychological impairment[41]. Moreover, it has been shown that cognitive dysfunction in OCD can improve in the course of treatment[42] and that the neuropsychological profile could also determine treatment outcome[43].

1.7. Treatment

Many OCD patients can achieve substantial improvement through the establishment of an adequate treatment. However, for approximately 50% of the individuals with OCD treatment response is incomplete[44]. The first-line treatment for OCD consists of cognitive behavioural therapy (CBT), selective serotonin-reuptake inhibitors (SSRIs) or a combination of the two[45,46]. The effectiveness of the available SSRIs has been studied during more than 20 years. Recently, Soomro *et al*.[47] conducted a meta-analysis (which included 17 studies and over 3000 samples) that demonstrated that SSRIs are nearly twice as likely as placebo to produce a clinical response. It should be noted that several studies have shown a decrease of the hypermetabolic rates in the OFC, the caudate and the ventrolateral prefrontal cortex after CBT or pharmacological treatment in OCD patients, which provides support for the current model of OCD[4].

Non-pharmacological treatment, such as electroconvulsive therapy (ET), repetitive transcranial magnetic stimulation (rTMS), deep brain stimulation (DBS), or ablative neurosurgery, is indicated in severe OCD patients that do not respond to pharmacological and CBT[4,48]. Results from these neuromodulation therapies are encouraging and give also support to the prevailing model of OCD, as these treatments target the circuitry implicated in the disorder[4]. In fact, the ultimate goal of the surgery is to interrupt this circuitry imbalance thought to be present in OCD.

1.8. Biological model of OCD

The prevailing model for the neural and pathophysiological basis of OCD, based on data from neurobiological, neuropsychological and treatment studies, is the cortico–striato–thalamo–cortical (CSTC) model (also called frontostriatal model or corticostriatal model)[49].

Briefly, the CSTC circuit projects from specific territories in frontal cortex to targets in the striatum, and then, via direct and indirect pathways, through the

basal ganglia to the thalamus. Finally, these structures project back to the frontal cortex[50]. In healthy individuals, the direct pathway is modulated by the indirect pathway[4] (Figure 2). Specifically, glutamatergic signals from the frontal cortex (OFC and the ACC) lead to excitation in the striatum, which increases inhibitory GABA signals to the globus pallidus interna (GPi) and the substantia nigra (SNr) through the direct excitatory pathway. This decreases the inhibitory GABA output from the GPi and SNr to the thalamus, resulting in excitatory glutamatergic output from the thalamus to the frontal cortex. In the indirect pathway, the striatum inhibits the globus pallidus externa (GPe), which decreases its inhibition of the subthalamic nucleus (STN). The STN excites then the GPi and SNr that inhibit the thalamus and its glutamatergic output. In OCD individuals, there may be an imbalance between the direct and indirect pathways[4]. The prevailing model describes hyperactivity of the direct pathway over the indirect pathway due to a lower threshold for activation of this system, and this results in a hyperactivation of the orbitofrontal–subcortical pathway. As a consequence, the OFC could mediate exaggerated and persistent concerns and fears leading to the development of obsessions and, subsequently, to the development of compulsions, which are performed to neutralize the anxiety caused by the obsessions.

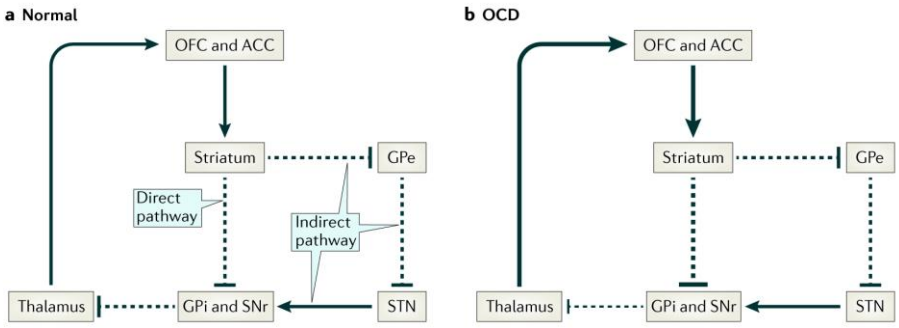


Figure 2. The cortico–striato–thalamo–cortical circuitry in healthy individuals (a) and OCD patients (b). From Pauls *et al*., 2014[4]

Recently, though, some investigators have proposed modifications of the model, with the involvement of the lateral and medial orbitofrontal cortices, the dorsal anterior cingulate cortex and the amygdalo-cortical circuitry, in addition to the cortico-striatal circuitry[4,50].

Pauls *et al.*[4] suggested an integrative model of genetics, environment and neurobiology for the expression of OCD. Individuals with OCD may have a genetic predisposition to the impact of environmental factors that may modify, through epigenetic mechanisms, the expression of genes involved in the serotonin, glutamate, catecholamine, and dopamine systems, which are involved in the OCD physiopathology. Neuroanatomical changes derived from these modifications may lead to an imbalance between the direct and indirect pathways involved in the CSTC circuit, and this imbalance, in turn, could result in the manifestation of clinical features of OCD (Figure 3).

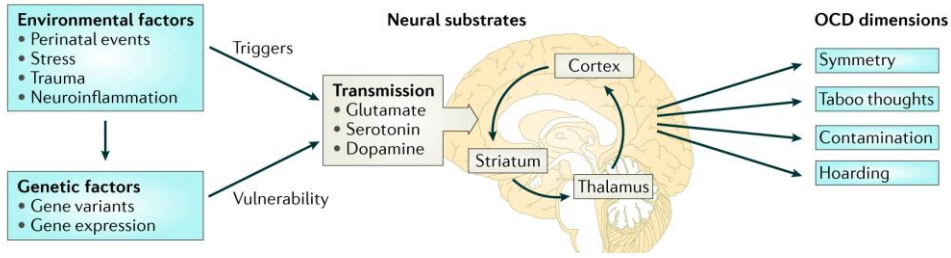


Figure 3. Biological model of OCD. From Pauls *et al.*, 2014[4]

2. Genetics of OCD

2.1. Genetic architecture of OCD

There is compelling evidence that OCD is a complex neuropsychiatric disorder that probably arises from a combination of environmental and genetic risk factors. The interaction of some genes with a mild to moderate effect increases

the vulnerability to OCD, while environmental factors contribute almost equally to OCD risk. However, the genetic architecture underlying OCD has not yet been well defined.

The genetic architecture of a complex disorder refers to a comprehensive description of how genes and the environment interact to produce the phenotype. It involves multiple factors: the number of genetic variants contributing to the phenotype, the size of their effects on the phenotype, the frequency of those variants in the population, and their interactions with each other and the environment[51,52] Is it not surprising, then, that defining the genetic architecture of complex disorders still represent an arduous task in human genetics research, although fundamental for understanding disease aetiology.

In the following sections I will provide an overview of all the genetic studies that have been done in OCD to help define its genetics architecture.

2.2. Heritability

OCD family studies published since 1930 have consistently reported that OCD is familial[53]: the rate of OCD among relatives of patients is significantly higher than either the rate of OCD among controls or the OCD estimated population prevalence.

There have been at least 18 studies on relatives of adult OCD cases, and only two of them did not report that OCD was familial[4]. However, these two[54,55] observed a higher rate of mental illness among OCD relatives. In addition, all studies performed on relatives of childhood-onset OCD concluded that OCD was familial[4]. These studies also observed a 10-fold higher risk of OCD for relatives of childhood-onset OCD patients. This risk is only two-fold higher in relatives of adult OCD cases[53] (Table 2).

Table 2. Recurrence rates among relatives in OCD family studies. From Stewart *et al*., 2010[56]

Publication years	Proband relatives		Control relatives	
	Obsessive-compulsive disorder	Obsessive-compulsive features		
Family history studies				
1930-1986	0 - 0.198	0.07 – 0.327	NA	
Adult family studies	**OCD**	**Subclinical OCD**	**OCD**	**Subclinical OCD**
1987-2006	0.007 – 0.117	0.046 – 0.156	0 – 0.027	0 – 0.030
Child family studies				
1990-2005	0.050 – 0.227	0.065	0.009 – 0.026	0.015

NA: not applicable.

Nevertheless, the fact that family studies showed that OCD is familial does not necessarily mean that it is transmitted within families through genetic factors. Twin studies, which allow an estimation of the extent to which genetic and environmental factors play a role in the aetiology of complex disorders comparing the concordance of monozygotic (MZ) and dizygotic (DZ) twins, did provide evidence that OCD familiality was influenced by both genetics and environment.

There have been numerous twin studies conducted to date investigating the role of additive and non-additive genetic effects, as well as shared and non-shared environment. Recently, Taylor S.[57] performed a meta-analysis that included 24,161 twin pairs from 14 published studies. The findings supported the hypothesis that genetic risk factors play an important role in the manifestation of obsessive-compulsive symptoms. Specifically, this meta-analysis showed that additive genetic variance accounts for 37 to 41% of variance of obsessive-compulsive behaviours and non-shared environmental factors account for 50 to 52%, while shared environmental factors and non-additive genetic effects made little or not contribution. These findings did not vary with sex or symptom severity, although variance due to non-shared environment increased with age, suggesting that environmental risk factors could be more important for the

manifestation of late-onset OCD. The study also described that interactions between non-shared environmental and genetic risk factors are crucial for developing the obsessive-compulsive behaviours, and that obsessive-compulsive symptoms are shaped by etiologic factors common to all types of obsessive-compulsive behaviours but also have symptom-specific aetiologies.

2.3. Genetic studies of OCD

2.3.1. Genetic linkage studies

As family and twin studies evidenced that there is a genetic basis for familial OCD, the first approaches to identity this genetic factors involved genetic linkage studies[58–63]. However, none of the genetic linkage studies performed reached accepted levels of statistical significance. This could be explained because genetic linkage studies are particularly useful to identify genes involved in Mendelian disorders (caused by alterations in a single gene) but not so much for finding risk alleles of complex disorders, which are thought to be caused by a large number of risk loci of small to moderate effect[64]. The fact that in almost all studies samples sizes were small could also explain the lack of significance achieved. The largest genetic linkage study[59] included 966 individuals from 219 families.

Nevertheless, it should be noted that two genomic regions on chromosomes 9 and 15 were identified in several studies[58,59,61,65] and that *SLC1A1*, a gene that encodes a glutamate transporter, is the closest gene to the linkage peak found in the chromosomal region 9p[65]. Moreover, several candidate genes studies corroborated a possible association of *SLC1A1* with OCD[4].

2.3.2. Candidate gene studies

To date, more than one hundred of candidate gene studies for OCD have been published, most of them focused on genes that are known to be involved in systems linked to the pathophysiology and pharmacology of OCD, as serotonin, glutamate and dopamine systems.

To integrate the results of so many studies, Taylor *et al*.[66] performed two meta-analyses of OCD. The first meta-analysis included 20 single nucleotide polymorphisms (SNPs) that were studied in 5 or more datasets. In this study, OCD was associated with variants in serotonin-related genes (*SLC6A4* and *HTR2A*) and, only in males, in the catechol-O-methyltransferase (COMT) and mono-amine oxidase A (MAOA) genes. There were, also, non-significant trends for one glutamate-related gene (*SLC1A1*) and two dopamine-related genes (*SLC6A3* and *DRD3*). The second meta-analysis, which was conducted for 210 polymorphisms that had been examined in less than five data sets, identified associations for polymorphisms in trophic factors (*BDNF, NGFR* and *NTRK2*), GABA (*GABRB3*), glutamate (*GRIK2*), serotonin (*HTR2A*), bradykinin (*BDKRB2*), acetylcholine (*CHMR5, CHRNA1*), glycine (*GLRB*), ubiquitin (*UBE3A*), immunological factors (*TNFA*) and myelinization (*OLIG*) genes.

a) Serotonin system

Serotonin-related genes have been extensively studied in OCD candidate gene studies[67] because OCD is commonly treated with drugs that act in the serotonin system, such as SSRIs. The serotonin transporter *SLC6A4* is probably one of the most widely studied genes in neuropsychiatry[68] and has been related to OCD in multiple studies[4].

b) Glutamate system

Glutamated-related genes were first found to be involved in OCD through imaging and animal model studies. Specifically, ^{1}H-MRS demonstrated that glutamate concentrations are altered in the caudate and ACC of individuals with OCD. Moreover, *Slitrk5* and *Dlgap3* knockout mice showed compulsive grooming behaviours related to glutamate signalling dysfunction[4]. Later, genetic association studies provided stronger evidence of glutamate involvement in the pathophysiology of OCD. Particularly, *SLC1A1* has been associated with OCD in several studies[69]. Other genes, such as *GRIN2B*, *GRIK2* and *GRIK3,* have also been implicated in the disorder[70].

c) Dopamine system

The dopamine hypothesis in OCD is based predominantly in pharmacological studies. As SSRIs are not effective in all individuals with OCD, it is thought that other neurotransmitters apart from serotonin must be involved in the pathophysiology of the disorder. Indeed, considerable improvement has been seen in individuals with OCD when SSRIs treatment is complemented with dopaminergic antagonists[71]. Besides, dopamine agonists have been found to provoke tic and repetitive behaviours in animal models[72]. Several genes, such as the dopamine transporter (*SLC6A3*) and the dopamine receptors (*DRD1*, *DRD2*, *DRD3* and *DRD4*), have yielded positive associations with OCD in several studies[4]. *DRD4*, in particular, has been extensively studied[73].

The *DRD4* gene encodes for the D4 subtype of the dopamine receptor, a member of the dopamine G-protein-coupled receptor family that also includes D1, D2, D3 and D5. *DRD4* is responsible for neuronal signalling in the mesolimbic system of the brain, an area of the brain that regulates emotion and complex behaviour. This receptor is located primarily in the frontal cortex, midbrain, amygdala and the cardiovascular system and is of great interest for research into neuropsychiatric disorders and psychopharmacology. Many mental disorders have been associated with mutations in this gene, including attention deficit hyperactivity disorder, autonomic nervous system dysfunction and the novelty seeking personality trait[73]. Moreover, it was found that density of DRD4 mRNA was 6-fold higher in brains of schizophrenic patients[74] and that it binds the antipsychotic drug clozapine with higher affinity than does any other dopamine receptor[75].

DRD4 has a high degree of genetic variation in the human population, containing a polymorphic number (from 2 to 10 copies) of 48 base pair (bp) repeats in the third intracytoplasmic loop of the receptor, a region that seems to be involved in G-coupling of the protein to its effector systems and that can influence clozapine binding[76]. It is important to highlight that in 1997, Cruz *et al*.[77] described an increased prevalence of the seven-repeat variant of *DRD4*

(DRD4*7R) in patients with OCD with tics, whereas Millet *et al*.[78] reported a protective effect of the DRDR*2R variant against OCD symptoms.

Nöthen *et al*.[79] reported a 13-bp deletion in the first exon of *DRD4*, which produces a stop at codon 99. They hypothesized that this variant consists of a null mutation that encodes a truncated non-functional protein, leading to a complete loss-of-function of the D4 receptor. They tested for association of this deletion with various psychiatric disorders (schizophrenia, bipolar affective disorder and Tourette's syndrome) and they showed that the mutation occurred in similar frequency in all psychiatric and control samples. It should be noted that the subjects included in the study had a German origin and that sample size was small (232 healthy volunteers, 118 individuals with schizophrenia, 99 with bipolar affective disorder and 91 with Tourette's syndrome, according the DSM-III-R criteria). Another study[80] also reported no association of the null mutation in *DRD4* in Italian patients with OCD, bipolar mood disorder and schizophrenia. Again, the sample size was small (157 OCD patients, 196 schizophrenics, 111 bipolars and 162 healthy controls of Italian descent). Nevertheless, a recent study[81] highlighted again the presence of this 13-bp exonic deletion in *DRD4* in one family after exome-sequencing of ten trios with OCD. In addition, an in-frame deletion of 21-bp affecting codons 36 to 42 of *DRD4* was also associated with OCD by Chichon *et al*.[82]

2.3.3. Genome-wide association studies

Once the human genome was almost entirely sequenced in 2003[83], it was possible to develop genotyping arrays, which provide genotype calls for thousands of SNPs and can also be used to detect some genomic structural variants (SVs), mostly copy number variants (CNVs). They have been widely used for genome-wide association studies (GWAS)[84].

GWAS are observational studies that test if any measured common variant (minor allele frequency, MAF >5%) is associated with a trait[85]. The latest SNP-arrays assess between half a million to two million common variants along the

human genome to detect differences in allele frequencies between cases and controls[86,87]. Nowadays, the biggest catalogue of published GWAs[88] contains over 3,300 publications and almost 60,000 unique SNP-trait associations, which are, mostly, at non-coding variants and enriched at regulatory sites. Almost all complex disease categories have been addressed by GWAS, including neuropsychiatric, neurodegenerative, cardiovascular, metabolic, autoimmune and musculoskeletal diseases, and several types of cancer. These include two OCD GWAS performed by independent OCD consortia[89,90].

The International OCD Foundation Genetics Collaborative (IOCDF-GC) published the first GWAS[89], comprising 1,465 cases, 5,557 ancestry-matched controls and 400 trios from 22 sites worldwide, and genotyping about 500,000 SNPs. In the case-control-trio analysis, which included all the samples together, no SNPs were found to be associated with OCD at a significant genome-wide level. The most significantly associated SNP was rs297941, near *FAIM2* (FAS apoptotic inhibitory molecule 2, P=4.99 x 10^{-7}). In the case-control analysis, two SNPs in *DLGAP1* (discs large-associated protein 1), a member of the postsynaptic scaffold in neuronal cells showed the strongest association with the phenotype (P=2.49 x 10^{-6} and P=3.44 x 10^{-6}). When only the data of the trio analysis was considered, a SNP near *BTBD3* (BTB (POZ) domain-containing 3), a key regulator of dendritic field orientation, yielded genome-wide significance (P=3.84 x 10^{-8}). However, this SNP had no genome-wide significance when all the data was analysed together (case-control-trio analysis).

The OCD Collaborative Genetics Association Study (OCGAS) published the second GWAS[90], which included 1,065 families that involved 1,406 patients with an early age of OCD onset and population-based samples resulting in a total sample of 5,061 individuals. A marker on chromosome 9, near the gene *PTPRD*, a member of the protein tyrosine phosphatase (PTP) family that is thought to have a role in promoting neurite growth and regulating neurons axon guidance, presented the smallest p-value (P=4.13 × 10^{-7}).

A meta-analysis of the two consortia[91], investigating a total of 2,688 OCD cases with European ancestry and 7,037 matched controls, was also performed. No SNP exceeded the genome-wide threshold for significance. The SNP with the lowest p-value (P=7.1 × 10^{-7}) was 87.2 kb 5' to *CASC8* (Cancer Susceptibility Candidate 8) and the second SNP with the lowest p-value (P=1.1 × 10^{-6}) lied entirely within *GRID2* (Glutamate Ionotropic Receptor Delta Type Subunit 2). Variants located in or near the genes *ASB13*, *RSPO4*, *DLGAP1*, *PTPRD*, *GRIK2*, *FAIM2* and *CDH20* were among the top signals.

2.3.4. Animal models of OCD

During the last 30 years, many attempts have been done to develop animal models of OCD that could help us understand the underlying biology of this disorder. Although the intrusive obsessional thoughts and fears about human topics could never be assessed through animal models, they could help us study other aspects of OCD, such as compulsions and ritualistic behaviours.

Animal models of OCD are based on behavioural similarity, as it is suggested that the behaviour of genetically modified animal models should be similar to some specific conducts in individuals with this disorder[92]. Some models are induced by genetic manipulation, but there are also pharmacological animal models of OCD (based on drug-induced OCD-like behaviours) and behavioural manipulation-based animal models of OCD (based on repetitive behaviours occurring naturally or under stressful events)[92].

Currently, there are eight mouse models of OCD that show compulsive-like behaviour due to genetic modifications (*5-HT2c* receptor, aromatase (*Cyp19a1*), *Slitrk5* and *Slc1a1* knock outs (KO), dopamine transporter (DAT) knockdown, and *DiCT-7*, *Hoxb8*, and *Dlgap3* mutant mice) and one mouse model of OCD that shows compulsive-like behaviour as a result of selective breeding[92,93]. These mouse models provided evidence that serotonin, glutamate and dopamine are related to the expression of OCD-like behaviours such as anxiety and excessive self-grooming[94,95].

In 2014, Tang *et al.*[96] published a canine model of OCD. They suggested that the limited genetic diversity of dog breeds facilitates the identification of genes, functional variants and regulatory pathways underlying complex psychiatric disorders. They identified four genes involved in synaptic function with variation present only in cases: neuronal cadherin (*CDH2*), catenin alpha2 (*CTNNA2*), ataxin-1 (*ATXN1*), and plasma glutamate carboxypeptidase (*PGCP*).

Recently, D'Amico *et al*.[97] suggested that functional validation of all the candidate genes and variants that are being discovered through next generation sequencing (NGS) methods will be cost and time consuming using traditional animal models (mostly rodents). For this reason, they proposed switching to zebrafish models of OCD, which have faster and cheaper genetic manipulation, phenotypic reproducibility of OCD-like behaviours and feasibility to develop high-throughput screenings for the discovery of novel OCD drug therapies.

3. Missing heritability in OCD

3.1. The missing heritability problem

Despite all the genetic studies performed in OCD, as well as in many other complex disorders, there is still a big difference between the proportion of phenotypic variance, predictable to be explained by genetic influences, and the heritability really explained by the genetic variants identified so far. This is known as the "missing heritability"[86].

Taking into account the *common disease/common variant* hypothesis[86], which assumes that the genetic component of complex diseases is the sum of the effects of common genetic variants with small to modest effect sizes, it was thought that, at least, GWAS would be an effective method to identify the genetic variation that contributes to the pathogenesis of OCD. However, GWAS seem to be insufficient to explain the heritability of, not only OCD, but most complex traits

and diseases[98]. So, where is the rest of the genetic variation underlying this type of disorders?

There are many plausible explanations (but no consensus) as to where this missing heritability is hiding. Some of the most supported hypotheses involve common genetic variants with small effect sizes, rare and low-frequency variants, structural variation, and gene-gene and gene-environment interactions.

3.1.1. Common genetic variants with small effect sizes

One possible explanation for the missing heritability issue is that there might be common genetic variants with such small effect sizes that they have not yet been identified by GWAS because of its inadequate statistical power[51,87]. To identify these kind of variants with enough statistical power it would be required to conduct GWAS with samples sizes of, al least, 60,000 individuals[87]. However, GWAS and meta-analyses of combined GWAS conducted to date using bigger sample sizes still only explain a small proportion of the heritability (20%)[84], so it is unclear how much more of the genetic variance underlying complex traits would be explained by increasing sample sizes[87].

3.1.2. Rare and low-frequency variants

Nowadays, much of the speculation about missing heritability is focused on the possible contribution of rare (MAF <1%) and low-frequency variants (MAF between 1% and 5%)[99]. Indeed, rare variation is already known to play an important role in human diseases, as many monogenic diseases are caused by highly penetrant and relatively rare variants. Moreover, loss-of-function (LoF) variants are very rare and they are less probable to occur than missense variants. Evolutionary theory predicts that deleterious variants are likely to be rare as a result of purifying selection[100].

The *common disease/rare variant* hypothesis[86] proposes that at least part of the genetic component of complex phenotypes may be due to the sum of the effects

of rare and low-frequency variants. This hypothesis posits that these variants are likely to have effect sizes larger than those of common variants, conferring a moderate but readily detectable increase in relative risk of complex phenotypes, without demonstrating clear Mendelian segregation and contributing substantially to the missing heritability (Figure 4)[99].

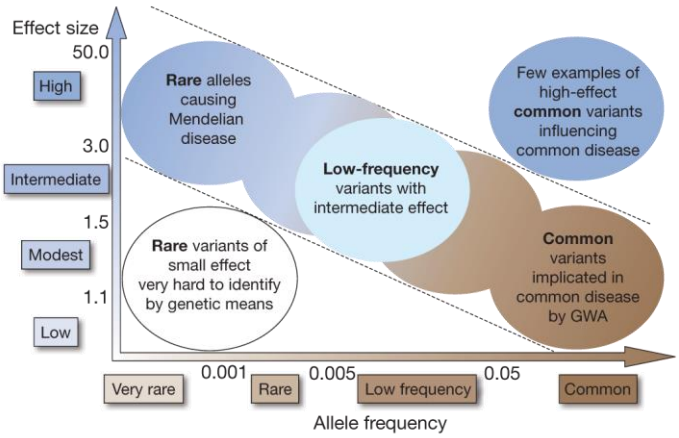


Figure 4. Classification of genetic variants by risk allele frequency and strength of genetic effect. From Manolio *et al.*, 2009[99]

Because rare variants are not usually captured in GWAS or candidate gene studies, they could be contributing substantially to missing heritability[99]. In fact, rare variants have been underrepresented on genome-wide genotyping arrays and are difficult to impute from common variants, as they are in low linkage disequilibrium[86,99]. Low-frequency variants are already being studied in GWAS with a high number of samples. In any case, both require extremely high sample sizes to be detected in enough samples for a sufficiently powered association test. Even so, once MAF falls below 0.5%, detection of associations becomes very unlikely[99]. Furthermore, rare variants involved in complex disorders cannot be studied by classical linkage analysis because they do not usually have enough large effect sizes. So, detection of rare and low-frequency variants is an essential step to evaluate the role of this variation in complex phenotypes.

3.1.3. Structural variation

Although there are several studies about SVs in complex traits and diseases, most of this kind of variation, such as, CNVs, novel sequence insertions, big deletions, inversions, translocations, and complex rearrangements, are incompletely assessed by most association studies. Hence, they could be contributing to some of the unexplained heritability of complex phenotypes.

Currently, much of the research in SVs is focused on CNVs, which have been shown to be associated with a few complex disorders. CNVs were first analysed through the array-based comparative genomic hybridization (CGH) technique, and later by GWAS. Disease-associated CNVs detected so far include both rare variants with large effect sizes and common variants with more modest effects[99]. In general, rare disease-associated CNVs are large (600 kb - 3 Mb), affecting many genes, whereas common disease-associated CNVs are smaller (20-45 kb)[99]. In particular, rare de novo CNVs have been shown to be of importance in several neuropsychiatric disorders[101].

In fact, there have been two studies analysing the involvement of CNVs in OCD. McGrath *et al*.[102] published a genome-wide investigation of large (>500 kb), rare (MAF <1%) CNVs in OCD and Tourette's syndrome, including 1,613 OCD patients. There was no global CNV burden difference between cases and controls, but there was a 3.3-fold increased burden of large deletions previously associated with other neurodevelopmental disorders. Moreover, OCD patients showed a 1.4% rate of *de novo* CNVs, slightly higher than estimates in controls (0.7%). Two regions stand out from this analysis: five cases shared deletions in chr16p13.11, three of them *de novo*, and most of them in individuals with OCD, and another four OCD patients presented SVs in chr22q11 (three duplications and one *de novo* deletion). Recently, Gazzellone *et al*.[81] published a high-resolution analysis of CNVs in a paediatric cohort of OCD, including 307 unrelated probands and 3,861 controls. They genotyped rare CNVs (MAF <0.5%) of at least 15 kb and they identified *de novo* CNVs in 4/174 trios. They also showed an enrichment of CNVs in genes that encode targets of the fragile

X protein (involved in intellectual disability), as well as deletions or duplications of exons in *ASTN2*, *NLGN1*, *PTPRD*, *DLGAP1*, *DLGAP2*, and *BTBD9*, all of them involved in neuronal processes. Furthermore, four individuals with OCD had CNVs involving known genomic disorder loci.

Despite the last improvements in SVs detection through SNP arrays, this type of variation still remains largely unexplored, especially for small indels and CNVs embedded within complex regions of the human genome. However, NGS approaches are evolving at a fast pace and are able to identify SVs with higher accuracy[103].

3.1.4. Gene-gene and gene-environment interactions

It has been proposed that another significant part of the missing heritability may be due to the interactions between genes or between genes and the environment.

Gene-gene interactions are also known as “epistasis”. Specifically, functional epistasis is the phenomenon where the effect of a particular variant on the phenotype depends on the genotype of another variant, while statistical epistasis is the effect of a combination of causal variants, where the sum of their effects are not independent[104]. It is thought that part of the missing heritability could be explained by epistatic genetic interactions between the already identified genetic variants. As an example, Zuk *et al.* reported that the amount of heritability explained by 71 risk loci associated with Crohn’s Disease was 21.5% under the assumption of additive genetic architecture, and 62.8% under the a model that considers epistatic interactions[51,105]. However, studying epistasis is an overflowing task and the magnitude of its actual contribution to the heritability of complex phenotypes still remains to be determined.

In addition to gene-gene interactions, gene-environment interactions could also influence the genetic architecture of complex traits through epigenetic changes of gene expression[51,86]. Although there is evidence for gene-environment

interactions in model organisms, human genetic-environmental interaction analyses still need to evolve and achieve robust results[51,106].

3.2. New approaches to decipher the missing heritability

Nowadays, many efforts are being made to identify the genetics behind the missing heritability in complex traits and diseases. Emerging research areas, methods and tools may help explain some of the underlying biology of complex phenotypes, including OCD, and to validate the proposed hypotheses for the missing heritability, which are not necessarily exclusive. We describe them in the following sections.

3.2.1. Next-generation sequencing

The emergence of NGS methods in the last decade has revolutionised genomic research as they started to become essential for the identification of human genetic variation in health and disease. These methods provide fast high throughput sequencing data, outperforming by several degrees of magnitude previous technologies. Nowadays, one single human genome can be sequenced in one day for approximately 1000 euros[107].

Despite the cost reductions in NGS, whole-genome sequencing (WGS), the process of determining the complete sequence of a genome at a single time, is still an expensive approach in the “genomics of disease” research field, where a lot of samples are needed to achieve statistically significant results. A cost-effective alternative consists of the enrichment of specific regions of interest, such as the exome (the protein-coding region of a genome), or a specific subset of genes or regions. In fact, although the exome represents less than 1-2% of the genome, whole-exome sequencing (WES) is a well-justified, and extensively used strategy for disease gene identification, because about 85% of known disease-related variants are located in exons[108].

NGS technologies include a variety of methods that are grouped broadly as template preparation, sequencing and imaging, and data analysis. The unique combination of specific protocols distinguishes one technology from another and determines the type of data produced from each platform[109]. Although Illumina/Solexa dominates the NGS market currently, some methods (e.g. Roche, Life/APG, Helicos BioSciences, Polonator and the near-term technology of Pacific Biosciences) have clear advantages for particular applications over others.

NGS has evolved in parallel with bioinformatics, as it relies heavily in bioinformatic algorithms for providing appropriate genotypes. Through these algorithms, NGS allows the detection of many different types of genetic variation, whether known or novel, such as nucleotide substitutions, small indels, and some SVs (inversions, translocations, CNVs, and novel sequence insertions). However, effective detection of SVs remains challenging and needs to improve in accuracy, sensitivity and specificity.

All this genetic variation can have a functional impact in the phenotype, which is usually well established when the variants lie in protein-coding DNA regions. Bioinformatic tools are used to annotate the location of the variants and to predict their functional effect. Variants in the coding region are classified as synonymous (no amino-acid change), missense (change of the amino acid encoded), nonsense (introduction of a premature stop codon) or stoploss (loss of a stop codon), while indels can be in-frame (multiple of three base-pairs) or frameshift if they lead to a change in the reading frame and thus the amino-acid composition of the protein. For missense variants, a prediction on the effect on the protein functionality is based on sequence conservation among species and on the possible effect of the amino acid changes to the protein structure. Variants outside of the coding region are annotated when they modify the sites at which splicing takes place, located at the extremes of exons and introns[110]. These variants may lead to abnormal splicing, by including intronic sequence or not including exons. Bioinformatic functional prediction is also possible for other non-coding variants that may affect regulatory elements controlling gene

expression[111]. Finally, the effect of VNTRs and SVs is usually not functionally annotated, as functional prediction is more complex. They may affect coding regions, inserting or deleting coding sequence, leading to in-frame or frameshift products. CNVs may increase or decrease gene copy numbers, and the breakpoints of SVs may occur within genes, leading to shortened or chimeric genes. Usually these effects are annotated manually.

In the case of OCD, there have been two published studies that have performed WES. Cappi *et al*.[112] published a WES study of 20 sporadic OCD cases and their unaffected parents to identify rare *de novo* SNVs conferring risk to the phenotype. They described that the rate of *de novo* SNVs in OCD was significantly higher than the rate of *de novo* SNVs in unaffected subjects. Moreover, several genes harbouring *de novo* SNVs were highly interconnected when a protein-protein interaction (PPI) network was constructed to analyse functional molecular interactions among them. These genes also ranked high when a Degree-Aware Disease Gene Prioritization (DADA) study was performed to observe relatedness to the candidate OCD genes reported by the two OCD GWAS published to date. Finally, they found enrichment in immunological and central nervous system functioning pathways. For instance, three of the most relevant genes were *WWP1*, *BAMBI* and *SMAD4*, all three involved in neurological processes. Gazzellone *et al.*[81] sequenced exomes of ten trios and identified a 13-bp exonic deletion in *DRD4*, supporting the hypothesis of the contribution of this gene in the OCD aetiology.

3.2.2. Rare variant association studies

Two types of rare variants association studies (RVAS) are used to explore the contribution of rare variants in the missing heritability of complex traits and diseases: variant-based tests and gene-based tests. Variant-based tests examine if a variant is enriched or depleted in cases versus controls from the general population (with individuals assumed to be unrelated). However, standard single variant association analyses are statistically underpowered to detect rare variant associations, except when sample sizes and/or effect sizes

are very large. To solve this problem, investigators have recently developed statistical methods based in aggregating or collapsing rare variants within biological units of association, defined using gene annotations, genomic coordinates or functional characterization[113]. Traditionally, rare variants are grouped by genes and, for this reason, these tests are also known as gene-based tests.

We can divide gene-based tests into four main categories: burden tests, adaptive burden tests, variance-component tests, and combined burden and variance-component tests (Table 3)[114]. These methods are based on different genetic architectures underlying complex phenotypes, and power for each test depends on the true disease architecture. As true genetic architectures of complex disorders are unknown, it is highly recommended to use more than one test when performing RVAS[114].

a) Burden Tests

Burden tests collapse information for multiple genetic variants into a single genetic score and test for association between this score and a trait[114]. The main limitation of these tests is that they assume that all rare variants in a set are causal and influence the phenotype in the same direction. However, it is known that the same gene can carry variants affecting the phenotype in opposite directions. So, in scenarios where only a small fraction of the rare variants are causal, or where both trait-increasing and trait-decreasing variants are present, these tests lose power.

b) Adaptive Burden Tests

These tests were developed to address the limitations of the original burden tests[114]. They are more robust because they allow for weighting of variants (by allele frequency or pre-estimated direction of the effect, for example) and for the inclusion of covariates when appropriate. However, most adaptive burden test are computationally intensive and simulation studies suggested that they have similar power to that of variance-component and combined tests.

The Kernel-Based Adaptive Cluster (KBAC)[115] method is an example of adaptive burden test that combines causal/non-causal variant classification and association testing. Moreover, it allows the incorporation of covariates to control for potential confounders including age, sex, and population substructure. This method was reported to perform particularly well in the presence of variant misclassification and gene interaction.

c) Variance-Component Tests

Variance-component methods test for association considering the distribution of genetic effects for a group of variants[114]. They are powerful in the presence of both trait-increasing and trait-decreasing variants or when there is only a small fraction of causal variants, but they lose power compared to burden tests when most variants are causal and have the same direction of effects.

The Sequence Kernel Association Test (SKAT)[116] is a computationally efficient variance component test that tests for association of common and rare variants and that offers flexibility in terms of covariate adjustment, study design, and different variant prioritization/weighting strategies.

d) Combined Burden and Variance-Component Tests

Some complex phenotypes may present a combination of scenarios. This means that it is possible that some regions present a large proportion of causal variants with effects in the same direction, while other regions present both trait-increasing and trait-decreasing variants, or a small fraction of causal variants. In these cases, it is better to use a method that combines burden and variance-component tests[114].

The SKAT Optimal test (SKAT-O)[117] and the Mixed effects Score Test (MiST)[118] are both combined methods. SKAT-O is an adaptive linear combination of unidirectional burden test and variance-component SKAT test, while MiST is a hierarchical regression model combining two independent test statistics that quantify variant effect sizes and directions of association.

Table 3. Summary of Statistical Methods for Rare Variant Association Testing. From Lee *et al.*, 2014[114]

	Description	Methods	Powerful when			
			A large proportion of variants are causal	**Only a small fraction of variants are causal**	**All or near all variants have the same direction of effect**	**There are both trait-increasing and trait-decreasing variants**
Burden tests	Collapse rare variants into genetic scores	ARIEL test, CAST, CMC method, MZ test, WSS	✓	✗	✓	✗
Adaptive burden tests	Use data-adaptive weights or thresholds	**KBAC method**, aSum, Step-up, EREC test, VT, RBT	✓	✗	✗	✓
Variance-component tests	Test variance of genetic effects	**SKAT**, SSU test, C-alpha test	✗	✓	✗	✓
Combined tests	Combine burden and variance-component tests	**SKAT-O**, **MiST**, Fisher method	✓	✓	✓	✓

ARIEL: accumulation of rare variants integrated and extended locus-specific; aSum: data-adaptive sum test; CAST: cohort allelic sums test; CMC: combined multivariate and collapsing; EC: exponential combination; EPACTS: efficient and parallelizable association container toolbox; EREC: estimated regression coefficient; GRANVIL: gene- or region-based analysis of variants of intermediate and low frequency; KBAC: kernel-based adaptive cluster; MiST: mixed-effects score test for continuous outcomes; MZ: Morris and Zeggini; RBT: replication-based test; Rvtests: rare-variant tests; SKAT: sequence kernel association test; SSU: sum of squared score; VAT: variant association tools; VT: variable threshold; and WSS: weighted-sum statistic.

Recently, Moutsianas *et al.*[119] evaluated the power of different gene-based tests under different genetic architectures, locus effect sizes, sample sizes, filters for neutral variation, and significance thresholds and reported that SKAT-O, KBAC and MiST have the highest individual mean power across simulated datasets. However, they observed wide architecture-dependent variability in the individual loci detected by each test, suggesting that inferences about disease architecture from analysis of sequencing studies can differ depending on which methods are used. Moreover, their results imply that tens of thousands of individuals, extensive functional annotation, or highly targeted hypothesis testing will be required to confidently detect or exclude rare variant signals at complex disease loci.

3.2.3. Transcriptomics

The effect of genomic and environmental factors can also be analysed through the study of the RNA, which can give us information about the effect of regulatory or splicing variants that we cannot detect by WES and/or that we do not know how to interpret them. The study of RNA can tell us which genes or pathways are involved in the phenotype and we can identify, later, the causal genetic or environmental factor.

About 88% of all the human genetic variation currently associated with complex traits and diseases by GWAS lie within intronic or intergenic regions and occur within putative regulatory elements far more often than expected by chance[120,121]. This suggests that this variation is likely to have causal effects by influencing gene expression rather than affecting protein function. Correspondingly, a growing number of studies are showing the relationship between genetic variants and gene expression variation, such as the latest effort of the GTEx consortium[122], describing the genetic effect on gene expression levels across 44 human tissues.

In fact, whole-transcriptome analysis (the study of the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological

condition[123]) is increasingly acquiring a pivotal role in the study of human complex traits and diseases. Currently, a valid approach in this field is focused on studying, first, the presence of differential gene expression between disease and healthy phenotypes, and then, examining the genome to identify genetic variants that could be responsible for the observed variation in gene expression. This approach is especially interesting in the case of regulatory variants, whose impact is difficult to predict from the DNA level. Indeed, it was demonstrated that combining WES and transcriptomics can provide a more comprehensive view of human diseases, increasing the overall diagnostic yield of exome-based studies up to 30%[124,125].

The first studies that investigated the role of RNA in specific phenotypes used microarrays and sequencing-based technologies, such as Serial Analysis of Gene Expression (SAGE), and Cap Analysis of Gene Expression (CAGE)[120]. But it was not until the development of NGS, and specifically RNA-sequencing (RNA-Seq), that we obtained significant progress in the resolution and analysis of different layers of transcriptome complexity[126]. When sequencing with high coverage, RNA-Seq allows to catalogue all species of transcripts, including mRNAs, small RNAs, microRNAs and non-coding RNAs; to determine the transcriptional structure of genes (their start sites, the splicing patterns, post-transcriptional editing and fusion transcripts); and to quantify gene expression levels (even when the levels of expression are low) in different conditions. It also allows us to analyse, at a single-nucleotide resolution, the allele-specific expression. Therefore, RNA-Seq provides a comprehensive view of the transcriptional landscape.

The RNA-Seq technology is very useful for differential expression (DE) analysis. Many software packages have been developed for the identification of differentially expressed genes between groups of samples[127]. They differ in the statistically design used. Some examples are shown in the Table 4.

Table 4. Summary of the top software packages developed for the identification of differentially expressed genes

Method	Description
baySeq[128]	Uses the Bayesian empirical approach to estimate a posteriori probability of each set of models, which defines differential expression patterns for each tuple.
limma+voom[129]	Based on the linear model and originally developed to analyse data from microarray and currently extended for RNA-Seq analysis. The limma user guide recommends the use of the TMM normalization of the edgeR package associated with the use of the voom conversion, which essentially transforms the normalized counts to logarithms base 2 and estimates the mean-variance relation to determine the weight of each observation made initially by a linear model.
edgeR[130]	A Poisson super dispersion model is used to account for technical and biological variation. Apply the Bayesian empirical method to moderate the degree of over dispersion against transcripts.
DESeq[131]	Based on a negative binomial distribution, with variance and mean bound by local regression.
DESeq2[132]	Firstly, it builds a model with observed counts. Secondly, it fits using the same method from the original DESeq, or fit in two steps: find the value of the parameter that makes the likelihood largest, which is called maximum likelihood estimation. Then, it takes all the gene values and moves these values towards an average value. It uses Bayes theorem to guides the amount of movement for each gene: if the information for the gene is low, its value is moved close to the average, if the information for the gene is high, its value is moved very little. Thus, the moved values are useful to evaluate different sets of genes as well as to apply a threshold.
NOIseq[133]	Empirically models the noise in the counting data and allows the data analysis without replication
SAMseq[134]	Uses re-sampling for sequencing counts with different depths. It can be applied to data with quantitative results, two-class, or multiple-class.

Several studies have evaluated these and other statistical methods for DE analysis, but no single method is clearly superior, since each has particular strengths that may be suitable for specific RNA-Seq datasets[127].

3.2.3.1. Transcriptomics and neuropsychiatric disorders

Due to the inability to access live brain tissues, which would be highly informative, transcriptomics studies of neuropsychiatric disorders have used, mostly, RNA samples from animal models and post-mortem brain tissues (Table 5)[120,135]. However, although they provide valuable information, they have some important inconveniences. For instance, animal models cannot reproduce the complete range of human neuropsychiatric symptoms. On the other hand, post-mortem brain tissues are subjected to changes in pH, hypoxia, dehydration and other factors that may affect stability of RNA products, and interfere in the

analysis. It is possible that differences in these factors between cases and controls may lead to false associations, while post-mortem samples of aged individuals may present symptoms of degenerative disease that could be confounded with the phenotype of interest. In addition, there is a scarcity of post-mortem brain samples and biobanks that hinder this approach.

Researchers have also used blood transcriptional profiles to discover possible disease biomarkers and investigate mechanisms relevant to mental disorders. Many studies have been done in different psychiatric disorders such as schizophrenia, autism or major depression, among others (Figure 5)[136–138]. This method represents a less-invasive and more feasible alternative to brain tissue. Further, there is a correspondence between blood and brain transcriptomic profiles: between 35% and 80% of known transcripts are present in both tissues, with correlations in the expression levels ranging from 0.25 to 0.64, and with stronger correlations observed among particular subsets of genes[139].

Another option is to perform transcriptomic studies on patient-derived cell lines. This is a very interesting field, especially since it is possible to generate iPSC-derived neuronal cell lines from patients and controls, which would help model neuropsychiatric disorders. Nevertheless, there remain many challenges in this type of studies.

To date, one study[140] has investigated the transcriptomics of OCD. It compares gene expression levels in various obsessive psychiatric disorders (which included OCD, obsessive-compulsive personality disorder or tics) and healthy subjects, using post-mortem brain tissue and microarrays. Doing so, they discovered 286 genes that were differentially expressed between cases and controls. However, they could not associate any known clinical risk SNV with gene expression differences observed in cases and controls.

Table 5. Summary of RNA sequencing studies in post-mortem human brains of psychiatric disorders. From Wu *et al*., 2017[135]

Studies	Neuropsychiatric disorder	Subjects		Post-mortem brain tissues	Results
		Cases	Controls		
Zhou *et al*. 2011	Cocaine and alcohol addiction	16	16	Hippocampus	Gene expression changes between both cocaine-addicted and alcoholic post-mortem brains and their respective controls
Wu *et al*. 2012	Schizophrenia	9	9	Superior temporal gyrus	Identification of three clusters strongly linked to schizophrenia: synaptic vesicle trafficking, neurotransmission-related functions, and neural development.
Sinclair *et al.* 2013	Schizophrenia	20	20	Prefrontal cortex	Abnormal expression of *FKBP5*, *PTGES3*, *BAG1*, and glucocorticoid receptor genes
Hwang *et al.* 2013	Schizophrenia	14	15	Hippocampus	144 differentially expressed genes in cases. Upregulation of immune/inflammation genes
Akula *et al.* 2014	Bipolar disorder	11	11	Dorsolateral prefrontal cortex	Altered expression of gene transcripts involved in neuroplasticity and circadian rythms
Kohen *et al.* 2014	Schizophrenia, Bipolar disorder and Major Depression disorder	50	29	Mid-hippocampus	Disrupted hippocampal miR-182 signalling
Cruceanu *et al.* 2015	Bipolar disorder	13	13	Anterior cingulate cortex	Dysregulation of G protein-coupled receptors
Farris *et al.* 2015	Alcoholism	16	15	Prefrontal cortex	Multiple ion channels and related processes in the human prefrontal cortex linked to extended alcohol abuse
Yin *et al.* 2016	Major Depression disorder and Suicide	30	29	Dorsal prefrontal cortex	Identification of dorsal striatum-specific immune response and oxidative phosphorylation pathways for BD
Pacifico & Davis 2017	Bipolar disorder	18	17	Dorsal striatum	Downregulation of *GABRG2* in suicide cases and identification of an SNP for association with suicide death

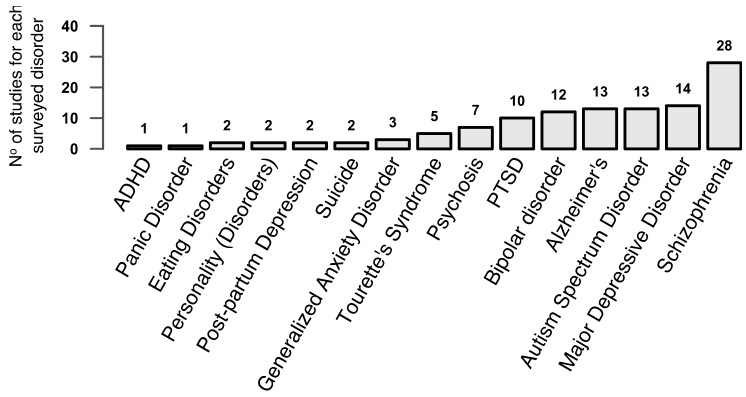


Figure 5. Classification of 108 human blood-based transcriptome gene expression studies performed between 2005 and 2015 by neuropsychiatric disease type. From Breen *et al.*, 2016[136]

3.2.4. The study of the microbiome

The human microbiome refers to the complex microbial ecosystem that resides on each person. It consists of about $3.8x10^{13}$ microbial cells and includes bacteria, archaea, fungi, protists and viruses[141]. The composition and abundance of the different microbial populations integrating the microbiome are variable not only between individuals but also within the different body parts. The metagenome refers to the collective genomes of the microbiome, although often, microbiome is used for both concepts[142].

In recent years, the compositional and functional diversity of the human microbiome has increasingly gained attention in the missing heritability problem[143]. Three main observations may explain the arising of this field: i) the microbiome has a strong impact in human health[144] and its composition is associated with many important traits, including obesity, cancer, and neurological disorders; ii) the human microbiome encodes 100 times more genes than the human genome[145], which may act as a rich source of genetic variation and phenotypic plasticity; and iii) human genotypes, host's behaviour, environment, and vertical and horizontal transmissions from other hosts can influence the composition and structure of the human microbiome, and this

microbiome can influence human phenotypes. Thus, if the development of a phenotype is related to the microbiome, looking only at human genetic variation produces a gap between phenotypic variance and observed genotypic variance, with high estimates of heritability values in family studies and low estimates in genetic studies.

Understanding the precise functionality and implications of the human microbiome is a complex task that is being carried out thanks to the development of metagenomics. There are two main metagenomic approaches: 16S ribosomal RNA (rRNA) gene amplicons sequencing and metagenomics shotgun sequencing.

To date, 16S-rRNA sequencing has been the most common approach to analyse the microbiome, although this technique is restricted to the identification of bacteria and archea. The 16S rRNA gene is present in all the species belonging to these two domains and is composed by an alternation of highly conserved and hypervariable regions (Figure 6). Genetic differences in the hypervariable regions have been considered to reflect, for most bacteria and archea, genome divergence. This method uses universal PCR primers complementary to highly-conserved regions to generate amplicons that contain hypervariable regions, which are then sequenced and used to infer taxonomic identifications based upon bioinformatic alignments against sequence databases[146].

Metagenomics shotgun sequencing is based in obtaining the completely set of genes that are present in the microbiome[147]. It provides better accuracy, and covers all taxonomical levels, including organisms such as viruses and fungi, which cannot be captured by 16S-rRNA sequencing. However, it is much more costly than 16S-rRNA sequencing, and the generated data presents numerous challenges in downstream analyses.

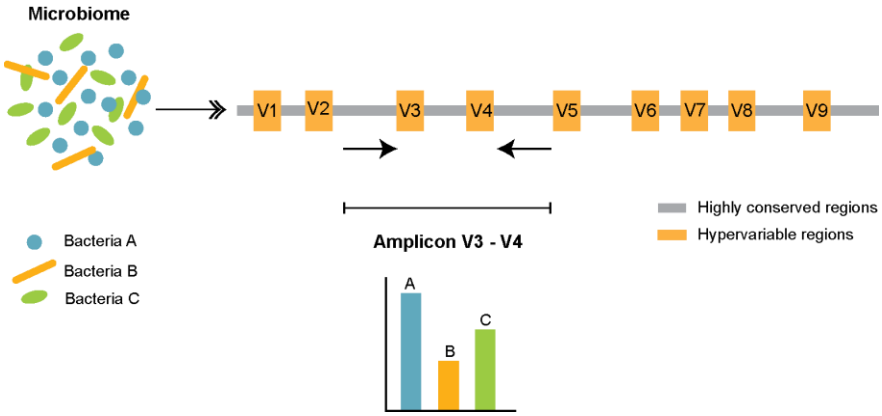


Figure 6. Amplification strategy of the hypervariable regions of the 16S rRNA gene

3.2.4.1. The gut-brain axis

The relationship between the gut microbiome and the brain has been one of the most studied systems in human metagenomics research. The existence of the gut-brain axis was proposed in 2004 by Sudo and colleagues[148], who discovered that germ-free mice had an impaired stress response, which could be reversed by reconstitution of a balanced microbiome at early developmental stages.

The gut–brain axis is a communication system that involves neural, hormonal and immunological signalling between the gut microbiome and the brain, with neural connections involving the central nervous (CNS), autonomic, and enteric nervous systems (ENS)[149]. This communication system is bidirectional: the brain can influence gastrointestinal and immune functions of the gut, and the gut microbiota and its metabolites can influence the brain (Figure 7).

Specifically, it is known that emotional factors can modulate gastrointestinal functions (such as motility, secretion and mucin production) and immune functions (such as modulation of cytokine production by cells of the mucosal immune system) of the gut. For example, it has been reported that stress can influence the chronic progression of some gastrointestinal illnesses, such as inflammatory bowel diseases (IBD)[150]. Stress and IBD have been associated, in

turn, with compositional changes in the gut microbiome, linking, so, the brain and the gut microbiota[151,152]. These modulations via the gut-brain axis are known as top-down modulations.

Bottom-up signalling has also been reported[149]. There is evidence that the gut microbiome influences brain development[153], neurogenesis[154], and brain function[149]. Some studies have shown that this influence may be explained via the gut-brain-axis by: i) the capability of the human gut microbiome to affect levels of excitatory and inhibitory neurotransmitters, such as serotonin, dopamine, norepinephrine and GABA, by producing and/or consuming them or by modulating host neurotransmitters and/or related pathways[155]; ii) the release of gut hormones from enteroendocrine cells; iii) the activation of the enteric nervous system and signalling of the brain via ascending neural pathways; and iv) activation of the immune system via cytokine release by the mucosal immune cells[156].

Interestingly, the greatest diversity of microbial genetic content resides in the human gut. Recently, Qin *et al.*[145] reported a human gut microbial gene catalogue, identifying 3.3 million non-redundant microbial genes from stool samples of 124 European individuals. They also showed that, although there is a considerable variability between individuals in the gut microbiome, there is a "core microbiome" shared between individuals, which can be defined as the set of genes present in a given habitat in all or the vast majority of humans[157]. Most of the microorganisms present in the gut are bacteria and there are already 1,000 species identified. The genus *Bacteroides* and the phylum *Firmicutes* account for over 90% of the total[149].

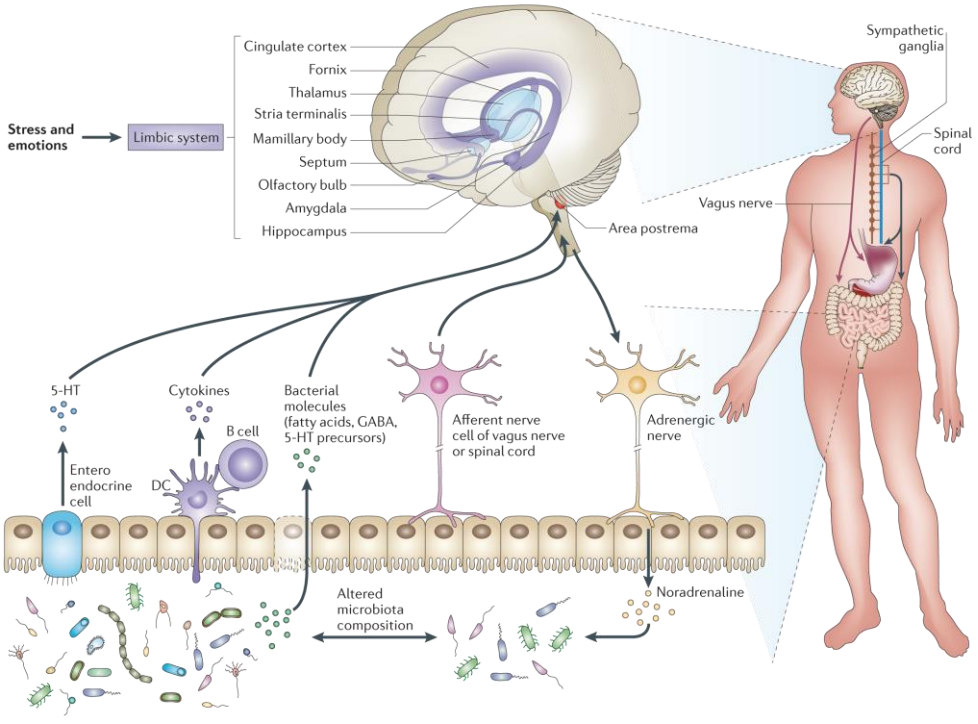


Figure 7. The neural, immunological, endocrine and metabolic pathways by which the microbiota influences the brain, and the proposed brain-to-microbiota component of this axis. Bacterial products gain access to the brain via the bloodstream and the area postrema, via cytokine release from mucosal immune cells, via the release of gut hormones from enteroendocrine cells, or via afferent neural pathways, including the vagus nerve. Stress and emotions can influence the microbial composition of the gut through the release of stress hormones or sympathetic neurotransmitters that influence gut physiology and alter the habitat of the microbiota. DC, dendritic cell; GABA, γ-aminobutyric acid. Adapted from Collins *et al*., 2012[149]

3.2.4.2. The human microbiome in psychiatric disorders

Based on the idea that gut microbiome can affect brain function, several studies have emerged focusing on variations in the microbiota and the effect on various neuropsychiatric disorders, including anxiety, depressive disorders, autism and schizophrenia, among others[158]. These studies, which have been done in both animals and humans, have mostly focused in the gut but also, to a lesser extent, in the oro-pharyngeal microbial composition, partly as a proxy to the gut microbiome, as it might be difficult to obtain stool samples from patients with psychiatric disorders.

While there are many differences in the microbial composition of the faecal and oral microbiome, Segata *et al.*[159] documented overlapping metabolic pathways of these two microbiome profiles, which supports the fact that many studies of microbiome and neuropsychiatric disorders have focused on the oral microbiome.

Several studies have observed the influence of the microbiome composition on the behaviour of animal models. For instance, there are several evidences in rodents that demonstrate the influence of the composition of the gut microbiota on anxiety and major depressive disorder. In these studies germ-free mice have demonstrated reduced anxiety-like behaviours[160,161] or depression-associated changes in stress response, accompanied by altered levels of monoamines and proinflammatory cytokines.

Human studies have reported a link to the microbiome for major depression disorder, autism, schizophrenia, anorexia and Alzheimer's disease, identifying differences in the composition of the microbiome in cases and controls. Specifically, two studies reported high-level differences at both phylum and genus levels between the oral microbiome of individuals with schizophrenia versus healthy controls[162,163]. Another study was able to classify patients with and without depression with 100% sensitivity and 97% specificity just by looking at the microbial genomes from stool swabs[164]. And a study on autistic patients

found a different microorganisms distribution in stool samples in cases versus healthy controls[165].

Other studies have evaluated response to probiotic administration. For instance, decreased cognitive reactivity to sad mood was observed in randomized controlled trials after a month of probiotics administration[166], and a fascinating study[167] showed a drastic drop to 0% in the rates of autistic spectrum disorders and ADHD in young teenagers supplemented with probiotics as infants compared to controls (individuals not supplemented).

Regarding OCD, there are still very few studies investigating the role of human microbiome in this disorder. Even so, there are some evidences that allow us to hypothesize that changes in human microbiota could explain some of the missing heritability in OCD. For example, there is one study that showed attenuation of obsessive-compulsive behaviour in mice after *Lactobacillus rhamnosus* probiotic administration[168]. Another study performed in healthy humans showed reduced obsessive-compulsive subscores on the Hopkins symptoms checklist after *Lactobacillus helveticus* and *Bifidobacterium longum* administration during 30 days[169]. Recently, just some months ago, Jung *et al.*[170] reported that compulsive checking in mice was accompanied by changes in several communities of bacteria belonging to the order *Clostridiales* (class *Clostridia*, phylum *Firmicutes*), and predominantly in *Lachnospiraceae* and *Ruminococcaceae* families of bacteria. Finally, a very recent study has shown that there is an altered bacterial community structure in the gut of PANDAS patients with respect to controls[171]. They suggested that this altered microbiome developed after streptococcal infections and could have lead to a pro-inflammatory status that may influence behaviour and brain functions and result in the sudden onset of tics, OCD, and other behavioural symptoms.

HYPOTHESES AND OBJECTIVES

OCD is a complex neuropsychiatric disorder that arises from a combination of environmental and genetic risk factors, as demonstrated by family and twin studies. However, the genetic architecture underlying this disorder has not been elucidated yet. The current model proposes that individuals with OCD may have a genetic predisposition to the impact of environmental factors that may modify the expression of genes involved in the serotonin, glutamate, catecholamine, and dopamine systems, which are involved in the OCD pathophysiology. These modifications, may, in turn, lead to an imbalance in the CSTC circuit, resulting in the manifestation of clinical features of OCD. Despite a number of genetic linkage, candidate genes, and association studies have been performed in order to decipher the biology underlying OCD, there is still a big gap between the proportion of phenotypic variance expected to be explained by genetic influences and the heritability explained by the genetic variants identified so far.

Nowadays, there are many plausible explanations as to where the missing heritability is hiding in complex disorders. Much of the speculation about it involves rare and low-frequency variants, common genetic variants with small effect sizes, structural variation, and gene-gene and gene-environmental interactions. In order to study these hypotheses, new genomic approaches are being developed. Moreover, transcriptomics is gaining increasing attention in the missing heritability problem, as the effect of genomic and environmental factors may be explained through the study of the RNA. Recently, metagenomics have also emerged as a new focus to study neuropsychiatric disorders based on the idea that variations in the human microbiome could affect brain function.

Based on these aforementioned facts, our hypotheses are:

1. Rare variants and common variants of small size effects contribute to OCD risk.

2. Some genes could be differentially expressed between OCD subjects and healthy individuals, which may be captured in blood transcriptomics.
3. There is an altered gut and oro-pharyngeal microbiome profile in OCD cases compared to healthy individuals, and this alteration may be modified by OCD treatment.

Arising from these hypotheses, our general objective is to gain insight into the genetic and environmental factors contributing to OCD risk. This general objective is assessed from two perspectives, which we separated in two different studies:

Objectives of *Study I: Deciphering OCD by whole-exome sequencing*

i. To identify genes and pathways with an enrichment of rare variants associated with OCD through:
   a) Rare variant association analysis from whole-exome sequencing data.
   b) Replication of the best candidates through targeted resequencing.
   c) Functional validation of candidate genes.
ii. To identify low-frequency and common, exonic, potentially damaging variants associated with OCD.

Objectives of *Study II: Multiomics longitudinal study of OCD*

iii. To identify differentially expressed genes in peripheral blood of OCD untreated cases compared to controls, reflecting an OCD specific transcriptome signature.
iv. To compare the gut and oro-pharyngeal microbiome profiles in OCD untreated cases and controls and evaluate potential differences.
v. To compare the transcriptomic and microbiome profile of the same OCD individuals after treatment and consider the potential effect of treatment.

METHODS

1. Subjects

To perform the work presented here, we had access to two independent OCD cohorts. Our main study cohort consisted of 668 adulthood-onset OCD patients that were recruited between 2004 and 2017 at the OCD Clinic and Research Unit of the Hospital Universitari de Bellvitge, in Barcelona. All patients met the DSM-IV[3] criteria for OCD for at least one year and their diagnosis was done by two experienced psychiatrists using the Structured Clinical Interview for DSM-IV Axis I Disorders - Clinician Version (SCID-IV)[172]. The severity of OCD symptoms was assessed using the Spanish clinical version of the Yale-Brown Obsessive–Compulsive Scale (Y-BOCS)[173,174]. All the patients were from Spain and had European ancestry. Exclusion criteria included: i) age under 18 or over 65; ii) presence or past history (in the previous six months) of psychoactive substance abuse or dependence; iii) mental retardation; iv) neurological disease comorbidity except tic disorder; v) present or past history of psychotic disorders; and vi) presence or past history of any other severe medical condition. Comorbidity with other Axis I disorders was not considered an exclusion criterion provided that OCD was the main diagnosis and the reason for seeking medical assistance. From these 668 OCD patients, 625 were included in the Study I (WES analyses) and were assessed at one time-point, where we collected blood to extract DNA (Table 6). They were all undergoing cognitive behavioural therapy (CBT) and pharmacological treatment. The remaining 40 patients were included in the longitudinal design of Study II, and were assessed at two different time-points: before and after at least three months of CBT and pharmacological treatment. For these patients we collected blood, pharyngeal swab and stool samples at the two time-points, as well as the dietary information (see questionnaire and dietary data in Supplementary Tables S1 and S2) from nearly all patients (Table 7). Three patients were included in both studies.

In addition, for replication analysis we included 117 children and adolescents aged between 8 and 19 years with a current diagnosis of OCD according to DSM-IV[3] criteria recruited by the Department of Child and Adolescent Psychiatry and Psychology at the Hospital Clínic i Provincial de Barcelona. The Spanish version of the semi-structured diagnostic interview K-SADS-PL (Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version)[175] was administered to both parents and the child as informant in order to establish the diagnosis of OCD and to assess past and current psychiatric comorbidity. OCD severity was measured at the time of admission using the Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)[176]. The age of onset of OCD was defined as the age at which patients first displayed significant distress or impairment associated with obsessive-compulsive symptoms. Exclusion criteria included intellectual disability and neurological disorders. Patients with psychiatric comorbidities were not excluded.

An age and gender matched control cohort comprising 105 unrelated healthy individuals from the same socio-demographic environment was also recruited by the OCD Clinic and Research Unit of the Hospital Universitari de Bellvitge. Prior to inclusion, each control participant underwent the Structured Clinical Interview for DSM-IV (non-patient version) to exclude presence or past history of any psychiatric disorder. Additional exclusion criteria were the same used for OCD patients. For 63 samples, only blood was collected, while for controls included in the longitudinal study we collected blood, pharyngeal swab and stool samples, as well as the dietary information at a single time-point for nearly all samples (Table 8 and Supplementary Table S3).

Written informed consent was obtained for all the participants after receiving a complete description of the study, which was done according to the principles of the Declaration of Helsinki after approval by the appropriate Ethic Committees (the Bellvitge University Hospital Ethical Committee, Barcelona, Spain; and the Hospital Clínic Ethical Committee, Barcelona, Spain).

In addition, exome sequencing data from 1896 additional in-house samples from other projects (involving controls and patients from different pathologies) were included in the bioinformatic analysis. From these, 567 were used as controls in the RVAS performed in the Study I. Finally, 1481 controls from the Multi Case-Control (MCC) – Spain cohort[177] were used in the targeted resequencing replication study.

Table 6 displays the collected samples for Study I and Study II. Tables 7 and 8 show all the analysis performed on samples collected in the Study II. Supplementary Table S4 displays the clinical information for nearly all OCD samples recruited and Supplementary Table S5 displays the clinical information for the OCD and control samples included in the metagenomics study.

Table 6. Samples included in the Study I and Study II

Analysis		OCD patients		Controls	
		Adulthood-onset OCD patients	**Childhood-onset OCD patients**	**Healthy individuals**	**Other controls**
Study I	WES	306	-	63 (BUH)	1896 (Various)
	Targeted	322 (BUH)	117 (CH)	8 (BUH)	1473 (MCC)
Study II		43* (BUH)	-	34 (BUH)	-

BUH: Hospital Universitari de Bellvitge, Barcelona; CH: Hospital Clínic i Provincial de Barcelona. Various: Various hospitals in Catalonia; MCC: Multi Case-Control (MCC) – Spain cohort. * Three of these samples were also included in Study I.

Table 7. Summary of samples and analyses performed on OCD cases included in Study II

OCD samples	Genomics	Transcriptomics				Metagenomics				Total
	WES	mRNA T0	mRNA T3	sRNA T0	sRNA T3	S T0	S T3	Ph T0	Ph T3	
493	✗	✗	✗	✗	✗	✗	✗	✓	✗	1
569	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
644	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
834	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
875	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
884	✓	✓	✓	✓	✓	✓	✗	✓	✓	8
885	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
893	✓	✓	✓	✓	✓	✗	✓	✗	✓	7
896	✓	✓	✓	✓	✓	✓	✗	✓	✗	7
897	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
898	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
899	✓	✓	✓	✓	✓	✓	✓	✗	✓	8
937	✓	✓	✓	✓	✓	✗	✗	✗	✗	5
953	✗	✗	✗	✗	✗	✗	✓	✓	✓	3
972	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
992	✓	✓	✓	✓	✓	✓	✓	✓	✗	8
994	✗	✗	✗	✗	✗	✓	✗	✗	✗	1
998	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1000	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1001	✗	✗	✗	✗	✗	✗	✗	✓	✓	2
1002	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1011	✓	✓	✓	✓	✓	✓	✗	✗	✗	6
1012	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1014	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1019	✓	✓	✓	✓	✓	✗	✗	✗	✗	5
1021	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1025	✓	✓	✓	✓	✓	✓	✓	✗	✓	8
1027	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1030	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
1031	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
2004	✓	✓	✓	✓	✓	✗	✗	✗	✗	5
2005	✓	✓	✓	✓	✓	✗	✗	✗	✗	5
2013	✓	✓	✓	✓	✓	✗	✗	✗	✗	5
2017	✗	✗	✗	✗	✗	✓	✓	✓	✓	4
2020	✓	✓	✓	✓	✓	✓	✓	✗	✓	8
2022	✓	✓	✓	✓	✓	✗	✗	✓	✓	7
2024	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
2026	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
2027	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
2029	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
2030	✓	✓	✓	✓	✓	✗	✓	✓	✓	8
2031	✓	✓	✓	✓	✓	✓	✓	✓	✗	8
2032	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
Total ✓	**38**	**38**	**38**	**38**	**38**	**32**	**31**	**32**	**32**	**317**
Total ✗	**5**	**5**	**5**	**5**	**5**	**11**	**12**	**11**	**11**	**70**

WES: whole-exome sequencing; mRNA: messenger RNA; sRNA: small RNA; S: stool samples; Ph: Pharyngeal swabs; T0: samples recruited before treatment; T3: samples recruited after, at least, three months of treatment. In grey, samples that have all analyses done.

Table 8. Summary of samples and analysis performed on controls included in Study II

Controls	Genomics	Transcriptomics		Metagenomics		Total
	WES	mRNA	sRNA	S	Ph	
C 91	✓	✓	✓	✓	✓	5
C 92	✓	✓	✓	✓	✓	5
C 93	✓	✓	✓	✓	✓	5
C 97	✓	✓	✓	✓	✓	5
C 98	✓	✗	✓	✗	✗	2
C 99	✓	✓	✓	✓	✓	5
C 100	✓	✗	✓	✓	✓	4
C 101	✓	✗	✗	✓	✓	3
C 102	✗	✗	✗	✓	✗	1
C 103	✓	✓	✓	✓	✓	5
C 104	✓	✓	✓	✓	✓	5
C 105	✓	✓	✓	✓	✓	5
C 106	✓	✓	✓	✓	✓	5
C 107	✓	✓	✓	✓	✓	5
C 108	✓	✓	✓	✓	✓	5
C 109	✓	✓	✓	✓	✓	5
C 110	✓	✓	✓	✓	✓	5
C 111	✓	✓	✓	✓	✓	5
C 112	✓	✓	✓	✓	✓	5
C 113	✓	✓	✓	✓	✓	5
C 114	✓	✓	✓	✓	✓	5
C 115	✓	✓	✓	✓	✓	5
C 116	✓	✓	✓	✓	✓	5
C 117	✓	✓	✓	✓	✓	5
C 118	✓	✓	✓	✓	✓	5
C 119	✓	✓	✓	✓	✓	5
C 120	✓	✓	✓	✓	✓	5
C 121	✓	✓	✓	✓	✓	5
C 122	✓	✓	✓	✓	✓	5
C 123	✓	✓	✓	✓	✓	5
C 124	✓	✗	✓	✓	✓	4
C 125	✓	✓	✓	✓	✓	5
C 126	✓	✓	✓	✓	✓	5
C 127	✓	✓	✓	✓	✓	5
Total ✓	33	29	32	33	32	159
Total ✗	1	5	2	1	2	11

C: control; WES: whole-exome sequencing; mRNA: messenger RNA; sRNA: small RNA; S: Stool samples; Ph: Pharyngeal swabs. In grey, samples that have all analyses done.

2. Study I: Genomics

2.1. DNA samples and quality control

For our main cohort, blood samples were collected in EDTA tubes and processed 24-48 hours after collection. Samples were first centrifuged at 2500 × g at room temperature for 10 minutes. Plasma was then removed and stored for future use. The remaining blood cells were subjected to DNA extraction using the Wizard® Genomic DNA Purification Kit (Promega Corporation) according to the manufacturer's instructions. DNA was quantified for each sample using the Qubit dsDNA BR Assay Kit (Invitrogen) and DNA integrity was assessed by running the samples on a 1% agarose gel stained with SYBR Safe DNA Gel Stain (Invitrogen).

2.2. Whole-exome capture and sequencing

We generated whole-exome libraries for 437 samples (306 OCD cases and 63 controls from Study I, and 35 OCD cases and 33 controls from Study II). The improvement of NGS capture kits during the project development has led to the usage of different capture kits for exome sequencing, to take advantage of the better capture options.

The initial 40 captures were performed with TruSeq DNA Sample Preparation Kit (Illumina) and Agilent SureSelect Human All Exon 35Mb Kit (Agilent 35; Agilent Technologies). From then on, all samples were captured with NimbleGen SeqCap EZ Library v3.0 (NimbleGen v3; Hoffmann-La Roche). Of these, 64 samples were prepared with TruSeq DNA Sample Preparation Kits (Illumina) and the remaining samples with NEXTflex™ Pre-Capture Combo Kit, NimbleGen SeqCap EZ Compatible (Bioo Scientific). A workflow for the whole-exome capture process is summarized in Figure 8.

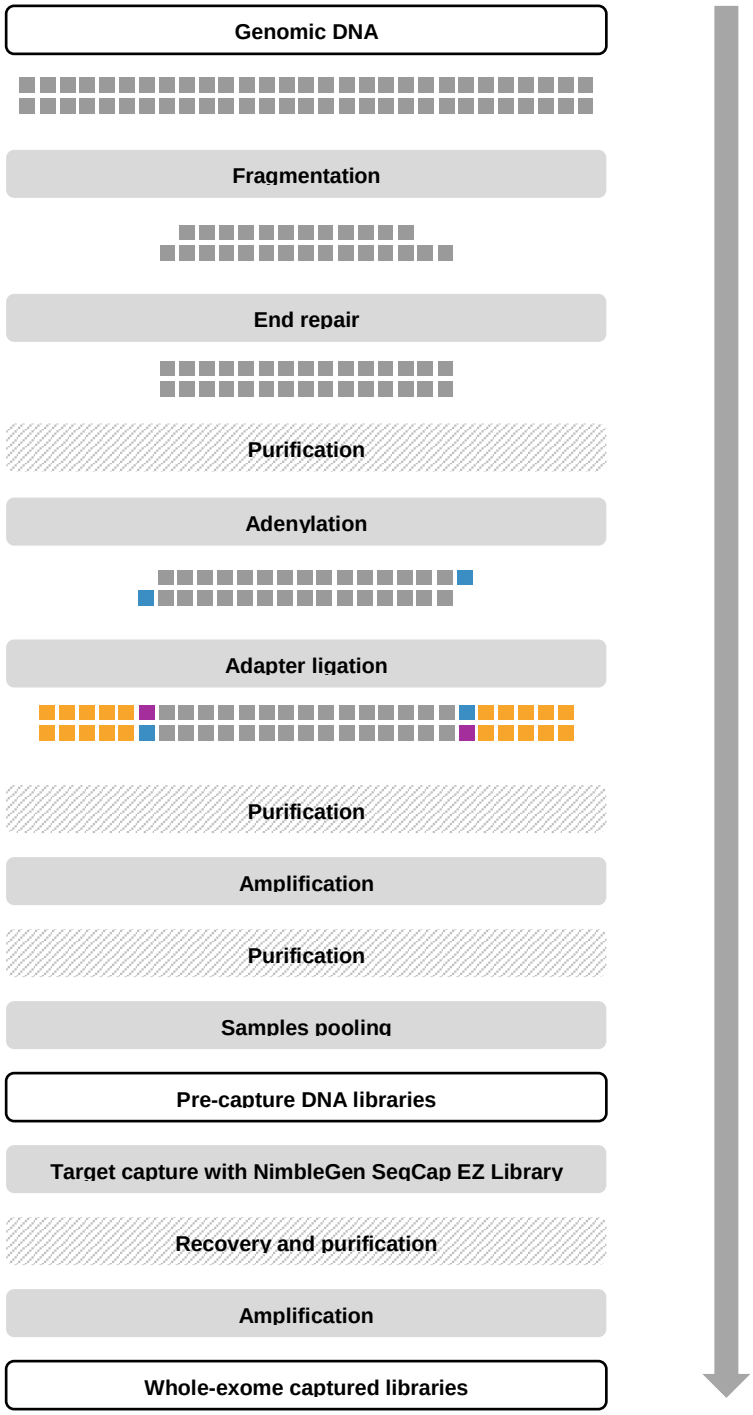


Figure 8. Whole-exome capture workflow with NEXTflex™ Pre-Capture Combo Kit, NimbleGen SeqCap EZ Compatible (Bioo Scientific) and NimbleGen SeqCap EZ Library v3.0 (Hoffmann-La Roche)

Briefly, for each sample, about 500 ng – 1µg of genomic DNA was fragmented using ultrasonication to generate double stranded DNA fragments of approximately 500 bp with 3' and 5' overhangs on a Covaris S2 instrument (Covaris). Then, fragments underwent under three enzymatic steps: end repair, adenylation, and ligation with specific paired-end indexed adapters (that will be used later for sequencing on the Illumina HiSeq). All purification steps between the enzymatic reactions were performed with AMPure XP beads (Beckman Coulter). Later, the fragments were amplified by polymerase chain reaction (PCR) with kit-specific PCR oligos to generate genomic libraries. PCR products were cleaned again using AMPure XP beads and quantified by a DNA High Sensitivity chip on a Bioanalyzer 2100 instrument (Agilent Technologies). Then, the DNA libraries were multiplexed in pools of 4 samples for a final combined mass of 1.1 µg, and the resulting library pools were hybridized to the biotin labelled probes of NimbleGen v3. A physical pull-down was then performed using streptavidin-bound T1 Dynabeads (Life Technologies) to purify the library-bait hybrids. After stringent washing (to remove nonspecific binding), each library pool was amplified by PCR with the kit-specific PCR oligos. After PCR amplification, the library pools were cleaned using QIAquick PCR Purification Kit (Qiagen), and quantified by a DNA High Sensitivity chip on a Bioanalyzer 2100 instrument. Each pool was then sequenced on one lane of an Illumina HiSeq2000 / Illumina Hiseq 3000 (Illumina) to generate 2 x 100 bp paired-end reads using SBS v3 chemistry (Illumina).

2.3. Bioinformatic analyses of DNA variants

After sequencing, we were provided fastq files for each sample by the sequencing facility. These were analysed with the in-house developed pipeline described below (Figure 9).

The analysis for Study I included data from 628 OCD cases, 63 controls and 1896 samples studied by our group. Inclusion of all these samples in the bioinformatic analyses provides more accurate results in variant calling as it

increases the sample size. Targeted data was analysed following the same workflow implemented for WES analyses.

2.3.1. Alignment

Reads were aligned to the GRCh37/hg19 version of human reference genome using the Burrows-Wheeler Alignment Maximal Exact Matches (BWA-MEM) algorithm version 0.7.10[178]. Alignment post-processing was performed according to GenomeAnalysisTK (GATK) best practice guidelines[179] by using picard-tools[180] and the GATK 3.2-2 version pipeline[181]. This included conversion of the mapping data (SAM file) to a sorted BAM file, PCR duplicate marking, local re-alignment around potential insertions/deletions, and base-quality recalibration. The resulting alignments were used as input for variant calling.

2.3.2. Variant calling

Variant calling was performed using GATK HaplotypeCaller v3.3[179,181]. The HaplotypeCaller algorithm calls SNPs and indels simultaneously via local de-novo assembly of haplotypes in active regions. We used HaplotypeCaller in its GVCF mode. In this mode an intermediate genomic gVCF file is generated per sample. These files are used for joint genotyping of multiple samples in a very efficient way.

2.3.3. Variant quality filtering

Potential false positive variant calls were filtered out based on six statistical annotation scores at the individual variant site and/or across samples: i) a minimum depth of coverage of 10 reads per variant; ii) a maximum allele balance bias (ABB) of 0.7; iii) Fisher strand bias (FS) in the top 10 percentile among all variants; iv) alternative allele frequency with thresholds at individual variant site of 0.2 and average across samples of 0.25; v) genotype quality score with a threshold of 20 at individual site and a minimum average across samples of 30; and vi) a minimum call rate across samples of 80%.

Variants that passed the hard-filtering step were then scored with the GATK Variant Quality Score Recalibration (VQRS) tool[179,181]. The VQSR filter uses annotation metrics, such as quality by depth, mapping quality, and variant position within reads, from a set of "true" variants (variants found in HapMap phase 3 release 3) to generate an adaptive error model. It then applies this model to the remaining variants to calculate a probability that each variant is real (and not a sequencing or data processing artefact). This probability is a recalibrated quality score called variant quality score log-odds (VQSLOD), which can be used to filter lower quality variants. We applied a GATK VQSLOD filter corresponding to a threshold that maintains 99.9% sensitivity for the "true" variants.

2.3.4. Annotation

Functional annotation of high quality variants was performed using ediva[182], which provides information from multiple databases, including SNP ID from the National Center for Biotechnology Information (NCBI) SNP Database (dbSNP) build 132, genomic annotation (exon/intron/UTR), gene annotation, variant type (synonymous, missense, nonsense, stopgain, splice variants), conservation around variants based on phastCons; segmental duplication filter, multiple estimates of the impact of amino acid substitution on the structure and function of proteins (tools: Sift, Polyphen2, Condel, LRT, PhyloP, MutationAssessor and MutationTaster), and frequencies of predicted variants in the 1000 Genomes project (1000G)[183], in the Exome Variant Server (EVS)[184] and in the Exome Aggregation Consortium (ExAC)[185].

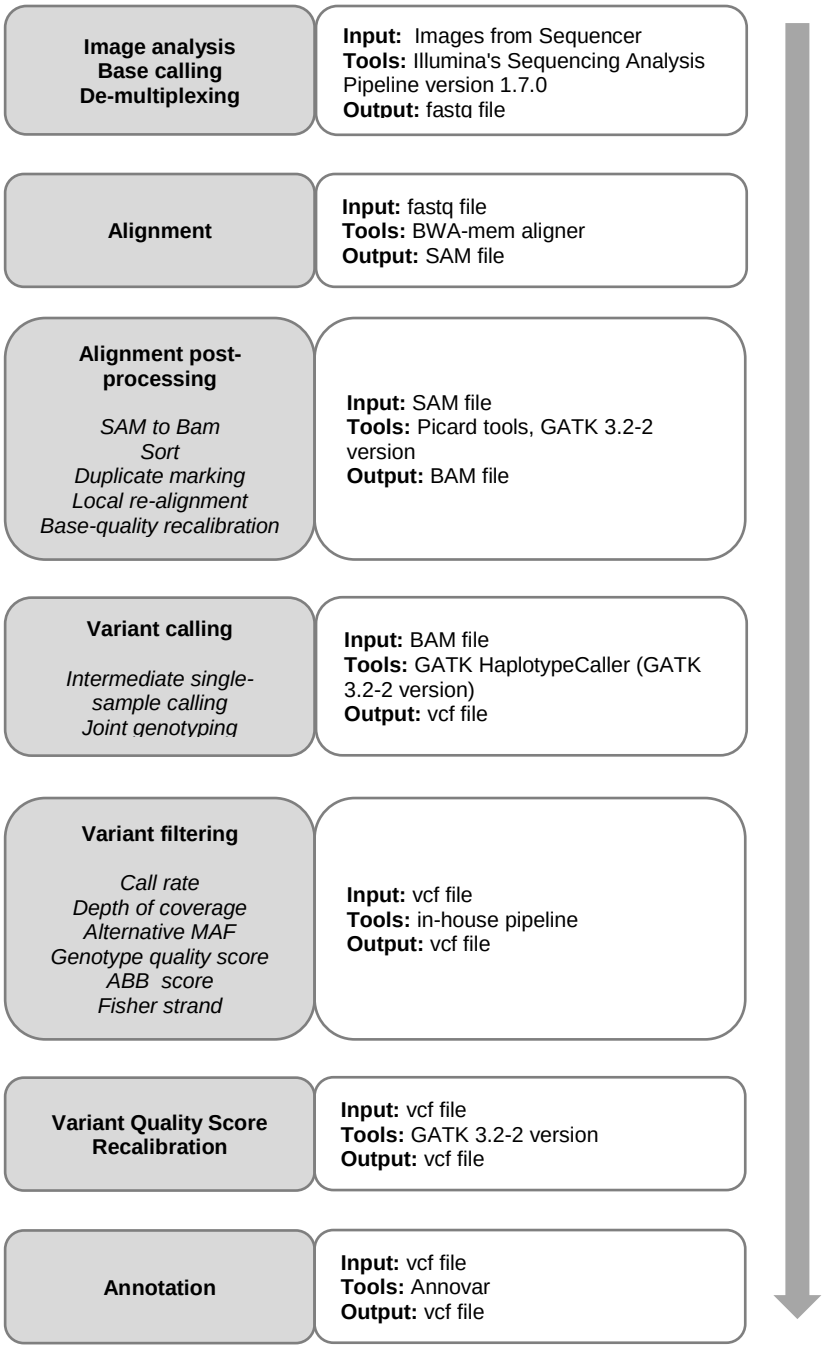


Figure 9. Flowchart used for bioinformatic analyses of DNA variants

2.4. Rare variant association study

Before running RVAS analysis, we used the software VCFtools[186] to exclude unknown related samples and an quality control (QC) in-house pipeline to filter samples for inclusion in the analysis based on different criteria: i) sample ID; ii) number of variants per sample; iii) sample transition to transversion (Ti/Tv) ratio; and iv) data stratification (technical or population based) (see Supplementary Methods S1 and Supplementary Figures S1-S8 for more details). Data stratification was computed through principal component analysis (PCA) of all synonymous SNVs without linkage disequilibrium, and outlier samples could be filtered out. PCA data was saved to be included as covariates in the RVAS method, allowing to correct possible batch effects. In addition, for studies including data obtained with different capture kits, variants that were not well covered by all kits were filtered out to remove kit-based stratification. Table 9 displays the number of variants before and after filtering for each QC analysis performed.

Table 9. Number of variants before and after QC analyses

		Study I RVAS	Study I RVAS	Study I Targeted	Study I WES analysis
Capture kit		Agilent 35 Agilent 50 NimbleGen v3	NimbleGen v3	Targeted	NimbleGen v3
OCD cases	**Before QC**	306	266	439	35
	After QC	292	206	1481	28
Controls	**Before QC**	630	253	427	33
	After QC	601	188	1474	28
Variants	**Before QC**	1,184,368	2,035,201	20,378	428,104
	After QC	624,516	490,150	13,751	403,973

QC: quality control; Agilent 35: Agilent SureSelect Human All Exon 35Mb Kit; Agilent 50: Agilent SureSelect Human All Exon 50Mb Kit; NimbleGen v3: NimbleGen SeqCap EZ Library v3.0.

The exome sequencing data was used to perform two different RVAS analyses. As the OCD samples were sequenced with two different kits (Agilent 35 and NimbleGen v3), a first RVAS included, after filtering steps, high-quality whole-exomes of 292 OCD samples and 601 controls, captured with Agilent 35, Agilent

SureSelect Human All Exon 50Mb Kit (Agilent 50) and NimbleGen v3. A second more homogenous analysis included high-quality whole-exomes of 206 OCD samples and 188 controls, all captured with NimbleGen v3. In both cases, the controls included the 63 healthy individuals recruited for this study, and additional samples selected from the 1896 samples analysed by our group. The selected samples for the first analysis consisted of: control samples, healthy parents from intellectual disability patients, and an even number of chronic lymphocytic leukaemia, fibromyalgia, cystic fibrosis and stroke patients. In the second analysis, only healthy parents from intellectual disability patients, fibromyalgia, and stroke patients were included.

For replication, a RVAS was done with high-quality targeted resequencing data from 427 OCD samples and 1474 controls (854 control samples were used to test association, whereas the rest was used to estimate the local AF).

RVAS was carried out using an in-house pipeline (Susak *et al.*, in revision), aggregating rare variants at the gene level. The pipeline includes four different RVAS methods: Burden test, KBAC[115], SKAT-O[187] and MiST[118]. The last version of this pipeline, which we used in the targeted resequencing analysis, included also a new Bayesian rare variant Association Test using Integrated Nested Laplace Approximation or INLA (BATI). This method is conceptually similar to the MiST approach in terms of statistical model specification but it is based on Bayesian inference (Susak *et al.*, in revision).

RVAS analysis was performed for all variants and considering separately all missense or all truncating mutations. We also applied two different MAF cut offs, of 0.01 and 0.005, filtering out variants above these cut-offs in our dataset and in the 1000 Genomes Project, EVS and ExAC databases. For missense variants, we also filtered out variants with a Cadd2 score below 15. When possible (SKAT-O, MiST, BATI (only for the replication data)), we included the first 10 principal components in the model as covariates (Supplementary Figure S9), and weighted the variants by MAF (giving more weight to rarer variants). In the MiST and BATI analysis we also included the Cadd2 score as additional variant

information. Finally, in the BATI analysis, we also included the number of variants per sample and the Ti/Tv ratio as covariates.

The RVAS output is a table including the number of unique variants considered for each gene, the number of unique variants observed in cases and in controls, the number of carriers in cases and in controls, the number of cases and controls that participated in the analyses, and the overall and FDR-adjusted p-values for the different Euclidean tests (in the case of frequentist tests, all but BATI). In the case of BATI, the Deviance Information Criteria (DIC)[188] is the Bayesian criteria applied for significance purposes (Susak *et al.*, in revision). To determine the DIC cutoff value corresponding to a certain FDR, we performed simulations. For each simulation we randomly shuffled cases and controls and then applied the BATI test in each gene. By doing so we expect that if a gene is found to be associated to the group of cases constructed artificially is due by random chance rather than by any true biological signal. The DIC threshold for a 0.01 FDR is then obtained from the 0.01 quantile of the empirical distribution of p-values across all genes.

We filtered out highly variable genes reported unlikely to be good candidates for disease causation[189].

2.5. Variant validation

Rare variants in significant genes that were found to be present in an unexpected high number of cases or controls were considered likely false positives and tested by conventional Sanger sequencing approaches.

Specific primers (Supplementary Table S6) were designed to surround the candidate variant with the software Primer3[190], tested *in silico* with Blat[191] and the USCS In-Silico PCR[192] tool, and ordered from Sigma-Aldrich Corporation. DNA was amplified using standard PCR amplification conditions and visualized on a 1.5% agarose gel. Unincorporated primers and dNTPs were removed with ExoSAP-IT (Affymetrix), and a 10 µl sequencing reaction was setup with 2 µl of

BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Byosistems), 10 μM of forward or reverse primer and 1-2 μl of PCR product. Sequences were purified by gel filtration using Sephadex G50 Gel-Filtration Resin (Sigma-Aldrich Corporation) and run on an ABI 3730XL DNA analyser (Applied Byosistems) at the UPF Genomics Core Facility. Data files were analysed with the software CLC Main Workbench (CLC bio).

2.5.1. *DRD4* deletion genotyping

We were particularly interested in a 13-bp frameshift deletion (p.78_82del) in *DRD4* for which we set-up a multiplex PCR to directly genotype all recruited OCD cases and a similar number of controls obtained from a cohort of samples from the Hospital Universitari Vall d'Hebron (Barcelona, Spain) and from a cohort of general population of school children recruited for the BRain dEvelopment and Air polluTion ultrafine particles in scHool childrEn (BREATHE) project[193] (Barcelona, Spain). Specifically, we genotyped 614 OCD patients and 664 controls.

The multiplex PCR design used two pairs of primers that amplified two different regions of *DRD4* simultaneously (Supplementary Tables S7 and S8). The first pair, called *DRD4* deletion primers, amplified a region of 674 bp only when the sample carried the deletion. We achieved this by designing a reverse primer on the deletion breakpoint. When the sample did not have the deletion, the primer could not amplify a PCR product, because the primer could not hybridize. The second pair, called *DRD4* control primers, amplified a region of 429 bp in all samples (Figure 10). All PCR runs included one positive control (a sample known to carry the deletion).

Figure 10. Schematic representation of the multiplex PCR design. Blue arrows represent the *DRD4* deletion primers (forward and reverse), designed to amplify only in samples that carry de *DRD4* deletion (represented in yellow). Purple arrows represent the *DRD4* control primers (forward and reverse), designed to amplify in both wild-type and deletion-carrying samples.

2.6. Targeted resequencing design and capture

For the validation of candidate OCD genes discovered by RVAS, we designed a SureSelect QXT Custom 1Kb-499kb library through the Agilent Sure Design tool (Agilent Technologies), in a shared capture array that included 20 candidate OCD genes (Supplementary Table S9).

Libraries for 439 OCDs (322 adults and 117 children) and 1481 MCC-Spain cohort control samples were prepared and captured at the CRG Genomics Core Facility, using an in-house pre-capture protocol. Briefly, DNAs underwent DNase I enzymatic fragmentation, adapter ligation, nick translation, pooling of samples and PCR amplification, with purification processes between each step, and with a final capture with the probes of the designed library.

2.7. Common variants analysis

On the Study I dataset, we also performed an association study of coding, damaging (Cadd2 score >15) common variants discovered by WES. We used, as input, the WES data resulting from the QC in-house pipeline. Moreover, variants that were not in Hardy-Weinberg equilibrium were filtered out.

The association analysis was based on logistic regression models, where each variant was analyzed at a time in contrast to the gene-based tests for rare variant association analysis, assuming a log-additive mode of inheritance. We included the first 10 PCA as in the RVAS analysis to correct for potential batch effects. The p-values were obtained from the likelihoods ratio tests derived from a comparison with the null model (i.e. the model without genetic information). We used the function WGassociation from R package SNPassoc[194].

2.8. Gene set enrichment analysis

Gene set enrichment analyses were performed with the ConsensusPathDB software[195]. Over-represented sets were searched among pathway-based sets and Gene Ontology (GO)-based sets using KEGG, Reactome, Wikipathways, Biocarta and GO terms as reference gene-sets. For each of the predefined sets, a p-value was calculated according to the hypergeometric test based on the number of genes present in both the predefined set and our list. The size of the tested predefined sets was corrected to the number of set members that were annotated with a gene ID. The p-values were corrected for multiple testing using the false discovery rate method (FDR), and results were available as q-values.

3. Study I: Functional analyses

Functional assays for the *DRD4* 13 bp frameshift deletion (p.78_82del) included functional tests on immortalized B-lymphoblastoid cell lines from OCD patients and controls, and the development of a *drd4* knockout zebrafish model by ZeClinics (Supplementary Methods S2).

3.1. Immortalization of lymphocytes by Epstein-Barr virus

We obtained blood samples collected in EDTA tubes from 5 OCD patients that carried the *DRD4* deletion and 7 healthy individuals. We then isolated mononuclear cells (MNCs) from peripheral blood using a density gradient

medium called Lymphoprep™ (Stemcell Technologies) and centrifuging samples at 800 × g for 20 minutes at room temperature. Granulocytes and erythrocytes have a higher density than MNCs and therefore sediment through the Lymphoprep™ layer. MNCs were recovered from the interphase layer and incubated with Epstein-Barr virus (EBV) for 2 hours at room temperature (1 ml of EBV per $6x10^6$ cells). The infected MNCs were cultured for at least 12-15 days in RPMI 1640 with L-Glutamine medium (Gibco, Thermo Fisher Scientific) supplemented with 15% heat inactivated fetal bovine serum (FBS) (Gibco, Thermo Fisher Scientific), 1% penicillin/streptomycin 10,000 U/mL (Thermo Fisher Scientific) and cyclosporin A (CsA) at the final concentration of 0.2 µg/ml. CsA was added because immortalization by EBV occurs with greater frequency if the lymphocytes T cells are functionally inactivated[196]. After this period, immortalized lymphocytes B cells were maintained in RPMI 1640 with L-Glutamine medium (Gibco, Thermo Fisher Scientific) supplemented with 10% heat inactivated FBS (Gibco, Thermo Fisher Scientific) and 1% penicillin/streptomycin 10,000 U/ml (Thermo Fisher Scientific).

3.2. Material extraction from B-lymphoblastoid cell lines

Genomic DNA and protein were extracted from B-lymphoblastoid cell lines using the Wizard® Genomic DNA Purification Kit (Promega Corporation), and the RIPA Lysis and Extraction Buffer (Thermo Fisher Scientific) with Halt™ Protease Inhibitor Cocktail (Thermo Fisher Scientific), respectively.

3.3. DRD4 expression levels in B-lymphoblastoid cell lines

Before testing DRD4 expression levels, we confirmed the carrier and wild-type status of all generated cell lines by multiplex PCR (see Supplementary Figure S10). Then, we used two approaches, western-blot and flow cytometry, to detect an effect of the *DRD4* deletion in protein expression levels, comparing DRD4 expression in B-lymphoblastoid cell lines from heterozygote carriers and controls.

3.3.1. Western-blot assay

Protein extracts were quantified with the Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific), using a BSA standard. Next, protein lysates (from 10 µg to 30 µg) were prepared in 4x NuPAGE™ LDS Sample Buffer (Thermo Fisher Scientific) and 10x NuPAGE™ Sample Reducing Agent (Thermo Fisher Scientific) and they were boiled for 10 min at 95ºC. We then separated the proteins by electrophoresis on NuPAGE™ 4-12% Bis-Tris protein gels (Thermo Fisher Scientific), using the NuPAGE™ MES SDS Running Buffer (Thermo Fisher Scientific) and the XCell SureLock™ Mini-Cell Electrophoresis system at 200 V for 1h. Transfer was performed by dry blotting of proteins onto nitrocellulose membranes using the iBlot™ Transfer Stacks (Thermo Fisher Scientific) and the iBlot® Gel Transfer Device at 20V for 7 min.

For the detection of DRD4, the blots were blocked with 3-6% of BSA in TBS-T 0.1% and incubated overnight at 4ºC with an anti-DRD4 monoclonal antibody (sc-136169; 1:2000 dilution; Santa Cruz Biotechnology). After washing with TBS-T 0.1%, blots were incubated with HRP-coupled anti-mouse IgG (1:2000 dilution; Sigma-Aldrich Corporation) during 1 h at room temperature and washed again. Later, bands were visualized using the Luminata™ Classico or Forte Western HRP substrates (EMD Millipore). DRD4 protein levels were quantified using ImageJ software[197] and normalize using the anti-α-Tubulin antibody (1:20,000 dilution; Sigma-Aldrich Corporation #T6199).

3.3.2. Flow-cytometry assay

Flow-cytometry analysis was performed to detect DRD4 surface expression in B-lymphoblastoid cell lines of OCD cases and patients. For each sample, cells were counted to a final concentration of $5x10^6$ cells/mL and pre-treated with human aggregated IgG (10 µg/ml) to block Fc receptors. Next, we performed indirect immunostaining: samples were incubated with the unlabelled anti-DRD4 monoclonal antibody (sc-136169; Santa Cruz Biotechnology) followed by PE-conjugated $F(ab')_2$ polyclonal rabbit anti-mouse IgG+IgM (Jackson Immunoresearch, West Grove, PA). Cell viability was assessed by incubating for

10 min at RT with 200 ng/mL DAPI (Sigma-Aldrich Corporation #D9542). A sample of cells stained only with the secondary antibody was included as control.

Data for 10,000 events per sample was acquired on LSR II flow cytometer (BD Biosciences) and analysed using FlowJo software (TreeStar). The PE median fluorescence intensity (MFI) was calculated for each sample. PE median fluorescence intensity ratio of stained/non-stained paired samples was evaluated as an indicator of the degree of DRD4 expression.

4. Study II: Transcriptomics

4.1. RNA samples and quality control

Blood samples for transcriptomic analyses were recruited from 38 OCD patients (at two time-points) and 32 healthy individuals in PAXgene Blood RNA Tubes (PreAnalytiX GmbH), which contain a reagent composition that protects RNA molecules from degradation by RNases and minimizes induction of gene expression. We then isolated and purified total intracellular RNA using the PAXgene Blood RNA Kit (PreAnalytiX GmbH). Quality control and quantification of RNA was done with the RNA 6000 Nano kit or the RNA 6000 Pico kit on a Bioanalyzer 2100 instrument (Agilent Technologies).

4.2. Total RNA sequencing

Total RNA sequencing (RNA-Seq) was done by the CRG Genomics Core Facility following the TruSeq Stranded Total RNA with Ribo-Zero Globin (Illumina) protocol. Briefly, about 500 ng of RNA were used to sequence whole-transcriptome of each sample. Ribosomal RNA and globin mRNA, which is present in high levels in whole blood, were depleted before library preparation. Next, whole-transcriptome sequencing libraries were prepared following these steps: i) RNA fragmentation; ii) cDNA synthesis; iii) adenylation; iv) adapters

ligation; and v) amplification of DNA fragments. Finally, RNA libraries were sequenced on a HiSeq2000 machine (Illumina), multiplexing 6 libraries per lane, with 50 bp single-end reads. This is expected to yield similar amount of information as an array-based expression level analysis.

4.3. RNA bioinformatic analyses

4.3.1. Quality control, alignment and estimation of transcript levels

The sequencing facility provided fastq files that were analysed with an extensive pipeline for RNA-Seq analyses developed by Roderic Guigo's group (CRG), called Grape RNA-Seq Analysis Pipeline Environment (Grape)[198]. After quality control, reads were aligned to the GRCh37/hg19 version of human reference genome using the GENCODE v19 annotations[199] with the Spliced Transcripts Alignment to a Reference (STAR) aligner[200] (version 2.4.0). After the alignment, Grape next estimates gene and transcript expression levels, calculates exon inclusion levels and identifies novel transcripts using bigwig[201] and RSEM[202].

We then constructed a count matrix containing the number of reads per transcript for each sample with htseq-count, a tool developed with HTSeq[203] that pre-processes RNA-Seq data for DE analysis by counting the overlap of reads with transcripts.

4.3.2. Normalization of read counts

DE analysis was performed with the help of Dr. Escaramís. Prior to DE analysis we first filtered out non-expressed genes, by requiring more than 4 reads in at least ten individuals for each gene of the count matrix. This was followed by upper-quartile normalization[204], which has been shown to perform better than scaling by total lane counts (e.g. RPKM).

To improve the normalization process, we later applied the strategy proposed by Risso *et al.*[205], the Remove Unwanted Variation (RUV) method from RNA-Seq data, that adjusts for nuisance technical effects by performing factor analysis on

suitable sets of control genes. The main assumption of RUV is that one can identify and use a set of synthetic negative control genes, which are genes whose expression is known *a priori* not to be influenced by the biological covariates under study (e.g., housekeeping genes or spike-in controls). Risso and colleagues also discuss the use of "*in-silico* empirical" controls if a good set of negative controls is not readily available, as in the case of our DE study. Thus, we used as the "*in-silico* empirical controls" the least significantly DE genes based on a first-pass DE analysis performed prior to RUV normalization. We decided to keep the first factors that showed correlation with our potential batch variables.

4.3.3. Differential expression analysis

DE was evaluated in three independent analyses: i) comparing OCD patients before treatment versus control individuals; ii) comparing the same OCD patients but after receiving the treatment (during, at least, three months) versus controls; and iii) OCD paired analysis, pre-treatment vs. post-treatment.

We performed DE analysis using the Bioconductor package edgeR[130] for DE analyses of read counts arising from RNA-Seq or similar technologies. edgeR functionality uses empirical Bayes methods that permit the estimation of gene-specific biological variation even for experiments with minimal levels of biological replication. For DE we used the statistical implementation of the package based on generalized linear models using likelihood ratio tests for inferential purposes.

5. Study II: Metagenomics

5.1. Microbiome samples

Stool samples from 28 OCD cases at two time-points, 7 OCD cases at a single time-point (before or after treatment), and 33 healthy subjects were collected with the Stool Collection Tube (Stratec Molecular), which has a liquid

stabilization buffer that inactivates DNases, preserves the microorganism titre, prelyses bacteria and prevents degradation of DNA. For subsequent DNA extraction we used the PSP Spin Stool DNA Basic Kit (Stratec Molecular).

Pharyngeal swab samples from 28 OCD cases at two time-points, 8 single time-point OCD cases and 32 healthy individuals were collected with the Catch-All sample collection swabs (Epicentre) and PowerBead tubes (MO BIO Laboratories). These tubes contain a buffer that protects nucleic acids from degradation. These samples were then processed with the PowerSoil DNA Isolation kit (MO BIO Laboratories) to extract genomic DNA from a variety of organisms.

5.2. 16S-rRNA sequencing

16S-rRNA sequencing was performed by the UPF Genomics Core Facility. For each set of samples (96 stool and 96 pharyngeal swab samples), the bacterial 16S ribosomal RNA (rRNA) gene was amplified using a specific primer set for the V3-V4 regions and the obtained PCR products were purified, quantified, and pooled in an equimolar way in a final amplicon library that was sequenced on a MiSeq System. One 2 x 300 bp paired-end sequencing run was performed for each analysis (stool and pharyngeal swab samples).

5.3. Metagenomics bioinformatics analyses

Metagenomics analyses were done in collaboration with Jesse Willis from Dr. Gabaldón's Group (CRG).

5.3.1. Processing of 16S rRNA sequence reads and taxonomy assignment

The DADA2 pipeline using version 1.6.0 of the DADA2 R package[206] was employed to obtain counts of amplicon sequence variants (ASV), and then assign taxonomy to the sequences using the silva database.[207,208] The parameters used in the pipeline were the same as in the DADA2 pipeline tutorial

(version 1.8), except for those that are particular to each dataset when using the filterAndTrim function. Those were as follows. For stool samples: truncLen=c(280,225), maxN=0, maxEE=c(5,10), truncQ=1, trimLeft=c(20,15), and for the pharyngeal swab samples: truncLen=c(250,230), maxN=0, maxEE=c(10,10), truncQ=1, trimLeft=c(20,30).

5.3.2. Microbiome composition profiling

The 16S rRNA ASV counts from the 96 stool and pharyngeal swab samples, along with clinical data and diet information collected for each individual, were stored and analysed in objects in R using the Phyloseq package (version 1.22.3)[209], which also has functions for filtering taxa, normalizing values and other calculations, as well as producing plots. The 16S counts were normalized per sample, obtaining the relative abundance of each taxon within a sample, with all values between 0 and 100.

5.3.3. Diversity measures

We estimated α- and β-diversity measures within samples using the Phyloseq[209], picante (version 1.6.2)[210] and vegan (version 2.4.6)[211] R packages (see Supplementary Methods S3 for details). α-diversity refers to species richness (number of taxa) within a single sample, while β-diversity refers to dissimilarity in taxonomic abundance profiles from different samples.

We estimated α-diversity using different indices, which give slightly different information. These include the Observed diversity, Chao1 index, Abundance-based Coverage Estimator (ACE), Shannon, Simpson, Inverse Simpson, and Fisher Diversity indices using the estimate_richness function from the Phyloseq package. We also calculated Faith's phylogenetic diversity and species richness using the pd function from the picante package (version 1.6.2)[210]. Boxplots were generated using ggplot2 (version 2.2.1)[212]. Statistical significance of α-diversity differences between groups was evaluated with Mann–Whitney U test when

samples were independent, and with Wilcoxon rank-sum test when samples were paired.

We estimated β-diversity as the weighted and unweighted UniFrac distance between samples with the Unifrac function, as well as the Jensen-Shannon Divergence (JSD) with the JSD function, both from the Phyloseq package. We also calculated the Bray-Curtis dissimilarity and Canberra index using the vegdist function in the vegan package (version 2.4.6)[211]. Furthermore, the adonis function in the vegan package was used to perform a PERMANOVA test on β-diversity with 999 permutations considering even dependence of samples (paired OCD samples after and before treatment) using the "strata" argument within the adonis function. We used a Principal Coordinate Analysis (PCoA) to visualize the clustering of the samples.

5.3.4. Statistical analyses

We performed the Kruskal-Wallis rank sum test between categorical variables (e.g. sample type, type of obsessions) and taxa abundances or other continuous variables. In all cases, we applied the Bonferroni correction to adjust the p-values by the number of comparisons. Boxplots were generated using ggplot2[212] and association plots were generated using the assoc function from the R package vcd (version 1.4.4)[213].

To identify possible taxa biomarkers associated with OCD, which differ in abundance and occurrence between OCD and control samples, we performed a linear discriminant effect size analysis (LEfSe)[214] via the Galaxy web application with the Huttenhower lab's tool. LEfSe combines Kruskal-Wallis test or pairwise Wilcoxon rank-sum test with linear discriminant analysis (LDA). It ranks features by effect size, which put features that explain most of the biological difference between sample groups at top. We used an α value for the statistical test equal to 0.05 and a logarithmic LDA score threshold of 2.0.

RESULTS

Study I: Deciphering OCD by whole-exome sequencing

In this section we present the results of the first part of the project, whose aim was to decipher the genetic architecture of OCD from WES data. We generated sequencing data from 306 OCD cases and 63 controls, and performed a joint variant calling together with data from 1896 Spanish individuals, thus increasing accuracy for rare variant calling. We analysed all the samples from alignment to annotation, placing special emphasis in the quality control and filtering steps to achieve highly accurate results. The generated genotypes were used for RVAS. Significant results were validated through Sanger sequencing validation, and we did gene set enrichment analysis. Further, we performed targeted resequencing to replicate some of the identified candidate genes associated with OCD. We also studied in detail a *DRD4* 13-bp frameshift deletion enriched in OCD cases compared to controls from our RVAS dataset. We tested its association with OCD in a larger cohort of cases and controls and performed some functional analyses. Exome data was also used to explore other scenarios, such as the association of common and low-frequency variants with functional effects.

1. Rare variant association analyses

1.1. Selection of well covered variants is an essential step for high accurate downstream analyses

Because our analysis included a large number of whole-exome samples belonging to different projects and captured with different kits, which could cause stratification of the data, we assessed this potential stratification by PCA. This analysis showed that samples clustered by the capture kit used (Figure 11).

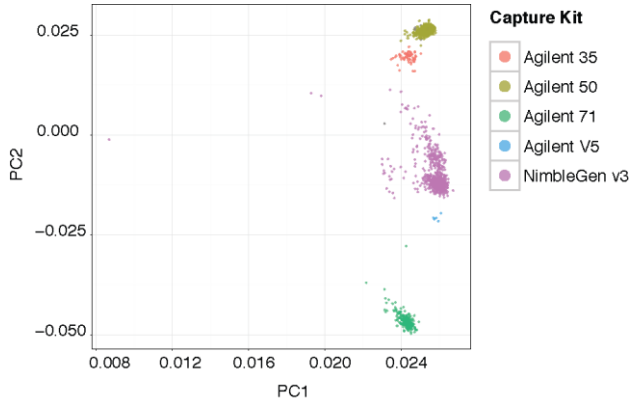


Figure 11. PCA of all 2265 whole-exome samples. We performed the PCA analysis selecting all synonymous SNVs without linkage disequilibrium of all samples.

In an attempt to reduce the stratification in the samples to be used in RVAS, we considered only the variants located within the intersection of the regions covered by the Agilent 35, Agilent 50 and NimbleGen v3 kits. However, after selecting those variants, we observed again stratification of our data. Finally, we selected variants that were effectively well covered (with at least 10 reads) by the three kits and were able to remove data stratification (Figure 12) despite losing a significant number of variants (Table 10). We also confirmed that samples did not cluster by project.

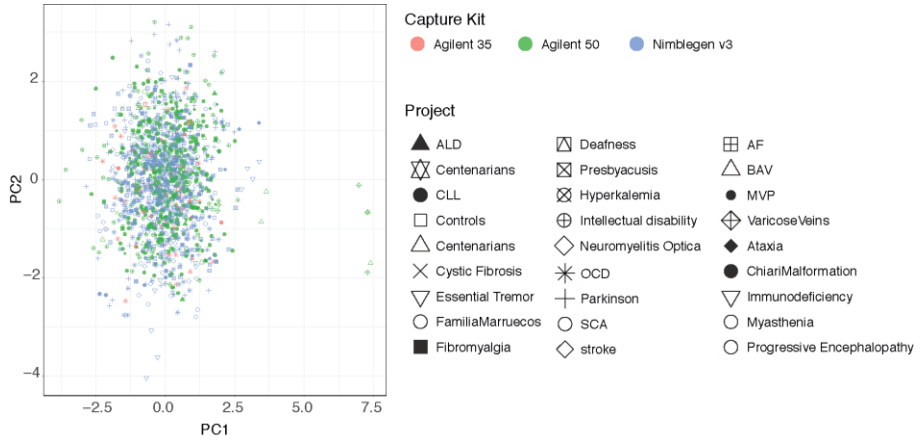


Figure 12. PCA of all 2265 whole-exome samples. We performed the PCA analysis selecting all synonymous SNVs without linkage disequilibrium of all samples.

Table 10. Number of SNVs and indels found in the resulting files of each process

Process	SNVs	Indels	Total
Pre-intersection (post-annotation)	2,163,981	147,786	2,311,767
Intersection (a)	1,360,558	81,760	1,442,318
Intersection (b)	1,122,164	62,204	1,184,368

Intersection (a) refers to the intersection of SNVs and indels with all the regions supposed to be covered by Agilent 35, Agilent 50 and NimbleGen v3. Intersection (b) refers to he intersection of SNVs and indels with all the regions effectively well covered by the three kits.

1.2. Results are highly dependent on the RVAS approximation

To find genes enriched in rare variants in OCD cases compared to controls, we performed two RVAS. The first one, with a larger sample size, included 292 OCD cases and 601 controls captured with Agilent 35, Agilent 50 and NimbleGen v3, whereas the second should include a larger number of variants by selecting only samples captured with NimbleGen v3 (253 cases and 187 controls) and thus skipping the filtering by well covered variants by all kits. In addition, we used two frequency cut-offs to determine rare variants (either MAF <0.01 or MAF <0.005) and considered separately all missense or all truncating variants.

We found little overlap in the results obtained in the different analyses. However, we observed an extremely high level of significance in the results by MiST, with many genes showing p-values of 0. We concluded that there was a likely error in the implementation of the MiST algorithm, and removed all genes with a p-value of 0. Table 11 summarizes the number of statistical significant genes found by all the analyses performed (nominal and adjusted p-value <0.05 with Benjamini-Hochberg correction). As can be seen from the data presented in this table, we consistently obtained a higher number of significant genes when we used samples captured with the three kits. On the other hand, the different methods applied found different significant genes, with little overlap between the four methods (Figure 13, Supplementary Table S10). Burden test and SKAT-O were the two methods that presented more similar results, while MiST was the method that presented higher differences with the other three tests. Finally, we compared the overlap between combining three capture kits and a larger number of samples or just one capture kit and a smaller number of cases and controls (Figure 14, Supplementary Table S11) and we saw little concordance.

Given the little overlap between tests, we decided to focus further analyses on a reduced set. Because the genetic architecture of OCD is unknown, we considered SKAT-O the best approach, as it considers a combination of scenarios, being able to detect associations both under the burden test and the variance-component method. MiST results were disregarded, as stated above. Burden test and KBAC were considered as further support for shared genes.

Quantile-quantile (Q-Q) plots of the association results (Figure 15) show the p-value distribution of the performed SKAT-O tests. From these plots, we can see that our study was underpowered to detect genome-wide significant associations. The analysis of truncating variants is clearly underpowered in contrast to the analysis of the missense ones, shown by the bigger deviation of the observed p-values towards the bottom side of the plot as well as the lower inflation factor estimates. This is mainly due to the fact of little amount of such types of variants.

Table 11. Number of statistical significant genes found by each approximation

	Agilent 35, Agilent 50 and NimbleGen v3								NimbleGen v3							
	MAF <0.01				MAF <0.005				MAF <0.01				MAF <0.005			
	Missense variants		Truncating variants		Missense variants		Truncating variants		Missense variants		Truncating variants		Missense variants		Truncating variants	
	n.pval	a.pval	n.pval	a.pval	n.pval	a.pval	n.pval	a.pval	n.pval	a.pval	n.pval	a.pval	n.pval	a.pval	n.pval	a.pval
Burden test	745	1*(A)	105	0	768	1* (A)	102	0	428	0	13	0	349	0	6	0
KBAC	563	0	63	0	529	0	61	0	242	0	7	0	248	0	5	0
SKAT-O	694	2* (A,B)	97	1* (C)	701	1* (A)	96	1* (C)	442	0	13	0	268	0	4	0
MiST	507	136	7	4	481	159	6	3	373	219	2	0	354	247	0	0

n.pval: nominal p-value <0.05; a.pval: adjusted p-value <0.05 with Benjamini-Hochberg correction. *These genes were false-positive associations (A:TIA1; B: MAGEF1; and C:ASPN). Highly significant genes identified by MiST with adjusted p-values of 0 were removed, as they are likely due to an error in the implementation of the algorithm.

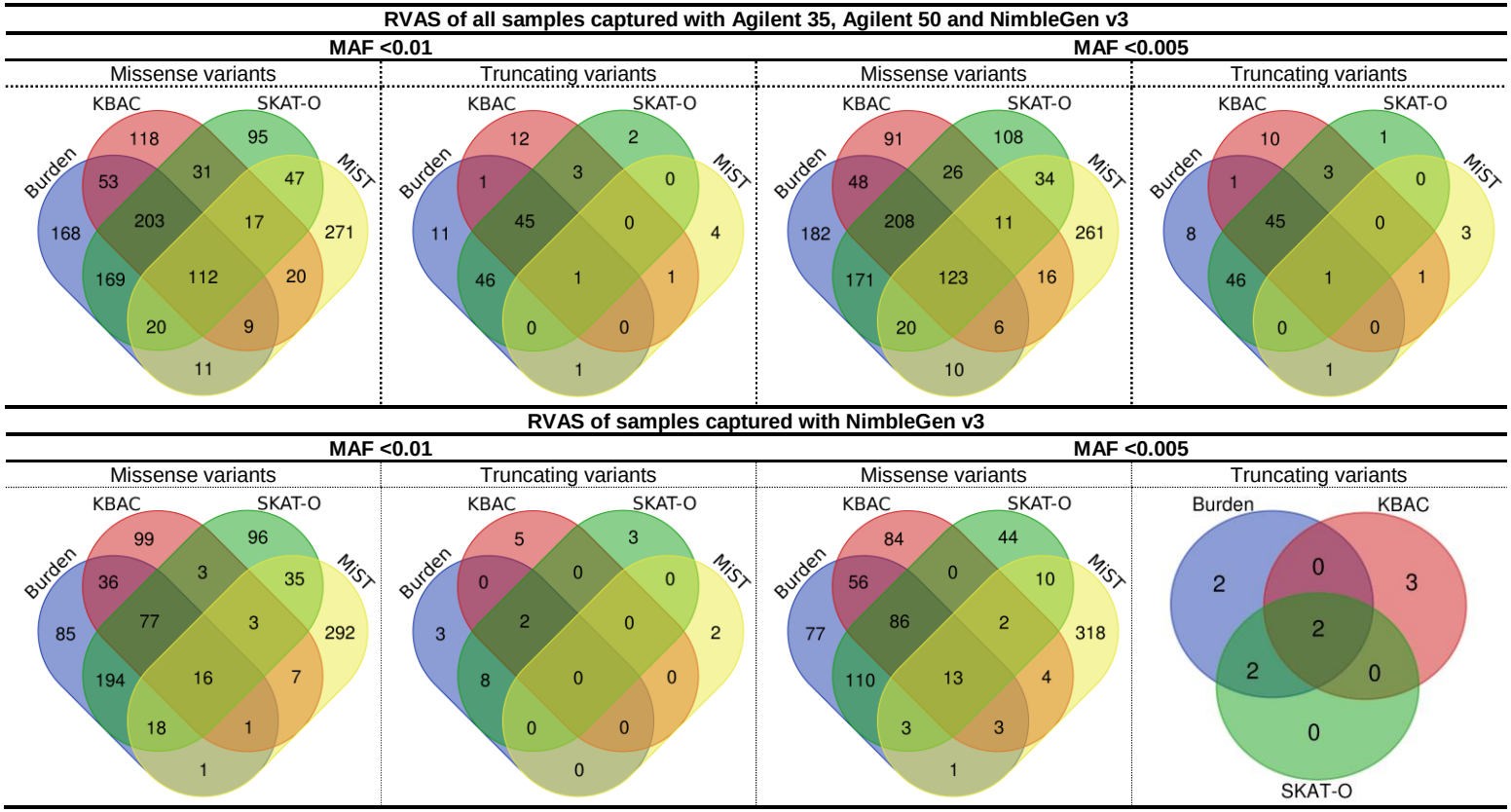


Figure 13. Representation of the concordance between the results of the distinct methods used in the different approximations tested. Highly significant genes identified by MiST with adjusted p-values of 0 were removed, as they are likely due to an error in the implementation of the algorithm.

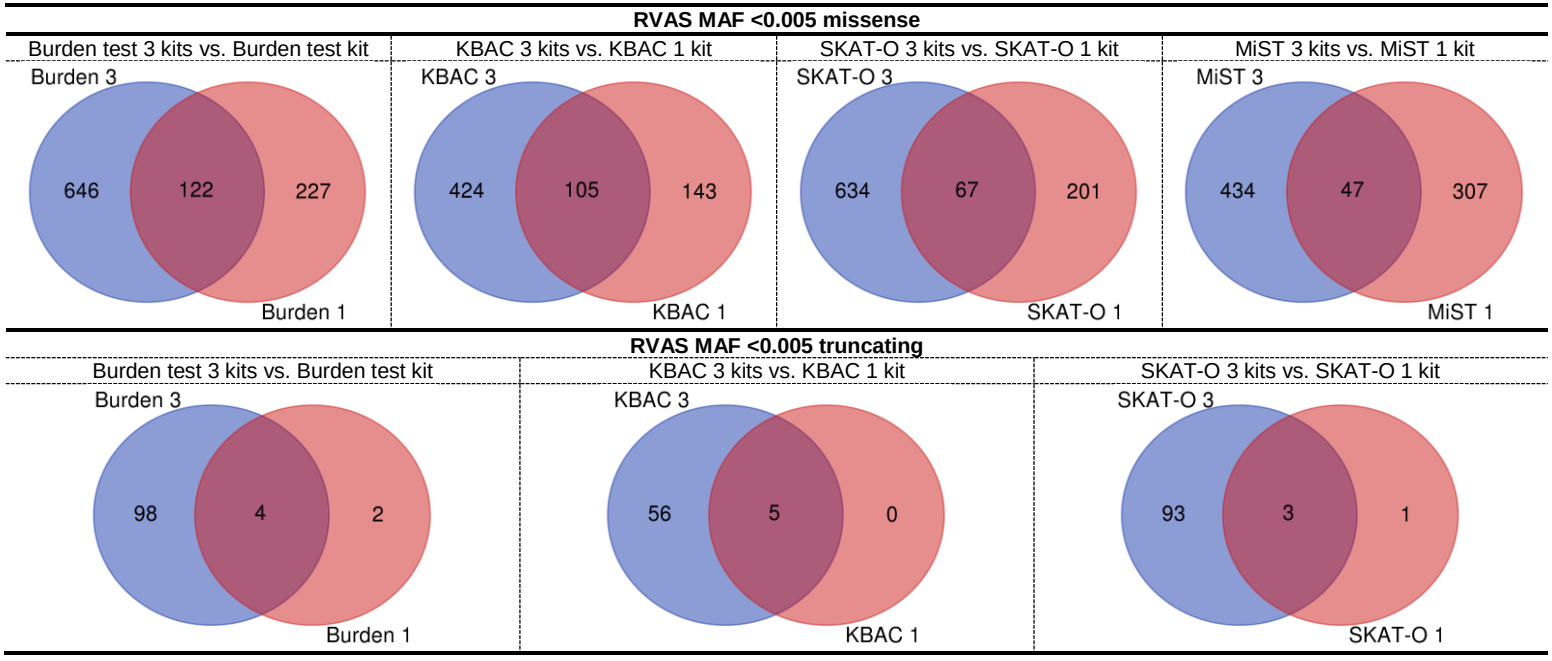


Figure 14. Representation of the concordance between RVAS results using whole-exome samples captured with Agilent 35, Agilent 50 and NimbleGen v3 and whole-exome samples captured with only NimbleGen v3. Highly significant genes identified by MiST with adjusted p-values of 0 were removed, as they are likely due to an error in the implementation of the algorithm.

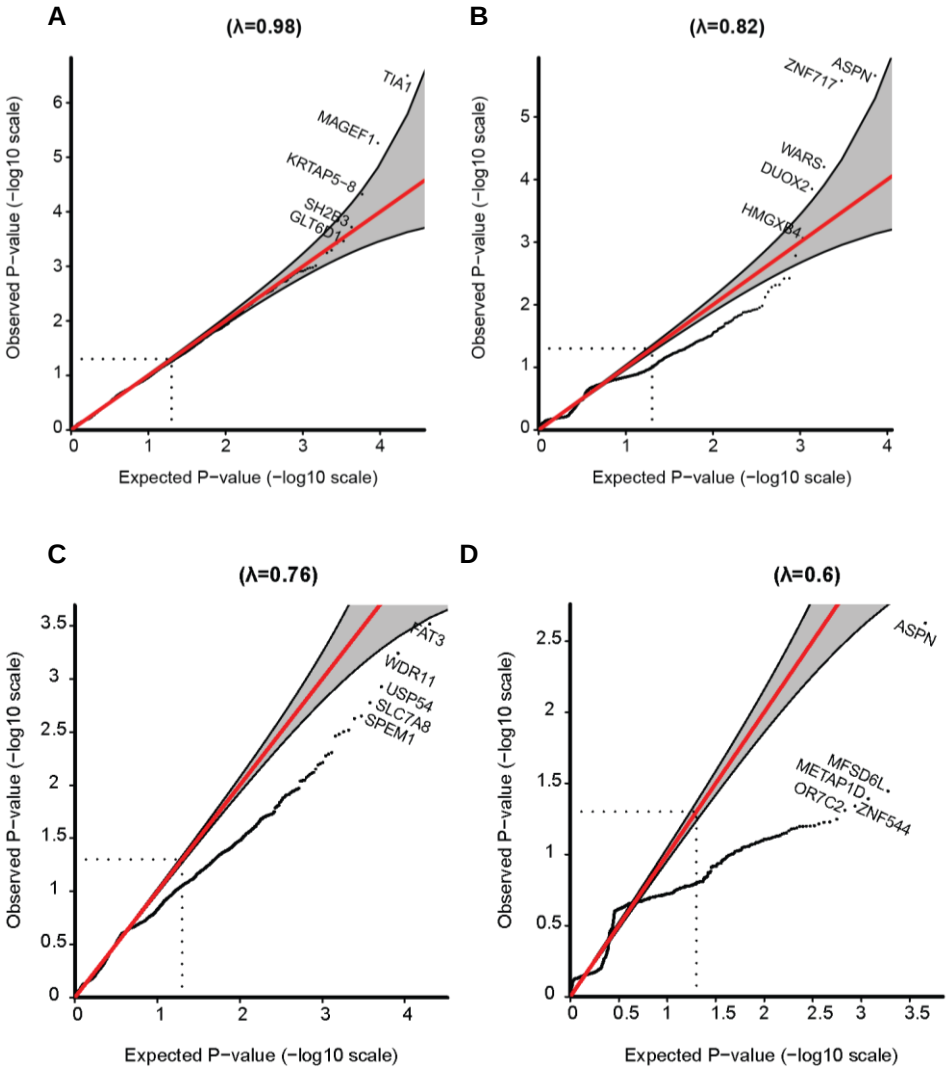


Figure 15. Quantile–quantile (Q-Q) plots of observed versus expected –log (*P*) p-values among genes of SKAT-O results for (A) missense variants with MAF <0.01 in samples captured with Agilent 35, Agilent 50 and NimbleGen v3, (B) truncating variants with MAF <0.01 01 in samples captured with Agilent 35, Agilent 50 and NimbleGen v3, (C) missense variants with MAF <0.005 in samples captured with NimbleGen v3, and (D) missense variants with MAF <0.005 in samples captured with NimbleGen v3. Grey area is delimited by 95% confidence bands of the expected p values. Genomic inflation factors (λ) are shown in each plot.

1.3. Top RVAS genes did not validate by Sanger sequencing

Some of the most significant genes were carrying, in each case, one specific variant that was present in around 10-28 OCD cases and between 0-5 control samples. This was unexpected, and we suspected errors in variant calling. To confirm these results, we tried to validate some of them by Sanger sequencing in 4-5 samples each. We sequenced variants in 11 suspicious genes (Table 12), and we included a variant in *GJA5*, which did not seem false positive. We could not detect any of the suspicious variants in any of the supposed carriers, while the *GJA5* variant was detected in all supposed carriers (Figure 16).

In many of these cases, the false-positive variants had an allele balance distribution deviating significantly from the expectation, a phenomenon we termed allele balance bias (ABB). At this point, our colleagues (Dr. Ossowski's group, CRG) developed a genotype callability score based on the ABB for all positions of the human exome, which detected false positive variant calls that passed state-of-the-art filters (Muyas *et al.,* in revision). We then re-ran the variant filtering step incorporating a maximum ABB threshold of 0.7, based on a probability model to belong to the recurrently deviated AB and on evaluation of ABB by Sanger sequencing. We also repeated all the downstream analyses.

Table 12. Results of variants validation by Sanger sequencing

Gene	Position	Variant	Samples tested	Positive samples
WARS	chr14:100835597	T>G	5	0
TIA1	chr2:70457911	T>G	5	0
RBM25	chr14:73572606	AAG>A	5	0
PPCS	chr1:42922346	T>G	4	0
GDE1	chr16:19519097	G>T	4	0
RBL1	chr20:35695248	G>A	4	0
SLC6A5	chr20:35695248	G>A	4	0
CHKA	chr11:67838254	G>T	4	0
MTOR	chr1:11188155	G>A	4	0
KCNA2	chr1:111146951	G>A	4	0
GRIN2B	chr12:13720095	A>C	4	0
GJA5	chr1:147230554	G>A	5	5

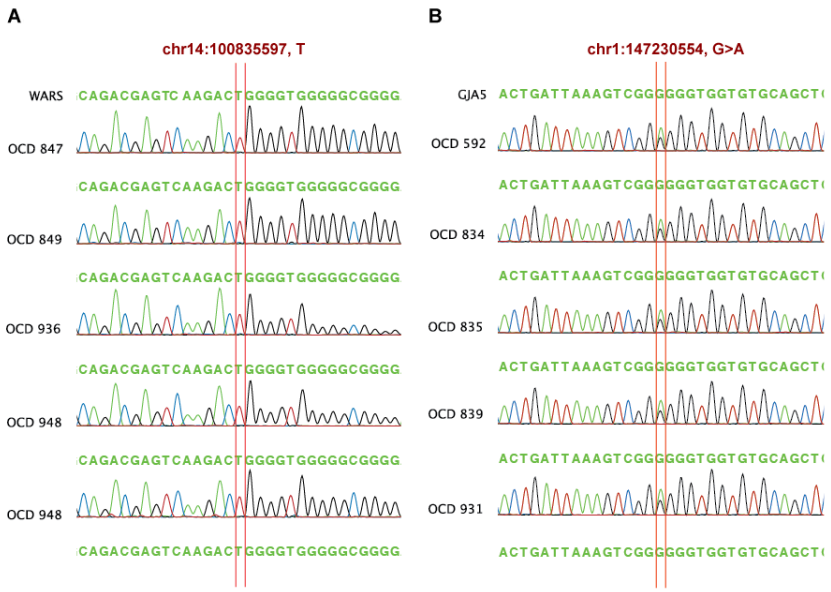


Figure 16. CLC Main Workbench capture of two examples of variant validation by Sanger Sequencing. (A) The suspicious variant detected in *WARS* was confirmed as a false positive, as it did not validate in 5 genomic DNA OCD samples that were supposed to carry it. (B) The variant detected in *GJA5* which we expected to be true, was detected in 5 genomic DNA OCD samples that were supposed to carry it.

1.4. RVAS results provided a set of novel OCD candidate genes

After analysing the ABB filtered results, and focusing on SKAT-O output, we identified a set of candidate genes potentially associated with OCD. These genes could present different scenarios: i) they were enriched in variants in OCD cases versus controls; ii) they had similar amount of variants between OCD cases and controls, but presented some specifics variants enriched in OCD cases; and iii) they were enriched in variants in controls, which means that they could have a protective effect. We considered the third scenario the most unlikely to be real, and did not follow up on those genes.

We selected the top OCD candidate genes derived from SKAT-O method (MAF <0.005), based on p-value scores, number and type of variants, function of the gene, brain expression (from GTEX[215] data), and its involvement in neurological

pathways. We also considered if the genes had been found in more than one analysis. Further, we searched for genes previously associated with OCD in the literature within the nominally significant enriched genes. These include genes linked to OCD through candidate gene studies, GWAS, functional evidence, animal models studies (such as canine-OCD genes) and latest NGS studies. We found four known OCD-related genes enriched in variants in OCD cases versus controls: *CHD8* (involved in OCD by Cappi *et al.*[112]), *ASTN2* and *USP54* (involved in OCD by Gazzellone *et al.*[81]), and *DHRS11* (involved in OCD by Stewart *et al.*[89]). Of note, *DHRS11* was nominally significant in the truncating analyses, which gave additional support to the association of this gene with OCD. The top 27 significant OCD candidate genes are listed in Table 13.

The results for these genes were mostly concordant for the assay with three kits or just with NimbleGen v3, but for some genes we observe that the analysis with three kits identifies a much larger number of control carriers, such as *FAT3*. This indicates that those genes might be false positives due to the small size of our control cohort in the NimbleGen v3 only analysis or that some samples added have some specific variants that had not too much weight in the analyses, as the p-values were still significant.

1.5. Gene set enrichment analyses highlighted neuronal development and function related pathways

We did pathway and Gene Ontology (GO) enrichment analyses with all statistically significant genes with nominal p-value <0.05 found by the SKAT-O method, performed selecting only those samples captured with NimbleGen v3 and a MAF <0.005. We selected this set of genes because this was the most homogeneous approximation performed and, hence, the one expected to have less false-positives. In total, we included 272 genes in the analyses (268 from missense variant analysis and 4 from truncating variant analysis).

Table 13. Top OCD candidate genes derived from SKAT-O method (MAF <0.005)

Gene	Kit	Total Variants	Unique variants cases	Unique variants controls	Affected cases	Affected controls	Total Cases	Total controls	p-value
Missense variants									
FAT3	1	34	28	10	44	9	253	188	0.0003
	3	47	28	29	40	44	281	597	0.0066
WDR11	1	9	9	0	11	0	253	188	0.0005
	3	11	9	5	11	7	291	601	0.0099
USP54	1	10	9	2	19	2	253	188	0.0012
	3	14	9	9	16	14	292	601	0.0226
SLC7A8	1	6	6	0	12	0	253	188	0.0017
	3	9	6	6	14	9	292	601	0.0083
MFSD6L	1	6	6	0	10	0	253	188	0.0030
	3	9	7	5	11	8	291	601	0.0093
TTLL4	1	9	8	1	13	1	253	188	0.0031
	3	16	10	8	17	9	291	601	0.0000
FAR1	1	2	2	0	12	0	253	188	0.0035
	3	1	1	0	1	0	292	601	0.2003
GLDN	1	7	7	0	9	0	252	188	0.0061
	3	7	7	5	10	12	291	601	0.1780
TRPM3	1	15	14	1	14	1	253	188	0.0064
	3	19	14	7	14	7	290	601	0.0041
CHD8	1	8	8	1	14	1	253	188	0.0080
	3	6	6	2	8	3	286	596	0.0178

Kit 1: NimbleGen v3; Kit 3: Agilent 35, Agilent 50 and NimbleGen v3. Total variants: total number of unique variants participating in analysis for this gene. Unique variants cases/controls: number of unique variants in cases/controls. Affected cases/controls: number of affected cases/controls. Total cases/controls: number of cases/controls that participated in the analysis.

Table 13 (continued). Top OCD candidate genes derived from SKAT-O method (MAF <0.005)

Gene	Kit	Total Variants	Unique variants cases	Unique variants controls	Affected cases	Affected controls	Total Cases	Total controls	p-value
Missense variants									
LRRK1	1	17	16	3	19	3	246	188	0.0114
	3	25	18	12	21	16	285	601	0.0052
PLXNA4	1	20	17	3	19	3	253	188	0.0130
	3	23	17	11	20	15	292	600	0.0020
PTPRF	1	9	8	1	9	1	252	187	0.0160
	3	15	11	5	12	5	289	598	0.0010
ASTN2	1	9	9	1	12	1	252	188	0.0185
	3	15	10	9	14	10	291	601	0.0092
RSPO2	1	3	3	0	8	0	253	188	0.0188
	3	5	3	3	8	3	292	601	0.0018
TMEM63A	1	5	5	0	9	0	212	166	0.0189
	3	6	6	2	10	3	250	574	0.0016
STXBP5L	1	10	9	1	13	1	253	188	0.0190
	3	16	9	9	14	9	292	601	0.0007
PAM	1	9	8	1	9	1	253	188	0.0217
	3	13	8	6	10	6	292	601	0.0213
CNPPD1	1	4	4	0	7	0	253	188	0.0245
	3	7	5	3	9	4	291	600	0.0123
PCDHAC1	1	6	6	0	7	0	253	188	0.0283
	3	10	6	5	7	6	292	601	0.0308

Kit 1: NimbleGen v3; Kit 3: Agilent 35, Agilent 50 and NimbleGen v3. Total variants: total number of unique variants participating in analysis for this gene. Unique variants cases/controls: number of unique variants in cases/controls. Affected cases/controls: number of affected cases/controls. Total cases/controls: number of cases/controls that participated in the analysis.

Table 13 (continued). Top OCD candidate genes derived from SKAT-O method (MAF <0.005)

Gene	Kit	Total Variants	Unique variants cases	Unique variants controls	Affected cases	Affected controls	Total Cases	Total controls	p-value
Missense variants									
SLC44A1	1	3	3	0	7	0	253	188	0.0441
	3	6	4	2	8	4	292	601	0.0013
ZNF883	1	5	5	0	5	0	253	188	0.1024
	3	5	5	1	5	1	288	600	0.0196
EPHA5	1	6	4	2	7	2	253	188	0.3081
	3	7	5	4	10	4	292	601	0.0013
Truncating variants									
DHRS11	1	2	2	0	2	0	253	188	0.0825
	3	2	2	0	2	0	292	601	0.0421
ZNF534	1	2	2	0	4	0	253	188	0.0846
	3	3	2	2	4	3	292	601	0.0516
MFSD6L	1	2	2	0	6	0	253	188	0.0358
	3	2	2	1	6	2	292	601	0.0099
METAP1D	1	1	1	0	5	0	253	188	0.0406
	3	2	1	1	5	1	292	601	0.0016

Kit 1: NimbleGen v3; Kit 3: Agilent 35, Agilent 50 and NimbleGen v3. Total variants: total number of unique variants participating in analysis for this gene. Unique variants cases/controls: number of unique variants in cases/controls. Affected cases/controls: number of affected cases/controls. Total cases/controls: number of cases/controls that participated in the analysis. Results that are not statistically significant are highlighted in blue.

We found six significantly enriched pathways (p-value < 0.01) (Table 14). These were: “TRP channels”, “Carboxyterminal post-translational modifications of tubulin”, “Other semaphorin interactions”, “Acyl chain remodelling of PS”, “Amine compound SLC transporters”, and “Acyl chain remodelling of PE”. All these pathways are related to neuronal development and function, which gives additional support to the genes highlighted in the RVAS analysis as good OCD candidate genes.

Gene Ontology enrichment identified twenty-one GO terms significantly enriched (p-value <0.01) (Supplementary Table S12). We found enrichment of some interesting GO terms, such as “neurotransmitter transporter activity” (with *SLC18A1*, *SLC6A9*, *SLC6A16, SLC36A2, SLC44A1* as gene members), “ion binding”, “transmembrane transporter activity”, and “microtubule-based process” (highlighted in grey).

Table 14. Enriched pathway-based sets

Pathway	p-value	q-value	Members	Size	Effective size
TRP channels	0.00046	0.17177	*TRPV5; TRPC3; TRPV3; TRPM3*	25	25
Carboxyterminal post-translational modifications of tubulin	0.00234	0.32172	*TTLL4; TTLL6; AGBL1; TTLL3*	38	38
Other semaphorin interactions	0.00261	0.32172	*ITGA1; PLXNA1; PLXNA4*	19	19
Acyl chain remodelling of PS	0.00458	0.42221	*MBOAT1; PLA2G4A; LPCAT4*	23	23
Amine compound SLC transporters	0.00972	0.43278	*SLC44A1; SLC6A9; SLC18A1*	30	30
Acyl chain remodelling of PE	0.00972	0.43278	*PLA2G4A; LPCAT4; MBOAT1*	30	30

p-value: p-value calculated according to the hypergeometric test based on the number of physical entities present in both the predefined set and user-specified list of physical entities; q-value: p-values corrected for multiple testing using the false discovery rate (FDR). Size of the predefined sets were also corrected to the number of set members that are annotated with an ID of the user-specified ID type.

1.6. *TMEM63A* association with OCD was confirmed in a targeted resequencing replication assay

The RVAS performed was part of a pilot approach for discovering OCD candidate genes and, although the sample size was relatively small (limited by the available budget), the results obtained were encouraging enough to develop a validation targeted resequencing in a bigger cohort of OCD patients. We performed targeted resequencing for 20 of the top OCD candidate genes identified in the RVAS (Table 15). These were sequenced in 439 OCDs (322 adults and 117 children) and 1481 MCC-Spain cohort control samples. We selected these genes based, again, on p-value scores, number of variants in cases versus controls, type of variants giving high score in the algorithms used, and function of the gene, as well as involvement in neurological function and pathways.

After performing the alignment, variant calling and filtering, annotation and QC (see Supplementary Methods S1.3), there remained 427 OCD cases, 1474 controls, and a total of 13,751 unique SNVs and indels.

Targeted resequencing data was analysed with BATI, the new method developed in our group, included in the latest version of our pipeline. Given that BATI outperforms all other tests, we decided to focus on this method on this and future analyses. We used 427 OCDs and 854 control samples to test association of the 20 candidate OCD genes with the disorder (the remaining control samples were used to estimate the local AF), and we found one statistical significant gene associated with OCD when we tested missense variants (Table 16). This gene was the Transmembrane Protein 63A (*TMEM63A*), an osmosensitive calcium-permeable cation channel. OCD samples presented almost double of variants on this gene than controls. The DIC value for *TMEM63A* was 7.12, above the empirical DIC threshold cut-off of 6, corresponding to a FDR of 0.01 calculated for this dataset.

Table 15. Description of the 20 OCD candidate genes included in the targeted resequencing study

Gene	Name description	Function	Brain expression	Linked to OCD	Linked to other neuropsychiatric / neurological disorder
ASTN2	Astrotactin 2	Neuronal migration	Yes	Yes[81]	Autism, Schizophrenia, ADHD, bipolar disease, intellectual disability, and global developmental delay[216]
CHD8	Chromodomain Helicase DNA Binding Protein 8	Transcriptional regulation, epigenetic remodeling, promotion of cell proliferation, and regulation of RNA synthesis	Yes	Yes[112]	Autism[217]
CNPPD1	Cyclin Pas1/PHO80 Domain Containing 1	Involved in cell cycle processes	Yes	No	-
DHRS11	Dehydrogenase/Reductase 11	Oxidoreductase activity and coenzyme binding	Yes	Yes[89]	Autism[218] Schizophrenia[218]
EPHA5	EPH Receptor A5	Axon guidance molecule during development; plays also a role in synaptic plasticity in adult brain through regulation of synaptogenesis	Yes	No	Bipolar disorder[219] ADHD[220] MDD[221]
FAT3	FAT Atypical Cadherin 3	May play a role in the interactions between neurites derived from specific subsets of neurons during development	Yes	No	-
LRRK1	Leucine Rich Repeat Kinase 1	Protein kinase activity	Yes	No	Parkinson[222]
PAM	Peptidylglycine Alpha-Amidating Monooxygenase	Catalyze the conversion of neuroendocrine peptides to active alpha-amidated products. Alters cooper, an essential trace element crucial for normal synaptic function.	Yes	No	-
PCDHAC1	Protocadherin Alpha Subfamily C, 1	Establishment and maintenance of specific neuronal connections in the brain	Yes	No	-
PLXNA4	Plexin A4	Plays a role in axon guidance in the developing nervous system	Yes	No	Autism[223] Alzheimer[224] Parkinson[225]

Table 15 (continued). Description of the 20 OCD candidate genes included in the targeted resequencing study

Gene	Name description	Function	Brain expression	Linked to OCD	Linked to other neuropsychiatric / neurological disorder
PTPRF	Protein Tyrosine Phosphatase, Receptor Type F	Possible cell adhesion receptor. It possesses an intrinsic protein tyrosine phosphatase activity (PTPase) and dephosphorylates EPHA2 regulating its activity	Yes	No	-
RSPO2	R-Spondin 2	Activator of the canonical Wnt signaling pathway and regulator of the canonical Wnt/beta-catenin-dependent pathway and non-canonical Wnt signalling	Yes	No	-
SLC44A1	Solute Carrier Family 44 Member 1	Choline transporter. May be involved in membrane synthesis and myelin production	Yes	No	-
STXBP5L	Syntaxin Binding Protein 5 Like	Inhibitor of synaptic transmission. May inhibit exocytosis in neurosecretory cells	Yes	No	Infantile-onset neurodegenerative disorder[226]
TMEM63A	Transmembrane Protein 63A	Acts as an osmosensitive calcium-permeable cation channel	Yes	No	--
TTLL4	Tubulin Tyrosine Ligase Like 4	Glutamylase which preferentially modifies beta-tubulin and non-tubulin protein	Yes	No	-
USP54	Ubiquitin Specific Peptidase 54	Thiol-dependent ubiquitinyl hydrolase activity	Yes	Yes[81]	-
WDR11	WD Repeat Domain 11	Involved in a variety of cellular processes, including cell cycle progression, signal transduction, apoptosis, and gene regulation	Yes	No	-
ZNF534	Zinc Finger Protein 534	May be involved in transcriptional regulation	Yes	No	-
ZNF883	Zinc Finger Protein 883	May be involved in transcriptional regulation	Yes	No	-

.

Table 16. Results from RVAS of the 20 OCD candidate genes, using the BATI algorithm and missense variants with a MAF <0.005

Genes	Total variants	Unique cases	Unique controls	Affected cases	Affected Controls	Total cases	Total Controls	% variants in cases	% variants in controls	DIC
TMEM63A	14	11	6	17	18	427	854	3.98	2.11	7.12
USP54	24	13	17	21	27	427	854	4.92	3.16	2.85
FAT3	66	36	42	41	73	427	854	9.60	8.55	1.69
PTPRF	28	13	18	16	21	427	854	3.75	2.46	0.81
RSPO2	3	2	2	5	6	427	854	1.17	0.70	-0.24
ZNF534	16	10	12	11	34	427	854	2.58	3.98	-0.27
STXBP5L	18	9	14	12	16	427	854	2.81	1.87	-0.52
TTLL4	29	11	23	22	40	427	854	5.15	4.68	-0.75
PCDHAC1	16	10	11	12	18	427	854	2.81	2.11	-0.81
CHD8	21	12	14	12	19	427	854	2.81	2.22	-1.36
EPHA5	13	4	10	7	9	427	854	1.64	1.05	-1.43
SLC44A1	9	5	6	5	7	427	854	1.17	0.82	-1.44
CNPPD1	10	4	9	8	13	427	854	1.87	1.52	-1.52
PAM	16	9	10	10	15	427	854	2.34	1.76	-1.55
LRRK1	33	13	25	19	43	427	854	4.45	5.04	-1.61
ZNF883	2	1	1	1	3	427	854	0.23	0.35	-1.69
PLXNA4	32	15	20	17	37	427	854	3.98	4.33	-1.70
WDR11	25	12	15	11	20	427	854	2.58	2.34	-1.89
ASTN2	20	10	13	10	18	427	854	2.34	2.11	-1.91
DHRS11	1	1	0	2	0	427	854	0.47	0.00	NA

Total variants: total number of unique variants participating in analysis for this gene. Unique cases/controls: number of unique variants in cases/controls. Affected cases/controls: number of affected cases/controls. Total cases/controls: number of cases/controls that participated in the analysis. DIC: deviance information criterion. NA: not available.

2. A rare deletion in *DRD4* might be associated with OCD

2.1. Identification of a deletion in *DRD4* and validation in a larger cohort of OCD patients and controls

By doing variant curation, we found a heterozygous 13-bp frameshift deletion in *DRD4* (p.78_82del) carried by seven OCD cases (7 alleles in 306*2 chromosomes, giving an MAF=0.0114) and absent in the controls of our RVAS sample dataset. When we explored all the whole-exome samples of our database (1959 samples, excluding OCDs) we only observed this deletion in sixteen samples (MAF of 0.0041). Only three of these samples were Spanish, while twelve belonged to the TwinsUK project and one was Italian. We validated by Sanger sequencing the *DRD4* deletion found in the seven OCD patients and we confirmed (sequencing a few selected samples) that controls were true negatives (Figure 17).

With this scenario, we decided to study the frequency of this deletion in a larger cohort of OCD cases and controls with a specifically designed multiplex PCR (Figure 18). Combining all data, we obtained a total MAF of 0.011 (13 carriers in 614 cases) in OCD cases versus a total MAF of 0.0016 (8 carriers in 2558 controls) in controls. This difference in allele frequencies was statistically significant (allelic association with Odds ratio (OR)=6.8; p-value <0.0001) (Table 17).

We compared the obtained frequencies to the reported frequencies for the *DRD4* deletion in different databases. The MAF in the CIBERER Spanish Variant Server (0.0012) was similar to that of our control dataset. However, it was higher (between 0.009 and 0.02) in the rest of databases (Table 18) and, in some databases, similar to our OCD cases.

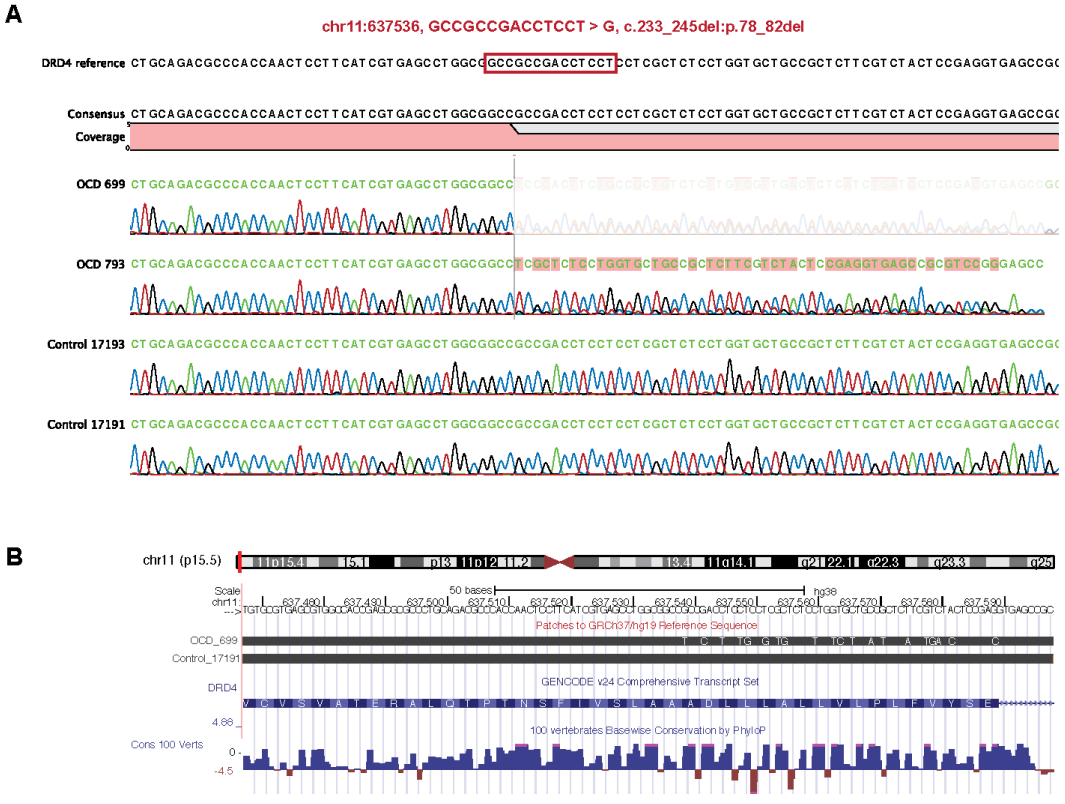


Figure 17. (A) CLC Main Workbench capture of the Sanger sequences of two OCD samples (OCD 699 and OCD 793) that carried the deletion (highlighted in a red box) and two controls (Control 17193 and Control 17191), aligned to the *DRD4* reference sequence (B) UCSC Genome Browser capture of the blat alignment of the Sanger sequences of one OCD (OCD 699) and one control (Control 17191) samples to the GRCh37 Reference sequence. The PhyloP plot showed sites predicted to be conserved (shown in blue with positive scores) or fast-evolving (shown in red with negative scores).

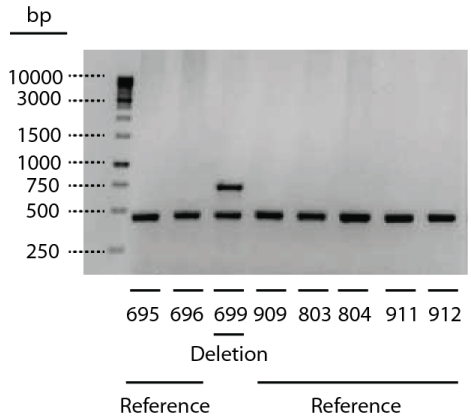


Figure 18. Gel electrophoresis of the multiplex PCR designed to detect the *DRD4* 13-bp frameshift deletion. Lanes correspond to 7 OCD samples homozygous for the reference allele (695, 696, 909, 803, 804, 911 and 912) and one (699) heterozygous carrier of the deletion. Primers amplified two different regions of *DRD4* simultaneously. A 674 bp fragment was amplified only when the sample had the deletion, and a 429 bp fragment was amplified in all samples as positive control.

Table 17. Allele frequencies of the *DRD4* 13-bp frameshift deletion in the different OCD and control cohorts tested

Cohort		Number of *DRD4* deletions	Total number of samples	MAF	Total MAF	OR
OCD	Initial OCD RVAS cohort	7	306	0.0114	0.011	6.8 (p-value <0.0001)
	Additional OCDs	6	308	0.0097		
Control	Initial control RVAS cohort	0	630	0	0.0016	
	Control in-house database cohort*	3	1264	0.0012		
	Additional VHIR control cohort	3	268	0.0056		
	Additional BREATHE control cohort	2	396	0.0025		

MAF: minor allele frequency. OR: Odds ratio. *Genotypes from WES. We selected only Spanish unrelated samples from the analysis.

Table 18. Allele frequencies of the *DRD4* 13-bp frameshift deletion in different databases

Database	MAF
CIBERER Spanish Variant Server	0.0012
1000 Genomes (Iberian populations in Spain)	0.009
ExAC (European Non-Finnish)	0.02016
EVS (EuropeanAmericans)	0.012
gnomAD(European Non-Finnish)	0.01321

2.2. Western blot and flow-cytometry did not show differences in DRD4 expression between OCD cases and controls

As the *DRD4* deletion was significantly associated with OCD, we performed functional studies to decipher if this variant could be affecting DRD4 expression, and thus, downstream pathways. We established immortalized human B-lymphoblastoid cell lines from blood samples of five OCD patients that carried the *DRD4* deletion and seven healthy individuals without this variant.

We studied DRD4 protein expression in B-lymphoblastoid cell lines by western blot in order to see if OCD patients carrying the *DRD4* deletion had lower levels of this protein than controls. We performed three independent experiments and we did not see differences between OCD cases and controls (Figure 19).

As we observed a high degree of variability between the biological replicates in the western blot assay, we also studied DRD4 protein expression in B-lymphoblastoid cell lines of OCD cases and controls by flow-cytometry (which has higher accuracy). We performed indirect immunostaining of DRD4 in all OCD cases and controls B-lymphoblastoid cell lines. After selecting only alive cells, we measured the median fluorescence intensity (MFI) for each sample and, after normalising with paired non anti-DRD4 stained samples, we obtained the MFI ratios, which were indicative of the degree of DRD4 expression. However, we did not see differences in MFI ratios between OCD cases and

controls, which indicates an absence of differences in DRD4 expression (Figure 20).

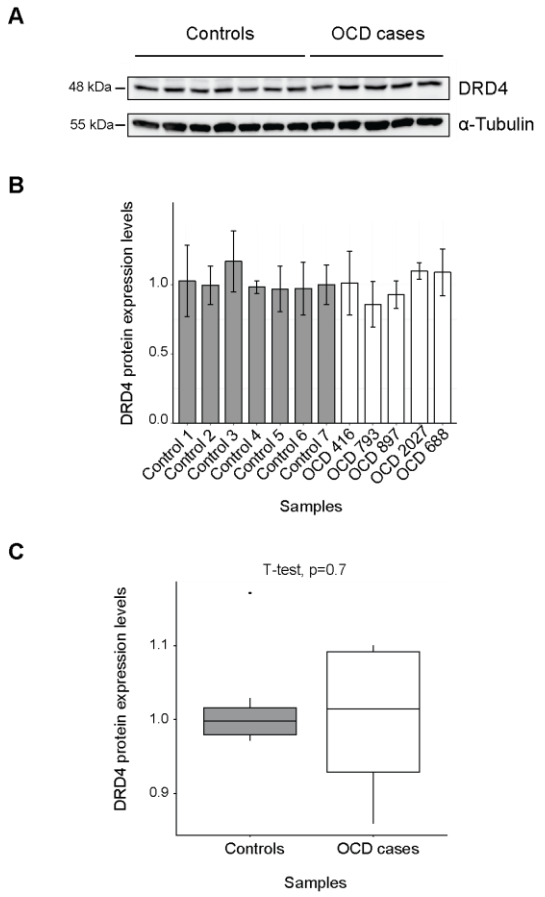


Figure 19. DRD4 expression in B-lymphoblastoid cell lines of OCD cases and controls. (A) Western blot analysis of DRD4 expression in OCD cases and controls. (B) Expression of DRD4 determined by western blot analysis in OCD cases and control samples (mean ± SD; n = 3 independent experiments). (C) Boxplots representing the distribution of the expression values of DRD4 in OCD cases and controls. p-value obtained from T-test is indicated.

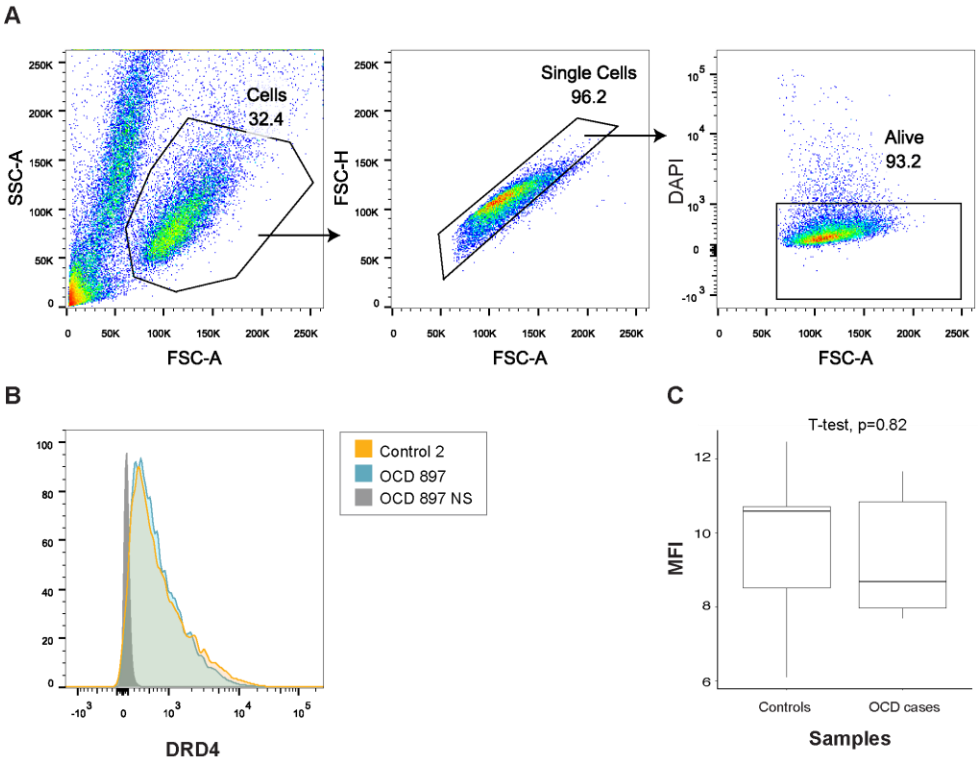


Figure 20. (A) Gating strategy for selection of B-lymphoblastoid cells. Single cells were selected using FSC and SSC and viable cells were further identified using DAPI. (B) Surface DRD4 labelling of B-lymphoblastoid cell lines by indirect immunofluorescence. Results of 3 representative experiments of 24 performed are shown (Orange histogram: Control 2 DRD4 labelled B-lymphoblastoid cell lines; Blue histogram: OCD 897 DRD4 labelled B-lymphoblastoid cell lines; Grey histogram: OCD 897 non stained). (C) Boxplots representing the distribution of the median fluorescence intensity (MFI) values in OCD cases and controls. Center lines show the medians and box limits indicate the 25th and 75th percentiles. p-value obtained from T-test is indicated.

2.3. The zebrafish *drd4* double mutant model did not show any behavioural phenotype

In parallel to the functional studies in B-lymphoblastoid cell lines, ZeClinics performed a *drd4* genetically modified zebrafish model to assess the potential role of *drd4* zebrafish orthologues (*drd4a* and *drd4rs*) in neural function and their potential role in OCD pathogenesis.

To evaluate this, a target validation approach was designed based on stable knockout *via* the CRISPR/Cas9 system. *DRD4* has three zebrafish orthologues. Of these, two have a clear CNS expression pattern (*drd4a* and *drd4rs*, Figure 21A). Therefore, we decided to focus the efforts on those two only. Interestingly, and as observed by CLUSTALW sequence alignment, the protein sequence conservation is extremely high between human *DRD4* and its zebrafish orthologues (Figure 21B). sgRNAs were selected targeting each gene on the first coding exon in order to generate knockout alleles for both genes (Figure 21C). Then, the single heterozygous knockouts were crossed to generate double heterozygous animals. Finally, single and double homozygous larvae were obtained through the cross of the double het animals for performing the proposed phenotypic analysis.

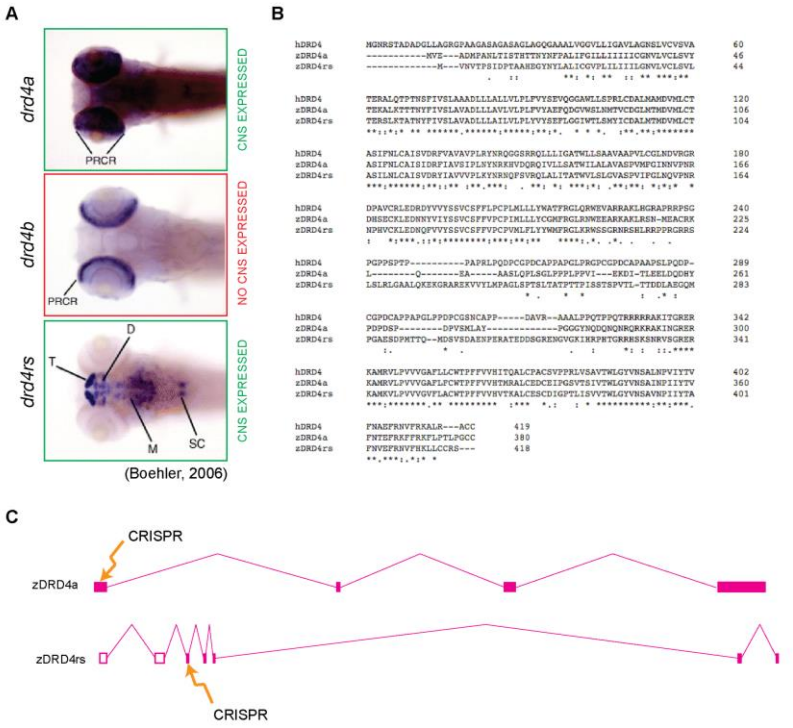


Figure 21. (A) From top to bottom, expression pattern in zebrafish larva CNS of *drd4a*, *drd4b* and *drd4rs*. (B) Protein sequence alignment among human *DRD4* and zebrafish *drd4a* and *drd4rs*. (C) Schematic representation of CRISPR targeting location for *drd4a* and *drd4rs*.

Thus, ZeClinics tested single and double homozygous larvae through a complete behavioural protocol, to address if mutant animals displayed any general neural dysfunction (total locomotion and locomotion per minute), anxiety (thigmotaxis) or learning/memory impairment. To this end they had separate groups divided as indicated in Table 19.

Table 19. Genotypes of the groups assessed

Genotype	n	Subgroups	n
ra+/+ rs+/+	83	ra+/+ rs+/+	1
		ra+/+ rs+/-	3
		ra+/- rs+/+	25
		ra+/- rs+/-	54
ra+/+ rs-/-	19	ra+/- rs-/-	18
		ra+/+ rs-/-	1
ra-/- rs+/+	54	ra-/- r s+/-	36
		ra-/- rs+/+	18
ra-/- rs-/-	18		

However, despite they generated early-truncated proteins for both tested genes, the results show that none of the mutant groups displayed any behavioural phenotype under the tested conditions, when compared to the wild-type group (Figure 22). This suggests that the CNS is perfectly functional for both single and double mutant animals. Additionally, they did not detect any developmental phenotype or important effect in other systems such as cardiac or metabolic.

The present data suggests that the experimental approaches used in this project are not appropriate to detect the expression of truncated form of DRD4.

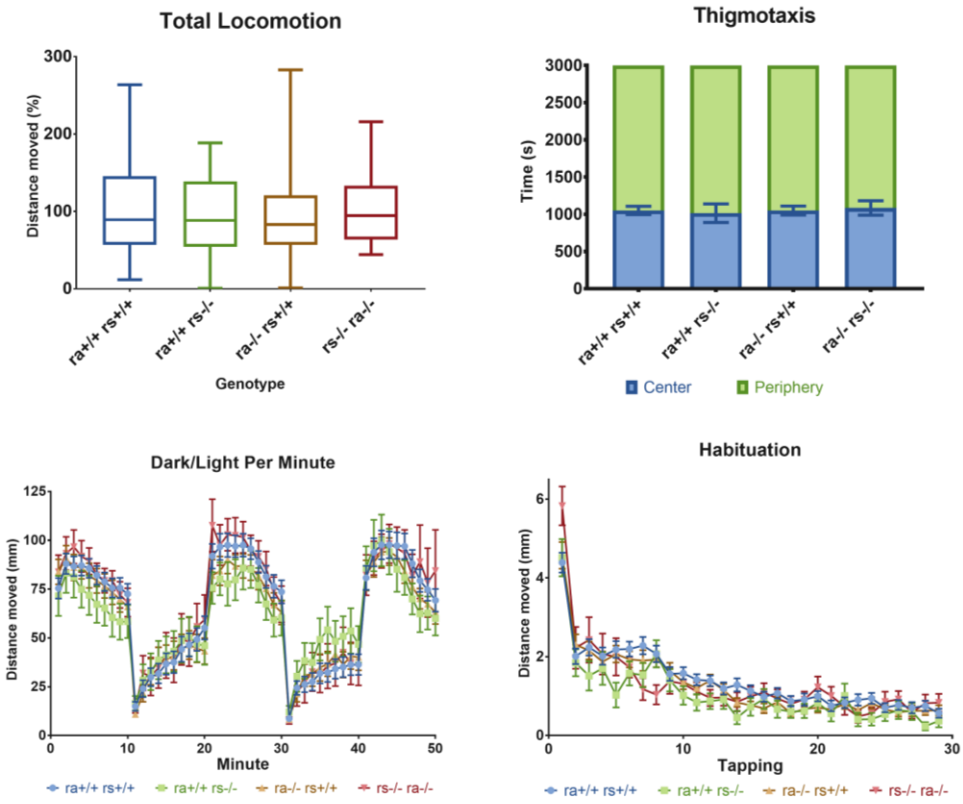


Figure 22. Phenotypic analyses of *Drd4* zebrafish orthologous genes. From left to right and up to bottom: total locomotion, thigmotaxis, locomotion per minute and habituation.

3. Common and low-frequency variants in genes involved in neuronal development and function showed association with OCD

To explore, with the available data, all the possible scenarios that could explain the genetic risk factors contributing to OCD, we also performed a common and low-frequency variant association study. As with the rare variant analysis we considered two sample sets: we tested association of potentially damaging, common exonic variants in (1) 292 OCD cases versus 601 controls (considering

samples captured with Agilent 35, Agilent 50 and NimbleGen v3); and (2) 253 cases and 187 controls (considering only NimbleGen v3 exomes). These studies detected 34 and 13 variants with statistical significance (Benjamini- Hochberg adjusted p-value <0.01) in analysis (1) and (2), respectively. Manhattan plots show the distribution of p-values along the chromosomes (Figure 23). In the analysis (1) 13 variants reached genome-wide significance, whereas in analysis (2) only a few were above the suggestive significance threshold. Associated variants included both risk (present in more cases than controls) and protective (present in more controls than cases) variants.

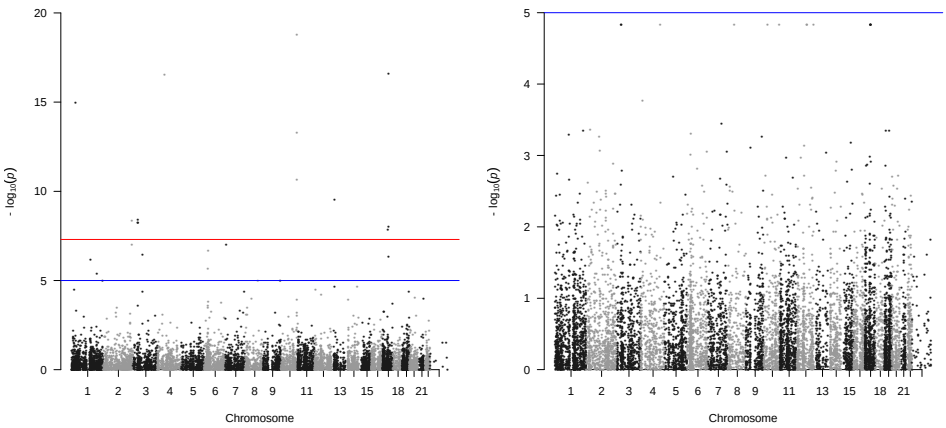


Figure 23. Manhattan plots for analysis (1) (left) and analysis (2) (right) of common variant analysis. A thin red line indicates level for genome-wide significant association ($-\log10(1 \times 10^{-8})$), whereas a thin blue line indicates level of suggestive evidence for association ($-\log10(1 \times 10^{-5})$).

As we were using controls from different projects we could have false-positive protective variants (variants associated to other disorders). Thus, we considered only those protective variants that were present in more than 90% of the controls Applying this criterion we found 1 and 17 variant in analysis (2) and (1), respectively (Tables 20 and 21). From the 17 variants, five variants reached the genome-wide significance, all of them within genes expressed in the brain.

As several of the identified variants were indels, we decided to check them in the ExAC database[185] as an additional quality control step (Supplementary Table S13). Nearly all of them were reported in ExAC and have a reference SNP (rs) identifier, although some of them did not pass their filters.

We also compared our results with the available data from the IOCDF-GC GWAS[89], the OCGAS GWAS[90], the meta-analysis from the two consortia[91] and the exon-focused GWAS done by Costas *et al*.[227]. However, none of our variants were reported by these studies nor were they located near the top hits of these analyses (within 2Mb).

Table 20. Common variants of the analysis (2) associated with OCD with statistical significance (adjusted p-value <0.001)

Variant	Variant type	Gene	p-value	adjusted p-value	% cases	% controls	Brain expr.
chr10; 126691552;t;g	Non-synonymous SNV	*CTBP2*	1,46e-05	1,46e-05	23,72	10,11	Yes

p-value: nominal p-value; adjusted p-value: adjusted p-value with Benjamini-Hochberg correction; % cases/controls: percent of cases/controls carrying this variant; Brain expr.: Brain expression from GTEX[215] or Human protein atlas data. This variant is not present in the ExAC database and is highlighted in blue.

Table 21. Common variants of the analysis (1) associated with OCD with statistical significance (adjusted p-value <0.001)

Variant	Variant type	Gene	p-value	adjusted p-value	% cases	% controls	Brain expr.	Neuropsychiatric/ neurological related disorders
chr4;54319247;cag;c	frameshift deletion	FIP1L1	2.61e-17	1.20e-16	18.84	5.32	Yes	
chr1;31905889;a;acag	inframe insertion	SERINC2	1.08e-15	4.12e-15	96.58	97.00	Yes	Alcohol dependence[228], ASD[229]
chr10;126691552;t;g	nonsynonymous SNV	CTBP2	2.24e-11	6.44e-11	21.58	6.49	Yes	
chr13;25670803;a;g	nonsynonymous SNV	PABPC3	2.70e-10	6.91e-10	12.33	3.00	Yes	
chr3;40503526;g;gctgctgctgctgcta	inframe insertion	RPL14	3.71e-09	8.54e-09	19.86	6.82	Yes	
chr17;39595484;g;a	stopgain SNV	KRT38	1.27e-08	1.95e-08	20.55	7.15	No	
chr2;233273011;c;g	nonsynonymous SNV	ALPPL2	8.96e-08	1.21e-07	16.78	5.32	No	
chr6;32489786;t;g	nonsynonymous SNV	HLA-DRB5	2.10e-07	2.69e-07	29.11	12.98	Yes	
chr3;75790880;a;g	nonsynonymous SNV	ZNF717	3.49e-07	4.22e-07	11.99	3.16	Yes	SCZ[230]
chr1;152280347;c;t	nonsynonymous SNV	FLG	6.64e-07	7.27e-07	14.04	4.66	No	
chr1;202407189;g;gt	frameshift insertion	PPP1R12B	4.07e-06	4.07e-06	8.56	3.16	Yes	
chr1;248113040;a;g	nonsynonymous SNV	OR2L8	1.00e-05	0.0017	100.00	100.00	No	
chr9;136037742;g;a	nonsynonymous SNV	GBGT1	1.00e-05	0.0017	22.95	8.49	Yes	
chr13;25670797;c;g	nonsynonymous SNV	PABPC3	2.00e-05	0.0031	6.85	2.16	Yes	
chr14;92537360;g;gctgctgctgctgctgctgctgctc	inframe insertion	ATXN3	2.00e-05	0.0031	14.38	6.16	Yes	SCA-3[231]
chr3;75786355;t;a	nonsynonymous SNV	ZNF717	4.00e-05	0.0057	7.53	2.33	Yes	SCZ[230]
chr7;149983566;g;a	stopgain SNV	ACTR3C	4.00e-05	0.0057	10.62	3.33	Yes	

p-value: nominal p-value; adjusted p-value: adjusted p-value with Benjamini-Hochberg correction; % cases/controls: percentage of cases/controls carrying this variant; Brain expr.: Brain expression from GTEX[215] or Human protein atlas data; Neuropsychiatric/neurological related disorders of the gene carrying the variant; ASD: autism; SCZ: schizophrenia; SCA-3: spinocerebellar ataxia type 3. Variants that are not present in the ExAC database are highlighted in blue. Variants that did not pass the ExAC filters are highlighted in grey.

Study II: Multiomics longitudinal study of OCD

In this section we present the results of the second part of the project, whose aim was to decipher the genetic architecture of OCD combining multiple omics analyses under a longitudinal study design. We analysed a total of 43 OCD patients and 34 healthy individuals, from whom we obtained blood, stool and/or pharyngeal swab samples. Samples from OCD patients were recruited at two time-points: before treatment (OCD T0 samples) and after at least 3 months of treatment (OCD T3 samples). The comparison of control versus OCD diagnosed and untreated patients should elucidate whether peripheral blood reflects an OCD specific transcriptomic signature. The study of the microbiome should explain dysbiotic signatures in OCD. The longitudinal design may address the treatment effect. Tables 7 and 8 shows all the analyses performed for each sample included in Study II.

1. Transcriptomics

We collected blood samples for transcriptomic analyses from 38 OCD patients, (at two time-points) and 32 healthy individuals. We extracted RNA and we did RNA-sequencing. After that, we processed the data and normalized the read counts to be able to perform DE analysis. We performed three independent comparisons for DE: (1) OCD T0 versus controls; (2) OCD T3 versus controls; and (3) OCD paired analysis (OCD T0 versus OCD T3).

1.1. RNA analysis is sensitive to batch effects

We processed the RNA-Seq output data to perform subsequent analyses with high levels of accuracy. First, we performed quality control, alignment and estimation of transcript levels. Next, we constructed a count matrix that contained the number of reads by transcript for each sample. We also filtered out non-expressed genes, by requiring more than 4 reads in at least ten individuals. Boxplots of relative log expression (RLE = log-ratio of read count to median read count across sample) and plots of principal components (PC) of raw data, showed a need for between-sample normalization (Figure 24). After upper-quartile normalization, results improved considerably but we still saw some remaining variability (Figure 25), probably due to some technical batch effects (e.g. RNA quality, library preparation day, pooling). We then applied the Remove Unwanted Variation (RUV) method for RNA-Seq data[205].

We used p-values from bivariate correlation analysis of 20 factors from RUV analysis versus different sets of potential batch effects derived from our study to decide which factors to include in the model (see Figure 26). There was a strong correlation of the variables blood extraction, RNA extraction, library preparation and personnel, sequencing lane, and pooling with the first few factors (OCD T0 versus controls, factors W_1 to W_9; OCD T3 versus controls factors W_1 to W_4; and OCD T0 versus OCD T3, factors W_1 to W_4). Including the specified number of unwanted factors as covariate regressors in our DE analyses was successful in reducing inflation of p-values, as can be seen in Figure 27.

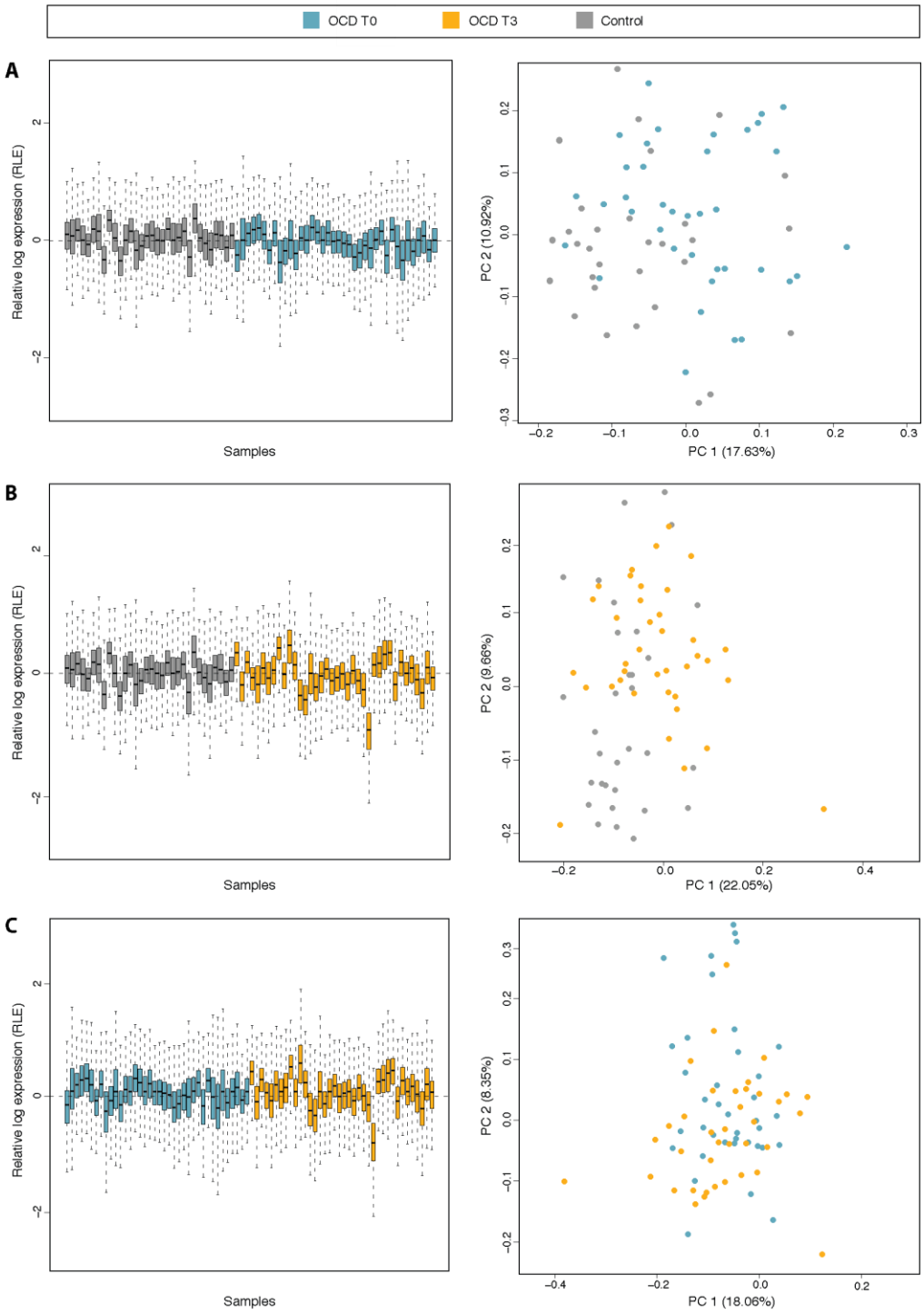


Figure 24. Pre-normalization quality control plots. Relative log ration expression (RLE) per sample (left images) and plot of principal components (PC) of raw data (right images) of (A) OCD T0 (blue colour) vs. controls (grey colour); (B) OCD T3 (orange colour) vs. controls; and (C) OCD T3 vs. OCD T0.

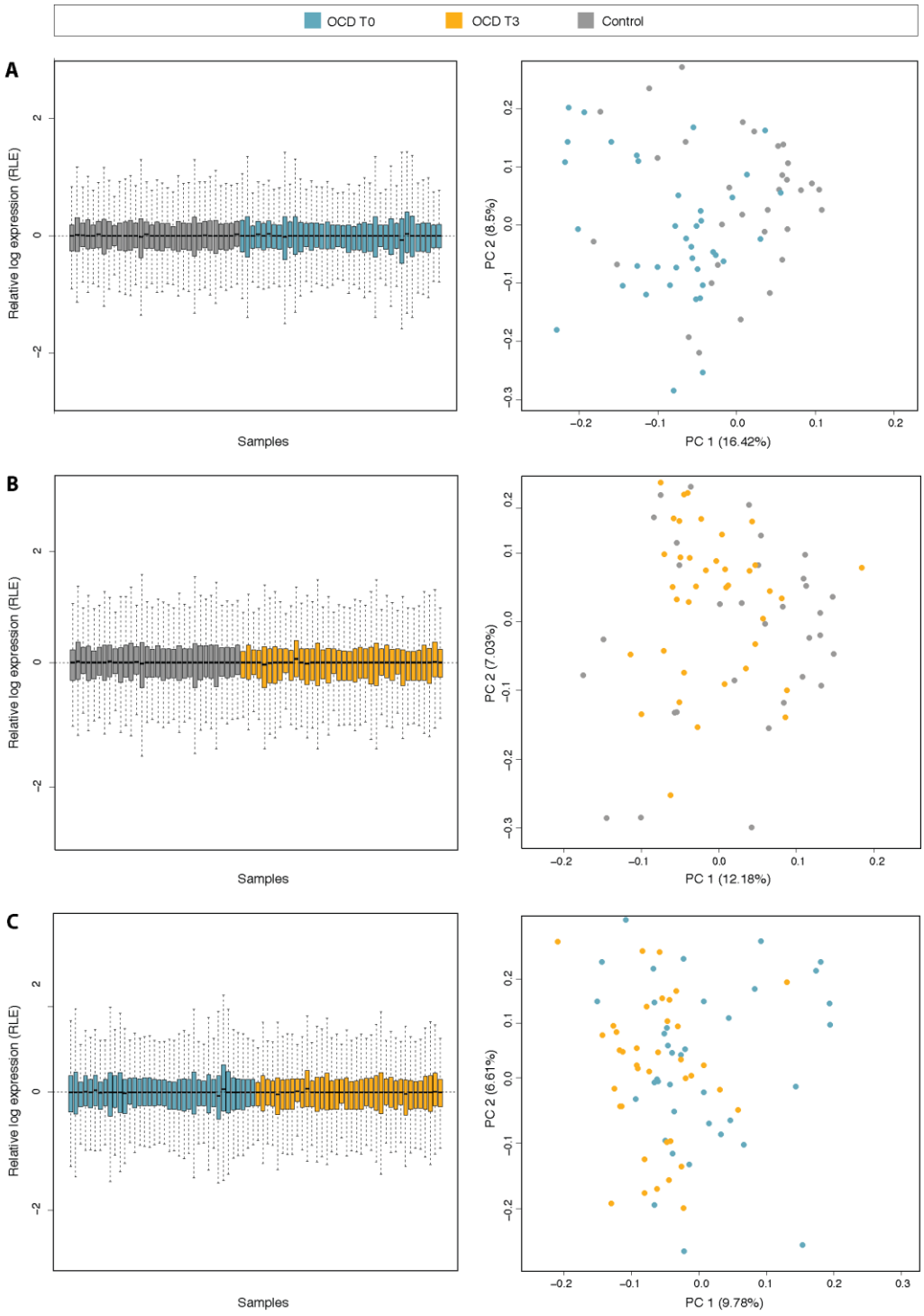


Figure 25. Upper quartile normalization quality control plots. Relative log ration expression (RLE) per sample (left images) and plot of principal components (PC) of processed data (right images) of (A) OCD T0 (blue colour) vs. controls (grey colour); (B) OCD T3 (orange colour) vs. controls; and (C) OCD T3 vs. OCD T0.

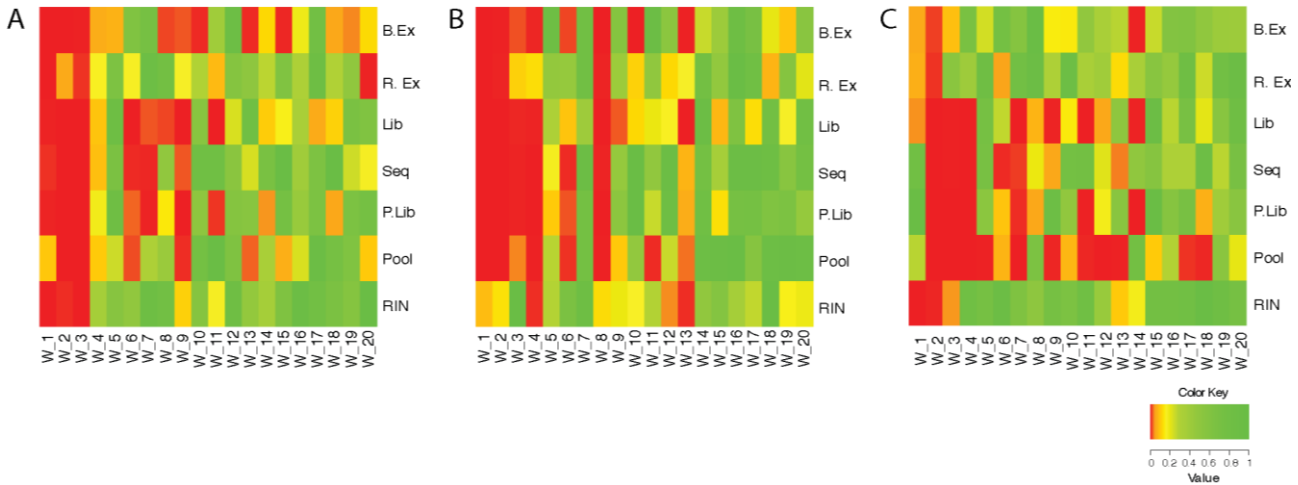


Figure 26. Heatmap of p-values from bivariate correlation analysis of 20 factors from RUV analysis versus different sets of potential batch effects derived from (A) OCD T0 vs. controls; (B) OCD T3 vs. controls; and (C) OCD T0 vs. OCD T3. p-values were calculated with ANOVA for categorical variables and with Spearman's rank correlation coefficient for quantitative variables. B.Ex: blood extraction day; R.Ex: RNA extraction day; Lib: library preparation; Seq: sequencing lane; P.Lib: person who did the library preparation; Pool: sequencing pool of the RNA sample; RIN: RNA Integrity number.

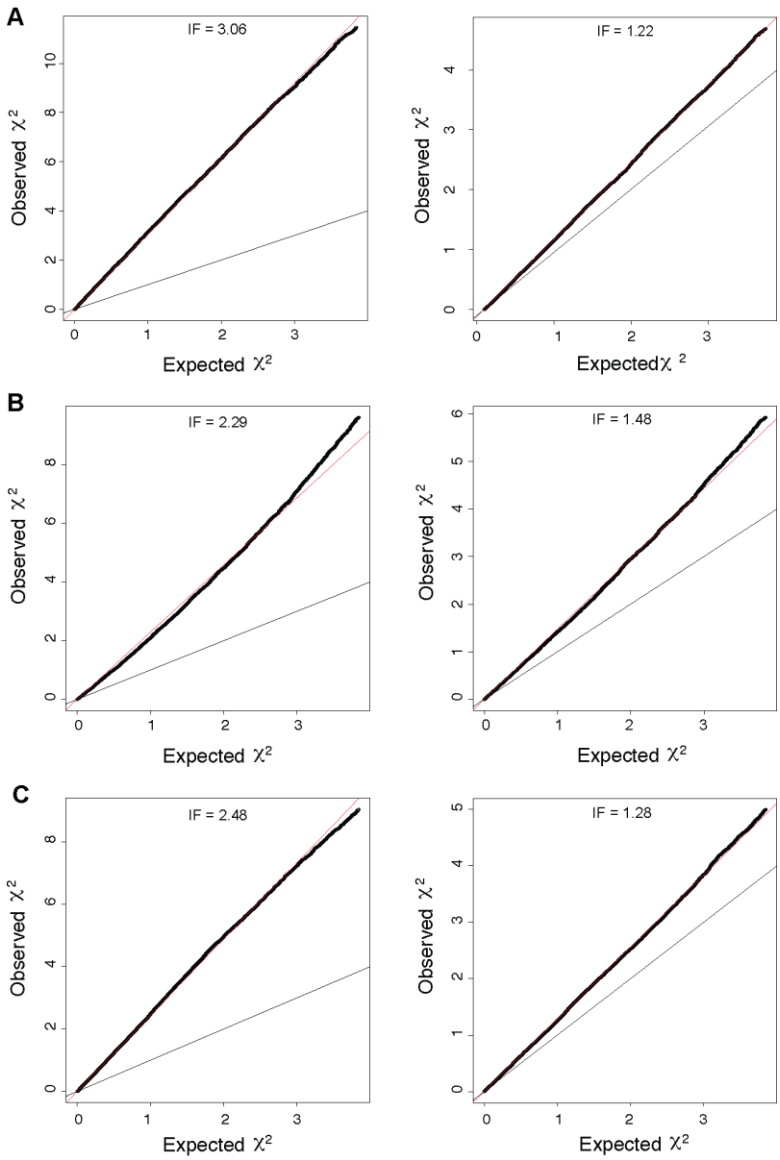


Figure 27. Q-Q plots of DE p-values pre RUV (left images) and post RUV analyses (right images). (A) OCD T0 vs. controls; (B) OCD T3 vs. controls; and (C) OCD T3 vs. OCD T0. Corresponding inflation factor values are indicated within each plot.

1.2. DE analyses identified genes specifically deregulated in OCD

DE analysis identified twenty-eight genes differentially expressed in OCD T0 vs. controls, 70 in OCD T3 vs. controls and 35 in OCD T0 vs. OCD T3 (nominal p-value <0.001, FC <0.83 (downregulated) or FC >1.2 (upregulated)) (Figure 28). In this project we will focus on the OCD T0 vs. control analysis, as this is the most relevant to identify factors contributing to OCD. The higher number of differentially expressed genes in OCD T3 vs. controls could reflect the effect of the pharmacological treatment.

Table 22 shows the top significant genes (nominal p-value <0.001, FC >1.2 or <0.83) from DE analysis of OCD T0 vs. controls, and how these genes behave in the OCD T3 vs. controls analysis, as well as in the OCD T0 vs. OCD T3. In general, genes overexpressed or underexpressed in OCD T0 vs. controls had a smaller FC in OCD T3 vs. control and did not have a significant FC in OCD T0 vs. OCD T3. The FC value between OCD T0 and OCD T3 was lower than the FC value between OCD T0 and controls or OCD T3 and controls. We looked at brain expression of these in the GTEX database[215].

From the top significant genes, there were five statistically significant with an adjusted p-value <0.05 (after Benjamini-Hochberg correction) and a FC >1.2 or FC <0.83: *NRCAM*, *AL583722.4*, *AC098935*, *KRTAP4-6*, and *HIST2H2BE*. From these genes, *AL583722* is an RNA gene of unknown function, *AC098935* a pseudogene, and KRTAP4-6 encodes a keratin-associated protein. Interestingly, *NRCAM* is a neuronal cell adhesion molecule and *HIST2H2BE* (Histone Cluster 2 H2B Family Member E) is a core component of nucleosome that may play a role in transcription regulation, DNA repair, DNA replication and chromosomal stability.

We looked for LoF variants in *HIST2H2BE* in WES data from the OCD cases and controls of this study, which could explain the underexpression of this gene, but we did not find any variant possible associated with the DE found.

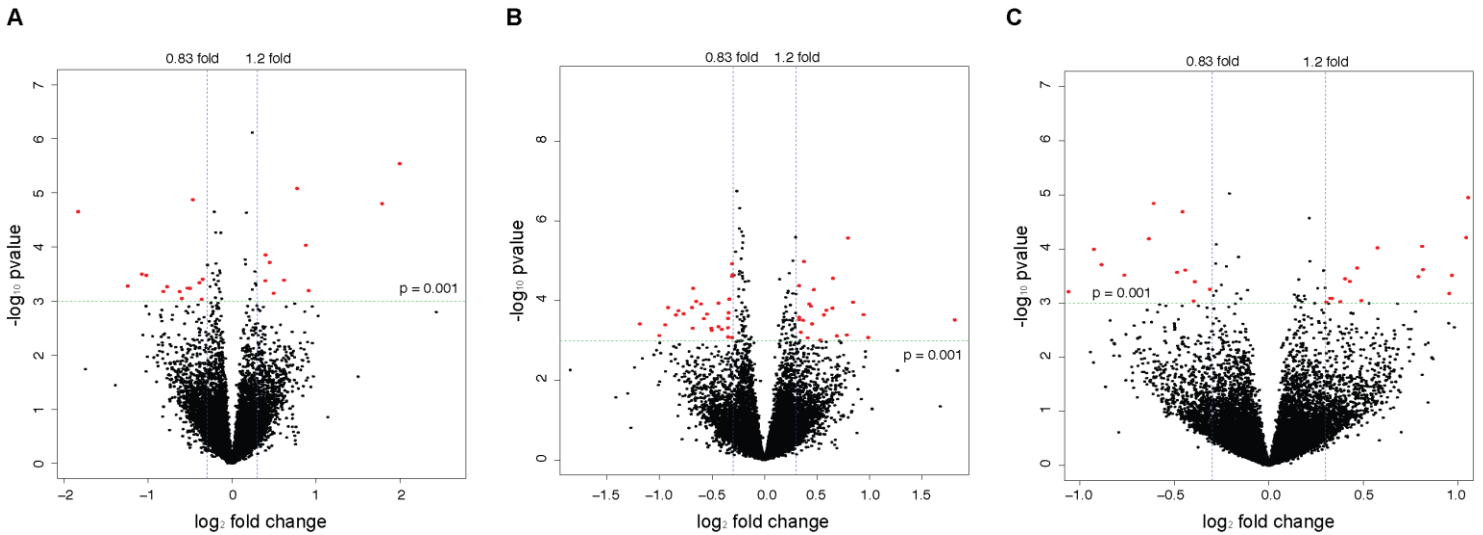


Figure 28. Volcano plot from the DE analysis of (A) OCD T0 vs. controls, (B) OCD T3 vs. controls and (C) OCD T0 vs. OCD T3. The horizontal green line corresponds to the 0.01 p-value threshold, whereas the vertical blue lines correspond to 0.83 and 1.2 fold change thresholds. Red dots correspond to genes with statistically significant differential expression. All values with a FC <0.83 and p-value <0.001 are highlighted as potential down-regulated in (A) OCD T0 vs. controls, (B) OCD T3 vs. controls or (C) OCD T3 vs. OCD T3. All values with a FC >1.2 and p-value <0.001 are highlighted as potential upregulated in (A) OCD T0 vs. controls, (B) OCD T3 vs. controls or (C) OCD T3 vs. OCD T3.

Table 22. Top significant genes from DE analysis of OCD T0 vs. controls, and comparison with results of OCD T3 vs. controls and OCD T0 vs. OCD T3

Ensemble ID	Gene Symbol	OCD T0 vs. Controls			OCD T3 vs. Controls			OCD T3 vs. OCD T0			Reads average	Reads range	>= 4 reads	B.e.
		logFC	p-val	padj	logFC	p-val	padj	log FC	p-val	padj				
ENSG00000091129.15	***NRCAM***	**1,99**	**2.88e-06**	**0.02**	1.81	3.06e-04	0.05	0.59	0.016	0.47	2.38	0-14	25	Yes
ENSG00000258858.1	***AL583722.4***	**0.77**	**8.30e-06**	**0.04**	0.54	0.001	0.10	-0.14	0.281	0.84	9.58	2-20	99	Yes
ENSG00000226945.1	***AC098935***	**-0.47**	**1.34e-05**	**0.04**	-0.33	0.007	0.22	-0.09	0.474	0.90	16.95	3-59	106	No
ENSG00000198090.3	***KRTAP4-6***	**1.78**	**1.58e-05**	**0.04**	NA	NA	NA	-0.10	0.712	0.96	1.81	0-12	18	No
ENSG00000184678.8	***HIST2H2BE***	**-1.83**	**2.23e-05**	**0.04**	-1.24	0.005	0.19	0.09	0.701	0.96	5.03	0-188	29	Yes
ENSG00000258754.3	*LINC01579*	0.88	9.32e-05	0.12	0.75	0.002	0.14	-0.21	0.088	0.68	27.57	3-117	107	Yes
ENSG00000251453.1	*HAUS1P1*	0.40	1.40e-04	0.17	0.35	0.002	0.12	0.02	0.779	0.97	70.45	24-195	108	Yes
ENSG00000268230.1	*AC012313*	0.45	1.93e-04	0.18	0.31	0.006	0.2	-0.04	0.638	0.95	25.76	7-67	108	NA
ENSG00000129484.9	*PARP2*	-0.29	2.15e-04	0.18	-0.08	0.376	0.82	0.26	0.001	0.23	48.09	13-86	108	Yes
ENSG00000179152.14	*TCAIM*	0.27	2.84e-04	0.20	0.23	0.009	0.23	-0.15	0.060	0.63	230.82	52-620	108	Yes
ENSG00000266644.1	*AC103810*	-1.07	3.18e-04	0.20	-1.19	3.84e-04	0.06	0.25	0.436	0.89	2.41	0-10	30	Yes
ENSG00000203615.2	*AC069200.1*	-1.02	3.34e-04	0.20	-0.09	0.735	0.95	0.74	0.012	0.43	2.55	0-14	29	Yes
ENSG00000270264.1	*NDUFB8P2*	-0.35	3.96e-04	0.21	-0.02	0.811	0.96	0.32	0.001	0.25	21.23	7-41	108	Yes
ENSG00000157766.11	*ACAN*	0.62	4.12e-04	0.21	0.35	0.076	0.53	-0.21	0.097	0.69	21.23	7-41	108	No
ENSG00000253908.1	*AC104115*	0.39	4.20e-04	0.21	0.27	0.138	0.64	0.04	0.723	0.96	16.67	6-37	108	Yes
ENSG00000129197.10	*RPAIN*	-0.39	4.58e-04	0.21	-0.41	5.29e-04	0.07	0.18	0.017	0.47	44.47	7-110	108	Yes
ENSG00000259177.1	*AC018946*	0.28	4.73e-04	0.21	0.21	0.008	0.23	-0.11	0.087	0.68	53.24	21-114	108	No

logFC: base 2logarithm of fold change; p-val: p-value; padj: adjusted p-value after FDR correction; Reads average: average number of reads per sample; Reads range: range of the number of reads across all samples; samples >=4 reads: number of samples containing at least 4 reads, from a total of 108 samples; B.e: brain expression; NA: not available. Top genes with significant adjusted p-value in the OCD T0 vs. controls comparison are shown in bold font. Overexpressed genes (log_2FC >0.263 (FC >1.2)) are highlighted in green, and underexpressed genes (log_2FC <-0.263 (FC <0.83)) are highlighted in orange.

Table 22 (continued). Top significant genes from DE analysis of OCD T0 vs. controls, and comparison with results of OCD T3 vs. controls and OCD T0 vs. OCD T3

Ensemble ID	Gene Symbol	OCD T0 vs. Controls			OCD T3 vs. Controls			OCD T3 vs. OCD T0			Reads average	Reads range	>= 4 reads	B.e.
		logFC	p-val	padj	logFC	p-val	padj	logFC	p-val	padj				
ENSG00000229048.4	*DUTP1*	0.29	5.11e-04	0.21	0.25	0.004	0.18	-0.05	0.492	0.91	45.79	16-87	108	Yes
ENSG00000187054.10	*TMPRSS11A*	-1.24	5.27e-04	0.21	-0.61	0.113	0.60	0.24	0.417	0.89	45.79	16-87	108	No
ENSG00000237781.2	*AL356356*	-0.77	5.42e-04	0.21	-0.47	0.073	0.53	0.31	0.154	0.74	5.81	0-38	48	No
ENSG00000175336.8	*APOF*	-0.53	5.78e-04	0.21	-0.38	0.013	0.28	0.09	0.637	0.95	8.08	1-17	97	No
ENSG00000258379.1	*AL355097*	-0.50	5.84e-04	0.21	-0.11	0.585	0.91	-0.06	0.635	0.94	14.22	2-43	100	No
ENSG00000140939.10	*NOL3*	0.91	6.40e-04	0.22	0.31	0.308	0.8	0.05	0.828	0.98	2.66	0-16	28	Yes
ENSG00000162368.9	*CMPK1*	-0.82	6.66e-04	0.22	0.05	0.803	0.96	0.62	0.006	0.37	4.31	0-18	55	Yes
ENSG00000105229.2	*PIAS4*	-0.62	6.70e-04	0.22	-0.44	0.030	0.38	0.35	0.062	0.63	5.73	0-18	77	Yes
ENSG00000234361.1	*AL391863*	0.49	7.18e-04	0.22	0.36	0.008	0.22	-0.06	0.587	0.93	16.55	1-45	107	No
ENSG00000254087.3	*LYN*	-0.60	8.97e-04	0.26	-0.50	0.006	0.20	0.06	0.750	0.96	16.55	1-45	107	Yes
ENSG00000257000.1	*AC137590*	-0.36	9.25e-04	0.26	-0.25	0.023	0.35	0.02	0.852	0.98	19.59	3-47	107	No

logFC: base 2logarithm of fold change; p-val: p-value; padj: adjusted p-value after FDR correction; Reads average: average number of reads per sample; Reads range: range of the number of reads across all samples; samples >= 4 reads: number of samples containing at least 4 reads, from a total of 108 samples; B.e: brain expression; NA: not available. Top genes with significant adjusted p-value in the OCD T0 vs. controls comparison are shown in bold font. Overexpressed genes (log_2FC >0.263 (FC >1.2)) are highlighted in green, and underexpressed genes (log_2FC <-0.263 (FC <0.83)) are highlighted in orange.

We also attempted to increase the FC threshold regardless of the adjusted p-value (FC >1.5 or FC <0.665, nominal p-value <0.05) considering that we had a small sample size to have real statistical power, and taking into account genes with an average of more than 10 reads across all samples. Following this approach, we observed upregulation of some interesting genes, such as *SYNGR1* (Synaptogrin 1), involved in synapse formation and function, and *MTRNR2L1* (MT-RNR2 Like 1), a neuroprotective molecule (Supplementary Table S14).

Finally, we also checked for overlap with the results from the study reported by Jaffe *et al.*[140], in which post-mortem brain tissue and microarrays were used to compare gene expression levels in various obsessive psychiatric disorders (which included OCD, obsessive-compulsive personality disorder or tics) and healthy subjects. We compared genes with FC >1.2 or FC <0.83, and we saw overlap of 23 genes. From these, 10 had the same direction of expression change (upregulated o downregulated) in both studies (Table 23), and three had nominal p-value <0.05 in our analysis: *ARPC3* (Actin Related Protein 2/3 Complex Subunit 3), *ZMAT2* (Zinc Finger Matrin-Type 2) and *PKD1* (Polycystin 1, Transient Receptor Potential Channel Interacting).

Table 23. Genes overlapped between our study and the study reported by Jaffe *et al.*[140]

Ensemble ID	Gene	Study II: transcriptomics			Jaffe *et al.*[140]		
		logFC	p-val	adj. p-val	logFC	p-val	adj. p.val
ENSG00000111229.11	ARPC3	0.880	0.010	0.487	0.338	0.026	0.209
ENSG00000146007.6	ZMAT2	-0.690	0.014	0.530	-0.305	5.4e-05	0.014
ENSG00000008710.13	PKD1	0.489	0.014	0.530	0.434	1.5e-05	0.009
ENSG00000184007.13	PTP4A2	-0.358	0.060	0.674	-0.354	8.3e-05	0.015
ENSG00000109670.9	FBXW7	-0.284	0.116	0.752	-0.379	6.6e-05	0.014
ENSG00000196850.4	PPTC7	-0.290	0.146	0.780	-0.272	1.6e-04	0.021
ENSG00000082701.10	GSK3B	-0.267	0.179	0.798	-0.404	1.2e-04	0.019
ENSG00000078177.9	N4BP2	-0.327	0.228	0.831	-0.276	0.001	0.036
ENSG00000125505.12	MBOAT7	0.294	0.290	0.863	0.291	2.6e-05	0.011
ENSG00000150656.10	CNDP1	-0.296	0.318	0.873	-0.454	0.002	0.063

logFC: base 2logarithm of fold change; p-val: p-value; padj: adjusted p-value after FDR correction.

1.2.1. Gene set enrichment analysis showed an overrepresentation of OCD associated genes belonging to axon guidance and semaphorin pathways

We did gene set enrichment analysis with all the statistically significant genes with nominal p-value <0.01 in the DE analysis of OCD T0 vs. controls (n=171). By doing pathway enrichment analysis, we found seven pathways significantly enriched (p-value <0.01) (Table 24). Among these pathways we found "axon guidance" and "Semaphorin interactions", which are related to neuronal development and function. Interestingly, *ARPC3* was present in the "axon guidance" pathway. By doing GO enrichment analyses we found twelve GO terms significantly enriched (p-value <0.01) (Supplementary Table S15), most of them related to cell function and organization.

Table 24. Enriched pathway-based sets

Pathway	p-value	q-value	Members	Size	Effective size
Fc gamma R-mediated phagocytosis	0.001	0.103	*ARPC3*; *SPHK1*; *AMPH*; *LYN*	91	90
Axon guidance	0.002	0.103	*LYN; DPYSL4; ARPC3; RHOC; VLDLR; NRCAM; TREM2*	357	356
Cilium Assembly	0.003	0.103	*PCM1; IFT74; ARL13B; HAUS6; ATAT1*	187	187
Semaphorin interactions	0.004	0.126	*DPYSL4; TREM2; RHOC*	64	64
Organelle biogenesis and maintenance	0.007	0.141	*PCM1; ARL13B; HAUS6; IFT74; ATAT1*	240	237
Cell Cycle	0.007	0.141	*LYN; PCM1; ARPP19; HIST2H2BE; HAUS6; PIAS4; STAG2; NHP2*	561	559
Dectin-2 family	0.009	0.152	*CLEC4E; LYN*	29	28

p-value: p-value calculated according to the hypergeometric test based on the number of physical entities present in both the predefined set and user-specified list of physical entities; q-value: p-values corrected for multiple testing using the false discovery rate (FDR). Size of the predefined sets were also corrected to the number of set members that are annotated with an ID of the user-specified ID type.

2. Metagenomics

We collected 56 OCD paired stool samples (OCD T0 and OCD T3), 7 OCD un-paired stool samples (4 OCD T0 and 3 OCD T3), and 33 stool samples from healthy individuals. We also collected 56 OCD paired pharyngeal swab samples, 8 un-paired OCD pharyngeal swab samples (4 OCD T0 and 4 OCD T3), and 32 pharyngeal swab samples from healthy individuals. We extracted DNA from all them and performed 16S-rRNA sequencing targeting the variable V3 and V4 regions of the 16S rRNA gene. Metagenomics analyses were done in collaboration with Jesse Willis from Gabaldón's Group (CRG). We analysed the microbiome composition profiling in each group of samples, estimated diversity measures, and performed several statistical analyses to study the association of taxa abundances with OCD and OCD subtypes.

2.1. Gut microbiome

2.1.1. OCD T0 samples showed a trend towards a decrease of α-diversity

The gut microbiome biodiversity for OCD T0, OCD T3 and controls was analysed via α- and β-diversity values. The OCD T0 group showed an overall lower level of all α-diversity indices considered in this study (Figure 29). Although these differences were not significant (the smallest p-value was 0.057) all tests showed the same trend. Moreover, the OCD T3 group showed more similarity to the control group in all measures, suggesting a possible interaction between the treatment and the gut microbiome.

We used five ecological indices to evaluate the compositional dissimilarity between groups both in terms of species abundance (i.e., Bray-Curtis distance, Canberra distance and Jensen-Shannon distance) and incorporating their phylogenetic relatedness (i.e., unweighted and weighted UniFrac). When analysing β-diversity we did not observe a strong separation in any measures, although some measures were statistically significant with p-values <0.05 (Canberra for OCD T0 vs. controls and Bray-Curtis, Canberra and Jensen-Shannon for OCD T3 vs. controls), as can be seen in Figure 30.

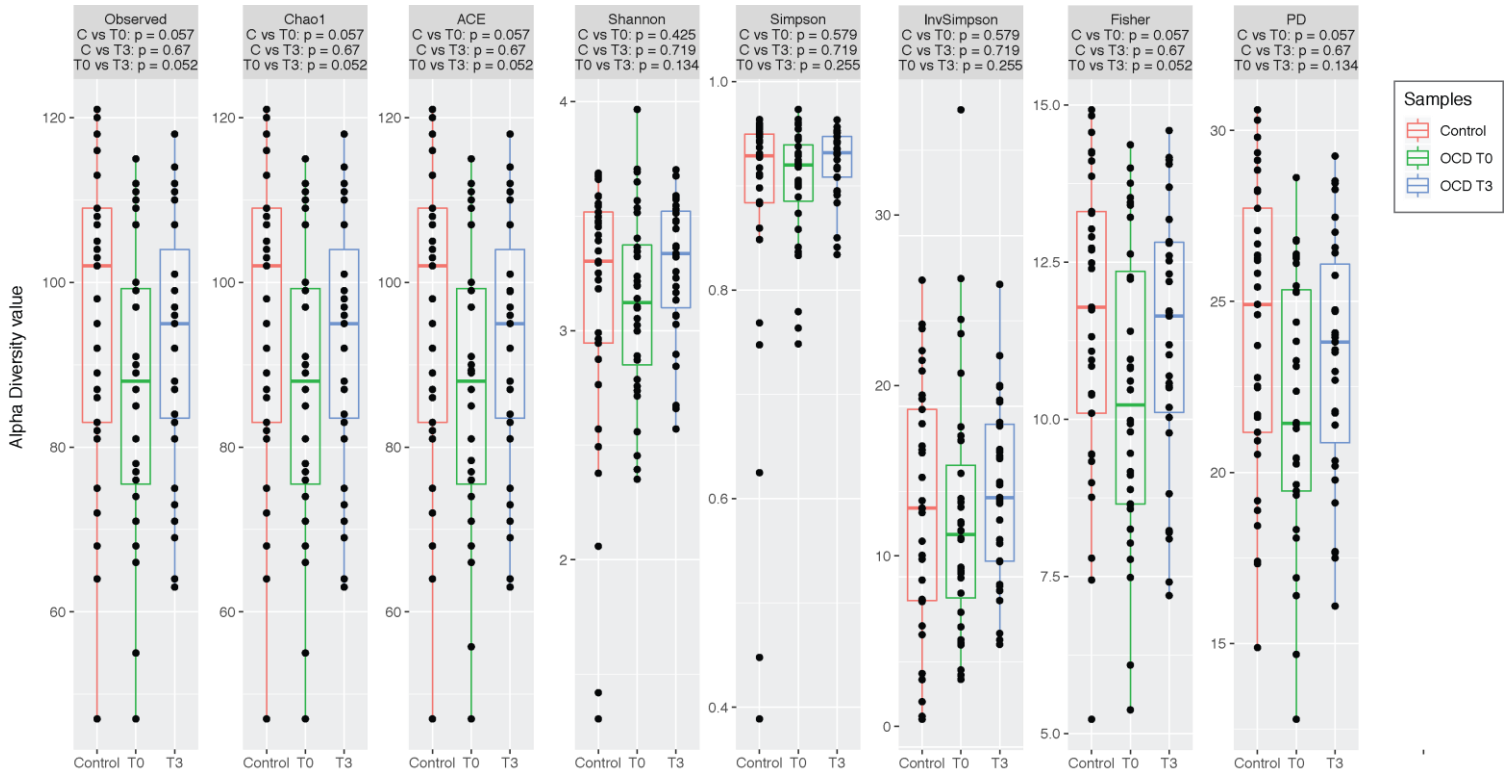


Figure 29. Boxplots representing α-diversity indices calculated for Control group (red), OCD T0 group (green) and OCD T3 group (blue) in stool samples: Observed, Chao1, ACE, Shannon, Simpson, Inverse Simpson, Fisher, and Faith's Phylogenetic Diversity. Center lines show the medians, box limits indicate the 25th and 75th percentiles, and outliers are represented by dots. The corresponding p-values are reported below each index (OCD T0 vs. controls and OCD T3 vs. controls calculated with Mann–Whitney U test; OCD T0 vs. 0CD T3 calculated with Wilcoxon rank-sum test).

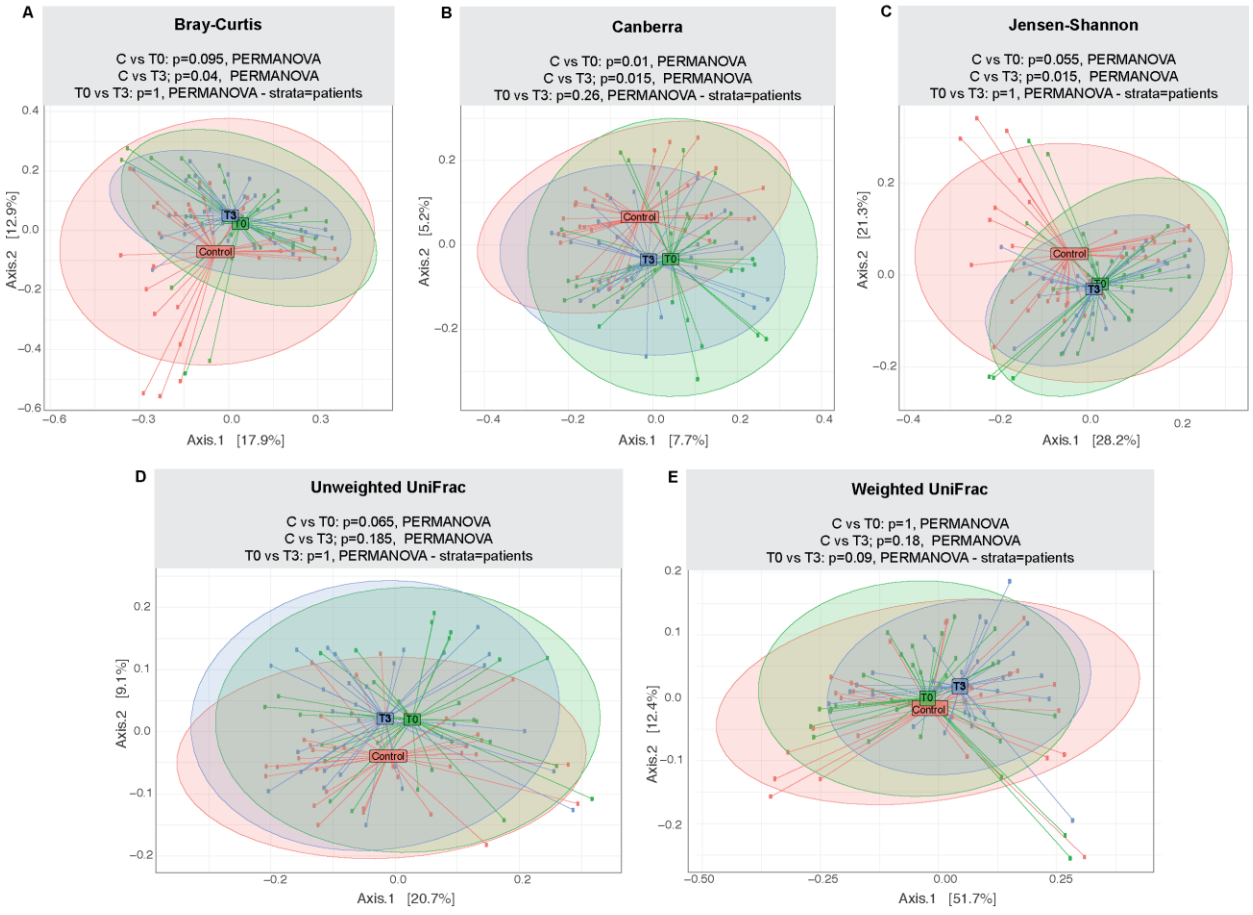


Figure 30. Principal coordinate analysis plot of OCD T0 (green), OCD T3 (red) and control (blue) groups in stool samples. The plots show the two principal coordinates for principal coordinates anallysis (PCoA) using Bray-Curtis (A), Canberra (B), Jensen-Shannon (C), unweigthed UniFrac (D) and weighted UniFrac (E) algorithms. The resulting p-values for PERMANOVA analyses are reported in the figures.

2.1.2. Specific bacterial families showed different abundances in OCD and control samples

We analysed the microbiome composition profiles in OCD T0, OCD T3 and control stool samples and we saw that the 20 most abundant species were similar within the three groups, although they differ in abundances. Gut bacterial abundances for each group at the genus and family level are shown in the Supplementary Figure S11.

To look for distinctive features in OCD T0 vs. control samples, taxa distribution was investigated at all taxonomical levels (species, genus, family, order, class and phylum). Results of Wilcoxon rank sum test highlighted statistically significant taxa abundance differences at the family level (Figure 31), with a higher percentage of *Rikenellaceae* and a lower level of *vadinBE97* in OCD T0 compared to controls. In addition, OCD cases with sexual obsessions presented, at the phylum level, lower percentage of *Firmicutes* and higher percentage of *Bacteroidetes*. Although there was some association between some of the diet variables and taxa abundances, there were no significant associations between the variables collected in the diet questionnaires and whether the individual belonged to the OCD or control group, which suggests that the taxa differences between OCD cases and controls may be influenced by the OCD phenotype, rather than by other variables.

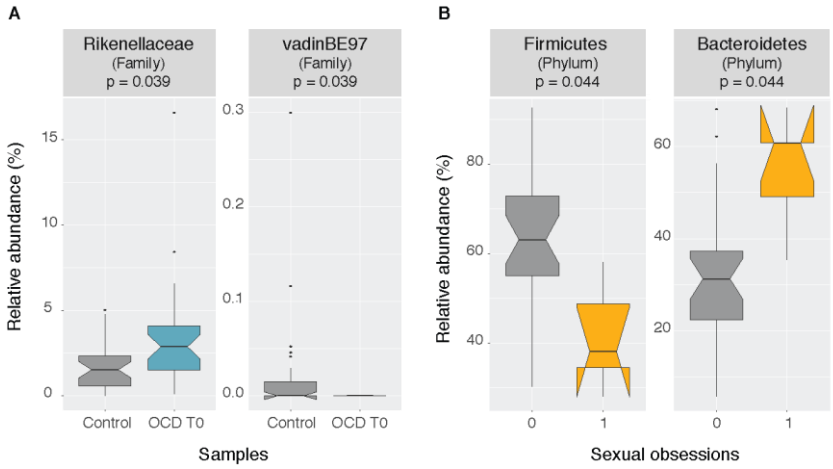


Figure 31. Boxplots representing significant Wilcoxon rank-sum test results of stool samples from (A) control (grey) and OCD T0 (blue) groups at the family level, and (B) OCD T0 without sexual obsessions symptoms (grey) and with sexual obsession symptoms (orange) at the phylum level (B). 0: absence of symptoms; 1: presence of symptoms. p-value is indicated.

2.1.3. LEfSe analysis found biomarkers of OCD

We further analysed the structure of the bacterial community associated with OCD by using LEfSe, which revealed a significant increase of the relative abundance of different bacterial taxa at different taxonomical levels (Figure 32). In concordance with the Wilcoxon rank sum test, the *Rikenellaceae* family (phylum *Bacteroidetes*) and the *Alistipes* genus, from this family, were found to be biomarkers of OCD. In addition, several genus in the *Clostridiales* order (phylum *Firmicutes*) were also found in higher levels in OCD than controls, including *Oscillibacter, Anaerostipes,* and *Flavonifractor,* as well as several *Costridiales* species: *Anaerostipes hadrus*, *Intestinimonas butyriciproducens* and *Clostridium hathewayi*. On the other hand, the control group had higher levels of the *Prevotellaceae* family (phylum *Bacteroidetes*), as well as a set of genus from the order *Clostridiales*: *Agathobacter*, *Coprococcus, Lachnospira, Howardella, Romboutsia, Butyricicoccus, Clostridium.* Moreover, control samples presented high levels of *Negativicutes* at the order level (class *Clostridia*, phylum *Firmicutes*). Figure 33 shows the phylogenetic relationship of significant bacterial taxa associated with each group.

Results obtained from LEfSe highlighted the differential composition of the microbiome in OCD vs. controls, especially of *Firmicutes* and *Bacteroidetes*. However, while it seems that there is an alteration in the composition of the bacterial gut community at the phylum level, there is not a differential *Firmicutes*/*Bacteroidetes* ratio in OCD T0 vs. controls (Figure 34).

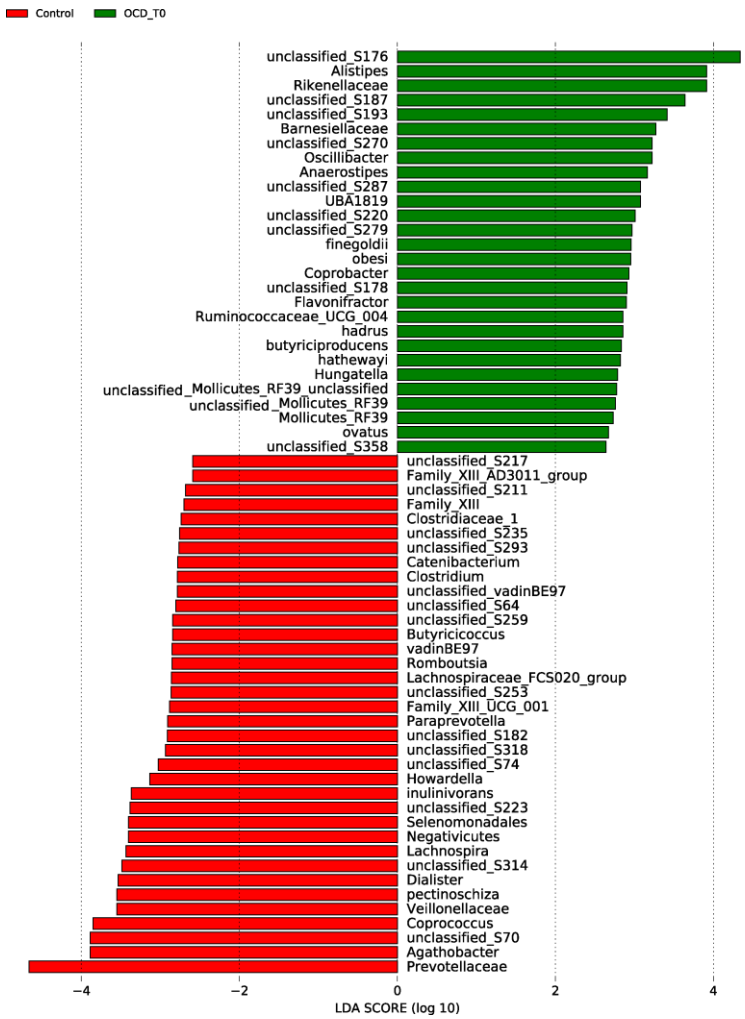


Figure 32. Biomarkers associated with OCD and control groups discovered by a linear discriminant effect size (LEfSe) analysis (α value=0.05, logarithmic LDA score threshold=2.0) in stool samples.

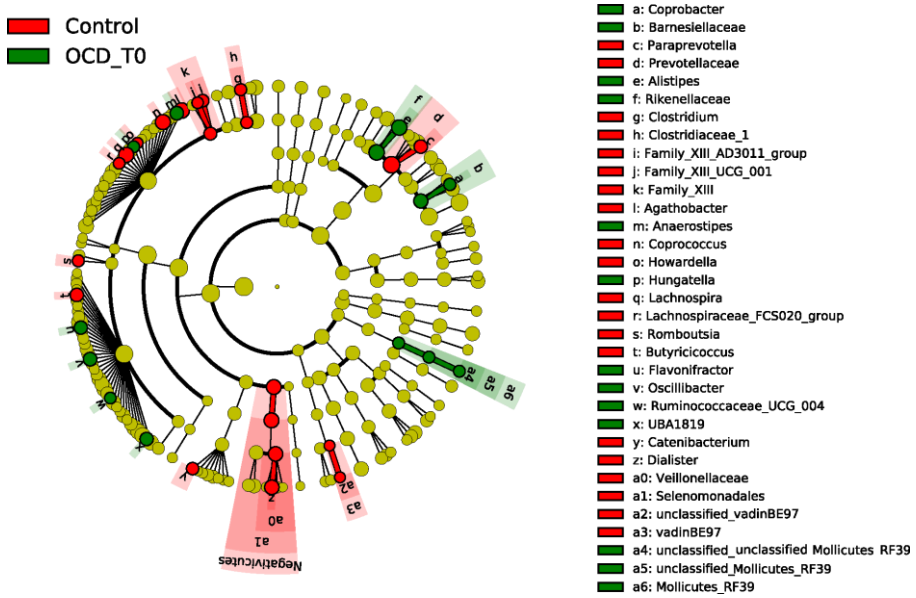


Figure 33. Cladogram representing the phylogenetic relationship of biomarkers associated with OCD and control groups through the linear discriminant effect size (LEfSe) analysis (α value=0.05, logarithmic LDA score threshold=2.0) in stool samples.

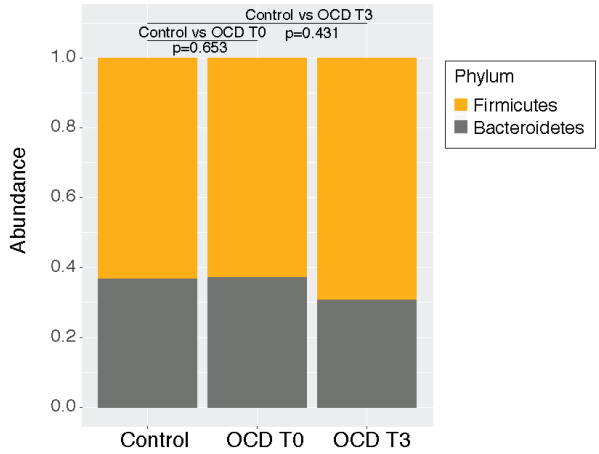


Figure 34. Mean relative abundances (%) of Firmicutes and Bacteroidetes in Control, OCD T0 and OCD T3 subjects. p-value of Wilcoxon rank-sum test between Controls and OCD T0, and controls and OCD T3 are indicated.

2.2. Oro-pharyngeal microbiome

2.2.1. The oro-pharyngeal microbiome showed little α- and β-diversity differences between OCD and controls

The oro-pharyngeal microbiome biodiversity for OCD T0, OCD T3 and controls was also analysed via α- and β-diversity values. In contrast with the gut microbiome biodiversity, in the oro-pharyngeal microbiome the OCD T0 group did not show any difference in any of the measures for α -diversity (neither statistically significant nor a trend) (Figure 35).

When analysing β-diversity with the five ecological indices (Bray-Curtis distance, Canberra distance, Jensen-Shannon distance, and unweighted and weighted UniFrac) we observed no separation in most of the measures (Figure 36). The only statistically significant (p-value <0.05) difference, between OCD T0 and controls, was observed for the weighted UniFrac index. So, there is only a separation of these two groups when we incorporate the phylogenetic relatedness of the taxa and weight for abundance of observed organisms.

2.2.2. OCD samples presented higher abundance of *Actinobacteria*

We analysed the microbiome composition profiles in OCD T0, OCD T3 and control pharyngeal samples and we saw that the 20 most abundant species were similar within the three groups, although they differ in abundances. Supplementary Figure S12 shows the distribution of relative abundances of oro-pharyngeal bacterial families and genera among the three groups.

We looked specifically for enrichment of *Streptococcus pyogenes* in OCD samples vs. controls, as this species was shown to be responsible of PANDAS[232]. However, the first 15 most abundant species did not include bacteria form this genus in either OCD T0, OCD T3, or controls.

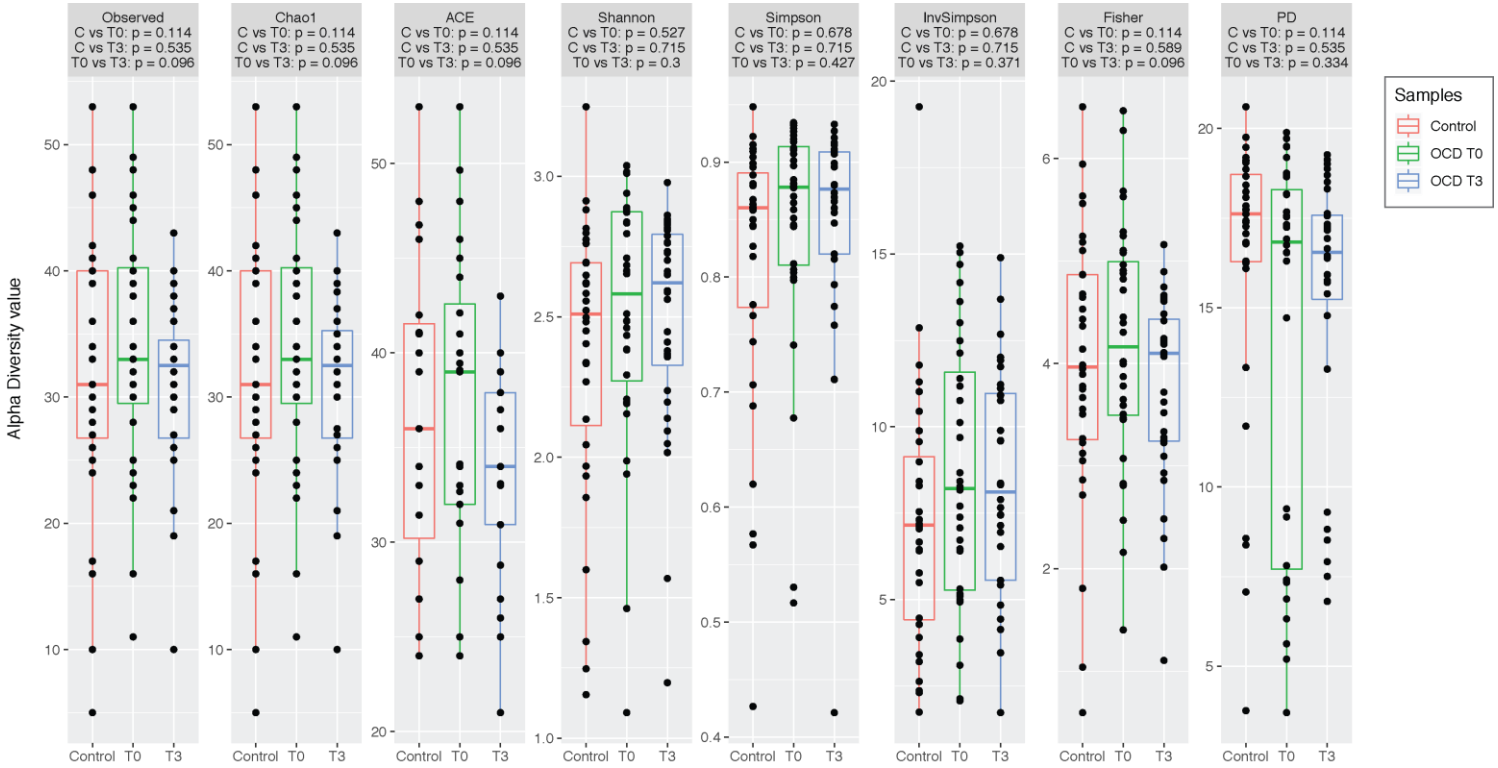


Figure 35. Boxplots representing α-diversity indices calculated for Control group (red), OCD T0 group (green) and OCD T3 group (blue) in pharyngeal samples: Observed, Chao1, ACE, Shannon, Simpson, Inverse Simpson, Fisher, and Faith's Phylogenetic Diversity. Center lines show the medians, box limits indicate the 25th and 75th percentiles, and outliers are represented by dots. The corresponding p-values are reported below each index (OCD T0 vs. controls and OCD T3 vs. controls calculated with Mann–Whitney U test; OCD T0 vs. 0CD T3 calculated with Wilcoxon rank-sum test).

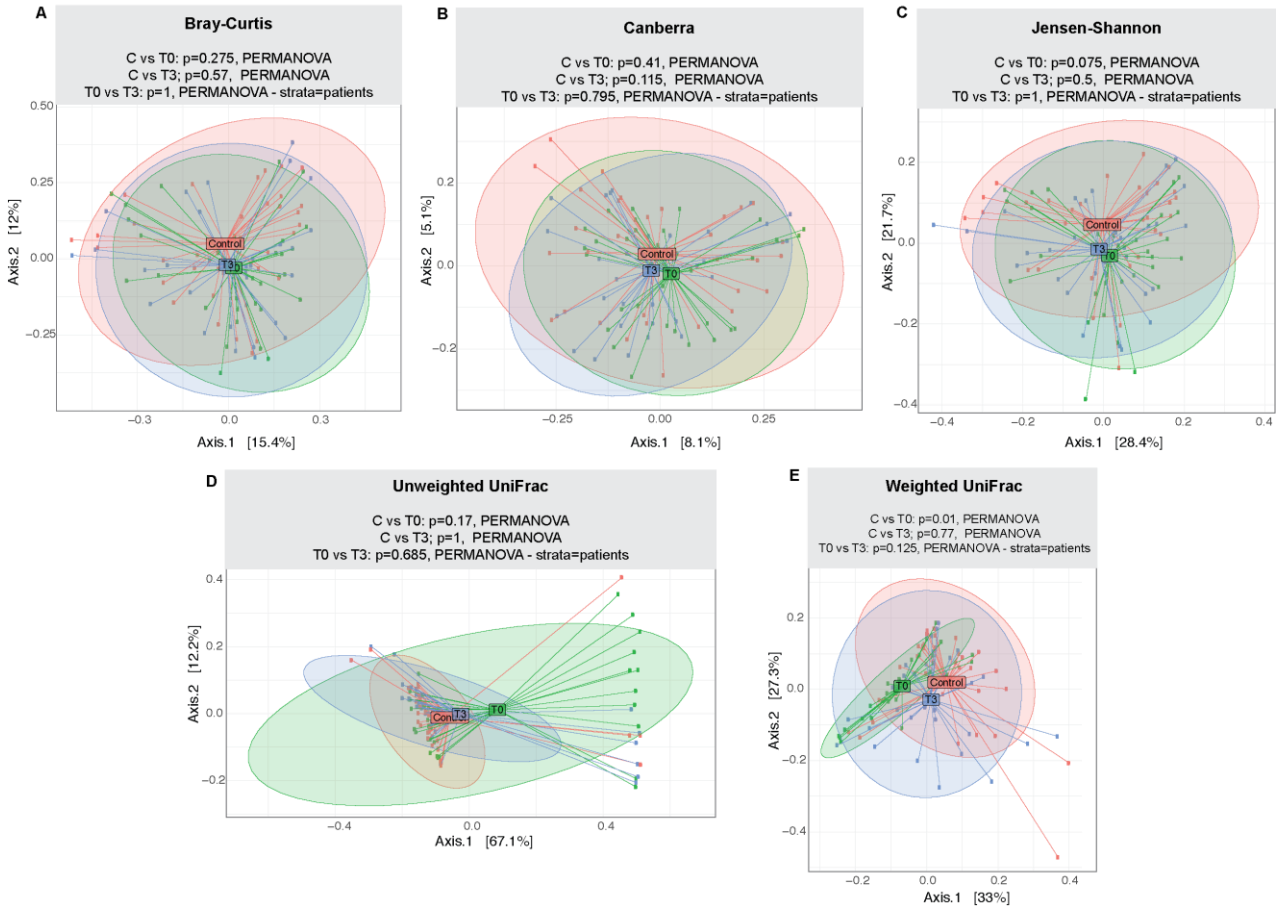


Figure 36. Principal coordinate analysis plot of OCD T0 (green), OCD T3 (red) and control (blue) groups in stool samples. The plots show the two principal coordinates for principal coordinates anallysis (PCoA) using Bray-Curtis (A), Canberra (B), Jensen-Shannon (C), unweigthed UniFrac (D) and weighted UniFrac (E) algorithms. The resulting p-values for PERMANOVA analyses are reported in the figures.

To look for distinctive features in OCD T0 and control samples, taxa distribution was investigated at all taxonomical levels. Results of Wilcoxon rank sum test highlighted taxa abundance differences (statistically significant) at the order, class and phylum level (Figures 37 and 38), with a higher percentage of *Coriobacteriales*, *Coriobacteriia* and *Actinobacteria*, respectively, in OCD T0 compared to controls. In addition, there is also an unclassified taxa less abundant in OCD T0 than in controls. Interestingly, OCD cases with ordering compulsions presented, at family and order levels, higher percentage of *Neisseriaceae* and *Betaproteobacteriales* (now called *Neisseriales*), respectively. As in gut, although there was a certain association between diet and some taxa, it did not correlate with belonging to the OCD or control group.

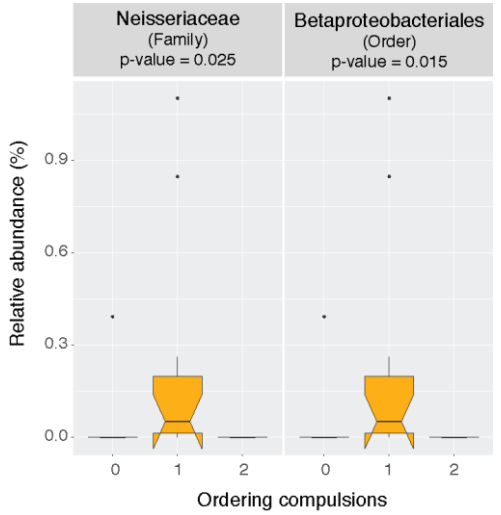


Figure 37. Boxplots representing significant Wilcoxon rank-sum test results of stool samples from OCD T0 without ordering compulsions (grey) and with ordering compulsions (orange) at the family (left) and order (right) level. 0: absence of symptoms; 1: presence of symptoms; 2: principal symptom. p-value is indicated.

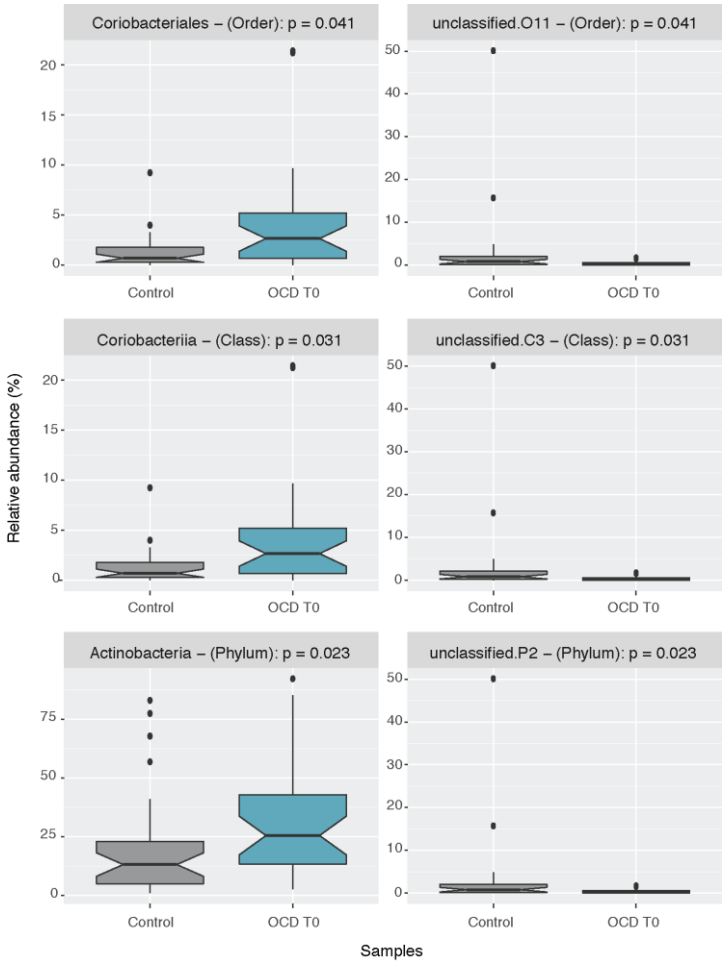


Figure 38. Boxplots representing significant Wilcoxon rank-sum test results of pharyngeal samples from the control (grey) and OCD T0 (blue) groups at different taxonomical levels. p-value is indicated.

2.2.3. LEfSe analyses identified OCD biomarkers

We further analysed the structure of the bacterial community associated with OCD by using LEfSe, which revealed a significant increase of the relative abundance of different bacterial taxa in different taxonomical levels (Figure 39). In concordance with the Wilcoxon rank sum test, we found *Actinobacteria* as an OCD biomarker at the phylum level, including *Actinobacteria* and *Coriobacteriia* at the class level, *Actinomycetales* and *Coriobacteriales* at order level,

Actinomycetaceae and *Atopobiaceae* at family level, *Actinomyces* and *Atopobium* at genus level, and *Actinomyces odontolyticus* and *Atopobium parvulum* at species level. Other taxa found in higher levels in OCD than controls were *Lachnospiraceae* (at the family level), including the genus *Lachnoanaerobaculum* and *Oribacterium*; *Mogibacterium* (at the genus level); and *Peptostreptococcus asaccharolyticus* (at the species level). All of them belong to the order *Clostridiales* (class *Clostridia* class, phylum *Firmicutes*). There were also some unclassified *Bacteroidales* as biomarkers of OCD. On the other hand, controls had higher levels of *Fusobacteria* at phylum level, *Fusobacteriia* at class level, and *Fusobacteriales* at order level. Figure 40 shows the phylogenetic relationship of significant bacterial taxa associated with each group.

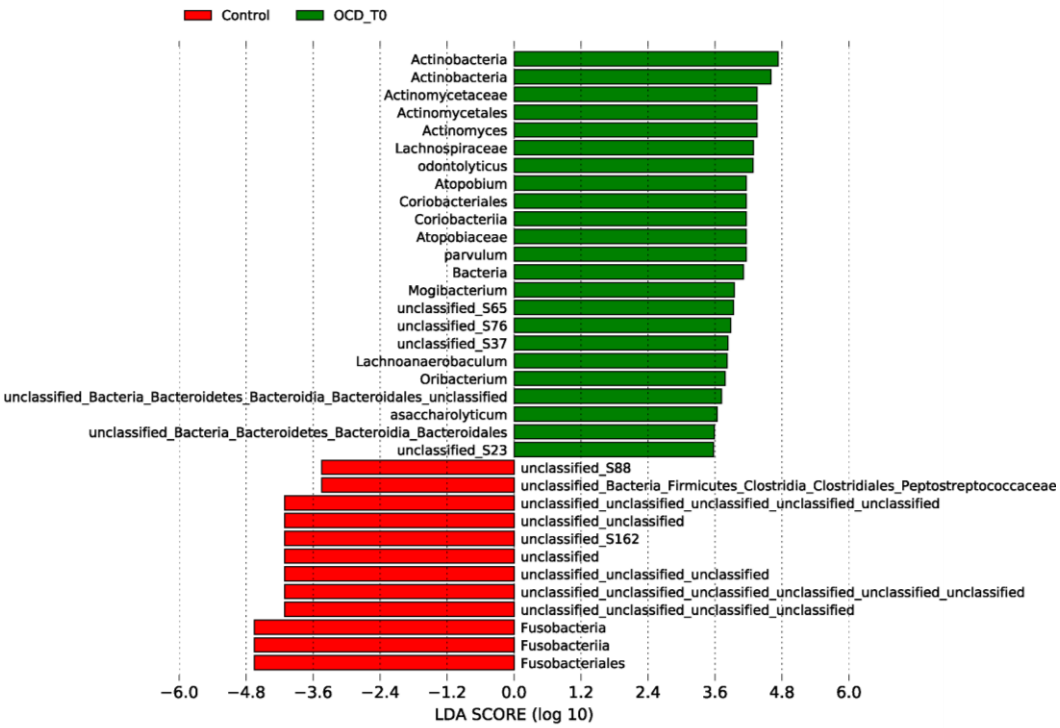


Figure 39. Biomarkers associated with OCD and control groups discovered by a linear discriminant effect size (LEfSe) analysis (α value = 0.05, logarithmic LDA score threshold = 2.0) in pharyngeal samples.

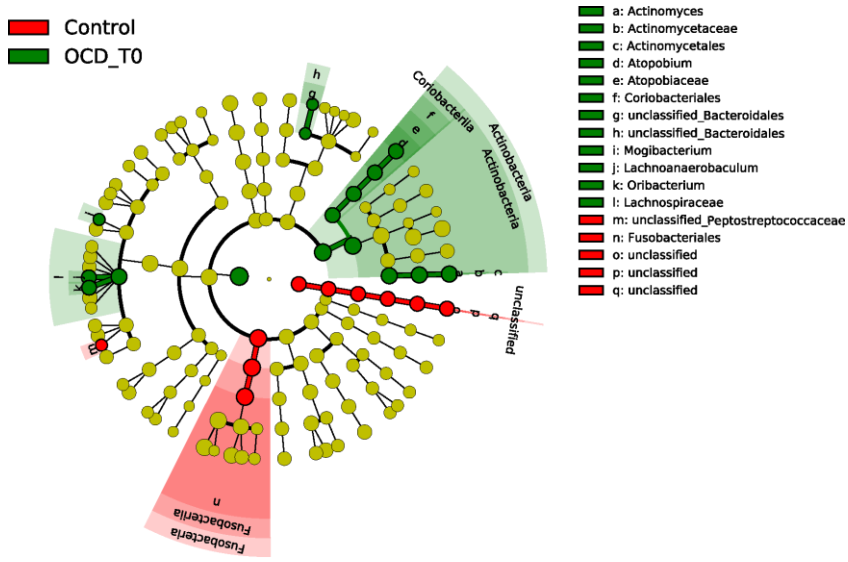


Figure 40. Cladogram representing the phylogenetic relationship of biomarkers associated with OCD and control groups through the linear discriminant effect size (LEfSe) analysis (α value=0.05, logarithmic LDA score threshold=2.0) in pharyngeal samples.

Results obtained from LEfSe highlighted the differential composition of the oro-pharyngeal microbiome in OCD vs. controls, especially of *Actinobacteria* and *Fusobacteria*. Phylum levels analysis showed a clear alteration of the bacterial pharynx community in OCD T0 characterized by a higher *Actinobacteria*/*Fusobacteria* ratio (p<0.004, Wilcoxon rank-sum test) in OCD T0 than in controls (Figure 41).

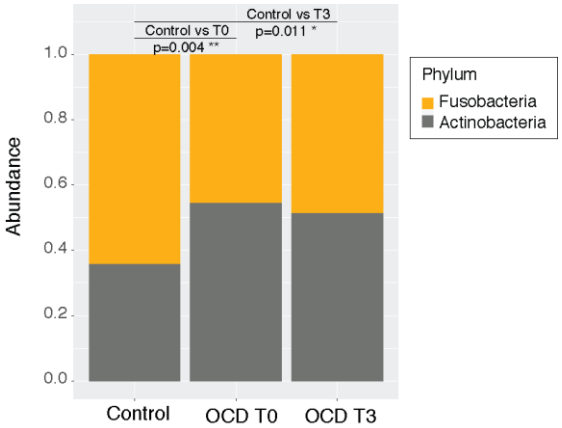


Figure 41. Mean relative abundances (%) of *Fusobacteria* and *Actinobacteria* in Control, OCD T0 and OCD T3 subjects. p-value of Wilcoxon rank-sum test between Controls and OCD T0, and controls and OCD T3 are indicated.

DISCUSSION

Most research in OCD has been focused on neurobiological, neuropsychological and treatment studies. In the last decades, numerous efforts have also been made in elucidating the genetic causes of this disorder, mostly through candidate genes, linkage studies, and GWAS, and highly successful international collaborative projects have enabled association studies to reach sizes of thousands samples[89,90]. However, despite the increases in statistical power afforded by these large-scale studies, there is still a big gap between the phenotypic variance and the genetic variance identified so far.

Considering this, we decided to explore new layers of complexity, relatively underexplored in OCD, which could explain part of the missing heritability observed in this disorder. We believe that understanding the heritability of complex neuropsychiatric disorders such as OCD requires a more comprehensive assessment of human genetic variation, including rare variation, common and low-frequency genetic variation with small effect sizes, structural variation, gene-gene and gene-environmental interactions, and taxonomic and functional changes to the composition of the human microbiome. Therefore, in this project, we have explored OCD through: i) the analysis of both rare and common and low-frequency variants from WES data of 306 OCD cases and 601 controls, and ii) the longitudinal analysis of the transcriptome and the gut and oro-pharyngeal microbiome in a small subset of samples (43 OCD cases and 32 controls).

Study I: Deciphering OCD by whole-exome sequencing

1. Association studies

1.1. Analysis of rare variants through WES and targeted resequencing identifies *TMEM63A* as a novel OCD candidate gene

We have searched for enrichment of rare coding variants in 306 OCD patients using WES and RVAS. This represents the first study in OCD following this approach. Moreover, this study also represents the first one that includes a WES cohort of OCD of this size as, to date, the two published studies[81,112] that sequenced the exome of individuals with OCD included only 20 and 10 sporadic trios, respectively.

To identify rare variants potentially implicated in this disorder, we used different RVAS methods and approximations, as recommended[114] when the genetic architecture underlying the disorder is not known. In this discovery phase we ended up with a large list of genes possibly associated with OCD with nominal p-value <0.05. None of these genes was significant after adjusting for multiple testing (FDR), but this was not unexpected, as our sample size was small and we did not have enough statistical power. Since each method assumes a different genetic basis of OCD, determining which of these genes are really associated with the disorder would have required validation of all the results in a larger and independent cohort of samples. However, at this point we could only perform a preliminary capture array of 20 genes in 439 OCD cases and 1481 controls, prioritizing risk genes based on statistical significance and biological relevance. Moreover, we decided to focus on SKAT-O results, as this method considers a combination of scenarios, being able to detect associations both under the burden test and the variance-component method.

Of the 20 genes included in the targeted resequencing replication, *TMEM63A* (Transmembrane Protein 63A) showed significant enrichment above the FDR

cut-off for missense variants in OCD cases. This gene is highly expressed in different parts of the brain, such as the prefrontal cortex, the occipital and parietal lobes, and the trigeminal ganglion. It encodes for a calcium-permeable cation channel that was described as osmosensitive for years, but, very recently, it has been also reported as a mechanically activated ion channel[233]. Mechanosensitive ion channels are involved in the regulation of axon guidance. Specifically, they detect tissue stiffness and regulate axon spreading (axons grew faster in stiffer substrates). It remains to be determined whether the rare variants found in *TMEM63A* affect the activity of this channel, which, in turn, could affect axon guidance during neuronal development and function and lead to OCD.

Among the remaining top OCD candidate genes that we found, four had been already linked to OCD and have important roles in neuronal development and/or function: *CHD8* (Chromodomain Helicase DNA Binding Protein 8), *ASTN2* (Astrotactin 2), *USP54* (Ubiquitin Specific Peptidase 54), and *DHRS11* (Dehydrogenase/Reductase 11).

CHD8 acts as a chromatin remodelling factor and a transcription regulator. It is a negative regulator of the Wnt signaling pathway, which plays an important role in developing neural circuits and in adult brain function[234]. Cappi *et al.*[112] identified a *de novo* missense variant within this gene in an OCD patient. Moreover, *de novo* truncating mutations of *CHD8* are amongst the strongest individual risk factors for ASD[217] and at least one of the reported ASD cases with a *de novo* balanced translocation disrupting *CHD8* presented OCD among the various symptoms described[217].

ASTN2 encodes for a protein that is expressed in the brain and plays a role in neuronal migration and modulation of synaptic activity[216]. CNVs of *ASTN2* have been identified in patients with different neurodevelopmental disorders, including autism, schizophrenia, ADHD, bipolar disease, intellectual disability, and global developmental delay[216]. Recently, Gazzellone *et al.*[81] genotyped OCD paediatric

probands and detected a rare deletion within this gene (of at least 15 kb) in two OCD patients, one of them presenting also ADHD.

In the same manuscript, Gazzellone *et al.*[81] also performed WES of OCD parent trios and found one individual with OCD carrying a nonsense variant in *UPS54*, which encodes for a member of the to the Ubiquitin Specific Peptidases (USP) family, which are involved in the Ubiquitin-Proteasome Pathway. This pathway is critical for normal function of the nervous system and is implicated in various neurological diseases[235].

Finally, in the IOCDF-GC GWAS, Stewart *et al*.[89] found association of a genome-wide significant SNP (rs6131295), which was an eQTL for *DHRS11.* This gene codes for a dehydrogenase/reductase that has been recently involved in neurosteroid metabolism[236]. Neurosteroids alter neuronal excitability through interaction with ligand-gated ion channels and other cell surface receptors, and can exert both excitatory and inhibitory actions on neurotransmission.

UPS54 was not replicated in the targeted resequencing assay, but a mild enrichment of rare variants in OCD cases compared to controls was observed. In the case of *DHSR11*, the RVAS performed with the targeted data only found one missense variant, present in two OCD patients and not in controls, and this gene was not significant, whereas the variants reported in the WES RVAS were truncating mutations. So, although interesting, this result cannot be evaluated. We should consider increasing our sample size to test association of these genes with enough statistical power. *ASTN2* and *CHD8* had negative DIC values, which is indicative of non-statistical significant association with OCD in our data.

We also did pathway enrichment analysis of all statistically significant genes (nominal p-value <0.05) found by the SKAT-O method and the most homogeneous approximation performed (i.e. selecting only those samples captured with NimbleGen v3 and a MAF <0.005), and we found six pathways significantly enriched (p-value <0.01): “TRP channels”, “Carboxyterminal post-

translational modifications of tubulin", "Other semaphorin interactions", "Acyl chain remodelling of PS", "Amine compound SLC transporters", and "Acyl chain remodelling of PE". All of them are related to neuronal development and function, which gives additional support to the RVAS results as good OCD candidate genes.

Of these pathways, we consider the first one especially interesting, as it relates to calcium signalling, like *TMEM63A*. Four genes from the TRP (transient receptor potential) channel family were enriched in rare variants in OCD cases: *TRPV5*, *TRPC3*, *TRPV3*, and *TRPM3.* These proteins form non-selective cation channels that can activate or inactivate voltage-gated ion channels, and regulate calcium signalling, which controls diverse cellular functions[237]. TRP channels are involved in many processes in the nervous system, such as the transduction of sensory stimulation, neuronal cell death, proliferation and differentiation of neural progenitor cells, nerve growth, synaptic transmission, and signal transduction of axon guidance during brain development[238]. TRPV5 and TRPV3 have been associated, so far, to neuronal functions in rats[239,240]. TRPC3 and TRPM3 were reported highly expressed in human brain, where they play important roles regulating diverse neuronal and glial functions[237]. In fact, recent studies are suggestive of potential roles of TRP channels in numerous neurological and psychiatric disorders, such as schizophrenia, autism, bipolar disorder, anxiety disorder, or Alzheimer's disease, among others[237].

Also interesting is the semaphorin related pathway, which included three semaphorin receptors (*ITGA1, PLXNA1* and *PLXNA4*). Semaphorins have an important role in the development of the nervous system and in axonal guidance[241]. Recent evidence points to additional roles in the development, function and reorganization of synaptic complexes[242]. In addition, mutations in semaphorin genes are linked to several human diseases associated with neurological changes[242], although their actual influence in the pathogenesis of these diseases remains to be demonstrated.

Semaphorin-triggered signalling also induces the rearrangement of the actin and microtubule cytoskeleton, which could be related to the enrichment of the "carboxyterminal post-translational modifications of tubulin" pathway (genes involved: *TTLL4*, *TTLL6*, *AGBL1* and *TTLL3*). Interestingly, anomalies of the microtubule and microtubule related proteins have also been associated with psychiatric diseases[243].

Three genes (*MBOAT1, PLA2G4A* and *LPCAT4*) participate in the two pathways related to lipid metabolism that showed enrichment: "Acyl chain remodelling of PS" and "Acyl chain remodelling of PE". Phosphatidylserine (PS) is the major anionic phospholipid class particularly enriched in the inner leaflet of the plasma membrane in neural tissues, and it is synthesized from phosphatidylcholine or phosphatidylethanolamine (PE)[244]. PS is necessary for the activation of Akt and Raf-1 and protein kinase C signalling, which are relevant for neuronal survival and differentiation[244]. In addition, lipid metabolic disorders or abnormalities can lead to a variety of neuropsychiatric disorders, such as bipolar disorder, schizophrenia or major depressive disorder[245].

Finally, three genes from the solute carrier (SLC) family (*SLC44A1, SLC18A1,* and *SLC6A9*) led to the enriched pathway "Amine compound SLC transporters". SLC transporters facilitate the transport of a wide array of substrates across biological membranes and have important roles in physiological processes[246]. In particular, these three genes are involved in the transport of choline (*SLC44A1*), the reuptake of glycine (*SLC6A9*) and the regulation of glycine levels in NMDA receptor-mediated neurotransmission[247], and the intracellular transport of monoamines, such as serotonin (*SLC18A1*), to the secretory vesicles of neuroendocrine and endocrine cells. Interestingly, several studies have reported genetic association between variants in *SLC18A1* and susceptibility to bipolar disorder and schizophrenia[248].

Despite we have found genes and pathways related to neuronal development and function, which would indicate that these genes might be good candidates for OCD involvement, our results are very preliminary and stem from a pilot

study with WES data from only 306 OCD cases. These associations would need further validation by additional studies in independent cohorts of OCD cases and controls, such as case-control studies of rare variants from WES data, studies of rare *de novo* variants in OCD parent trios, and targeted resequencing replication assays of these genes, as well as by gene expression and functional studies.

1.2. Analysis of common and low-frequency variants points towards novel OCD candidate genes

We also performed an association study of common and low-frequency variants discovered by WES. These study detected 34 and 13 variants (in samples captured with Agilent 35, Agilent 50 and NimbleGen v3 or only with NimbleGen v3, respectively) associated with OCD with statistical significance (Benjamini-Hochberg adjusted p-value <0.01). Five variants reached genome-wide significance in the analysis that included samples captures with the three kits, after removing those protective variants that were present in less than 90% of the controls. However, the p-values obtained were unexpectedly high, especially considering our sample size and significant levels achieved in previous OCD GWAS[89,90]. This may indicate a variant calling bias and validation by Sanger sequencing is required, since they could be false-positives. Of note, we did not see genome-wide significant variants in the analysis performed with only NimbleGen v3 samples, which could be indicative of a batch effect in the Agilent libraries. It would be necessary to validate all these variants by Sanger sequencing. We also checked for overlap with regions with suggestive evidence of association with OCD (p-values $<10^{-05}$), but we did not identify any coincidence.

Some of the variants for which we observed association with OCD were within genes involved in neuronal development and function. Nevertheless, our top hit corresponded to a variant in *FIP1L1* (Factor Interacting With PAPOLA And CPSF1), which has no neuronal function known. We also found missense variants in *CTBP2* (C-Terminal Binding Protein 2), *HLA-DRB5* (Major

Histocompatibility Complex, Class II, DR Beta 5), *GBGT1* (Globoside Alpha-1,3-N-Acetylgalactosaminyltransferase 1), *PABPC3* (Poly(A) Binding Protein Cytoplasmic 3) and *ACTR3C* (ARP3 Actin Related Protein 3 Homolog C), an inframe insertion in *SERINC2* (Serine Incorporator 2) and a frameshift insertion in *PPP1R12B* (Protein Phosphatase 1 Regulatory Subunit 12B, MYPT2). Interestingly, *CTBP2* encodes a protein from the CTBPs family that is suggested to regulate neuronal differentiation and to be involved in synaptic functions[249]. *SERINC2* encodes for a transmembrane protein that facilitates incorporation of serine into phosphatidylserine and sphingolipids, which play important roles in neural plasticity, signalling and axonal guidance. Moreover, *SERINC2* has been linked to alcohol dependence[228] and autism[229]. *PPP1R12B* encodes for the regulatory subunit of the myosin phosphatase and its expression is specific to heart, skeletal muscle, and brain. It was reported that myosins have specific pre- and postsynaptic roles that are required for synapse function and synaptic plasticity[250].

We found other variants statistically significant that we did not consider because they did not pass the ExAC filters. The variants in *CTBP2* and *PABPC3* were not reported in ExAC, and it would be necessary to validate them by Sanger sequencing to know if they are population specific.

It would be interesting to genotype the top variants, after validating them by Sanger sequencing, in a larger cohort of OCD cases and controls.

1.3. Limitations and considerations of rare, low-frequency and common variant analyses from WES data

Here, we assessed the feasibility of WES analysis to identify rare, low-frequency and common genetic variation associated with OCD. In addition to the inherent sample size limitation, which was known before hand, our study has uncovered some limitations that should be considered in future studies.

The improvement of NGS capture methods along the project involved the inclusion of samples sequenced at different time points, with diverse sequencing platforms and enrichment technologies. These differences could influence findings in the subsequent association analysis, generating false associations. Aware of this, we removed all possible biases of our data re-analysing all the whole-exome samples with the same pipeline, from alignment to variant calling, filtering, and annotation. We observed that the main factor leading to different clustering in PCA was the set of kit-specific variants originated from different captures, while there was practically no detectable impact from sequencing date or run, or from the samples' origin (considering that all samples were collected at Spanish hospitals, most around Barcelona). This led us to select only those regions well covered by all the kits used to sequence the samples included in the association analyses. This meant analysing the coverage in our real data and establishing a read depth threshold (10 reads) for each position covered by all the kits, rather than selecting the regions targeted by the intersection of them, independently of the coverage. Sometimes, different libraries can target the same regions, but with different coverage results. If allele frequencies are compared from variants sequenced with different read depth in case and control cohorts, false associations may be generated and/or true ones masked[251].

However, although we tried to diminish at minimum the possible bias, there were still confounders that could influence findings, such as the usage of different sequencing platforms (case-control imbalances across different sequencing platforms might increase type I error rates). For this reason, we conclude that the best strategy for NGS association analysis requires both cases and controls to be sequenced together using a common experimental design, with the same library capture array, library capture kit, platform and sequencing parameters.

In addition, a particular feature of our analysis is that we included as controls whole-exome data from various projects. Although the data included corresponded to individuals with non-neuropsychiatric or neurological related disorders, this is not the optimal approach. It would have been preferable to use OCD specific controls. Nevertheless, we only had data for 63 control samples

recruited as controls for OCD, which would have not given any statistical power to the analysis. Therefore, we reasoned that having a mixture of samples from a variety of projects as controls would considerably increase sample size and statistical power without having much impact on false associations.

We also need to consider the RVAS methods used in this project. As described above, standard single variant association analyses are statistically underpowered to detect rare variant associations, except when sample and/or effect sizes are large, and RVAS methods can overcome this problem by testing the cumulative effects of multiple variants in a genomic region. However, these tests have still important limitations. First, gene-based tests should be optimized for a specific genetic architecture. In the case of OCD, as in other complex diseases, the genetic architecture is not known, and we solved these issue by using different RVAS methods, which accounted for different genetic architectures. However, this translated in a huge amount of data from the different tests that should be validated by a large targeted resequencing study. Second, most of the genetic variants identified through WES studies may have no discernible effect on OCD, and the inclusion of large numbers of variants with no effect in a gene-based test could reduce power. We tried to reduce this problem by selecting only exonic variants. In addition, even though some of the algorithms used can deal with different variants in the same gene having different direction of effects, power would be increased if all variants had the same direction, so in order to increase power, we separated them in missense and truncating, and performed separate analyses. Missense variants can have different effects on protein function, while LoF variants are more likely to all have the same type of effect.

Although their capacity to improve statistical power compared to single variant association analyses, RVAS methods still need large sample sizes to find significant associations. This is especially relevant in the analyses of truncating variants. Indeed, we observed that our study was underpowered to detect genome-wide significant associations especially in the case of truncating variants (Figure 15). The sample size limitation, as mentioned before, was a

known limitation of the proposed pilot study, and we intended to follow our results with a targeted resequencing analysis of the more relevant genes. This kind of two-phase approach has already been proved effective in GWAS and gene-disease association studies, as a means to improving power while minimizing the cost of the study[252–254].

Finally, we have also to take into account that association studies are highly dependent on the degree of accuracy of the data. Inclusion of false positive variants may lead to false positive associations. We experienced this issue in our first implementations of the RVAS methods and we dealt with it by adding steps and measures in the quality control and variant filtering processes. Nevertheless, it might be important to validate through Sanger sequencing all significant associations.

1.4. Future approaches

In our study we considered the different algorithms available for the analysis of rare variants associations. We compared their approaches and concluded that the concept developed by the MiST algorithm was likely the best, but its implementation could be improved. Based on this, Dr. Escaramís, in collaboration with the group of Dr. Ossowski, developed a new RVAS method, called BATI (Susak *et al.,* in revision), which is based in Bayesian inference. BATI allows the inclusion of prior knowledge about the variants (such as functionality or damaging scores) and incorporation of confounders at patient level (such as population stratification). In simulated data, BATI substantially outperformed existing methods, especially when the information about the variants contributes to the development of the disorder.

We think that it would be advantageous to reanalyse our OCD WES dataset using BATI. We will also increase the sample size by adding the whole-exome samples of the 38 OCD cases and the 33 healthy individuals from Study II, which were not available at the time we did the RVAS analyses. Moreover, we have recently detected and solved a problem in the QC step of the NimbleGen

v3 dataset that had led to the unnecessary removal of some variants, which would now be considered in the reanalysis. We hope that BATI will improve the power of our analysis, and we will still be able to detect true associations with OCD while reducing the number of false results.

Nevertheless, the results of BATI analysis would still suffer from some of the abovementioned limitations, such as false positives due to inaccurate data and limited power, and, obviously, any result would require further validation by targeted resequencing replication. The optimal capture would include a larger number of genes, including those passing FDR (if any) and at least 100 of the top candidate genes based on their DIC value. These would be assessed in a larger cohort of OCD cases and controls.

2. The analysis of the consequences of the *DRD4* 13-bp frameshift deletion needs additional functional approaches

DRD4 has been associated with several neuropsychiatric disorders[73–75] and, in particular, with OCD[77,78]. Specifically, a 48-bp VNTR has been associated with this disorder, with increased prevalence of the seven-repeat variant (DRD4*7R) in patients with OCD and tics[77] and a protective effect of the DRDR*2R variant against OCD symptoms[78].

Although *DRD4* did not show up in the list of significant genes in the RVAS analyses, we noticed a high MAF difference of a heterozygous 13-pb frameshift deletion between cases and controls. In fact, this variant had already been noticed, and some studies, with small sample sizes, had reported no OCD association of the variant in samples from German and Italian origin[80,255]. Nevertheless, a recent study of Gazzellone *et al.*[81] highlighted the presence of this deletion in one OCD patient, which prompted us to consider re-evaluating this variant. We decided to genotype this deletion in a larger cohort of OCD cases and controls and test for association, obtaining a total MAF of 0.011 in

OCD cases versus an MAF of 0.0016 in controls (OR 6.8; p-value <0.0001), which gave support to this variant as possible associated to OCD.

We compared the obtained frequencies to the reported frequencies for the *DRD4* deletion in different databases and we found that only in the CIBERER Spanish Variant Server, the MAF was similar to that of our control dataset, but higher in the rest of databases (1000G, ExAC, EVS, and gnomAD). Our first thought was that this might be a population specific variant, although this would not be supported by the 100 IBS samples in 1000G (MAF 0.009). While it is possible that the association detected was spurious, we considered that the potential functional effect of the variant merited further investigation.

Nöthen *et al*.[79] hypothesized that this variant consists of a null mutation that encodes a truncated non-functional protein, leading to a complete loss-of-function of the D4 receptor. Thus, we expected to find lower expression levels of DRD4 in OCD cases carrying this deletion compared to controls. To test this, we generated immortalized B-lymphoblastoid cell lines from carriers and non-carriers of the variant. However, western-blot and flow-cytometry analyses did not show consistent differences between deletion carriers and wild-type cell lines. The fact that we did not observe changes in DRD4 expression may indicate a compensatory effect of the intact allele or a consequence of the B-lymphoblastoid cell lines immortalization process. Analogous studies in post-mortem brain tissue of carriers and wild type individuals should help to understand whether carriers show decreased expression of DRD4 in the most relevant structure.

In parallel to the analysis of expression levels in cell lines, ZeClinics performed a *drd4* genetically modified zebrafish model to assess the potential role of *drd4* zebrafish orthologues (*drd4a* and *drd4rs*) in neural function and their potential role in OCD pathogenesis. The idea was to generate, later, mutants carrying the DRD4 deletion and study their phenotype. However, no behavioural or neurodevelopmental abnormalities were found, nor any effect in other systems, when comparing single and double homozygous larvae to the wild-type group.

Nevertheless, it has been shown that *Drd4* knockout decreases life span in mice[256] and that the *DRD4* genotype predicts longevity in humans[256]. Moreover, exploratory behaviour was reduced in mice lacking *Drd4*[257], as a consequence of increased anxiety levels. Thus, although, our present data suggest that *drd4* knockout does not affect natural locomotor behaviour, anxiety state, or defects in short memory and learning in zebrafish, *DRD4* influence in OCD may require specific tests (analogous to the ones performed in the mice knockout study[257]) and functional outcomes that have not been evaluated in the present study. It would be also interesting to perform the behavioural tests in adult zebrafish, rather than in larvae, and simulate the same design performed in the *Drd4* lacking mice study (8-12 weeks old)[257]. We should also consider that zebrafish may not be the adequate model to see the outcome of the *DRD4* 13-bp frameshift deletion found in humans.

Finally, it is also possible that in the Spanish population this variant is in LD with the DRD4*7R variant, associated with OCD. It would be interesting to assess the DRD4 48-bp VNTR genotype in the samples carrying the DRD4 deletion, both OCD and controls, and study any potential LD.

Study II: Multiomics longitudinal study of OCD

1. Implications of transcriptomic signatures in OCD patients

1.1. Differential gene expression between OCD cases and controls

To date, only one study has investigated transcriptomics profiles of OCD[140] using post-mortem brain tissue and microarrays to compare gene expression levels in various obsessive psychiatric disorders (which included OCD, obsessive-compulsive personality disorder or tics) and healthy subjects. Here, we have searched for differentially expressed genes in peripheral blood of OCD cases compared to controls, being the first study with this approach in OCD.

Twenty-eight genes showed differences in expression with a nominal p-value <0.001 and a FC >1.2 or FC <0.83 between OCD T0 and control samples. Of these, five were statistically significant after FDR correction (p-value <0.05): *NRCAM*, *AL583722.4*, *AC098935*, *KRTAP4-6*, and *HIST2H2BE.* Moreover, we found 70 genes differentially expressed in OCD T3 vs. controls and 35 in OCD T0 vs. OCD T3 (nominal p-value <0.001, FC >1.2 or FC <0.83). In general, genes overexpressed or underexpressed in OCD T0 vs. controls had a smaller fold change in OCD T3 vs. controls and did not have a significant FC in OCD T0 vs. OCD T3. The lower FC in OCD T3 samples vs. controls could be explained by a treatment effect.

Of the five significant transcripts in OCD T0 vs. controls, two, *AL583722 and* AC098935, are non-coding. *AL583722.4* is a lincRNA and AC098935 is a processed pseudogene affiliated to the antisense RNA class. Both presented relatively low levels of expression (average number of reads per sample = 9.6 and 16.9, respectively), but were detectable in most samples. *AL583722* was upregulated, whereas *AC098935* was downregulated, and these differences were maintained in the OCD T3 vs. controls analysis, although with a lower FC. The other three are coding genes, and two of them might be relevant to OCD.

NRCAM (Neuronal Cell Adhesion Molecule) encodes for a neuronal cell adhesion molecule that plays a wide variety of roles in neural development, axon growth and guidance, synapse formation, and formation of the myelinated structure[258]. *NRCAM* has been implicated in neuropsychiatric disorders including addiction-related behaviours and autism. Indeed, Ishiguro *et al*.[259,260] showed that an haplotype linked to decreased *NRCAM* expression in post-mortem brain samples was protective against addiction vulnerability for polysubstance abuse in humans. There is also a study[261] reporting association of particular haplotypes of *NRCAM*, which may relate to the expression level of NRCAM in the brain, with a subset of autism patients with severe obsessive-compulsive behaviour, but not with the full cohort of autism patients. In our cohort of OCD cases we observed an increase in the levels of NRCAM expression in the OCD cases (both OCD T0 and OCD T3), although the general levels of expression were very low (average number of reads per sample = 2.38; range of reads across samples = 0-14). 29% of the OCD cases presented at least four reads or more, in front of a 9.4% of the controls. Despite of these low numbers, the FC value of 3.99 is relevant enough to validate the expression of this gene in our dataset by real-time PCR and/or replicate it in an independent cohort of OCD cases and controls.

HIST2H2BE (Histone Cluster 2 H2B Family Member E) encodes for a core component of the nucleosome that is expressed in brain and that may play a role in transcription regulation, DNA repair, DNA replication, and chromosomal stability. Histone modifications can induce lasting and stable changes in gene expression, contributing to functional changes within cells that impact circuit level changes in brain and ultimately behavior[262]. In addition, *HIST2H2BE* has been reported significantly upregulated in schizophrenia as compared to control fibroblasts[263]. In this dataset, we observed downregulation in both OCD T0 and OCD T3 samples, although there is much variability in the expression levels (number of reads range from 0 to 188), and it is detected in only 29 samples (18/76 OCD samples and 11/32 controls). However, we have observed that the increased expression detected in controls is mostly due to two samples, which have higher number of reads than the rest of samples (188 and 88 reads, while the rest of controls samples have between 0 and 8 reads). Thus, to know if this

differential expression is not a consequence of two possible outlier samples, we should increase the sample size of the analysis or validate it by RT-PCR.

We also searched if the observed differences in expression can be associated with genetic variants. Based on WES data from the samples included in these analyses, we looked for LoF exonic variants present in the OCD cases that might lead to downregulation. However, we couldn't associate any of the significant changes in gene expression with genetic variation. This could be explained because gene expression changes can be a consequence of genetic variation in non-coding regions (e.g., transcriptional regulatory elements, non-coding RNAs), or caused by epigenetics mechanisms, such as DNA methylation or histone modifications.

We performed pathway and GO enrichment analysis with all the statistically significant genes with nominal p-value <0.01 in the DE analysis of OCD T0 vs. controls, and we found enrichment of the "axon guidance" and "Semaphorin interactions" pathways, which are related to the pathways found in the RVAS studies, suggesting a link between changes in rare variation and gene expression.

Finally, we also checked for overlap with the results from the study reported by Jaffe *et al.*[140], in which they analysed differentially expressed genes from post-mortem brain tissue in various obsessive psychiatric disorders. We compared genes with FC >1.2 or FC <0.83, and we saw 10 genes that had the same direction of expression (upregulated o downregulated) in both studies, but only three (*ARPC3*, upregulated, and *ZMAT2* and *PKD1,* downregulated) had nominal p-value <0.01 in our study. It would be of interest to check these genes in an independent and larger cohort of OCD, in order to identify their relationship with this phenotype.

ARPC3 encodes one of seven subunits of the human Arp2/3 protein complex, which is implicated in the control of actin polymerization in cells. This complex is essential at multiple stages of neural development, such as neurogenesis and

neuronal migration, and a role in axon guidance has been suggested[264]. Furthermore, *ARPC3* is required for actin polymerization in dendritic spines and Kim *et al.*[265] showed that disruption of actin dynamics in the frontal cortex of mice by knockout of *Arpc3* resulted in abnormal dendritic spine morphology leading to dysregulation of the psychomotor circuit, a phenotype related to positive symptoms of psychosis in humans. It has not been determined whether an overexpression of *ARPC3* could lead to abnormal actin polymerization in dendritic spines.

ZMAT2 (Zinc Finger Matrin-Type 2) has non neuronal function known, whereas *PKD1* (Polycystin 1, Transient Receptor Potential Channel Interacting) has been involved in synapse development[266]. Thus, a downregulation of *PKD1* may affect synapse formation in OCD patients.

1.2. Limitations and considerations of the transcriptomic analyses

RNA-Seq can overcome some limitations associated with array-based technologies in DE analyses, such as the requirement of information about the sequences being interrogated, the cross-hybridization of highly related sequences, the hybridization saturation for highly abundant genes, or the difficulty to confidently detect and quantify low-abundance species due to the analogue nature of the signal[126]. However, RNA-Seq has still some limitations, such as biases inherent in technical RNA-Seq library preparation and sequencing[205].

Most of the efforts regarding biases in RNA-Seq have been focused on the normalization of sequencing depth. However, these approaches did not usually correct all the unwanted biological or technical effects in the data, as we saw in our first attempts to normalize our dataset (Figures 24 and 25). We removed these biases (variation in blood and RNA extraction, library preparation, personnel, pooling and sequencing lane) using the RUV method, ensuring a more accurate inference of gene expression levels. Part of this variation may have been solved by extraction of blood RNA in parallel, in all samples, and

using exactly the same extraction methodology and personnel. Although we processed all RNA samples at once for library preparation, the number of samples required several experimental batches.

We did RNA-Seq with 50-bp single-end reads multiplexing 6 libraries per lane. The total reads per sample ranged from 45,486,470 to 113,250,744, with a mean of 73,546,964. Usually, the amount of sequencing required is determined by the goals of the experiment and the RNA sample nature and a lower number of reads than required can have an impact on the interpretation of the statistical analysis results. In our data, some samples and regions have especially low number of reads. So, it is not clear whether the observed gene counts are representative of the true gene counts. In fact, about 79% of the genes had an average number of reads below 10 over all samples. This can lead to false positive associations and, thus, we should perform and additional round of sequencing of the generated libraries to increase the number of reads.

It would be interesting to analyse treatment response regarding transcriptomic profile of OCD samples, comparing those OCD T0 samples that responded to the treatment (27.9% of our dataset) with the ones that did not respond, as well as to study treatment effect comparing OCD T0 samples that responded to the treatment with their paired OCD T3 samples. As it is unlikely to have enough statistical power with our sample size, we should consider adding OCD samples from patients that responded to the treatment.

2. Implications of altered microbiome in OCD patients

This is the first study on the microbiome of adulthood onset OCD. As a pilot project, we have preliminarily identified gut and oro-pharyngeal microbial taxa associated with OCD and have described a dysbiosis in the gut and oro-pharyngeal community in OCD patients, suggesting a potential role for specific microorganisms in the progression of the disorder. Nonetheless, we cannot establish a causal relationship. Changes in the microbiome could be a

consequence of OCD, and not a cause, or they could be a consequence of both bottom-up and top-down modulation of the gut-brain axis, as some animal studies suggest[156]. Thus, most of the discussion is speculative and further research is needed to replicate these findings and further interrogate the role of all these bacteria in OCD, as well as the direction of the effect.

2.1. The gut microbiome

In this study, an initial comparative analysis of gut microbiome in OCD cases (before and after treatment) with that of healthy controls was conducted via 16S rRNA sequencing. We observed that there was a trend of a reduced α-diversity and a microbiome composition shift in OCD samples, which presented an enrichment of some species and an impoverishment of others.

In the analysis of microbial diversity, OCD T0 samples showed a tendency of lower levels of all α-diversity indices measured, whereas OCD T3 group was more similar to the control group, suggesting a possible effect of the treatment on the gut microbiome. The decrease of α-diversity has also been shown in studies of PANDAS[171] and ADHD[267], which suggests that neuropsychiatric patients share changes in the microbiome in the same direction.

There is evidence that the gut microbiome may influence brain development[153], neurogenesis[154], and brain function[149] by, for example, its capability to affect levels of excitatory and inhibitory neurotransmitters by producing and/or consuming them or by modulating host neurotransmitters and/or related pathways[155], or its capability to activate the immune system via cytokine release by the mucosal immune cells[156]. Thus, it is possible that the reduced α-diversity found in the OCD T0 group could reflect an abnormal microbiome community that may lead to an anomalous gut-brain communication and deviant levels of neurotransmitters, which could be involved in OCD pathophysiology. In fact, some studies in germ free and specific pathogen-free (SPF) mice reported differences in neurotransmitters and response to stress[268–270]. Moreover, Huo *et al.*[270] suggested that imbalances of the gut-brain axis caused by

presence/absence of specific intestinal microbes could affect the neuroendocrine system in the brain, resulting in an anxiety-like behavioural phenotype. An anomalous microbial diversity could also lead to alterations in the immune system and inflammatory reaction, which could be related with the association of OCD with inflammatory markers present in the CNS and the periphery[156].

On the other hand, the brain can also modulate the gut by a top-down function of the gut-brain axis. So, it is also possible that the lower α-diversity observed could be a consequence of the stress and anxiety provoked by obsessions. Indeed, some animal studies[271,272] have demonstrated how stressful events alter abundance and composition of gut microbiome.

Regarding the specifics of microbial composition, we observed a higher percentage of the *Rikenellaceae* family (phylum *Bacteroidetes*) and a lower level of the *vadinBE97* family (phylum *Lentisphaerae*) in OCD T0 as compared to controls. In addition, LEfSe analysis revealed the *Rikenellaceae* family as a biomarker of OCD. This increase of *Rikenellaceae* in OCD T0 is consistent with the results of a recent study that identified *Rikenellaceae* as a biomarker of PANDAS[171], among others. *Rikenellaceae* is also positively associated with pro-inflammatory status in several metabolic and autoimmune diseases[273,274] and neuroinflammation in the basal ganglia as an autoimmune response to infections was proposed in a subset of OCD cases[275]. Moreover, a recent study[276] also reported OCD cases with neuroinflammation throughout the CSTC circuit of OCD. Furthermore, higher levels of *Rikenellaceae* have also been reported in ADHD[277], Alzheimer's disease[278] and major depressive disorder[279]. LEfSe analysis also revealed *Alistipes* genus (which belongs to the *Rikenellaceae* family) as a biomarker of OCD. This genus has been found as a biomarker of MDD[279]. However, a decrease in the relative abundance of its genus was reported in autism patients[280].

LEfSe analysis also found specific members of the *Firmicutes* phylum (all within the *Clostridiales* order) in higher levels in OCD T0 cases compared to controls:

Oscillibacter, Anaerostipes, and *Flavonifractor* (at genus level) and *Anaerostipes hadrus*, *Intestinimonas butyriciproducens* and *Clostridium hathewayi* (at species level). Of note, *Oscillobacter* have been related with prenatal stress[281], one of the possible environmental risk factors for OCD. The genera *Anaerostipes* and *Flavonifractor* belong to the *Lachnospiraceae* and *Ruminococcaceae* families (respectively), which have been associated with compulsion-checking behaviour[170]. Moreover, some studies suggest a relationship between these bacterial families and changes in dopamine activity, which is thought to play an important role in OCD[282–284]. *Lachnospiraceae* and *Ruminococcaceae* include butyrate-producing species, a short chain fatty acid (SCFA) that provides energy for other microbes and host cells and promotes energy expenditure, as well as facilitating fatty acid oxidation and lypolisis[170]. Interestingly, *Intestinimonas butyriciproducens,* represents another butyrate-producing bacteria belonging to an unclassified family. Jung *et al*.[170] hypothesized that, if the changes in these bacteria are consequences and not causes of OCD, the *Costridiales* members may serve to support the energy needs in some compulsive behaviour.

On the other hand, OCD T0 patients had lower levels of *Prevotellaceae* at family level (which belongs to *Bacteroidales* order and *Bacteroidetes* phylum). Decrease of *Prevotellaceae* has also been reported in ADHD[267] and Parkinson's disease[285], as well as reduction of *Prevotella* in autism[286,287], supporting the relevance of this bacterium in CNS disorders. *Prevotella* interacts with the immune system and plays a key role in degrading a broad spectrum of saccharides[288]. OCD T0 also presented a decrease of other *Clostridiales* genus (*Agathobacter*, *Coprococcus, Lachnospira, Howardella, Romboutsia, Butyricicoccus, Clostridium*), as well as reduced levels of *Negativicutes* at order level (which belongs to *Clostridia* class and *Firmicutes* phylum). Lower levels of *Coprococcus* has also been reported in autism[286].

Interestingly, there was not a differential ratio of *Firmicutes*/*Bacteroidetes* in OCD T0 cases and controls. These two groups are dominant in the human gut microbiome and the *Firmicutes* to *Bacteroidetes* ratio is regarded to be of significant relevance in human gut microbiota composition[289]. *Firmicutes* are

primarily associated with energy harvest from food, while *Bacteroidetes* are linked with degradation of complex sugars and proteins into metabolizable SCFAs[290]. Higher *Bacteroidetes*/*Firmicutes* ratios have been observed in autoimmune and psychiatric disorders, including PANDAS[171]. It is interesting to note that we did observe a higher *Bacteroidetes*/*Firmicutes* ratio in OCD cases with sexual obsessions compared to OCD cases with other type of obsessions, although it involved only six OCD cases with this symptomatology.

2.2. The oro-pharyngeal microbiome

We also performed a pilot comparative analysis of oral microbiome in OCD cases (before and after treatment) with that of healthy controls. The first goal was to identify an enrichment of *Streptococcus pyogenes* in OCD samples vs. controls, which had been associated with PANDAS[232]. However, we did not see it.

In contrast to the results from the gut microbiome, we did not observe a difference in α-diversity in the oro-pharyngeal microbiome. In the analysis of β-diversity, only one of the tests showed a difference between OCD T0 and controls. In this case it was the weighted UniFrac distance, which considers both abundance and phylogenetic relatedness of the taxa.

Interestingly, there was a significant higher *Actinobacteria*/*Fusobacteria* ratio in OCD T0 than in controls, two of the phyla more abundant in the oropharynx[159]. This was supported both by an increase of *Actinobacteria* by Wilcoxon test and by the identification of this phylum as a biomarker of OCD by LEfSe analysis. LEfSe analysis also highlighted an increase of *Actinomycetales* and a decrease of *Fusobacteriales*, all in accordance with this distorted ratio.

Qiao *et al.*[288] reported a reduced abundance of *Fusobacterium* in autism samples, although results regarding *Actinobacteria* are in the opposite direction.

This may suggest a characteristic dysbiotic signature in OCD. The biological consequences of this difference are worthy of further investigation.

Other taxa found through LEfSe analysis in higher levels in OCD T0 samples compared with controls were *Lachnospiraceae* (at family level), and *Lachnoanaerobaculum*. The increase in abundance of *Lachnospiraceae* is consistent with the results reported recently by Jung *et al.*[170], showing increased levels of this family in mice with compulsive checking.

Interestingly, OCD T0 cases with ordering compulsions presented higher percentage of *Neisseriaceae*. This association would need further support both from replication in metagenomic analysis in an independent cohort of patients with this subtype of OCD.

2.3. Limitations and considerations of the metagenomics studies

It is important to highlight that this is a pilot project and the results are preliminary. All the hypotheses performed are based in a samples size of 43 OCD patients and 33 controls, which is small. The fact that the samples sizes of other studies in microbiome and neuropsychiatric disorders are similar or even smaller is showing that this is an emerging field, almost unexplored until a few years ago.

Regarding microbial composition, we decided to compare only OCD T0 samples to controls, as OCD T3 could present treatment effect bias. To date, no information is available as to whether SSRIs affect the bacterial composition in the gut, although it is known that they possess antimicrobial activity. Therefore, we cannot exclude that SSRIs have an impact on gut microbiota. However, it would be interesting to further analyse these data comparing OCD T0 samples of patients that responded to treatment successfully and those that did not, and investigating the microbiome profile before and after treatment of the OCD patients that responded.

Some findings from this study overlapped with the observations from the microbiota profiling study of PANDAS[171] and the study performed in rats with checking compulsions[170], giving support to our results, although further validation is required. Moreover, the shared increase in specific microbial taxa between OCD and PANDAS supports the idea that these two disorders should share common etiological mechanisms, as they also share some symptomatology.

Study I and II: Discussion remarks

In this project we have applied a multiomics approach towards elucidating the aetiology of OCD. Each type of omics study has provided a list of differences associated with the disorder and the data generated, despite the limitations described above, can provide insight on the biological pathways involved in its pathophysiology. The analysis of the different omics studies has enabled us to characterize biological processes across different layers and to understand the interaction of different levels of complexity underlying the disorder. In Figure 42 we show a modified model of OCD incorporating our results and including additional biological levels where the missing heritability could be hiding.

By rare variant association in WES and targeted resequencing data we have identified *TMEM63A*, a calcium-permeable cation channel. Moreover, we have observed an overrepresentation of TRP channels enriched in rare variants in OCD cases, suggesting a potential role of calcium signalling in the aetiology of OCD.

We have also found other interesting genes related to neuronal development and function, which require further validation, such as neurotransmitter transporters, the *CTBP2* gene, suggested to regulate neuronal differentiation and to be involved in synaptic functions[249], or the *SERINC2* gene, which play important roles in neural plasticity, signalling and axonal guidance.

Transcriptomic studies have identified differential expression of genes involved in neuronal function, such as *NRCAM*, which encodes for a neuronal cell adhesion molecule. We have also replicated the upregulation of the genes *ARPC3, ZMAT2* and *PKD1*, found by Jaffe *et al.*[140].

Integration of our RVAS and transcriptomic results converge in an overrepresentation of genes belonging to semaphorin pathways, which play a

role in development of the nervous system, function and reorganization of synaptic complexes, and axonal guidance. In fact, *NRCAM*, *ARPC3*, *SERINC2* and the *TRP channels* have been implicated in axonal guidance, and we have suggested a possible role of *TMEM63A* in this process, as it was recently described as a mechanosensitive channel highly expressed in brain. Semaphorin-triggered signalling also induces the rearrangement of the actin and microtubule cytoskeleton, and we have found several OCD associated genes belonging to the "carboxyterminal post-translational modifications of tubulin" pathway. Moreover, *ARPC3*, the gene replicated from Jaffe *et al.*[140], is required for actin polymerization in dendritic spines. Interestingly, the actin-microtubule cytoskeleton system is essential for correct pathfinding. With this, we suggest an important role of axon guidance in OCD and we encourage further studies in this direction.

Finally, metagenomics studies have confirmed the increase of the *Rikenellaceae* bacterial family in the gut microbiome as a potential biomarker of OCD and have shown a specific oro-pharyngeal dysbiotic signature in OCD, characterised by a significant higher *Actinobacteria*/*Fusobacteria* ratio compared to controls. As mentioned before, the changes in the human microbiome can be a cause or a consequence of OCD, and specific studies to elucidate the microbiome-OCD relationship are needed.

Although this study represents a pilot project and has several limitations that must be considered, our results support the high complexity of OCD and actively encourage further research in these areas through multiomics approaches. We also consider that it would be interesting to study structural variation and genetic variation in non-coding regions through WGS, as well as gene-gene and gene-environment interaction.

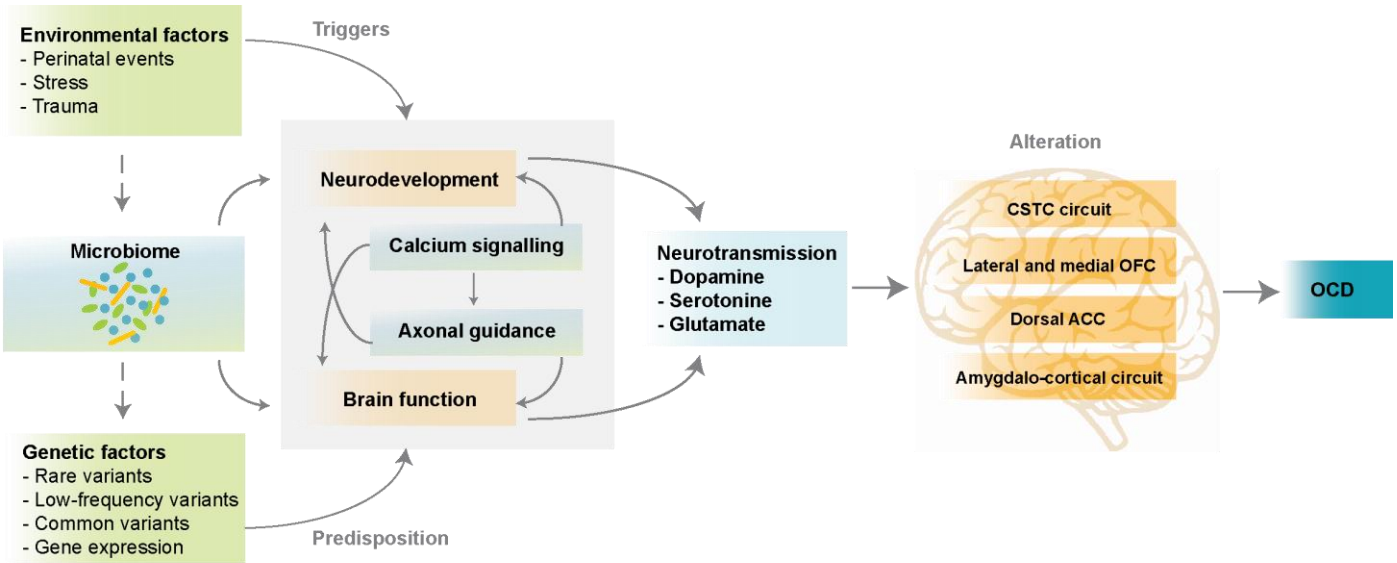


Figure 42. An integrative model of genetics, microbiome, environment and neurobiology for the expression of OCD. Modified from Pauls *et al.*[4] Individuals with OCD may have genetic predisposition to the impact of environmental factors that may trigger alterations in neurodevelopment and brain function, two processes that may be affected by dysregulation of the calcium signaling and axonal guidance pathways caused by modification of the expression of genes involved in these systems. This, in turn, may modify the expression of glutamate-, serotonin- and dopamine-system-related genes. Neuroanatomical expression of these modifications may result in an alteration of the brain circuits involved in OCD, leading to the OCD symptomatology. The human microbiome, which can be modified by environmental factors, may influence brain development and function by itself or modifying the expression of specific host genes.

CONCLUSIONS

1. Selection of genetic variants well covered by all exome capture kits is an essential step for downstream analyses, as it allows removing PCA stratification due to library capture kit batch effect.

2. We have identified *TMEM63A* as a novel OCD candidate gene by analysis of rare variants through WES and targeted resequencing.

3. We have observed an overrepresentation of calcium related genes enriched in rare variants in OCD cases, such as *TMEM63A* and TRP channels, suggesting a potential role of calcium signalling in the aetiology of OCD.

4. The analyses performed in B-lymphoblastoid cell lines and zebrafish suggest that these model systems are not adequate to validate the functional consequences of the *DRD4* 13-bp frameshift deletion. The influence of the *DRD4* deletion in OCD may require specific tests and functional outcomes that have not been evaluated in the present study.

5. We have observed some common and low-frequency variants associated to OCD with genome-wide significance, which do not overlap with previous reported associated regions. However, validation and replication studies are needed to confirm the relationship of these variants with OCD.

6. We have identified differential expression of genes involved in neuronal function in OCD patients, such as *NRCAM*, a neuronal cell adhesion molecule. However, these results require validation and/or replication, since the average number of reads obtained is very low.

7. RVAS and DE analyses results converge in an overrepresentation of OCD associated genes belonging to semaphorin pathways, which are relevant in neuronal development and function, and suggests an important role of axon guidance in OCD.

8. We have noticed a trend towards a decrease of α-diversity in the gut microbiome of OCD patients, and validated the previously reported increase of the *Rikenellaceae* family in OCD individuals as a potential OCD biomarker.

9. We have identified a significant higher *Actinobacteria*/*Fusobacteria* ratio in the oro-pharyngeal microbiome in OCD patients as compared to controls, suggesting a characteristic oro-pharyngeal dysbiotic signature in OCD.

BIBLIOGRAPHY

1. Osborn, I. *Tormenting Thoughts and Secret Rituals: the hidden epidemic of obsessive-compulsive disorder*. (Dell Publishing, 1998).
2. Berrios, G. E. Obsessive-compulsive disorder: its conceptual history in France during the 19th century. *Compr. Psychiatry* **30,** 283–95 (1989).
3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. (Washington, DC, 1994).
4. Pauls, D. L., Abramovitch, A., Rauch, S. L. & Geller, D. A. Obsessive–compulsive disorder: an integrative genetic and neurobiological perspective. *Nat. Rev. Neurosci.* **15,** 410–424 (2014).
5. Mataix-Cols, D., do Rosario-Campos, M. C. & Leckman, J. F. A Multidimensional Model of Obsessive-Compulsive Disorder. *Am. J. Psychiatry* **162,** 228–238 (2005).
6. Mataix-Cols, D. *et al.* Distinct Neural Correlates of Washing, Checking, and Hoarding Symptom Dimensions in Obsessive-compulsive Disorder. *Arch Gen Psychiatry* **61,** 564–576 (2004).
7. Bloch, M. H., Landeros-Weisenberger, A., Rosario, M. C., Pittenger, C. & Leckman, J. F. Meta-Analysis of the Symptom Structure of Obsessive-Compulsive Disorder. *Am. J. Psychiatry* **165,** 1532–1542 (2008).
8. Goodman, W. K. *et al.* The Yale-Brown Obsessive Compulsive Scale. I. Development, Use, and Reliability. *Arch. Gen. Psychiatry* **46,** 1006 (1989).
9. Goodman, W. K. *et al.* The Yale-Brown Obsessive Compulsive Scale. II. Validity. *Arch. Gen. Psychiatry* **46,** 1012 (1989).
10. Kessler, R. C. *et al.* Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62,** 593 (2005).
11. Zohar, J., Greenberg, B. & Denys, D. Obsessive-compulsive disorder. *Handb. Clin. Neurol.* **106,** 375–390 (2012).

12. Burke, K. C., Burke, J. D., Regier, D. A. & Rae, D. S. Age at Onset of Selected Mental Disorders in Five Community Populations. *Arch. Gen. Psychiatry* **47,** 511 (1990).
13. Weissman, M. M. *et al.* The cross national epidemiology of obsessive compulsive disorder. *J. Clin. Psychiatry* **55 Suppl,** 5–10 (1994).
14. Geller, D. *et al.* Is juvenile obsessive-compulsive disorder a developmental subtype of the disorder? A review of the pediatric literature. *J. Am. Acad. Child Adolesc. Psychiatry* **37,** 420–7 (1998).
15. Nakatani, E. *et al.* Children with very early onset obsessive-compulsive disorder: clinical features and treatment outcome. *J. Child Psychol. Psychiatry* **52,** 1261–1268 (2011).
16. Chabane, N. *et al.* Early-onset obsessive-compulsive disorder: a subgroup with a specific clinical and familial pattern? *J. Child Psychol. Psychiatry* **46,** 881–887 (2005).
17. Ruscio, A., Stein, D., Chiu, W. & Kessler, R. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Mol. Psychiatry* **15,** 53–63 (2008).
18. Karno, M., Golding, J. M., Sorenson, S. B. & Burnam, M. A. The epidemiology of obsessive-compulsive disorder in five US communities. *Arch. Gen. Psychiatry* **45,** 1094–9 (1988).
19. Castle, D. J., Deale, A. & Marks, I. M. Gender Differences in Obsessive Compulsive Disorder. *Aust. New Zeal. J. Psychiatry* **29,** 114–117 (1995).
20. Noshirvani, H. F., Kasvikis, Y., Marks, I. M., Tsakiris, F. & Monteiro, W. O. Gender-divergent aetiological factors in obsessive-compulsive disorder. *Br. J. Psychiatry* **158,** 260–3 (1991).
21. Geller, D. A. *et al.* Perinatal Factors Affecting Expression of Obsessive Compulsive Disorder in Children and Adolescents. *J. Child Adolesc. Psychopharmacol.* **18,** 373–379 (2008).
22. Murphy, T. K., Kurlan, R. & Leckman, J. The immunobiology of Tourette's disorder, pediatric autoimmune neuropsychiatric disorders associated with Streptococcus, and related disorders: a way forward. *J. Child Adolesc. Psychopharmacol.* **20,** 317–31 (2010).

23. Brander, G., Pérez-Vigil, A., Larsson, H. & Mataix-Cols, D. Systematic review of environmental risk factors for Obsessive-Compulsive Disorder: A proposed roadmap from association to causation. *Neurosci. Biobehav. Rev.* **65,** 36–62 (2016).
24. Swedo, S. E. *et al.* Pediatric Autoimmune Neuropsychiatric Disorders Associated With Streptococcal Infections: Clinical Description of the First 50 Cases. *Am J PsychiatryAm J Psychiatry* **155,** 264–271 (1998).
25. Eisen, J. L. *et al.* Impact of obsessive-compulsive disorder on quality of life. *Compr. Psychiatry* **47,** 270–5 (2006).
26. Kamath, P., Reddy, Y. C. J. & Kandavel, T. *Suicidal Behavior in Obsessive-Compulsive Disorder*. *The Journal of Clinical Psychiatry* **68,** ([Physicians Postgraduate Press], 2007).
27. Skoog, G. & Skoog, I. A 40-year follow-up of patients with obsessive-compulsive disorder [see commetns]. *Arch. Gen. Psychiatry* **56,** 121–7 (1999).
28. Rotge, J.-Y. *et al.* Meta-Analysis of Brain Volume Changes in Obsessive-Compulsive Disorder. *Biol. Psychiatry* **65,** 75–83 (2009).
29. Menzies, L. *et al.* Integrating evidence from neuroimaging and neuropsychological studies of obsessive-compulsive disorder: The orbitofronto-striatal model revisited. *Neurosci. Biobehav. Rev.* **32,** 525 (2008).
30. Kwon, J. S., Jang, J. H., Choi, J.-S. & Kang, D.-H. Neuroimaging in obsessive–compulsive disorder. *Expert Rev. Neurother.* **9,** 255–269 (2009).
31. Koch, K. *et al.* White matter structure and symptom dimensions in obsessive-compulsive disorder. *J. Psychiatr. Res.* **46,** 264–70 (2012).
32. van den Heuvel, O. A. *et al.* The major symptom dimensions of obsessive-compulsive disorder are mediated by partially distinct neural systems. *Brain* **132,** 853–868 (2008).
33. Fitzgerald, K. D. *et al.* Developmental alterations of frontal-striatal-thalamic connectivity in obsessive-compulsive disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **50,** 938–948.e3 (2011).

34. Pittenger, C. *Obsessive-compulsive Disorder: Phenomenology, Pathophysiology, and Treatment*. (New York, NY: Oxford University Press, [2017], 1972).
35. Harrison, B. J. *et al.* Altered Corticostriatal Functional Connectivity in Obsessive-compulsive Disorder. *Arch. Gen. Psychiatry* **66,** 1189 (2009).
36. An, S. K. *et al.* To discard or not to discard: the neural basis of hoarding symptoms in obsessive-compulsive disorder. *Mol. Psychiatry* **14,** 318–331 (2009).
37. Gilbert, A. R. *et al.* Neural correlates of symptom dimensions in pediatric obsessive-compulsive disorder: a functional magnetic resonance imaging study. *J. Am. Acad. Child Adolesc. Psychiatry* **48,** 936–44 (2009).
38. Abramovitch, A., Abramowitz, J. S. & Mittelman, A. The neuropsychology of adult obsessive–compulsive disorder: A meta-analysis. *Clin. Psychol. Rev.* **33,** 1163–1171 (2013).
39. Chamberlain, S. R., Blackwell, A. D., Fineberg, N. A., Robbins, T. W. & Sahakian, B. J. The neuropsychology of obsessive compulsive disorder: the importance of failures in cognitive and behavioural inhibition as candidate endophenotypic markers. *Neurosci. Biobehav. Rev.* **29,** 399–419 (2005).
40. Hashimoto, N. *et al.* Distinct neuropsychological profiles of three major symptom dimensions in obsessive-compulsive disorder. *Psychiatry Res.* **187,** 166–73 (2011).
41. Abramovitch, A., Dar, R., Schweiger, A. & Hermesh, H. Neuropsychological Impairments and Their Association with Obsessive-Compulsive Symptom Severity in Obsessive-Compulsive Disorder. *Arch. Clin. Neuropsychol.* **26,** 364–376 (2011).
42. Katrin Kuelz, A. *et al.* Neuropsychological Impairment in Obsessive-Compulsive Disorder—Improvement Over the Course of Cognitive Behavioral Treatment. *J. Clin. Exp. Neuropsychol.* **28,** 1273–1287 (2006).
43. D'alcante, C. C. *et al.* Neuropsychological predictors of response to randomized treatment in obsessive–compulsive disorder. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **39,** 310–317 (2012).

44. Fineberg, N. A. *et al.* Obsessive–compulsive disorder (OCD): Practical strategies for pharmacological and somatic treatment in adults. *Psychiatry Res.* **227,** 114–125 (2015).
45. Koran, L. M. *et al.* Practice guideline for the treatment of patients with obsessive-compulsive disorder. *Am. J. Psychiatry* **164,** 5–53 (2007).
46. Bandelow, B. *et al.* World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for the Pharmacological Treatment of Anxiety, Obsessive-Compulsive and Post-Traumatic Stress Disorders. *Austria Dan J. Stein (South Africa) World J. Biol. Psychiatry* **9,** 248–312 (2008).
47. Soomro, G. M., Altman, D. G., Rajagopal, S. & Oakley Browne, M. Selective serotonin re-uptake inhibitors (SSRIs) versus placebo for obsessive compulsive disorder (OCD). *Cochrane Database Syst. Rev.* CD001765 (2008).
48. Fineberg, N. A. *et al.* Obsessive–compulsive disorder (OCD): Practical strategies for pharmacological and somatic treatment in adults. *Psychiatry Res.* **227,** 114–125 (2015).
49. Saxena, S. & Rauch, S. L. Functional neuroimaging and the neuroanatomy of obsessive-compulsive disorder. *Psychiatr. Clin. North Am.* **23,** 563–586 (2000).
50. Milad, M. R. & Rauch, S. L. Obsessive-compulsive disorder: beyond segregated cortico-striatal pathways. *Trends Cogn. Sci.* **16,** 43–51 (2012).
51. Fu, W., O'Connor, T. D. & Akey, J. M. Genetic architecture of quantitative traits and complex diseases. *Curr. Opin. Genet. Dev.* **23,** 678–683 (2013).
52. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **17,** 782–790 (2014).
53. Pauls, D. L. The genetics of obsessive-compulsive disorder: a review. *Dialogues Clin. Neurosci.* **12,** 149–63 (2010).
54. Rosenberg, C. M. Familial aspects of obsessional neurosis. *Br. J. Psychiatry* **113,** 405–13 (1967).

55. McKeon, P. & Murray, R. Familial aspects of obsessive-compulsive neurosis. *Br. J. Psychiatry* **151,** 528–34 (1987).
56. Stewart, S. E. & Pauls, D. L. The Genetics of Obsessive-Compulsive Disorder. *Focus (Madison).* **8,** 350–357 (2010).
57. Taylor, S. Etiology of obsessions and compulsions: A meta-analysis and narrative review of twin studies. *Clin. Psychol. Rev.* **31,** 1361–1372 (2011).
58. Hanna, G. L. *et al.* Genome-wide linkage analysis of families with obsessive-compulsive disorder ascertained through pediatric probands. *Am. J. Med. Genet.* **114,** 541–52 (2002).
59. Shugart, Y. Y. *et al.* Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q. *Mol. Psychiatry* **11,** 763–70 (2006).
60. Hanna, G. L. *et al.* Evidence for a susceptibility locus on chromosome 10p15 in early-onset obsessive-compulsive disorder. *Biol. Psychiatry* **62,** 856–62 (2007).
61. Ross, J. *et al.* Genomewide Linkage Analysis in Costa Rican Families Implicates Chromosome 15q14 as a Candidate Region for OCD HHS Public Access. *Hum Genet. Hum Genet* **130,** 795–805 (2011).
62. Mathews, C. A. *et al.* Genomewide Linkage Analysis of Obsessive Compulsive Disorder Implicates Chromosome 1p36. *Biol Psychiatry* **72,** 629–36 (2012).
63. Samuels, J. *et al.* Significant Linkage to Compulsive Hoarding on Chromosome 14 in Families With Obsessive-Compulsive Disorder: Results From the OCD Collaborative Genetics Study. *Am J Psychiatry* **164,** 493–9 (2007).
64. Davis, L. K. *et al.* Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet.* **9,** e1003864 (2013).
65. Willour, V. L. *et al.* Replication Study Supports Evidence for Linkage to 9p24 in Obsessive- Compulsive Disorder. *Am. J. Hum. Genet* **75,** 508–513 (2004).

66. Taylor, S. Molecular genetics of obsessive–compulsive disorder: a comprehensive meta-analysis of genetic association studies. *Mol. Psychiatry* **18,** 799–805 (2013).
67. Sinopoli, V. M., Burton, C. L., Kronenberg, S. & Arnold, P. D. A review of the role of serotonin system genes in obsessive-compulsive disorder. *Neurosci. Biobehav. Rev.* **80,** 372–381 (2017).
68. Grünblatt, E. *et al.* Combining genetic and epigenetic parameters of the serotonin transporter gene in obsessive-compulsive disorder. *J. Psychiatr. Res.* **96,** 209–217 (2018).
69. Stewart, S. E. *et al.* Association of the SLC1A1 glutamate transporter gene and obsessive-compulsive disorder. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **144,** 1027–1033 (2007).
70. Wu, K. *et al.* The Role of Glutamate Signalling in the Pathogenesis and Treatment of Obsessive-Compulsive Disorder. *Pharmacol Biochem Behav* **100,** 726–735 (2012).
71. McDougle, C. J., Goodman, W. K. & Price, L. H. Dopamine antagonists in tic-related and psychotic spectrum obsessive compulsive disorder. *J. Clin. Psychiatry* **55 Suppl,** 24–31 (1994).
72. Goodman, W. K. *et al.* Beyond the serotonin hypothesis: a role for dopamine in some forms of obsessive compulsive disorder? *J. Clin. Psychiatry* **51 Suppl,** 36-43; discussion 55–8 (1990).
73. Aguirre-Samudio, A. J. & Nicolini, H. DRD4 polymorphism and the association with mental disorders. *Rev. Invest. Clin.* **57,** 65–75
74. Seeman, P., Guan, H.-C. & Van Tol, H. H. M. Dopamine D4 receptors elevated in schizophrenia. *Nature* **365,** 441–445 (1993).
75. Van Tol, H. H. M. *et al.* Cloning of the gene for a human dopamine D4 receptor with high affinity for the antipsychotic clozapine. *Nature* **350,** 610–614 (1991).
76. Tol, H. H. M. Van *et al.* Multiple dopamine D4 receptor variants in the human population. *Nature* **358,** 149–152 (1992).
77. Cruz, C. *et al.* Increased prevalence of the seven-repeat variant of the dopamine D4 receptor gene in patients with obsessive-compulsive disorder with tics. *Neurosci. Lett.* **231,** 1–4 (1997).

78. Millet, B. *et al.* Association between the dopamine receptor D4 (DRD4) gene and obsessive-compulsive disorder. *Am. J. Med. Genet.* **116B,** 55–59 (2003).
79. Nöthen, M. M. *et al.* Human dopamine D4 receptor gene: frequent occurrence of a null allele and observation of homozygosity. *Hum. Mol. Genet.* **3,** 2207–2212 (1994).
80. Di Bella, D., Catalano, M., Cichon, S. & Nöthen, M. M. Association study of a null mutation in the dopamine D4 receptor gene in Italian patients with obsessive-compulsive disorder, bipolar mood disorder and schizophrenia. *Psychiatr. Genet.* **6,** 119–21 (1996).
81. Gazzellone, M. J. *et al.* Uncovering obsessive-compulsive disorder risk genes in a pediatric cohort by high-resolution analysis of copy number variation. *J. Neurodev. Disord.* **8,** 36 (2016).
82. Cichon, S. *et al.* Identification of two novel polymorphisms and a rare deletion variant in the human dopamine D4 receptor gene. *Psychiatr. Genet.* **5,** 97–103 (1995).
83. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
84. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
85. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19,** 110–124 (2018).
86. Blanco-Gómez, A. *et al.* Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes. *BioEssays* **38,** 664–673 (2016).
87. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10,** 241–251 (2009).
88. GWAS Catalog. Available at: http://www.ebi.ac.uk/gwas/home. (Accessed: 5th July 2018)
89. Stewart, S. E. *et al.* Genome-wide association study of obsessive-compulsive disorder. *Mol. Psychiatry* **18,** 788–98 (2013).

90. Mattheisen, M. *et al.* Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Mol. Psychiatry* **20,** 337–344 (2015).
91. International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS), P. D. *et al.* Revealing the complex genetic architecture of obsessive–compulsive disorder using meta-analysis. *Mol. Psychiatry* **23,** 1181–1188 (2018).
92. Alonso, P., López-Solà, C., Real, E., Segalàs, C. & Menchón, J. M. Animal models of obsessive-compulsive disorder: utility and limitations. *Neuropsychiatr. Dis. Treat.* **11,** 1939–55 (2015).
93. Moya, P. *et al.* Transgenic Mouse Overexpressing EAAT3 (Neuronal Glutamate Transporter): A Novel Genetic Model of Obsessive-Compulsive Disorder. *Eur. Neuropsychopharmacol.* **27,** S425–S426 (2017).
94. Albelda, N. & Joel, D. Current animal models of obsessive compulsive disorder: an update. *Neuroscience* **211,** 83–106 (2012).
95. Ting, J. T. & Feng, G. Neurobiology of obsessive-compulsive disorder: insights into neural circuitry dysfunction through mouse genetics. *Curr. Opin. Neurobiol.* **21,** 842–8 (2011).
96. Tang, R. *et al.* Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. *Genome Biol.* **15,** R25 (2014).
97. D'Amico, D., Estivill, X. & Terriente, J. Switching to zebrafish neurobehavioral models: The obsessive–compulsive disorder paradigm. *Eur. J. Pharmacol.* **759,** 142–150 (2015).
98. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–266 (2008).
99. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
100. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–45 (2011).

101. Levy, R. J., Xu, B., Gogos, J. A. & Karayiorgou, M. in *Methods in molecular biology (Clifton, N.J.)* **838,** 97–113 (2012).
102. LM, M. Copy Number Variation in Obsessive-Compulsive Disorder and Tourette Syndrome: A Cross-Disorder Study. *J Am Acad Child Adolesc Psychiatry* **53,** 910–919 (2014).
103. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6,** S13–S20 (2009).
104. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15,** 722–733 (2014).
105. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1193–8 (2012).
106. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6,** 287–298 (2005).
107. The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI). Available at: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/. (Accessed: 30th June 2018)
108. Stylianos E. Antonarakis; Michael Krawczak; David N. Cooper. *The Nature and Mechanisms of Human Gene Mutation*. (McGraw-Hill, New York, 2002).
109. Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
110. Scacheri, C. A. & Scacheri, P. C. Mutations in the noncoding genome. *Curr. Opin. Pediatr.* **27,** 659–64 (2015).
111. Khurana, E. *et al.* Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science (80-.).* **342,** 1235587 (2013).
112. Cappi, C. *et al.* Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways. *Transl. Psychiatry* **6,** e764 (2016).
113. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7,** 16 (2015).

114. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).
115. Liu, D. J. & Leal, S. M. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet.* **6,** e1001156 (2010).
116. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
117. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–775 (2012).
118. Sun, J., Zheng, Y. & Hsu, L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genet. Epidemiol.* **37,** 334–344 (2013).
119. Moutsianas, L. *et al.* The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLOS Genet.* **11,** e1005165 (2015).
120. Costa, V., Aprile, M., Esposito, R. & Ciccodicola, A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.* **21,** 134–42 (2013).
121. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–5 (2012).
122. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
123. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
124. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9,** eaal5209 (2017).
125. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8,** 15824 (2017).
126. Shendure, J. The beginning of the end for microarrays? *Nat. Methods* **5,** 585–587 (2008).

127. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* **12,** e0190152 (2017).
128. Hardcastle, T. J. & Kelly, K. A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11,** 422 (2010).
129. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).
130. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
131. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
132. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
133. Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **43,** gkv711 (2015).
134. Li, J. & Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **22,** 519–536 (2013).
135. Wu, C., Bendriem, R. M., Garamszegi, S. P., Song, L. & Lee, C.-T. RNA sequencing in post-mortem human brains of neuropsychiatric disorders. *Psychiatry Clin. Neurosci.* **71,** 663–672 (2017).
136. Breen, M. S., Stein, D. J. & Baldwin, D. S. Systematic review of blood transcriptome profiling in neuropsychiatric disorders: guidelines for biomarker discovery. *Hum. Psychopharmacol. Clin. Exp.* **31,** 373–381 (2016).
137. Xu, Y. *et al.* Altered expression of mRNA profiles in blood of early-onset schizophrenia. *Sci. Rep.* **6,** 16767 (2016).

138. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5,** 5748 (2014).
139. Tylee, D. S., Kawaguchi, D. M. & Glatt, S. J. On the outside, looking in: A review and evaluation of the comparability of blood and brain "-omes". *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **162,** 595–603 (2013).
140. Jaffe, A. E. *et al.* Genetic neuropathology of obsessive psychiatric syndromes. *Transl. Psychiatry* **4,** e432–e432 (2014).
141. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol.* **14,** e1002533 (2016).
142. Morgan, X. C., Segata, N. & Huttenhower, C. Biodiversity and functional genomics in the human microbiome. *Trends Genet.* **29,** 51–58 (2013).
143. Sandoval-Motta, S., Aldana, M., Martínez-Romero, E. & Frank, A. The Human Microbiome and the Missing Heritability Problem. *Front. Genet.* **8,** 80 (2017).
144. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148,** 1258–70 (2012).
145. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
146. Rosselli, R. *et al.* Direct 16S rRNA-seq from bacterial communities: a PCR-independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon. *Sci. Rep.* **6,** 32165 (2016).
147. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469,** 967–977 (2016).
148. Sudo, N. *et al.* Postnatal microbial colonization programs the hypothalamic-pituitary-adrenal system for stress response in mice. *J. Physiol.* **558,** 263–75 (2004).
149. Collins, S. M., Surette, M. & Bercik, P. The interplay between the intestinal microbiota and the brain. *Nat. Rev. Microbiol.* **10,** 735–742 (2012).

150. Mawdsley, J. E. & Rampton, D. S. Psychological stress in IBD: new insights into pathogenic and therapeutic implications. *Gut* **54,** 1481–1491 (2005).
151. DuPont, A. W. & DuPont, H. L. The intestinal microbiota and chronic disorders of the gut. *Nat. Rev. Gastroenterol. Hepatol.* **8,** 523–531 (2011).
152. Collins, S. M. & Bercik, P. The Relationship Between Intestinal Microbiota and the Central Nervous System in Normal Gastrointestinal Function and Disease. *Gastroenterology* **136,** 2003–2014 (2009).
153. Heijtz, R. D. *et al.* Normal gut microbiota modulates brain development and behavior. *Proc. Natl. Acad. Sci.* **108,** 3047–3052 (2011).
154. Ogbonnaya, E. S. *et al.* Adult Hippocampal Neurogenesis Is Regulated by the Microbiome. *Biol. Psychiatry* **78,** e7–e9 (2015).
155. Strandwitz, P. Neurotransmitter modulation by the gut microbiota. *Brain Res.* **1693,** 128–133 (2018).
156. Turna, J., Grosman Kaplan, K., Anglin, R. & Van Ameringen, M. "What's bugging the gut in OCD?" A review of the gut microbiome in obsessive-compulsive disorder. *Depress. Anxiety* **33,** 171–178 (2016).
157. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449,** 804–810 (2007).
158. Clapp, M. *et al.* Gut microbiota's effect on mental health: The gut-brain axis. *Clin. Pract.* **7,** 987 (2017).
159. Segata, N. *et al.* Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13,** R42 (2012).
160. Neufeld, K.-A. M., Kang, N., Bienenstock, J. & Foster, J. A. Effects of intestinal microbiota on anxiety-like behavior. *Commun. Integr. Biol.* **4,** 492–4 (2011).
161. Neufeld, K. M., Kang, N., Bienenstock, J. & Foster, J. A. Reduced anxiety-like behavior and central neurochemical change in germ-free mice. *Neurogastroenterol. Motil.* **23,** 255-e119 (2011).

162. Castro-Nallar, E. *et al.* Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ* **3,** e1140 (2015).
163. Yolken, R. H. *et al.* Metagenomic Sequencing Indicates That the Oropharyngeal Phageome of Individuals With Schizophrenia Differs From That of Controls. *Schizophr. Bull.* **41,** 1153–61 (2015).
164. Naseribafrouei, A. *et al.* Correlation between the human fecal microbiota and depression. *Neurogastroenterol. Motil.* **26,** 1155–1162 (2014).
165. Wang, L. *et al.* Low relative abundances of the mucolytic bacterium Akkermansia muciniphila and Bifidobacterium spp. in feces of children with autism. *Appl. Environ. Microbiol.* **77,** 6718–21 (2011).
166. Steenbergen, L., Sellaro, R., van Hemert, S., Bosch, J. A. & Colzato, L. S. A randomized controlled trial to test the effect of multispecies probiotics on cognitive reactivity to sad mood. *Brain. Behav. Immun.* **48,** 258–264 (2015).
167. Pärtty, A., Kalliomäki, M., Wacklin, P., Salminen, S. & Isolauri, E. A possible link between early probiotic intervention and the risk of neuropsychiatric disorders later in childhood: a randomized trial. *Pediatr. Res.* **77,** 823–828 (2015).
168. Kantak, P. A., Bobrow, D. N. & Nyby, J. G. Obsessive–compulsive-like behaviors in house mice are attenuated by a probiotic (Lactobacillus rhamnosus GG). *Behav. Pharmacol.* **25,** 71–79 (2014).
169. Messaoudi, M. *et al.* Beneficial psychological effects of a probiotic formulation (Lactobacillus helveticus R0052 and Bifidobacterium longum R0175) in healthy human volunteers. *Gut Microbes* **2,** 256–261 (2011).
170. Jung, T. D. *et al.* Changes in gut microbiota during development of compulsive checking and locomotor sensitization induced by chronic treatment with the dopamine agonist quinpirole. *Behav. Pharmacol.* **29,** 1 (2017).
171. Quagliariello, A. *et al.* Gut Microbiota Profiling and Gut–Brain Crosstalk in Children Affected by Pediatric Acute-Onset Neuropsychiatric Syndrome and Pediatric Autoimmune Neuropsychiatric Disorders Associated With Streptococcal Infections. *Front. Microbiol.* **9,** 675 (2018).

172. First, M. B., Spitzer, R. L., Miriam, G. & Williams, J. B. W. Structured Clinical Interview for the DSM-IV Axis I Disorders - Clinical Version. (1996).
173. López-Pina, J. A. *et al.* The Yale–Brown Obsessive Compulsive Scale. *Assessment* **22,** 619–628 (2015).
174. Vega-Dienstmaier, J. M. *et al.* [Validation of a version in Spanish of the Yale-Brown Obsessive-Compulsive Scale]. *Actas Esp. Psiquiatr.* **30,** 30–5
175. Kaufman, J. *et al.* Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *J. Am. Acad. Child Adolesc. Psychiatry* **36,** 980–988 (1997).
176. Scahill, L. *et al.* Children's Yale-Brown Obsessive Compulsive Scale: Reliability and Validity. *J. Am. Acad. Child Adolesc. Psychiatry* **36,** 844–852 (1997).
177. MCC Spain – Multi-Caso Control Spain. Available at: http://www.mccspain.org/?q=es. (Accessed: 30th July 2018)
178. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
179. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).
180. Picard. Available at: http://picard.sourceforge.net/. (Accessed: 12th August 2013)
181. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
182. GitHub - mbosio85/ediva: eDiVA : pipeline to process WES and targeted sequencing data. Fully portable with docker and available at http://www.ediva.crg.eu/. Available at: https://github.com/mbosio85/ediva. (Accessed: 10th September 2018)
183. 1000 Genomes. Available at: http://www.1000genomes.org/. (Accessed: 12th August 2013)

184. Exome Variant Server. Available at: http://evs.gs.washington.edu/EVS/. (Accessed: 12th August 2013)
185. ExAC Browser. Available at: http://exac.broadinstitute.org/. (Accessed: 4th August 2018)
186. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
187. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–775 (2012).
188. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **64,** 583–639 (2002).
189. Fuentes Fajardo, K. V. *et al.* Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33,** 609–613 (2012).
190. Primer3. Available at: http://biotools.umassmed.edu/bioapps/primer3_www.cgi. (Accessed: 12th August 2013)
191. Human BLAT Search. Available at: http://genome.ucsc.edu/cgi-bin/hgBlat?command=start. (Accessed: 20th August 2013)
192. Chen, Y. *et al.* In silico gene prioritization by integrating multiple data sources. *PLoS One* **6,** e21137 (2011).
193. Breathe. Available at: https://breathe.isglobal.org/. (Accessed: 1st August 2018)
194. Gonzalez, J. R. *et al.* SNPassoc: an R package to perform whole genome association studies. *Bioinformatics* **23,** 654–655 (2007).
195. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41,** D793–D800 (2013).
196. Tosato, G., Pike, S. E., Koski, I. R. & Blaese, R. M. Selective inhibition of immunoregulatory cell functions by cyclosporin A. *J. Immunol.* **128,** 1986–91 (1982).
197. Schindelin, J., Rueden, C. T., Hiner, M. C. & Eliceiri, K. W. The ImageJ ecosystem: An open platform for biomedical image analysis. *Mol. Reprod. Dev.* **82,** 518–529 (2015).

198. Knowles, D. G., Roder, M., Merkel, A. & Guigo, R. Grape RNA-Seq analysis pipeline environment. *Bioinformatics* **29,** 614–621 (2013).
199. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
200. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
201. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26,** 2204–7 (2010).
202. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
203. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
204. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
205. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32,** 896–902 (2014).
206. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13,** 581–3 (2016).
207. Glöckner, F. O. *et al.* 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J. Biotechnol.* **261,** 169–176 (2017).
208. Callahan, B. Silva taxonomic training data formatted for DADA2 (Silva version 132). (2018).
209. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8,** e61217 (2013).
210. Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26,** 1463–4 (2010).

211. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, E. S. and H. W. CRAN - Package vegan. (2018).
212. Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).
213. Zeileis, A., Meyer, D. & Hornik, K. Residual-Based Shadings for Visualizing (Conditional) Independence. *J. Comput. Graph. Stat.* **16,** 507–525 (2007).
214. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12,** R60 (2011).
215. GTEx Portal. Available at: https://gtexportal.org/home/. (Accessed: 17th September 2018)
216. Behesti, H. *et al.* ASTN2 modulates synaptic strength by trafficking and degradation of surface proteins. *Proc. Natl. Acad. Sci. U. S. A.* 201809382 (2018). doi:10.1073/pnas.1809382115
217. Sugathan, A. *et al. CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc. Natl. Acad. Sci.* **111,** E4468–E4477 (2014).
218. Moreno-De-Luca, D. *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am. J. Hum. Genet.* **87,** 618–30 (2010).
219. Noor, A. *et al.* Copy number variant study of bipolar disorder in Canadian and UK populations implicates synaptic genes. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **165,** 303–313 (2014).
220. Matoso, E. *et al.* Insertional translocation leading to a 4q13 duplication including the *EPHA5* gene in two siblings with attention-deficit hyperactivity disorder. *Am. J. Med. Genet. Part A* **161,** 1923–1928 (2013).
221. Chang, L.-C. *et al.* A Conserved BDNF, Glutamate- and GABA-Enriched Gene Module Related to Human Depression Identified by Coexpression Meta-Analysis and DNA Variant Genome-Wide Association Studies. *PLoS One* **9,** e90980 (2014).

222. Schulte, E. C. *et al.* Rare variants in LRRK1 and Parkinson's disease. *Neurogenetics* **15,** 49–57 (2014).
223. Suda, S. *et al.* Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism. *Mol. Autism* **2,** 14 (2011).
224. Jun, G. *et al. PLXNA 4* is associated with Alzheimer disease and modulates tau phosphorylation. *Ann. Neurol.* **76,** 379–392 (2014).
225. Schulte, E. C. *et al.* Rare Variants in PLXNA4 and Parkinson's Disease. *PLoS One* **8,** e79145 (2013).
226. Kumar, R. *et al.* Homozygous mutation of STXBP5L explains an autosomal recessive infantile-onset neurodegenerative disorder. *Hum. Mol. Genet.* **24,** 2000–2010 (2015).
227. Costas, J. *et al.* Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. *Transl. Psychiatry* **6,** e768–e768 (2016).
228. Zuo, L. *et al.* Rare SERINC2 variants are specific for alcohol dependence in individuals of European descent. *Pharmacogenet. Genomics* **23,** 395–402 (2013).
229. Hnoonual, A. *et al.* Chromosomal microarray analysis in a cohort of underrepresented population identifies SERINC2 as a novel candidate gene for autism spectrum disorder. *Sci. Rep.* **7,** 12096 (2017).
230. Castellani, C. A. *et al.* Post-zygotic genomic changes in glutamate and dopamine pathway genes may explain discordance of monozygotic twins for schizophrenia. *Clin. Transl. Med.* **6,** 43 (2017).
231. Matos, C., Pereira de Almeida, L. & Nóbrega, C. Machado-Joseph disease / Spinocerebellar ataxia type 3:lessons from disease pathogenesis and clues into therapy. *J. Neurochem.* (2018). doi:10.1111/jnc.14541
232. Orefici, G., Cardona, F., Cox, C. J. & Cunningham, M. W. *Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections (PANDAS)*. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations* (University of Oklahoma Health Sciences Center, 2016).

233. Murthy, S. *et al.* OSCA/TMEM63 are an Evolutionarily Conserved Family of Mechanically Activated Ion Channels. *bioRxiv* 408732 (2018). doi:10.1101/408732

234. Durak, O. *et al.* Chd8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. *Nat. Neurosci.* **19,** 1477–1488 (2016).

235. Ristic, G., Tsou, W.-L. & Todi, S. V. An optimal ubiquitin-proteasome pathway in the nervous system: the role of deubiquitinating enzymes. *Front. Mol. Neurosci.* **7,** 72 (2014).

236. Endo, S., Miyagi, N., Matsunaga, T., Hara, A. & Ikari, A. Human dehydrogenase/reductase (SDR family) member 11 is a novel type of 17β-hydroxysteroid dehydrogenase. *Biochem. Biophys. Res. Commun.* **472,** 231–236 (2016).

237. Sawamura, S., Shirakawa, H., Nakagawa, T., Mori, Y. & Kaneko, S. *TRP Channels in the Brain: What Are They There For? Neurobiology of TRP Channels* (CRC Press/Taylor & Francis, 2017).

238. Cui, K. & Yuan, X. *TRP Channels and Axon Pathfinding*. *TRP Ion Channel Function in Sensory Transduction and Cellular SignalingCascades* (CRC Press/Taylor & Francis, 2007).

239. Kumar, S., Singh, U., Goswami, C. & Singru, P. S. Transient receptor potential vanilloid 5 (TRPV5), a highly Ca^{2+} -selective TRP channel in the rat brain: relevance to neuroendocrine regulation. *J. Neuroendocrinol.* **29,** (2017).

240. Singh, U. *et al.* Transient receptor potential vanilloid 3 (TRPV3) in the ventral tegmental area of rat: Role in modulation of the mesolimbic-dopamine reward pathway. *Neuropharmacology* **110,** 198–210 (2016).

241. Yazdani, U. & Terman, J. R. The semaphorins. *Genome Biol.* **7,** 211 (2006).

242. Mann, F., Chauvet, S. & Rougon, G. Semaphorins in development and adult brain: Implication for neurological diseases. *Prog. Neurobiol.* **82,** 57–79 (2007).

243. Marchisella, F., Coffey, E. T. & Hollos, P. Microtubule and microtubule associated protein anomalies in psychiatric disease. *Cytoskeleton* **73,** 596–611 (2016).
244. Kim, H.-Y., Huang, B. X. & Spector, A. A. Phosphatidylserine in the brain: Metabolism and function. *Prog. Lipid Res.* **56,** 1–18 (2014).
245. Sethi, S., Hayashi, M. A., Sussulini, A., Tasic, L. & Brietzke, E. Analytical approaches for lipidomics and its potential applications in neuropsychiatric disorders. *World J. Biol. Psychiatry* **18,** 506–520 (2017).
246. Lin, L., Yee, S. W., Kim, R. B. & Giacomini, K. M. SLC transporters as therapeutic targets: emerging opportunities. *Nat. Rev. Drug Discov.* **14,** 543–60 (2015).
247. Kim, K. M. *et al.* Cloning of the human glycine transporter type 1: molecular and pharmacological characterization of novel isoform variants and chromosomal localization of the gene in the human and mouse genomes. *Mol. Pharmacol.* **45,** (1994).
248. Lohoff, F. W. in *Methods in molecular biology (Clifton, N.J.)* **637,** 165–180 (2010).
249. Hübler, D. *et al.* Differential Spatial Expression and Subcellular Localization of CtBP Family Members in Rodent Brain. *PLoS One* **7,** e39710 (2012).
250. Kneussel, M. & Wagner, W. Myosin motors at neuronal synapses: drivers of membrane transport and actin dynamics. *Nat. Rev. Neurosci.* **14,** 233–247 (2013).
251. Derkach, A. *et al.* Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics* **30,** 2179–88 (2014).
252. Stanhope, S. A. & Skol, A. D. Improved Minimum Cost and Maximum Power Two Stage Genome-Wide Association Study Designs. *PLoS One* **7,** e42367 (2012).
253. Wason, J. M. S. & Dudbridge, F. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. *Am. J. Hum. Genet.* **90,** 760–73 (2012).

254. Satagopan, J. M., Venkatraman, E. S. & Begg, C. B. Two-Stage Designs for Gene-Disease Association Studies with Sample Size Constraints. *Biometrics* **60,** 589–597 (2004).
255. Nöthen, M. M. *et al.* Human dopamine D4 receptor gene: frequent occurrence of a null allele and observation of homozygosity. *Hum. Mol. Genet.* **3,** 2207–12 (1994).
256. Grady, D. L. *et al.* DRD4 genotype predicts longevity in mouse and human. *J. Neurosci.* **33,** 286–91 (2013).
257. Falzone, T. L. *et al.* Absence of dopamine D4 receptors results in enhanced reactivity to unconditioned, but not conditioned, fear. *Eur. J. Neurosci.* **15,** 158–64 (2002).
258. Sakurai, T. The role of NrCAM in neural development and disorders—Beyond a simple glue in the brain. *Mol. Cell. Neurosci.* **49,** 351–363 (2012).
259. Ishiguro, H. *et al.* NrCAM in Addiction Vulnerability: Positional Cloning, Drug-Regulation, Haplotype-Specific Expression and Altered Drug Reward in Knockout Mice. *Neuropsychopharmacology* **31,** 572–584 (2006).
260. Ishiguro, H. *et al.* NrCAM-regulating neural systems and addiction-related behaviors. *Addict. Biol.* **19,** 343–53 (2014).
261. Sakurai, T. *et al.* Association analysis of the NrCAM gene in autism and in subsets of families with severe obsessive-compulsive or self-stimulatory behaviors. *Psychiatr. Genet.* **16,** 251–257 (2006).
262. Mahgoub, M. & Monteggia, L. M. Epigenetics and Psychiatry. *Neurotherapeutics* **10,** 734–741 (2013).
263. Cattane, N. *et al.* Altered Gene Expression in Schizophrenia: Findings from Transcriptional Signatures in Fibroblasts and Blood. *PLoS One* **10,** e0116686 (2015).
264. Chou, F.-S. & Wang, P.-S. The Arp2/3 complex is essential at multiple stages of neural development. *Neurogenesis* **3,** e1261653 (2016).
265. Kim, I. H. *et al.* Spine pruning drives antipsychotic-sensitive locomotion via circuit control of striatal dopamine. *Nat. Neurosci.* **18,** 883–891 (2015).

266. Cen, C. *et al.* PKD1 promotes functional synapse formation coordinated with N-cadherin in hippocampus. *J. Neurosci.* **38,** 1640–17 (2017).
267. Prehn-Kristensen, A. *et al.* Reduced microbiome alpha diversity in young patients with ADHD. *PLoS One* **13,** e0200728 (2018).
268. O'Mahony, S. M., Clarke, G., Borre, Y. E., Dinan, T. G. & Cryan, J. F. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. *Behav. Brain Res.* **277,** 32–48 (2015).
269. Holzer, P. & Farzi, A. in *Advances in experimental medicine and biology* **817,** 195–219 (2014).
270. Huo, R. *et al.* Microbiota Modulate Anxiety-Like Behavior and Endocrine Abnormalities in Hypothalamic-Pituitary-Adrenal Axis. *Front. Cell. Infect. Microbiol.* **7,** 489 (2017).
271. Bailey, M. T. *et al.* Exposure to a social stressor alters the structure of the intestinal microbiota: Implications for stressor-induced immunomodulation. *Brain. Behav. Immun.* **25,** 397–407 (2011).
272. O'Mahony, S. M. *et al.* Early Life Stress Alters Behavior, Immunity, and Microbiota in Rats: Implications for Irritable Bowel Syndrome and Psychiatric Illnesses. *Biol. Psychiatry* **65,** 263–267 (2009).
273. Costello, M.-E. *et al.* Brief Report: Intestinal Dysbiosis in Ankylosing Spondylitis. *Arthritis Rheumatol. (Hoboken, N.J.)* **67,** 686–691 (2015).
274. Giannelli, V. *et al.* Microbiota and the gut-liver axis: bacterial translocation, inflammation and infection in cirrhosis. *World J. Gastroenterol.* **20,** 16795–810 (2014).
275. Leonard, H. L. & Swedo, S. E. Paediatric autoimmune neuropsychiatric disorders associated with streptococcal infection (PANDAS). *Int. J. Neuropsychopharmacol.* **4,** 191–8 (2001).
276. Attwells, S. *et al.* Inflammation in the Neurocircuitry of Obsessive-Compulsive Disorder. *JAMA Psychiatry* **74,** 833 (2017).
277. Aarts, E. *et al.* Gut microbiome in ADHD and its relation to neural reward anticipation. *PLoS One* **12,** e0183509 (2017).
278. Vogt, N. M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Sci. Rep.* **7,** 13537 (2017).

279. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain. Behav. Immun.* **48,** 186–194 (2015).
280. Strati, F. *et al.* New evidences on the altered gut microbiota in autism spectrum disorders. *Microbiome* **5,** 24 (2017).
281. Golubeva, A. V *et al.* Prenatal stress-induced alterations in major physiological systems correlate with gut microbiota composition in adulthood. *Psychoneuroendocrinology* **60,** 58–74 (2015).
282. Ning, T., Gong, X., Xie, L. & Ma, B. Gut Microbiota Analysis in Rats with Methamphetamine-Induced Conditioned Place Preference. *Front. Microbiol.* **8,** 1620 (2017).
283. Keshavarzian, A. *et al.* Colonic bacterial composition in Parkinson's disease. *Mov. Disord.* **30,** 1351–1360 (2015).
284. Hill-Burns, E. M. *et al.* Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov. Disord.* **32,** 739–749 (2017).
285. Scheperjans, F. *et al.* Gut microbiota are related to Parkinson's disease and clinical phenotype. *Mov. Disord.* **30,** 350–358 (2015).
286. Kang, D.-W. *et al.* Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children. *PLoS One* **8,** e68322 (2013).
287. Finegold, S. M. *et al.* Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe* **16,** 444–453 (2010).
288. Qiao, Y. *et al.* Alterations of oral microbiota distinguish children with autism spectrum disorders from healthy controls. *Sci. Rep.* **8,** 1597 (2018).
289. Mariat, D. *et al.* The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol.* **9,** 123 (2009).
290. Greenhalgh, K., Meyer, K. M., Aagaard, K. M. & Wilmes, P. The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environ. Microbiol.* **18,** 2103–2116 (2016).

SUPPLEMENTARY METHODS

S1. Quality control

S1.1. Quality control of samples captured with Agilent 35, Agilent 50 and NimbleGen v3

From our in-house WES dataset, we selected only those samples that we were going to use in subsequent analysis: 306 OCD cases and 630 unrelated Spanish samples that did not present any neuropsychiatric disorder or related, which we used as controls in the RVAS. All these samples had been whole-exome captured with Agilent 35, Agilent 50 and NimbleGen v3

We excluded samples with 0 variants and variants that had more then 20 % of NA calls in the samples, which made a total of 116,856 excluded variants (9.87%). We then filtered by a minimum number of variants per sample of 42,500 (Supplementary Figure S1), excluding 5 samples.

We did not filter by transition to transversion (Ti/Tv) ratio per sample because we considered that all samples met the standards: the ratio of transitions to transversions is typically around 2 across the entire genome and higher in protein coding regions (Supplementary Figure S2).

Then, we performed a PCA and filtered out outliers based on the first two principal components, PC1 and PC2 (Supplementary Figure S3). After selecting a maximum threshold of 3 for PC1 and of 5 for PC2 and a minimum threshold of -10 for PC2, we excluded 26 samples. Next, we excluded 442,995 positions with any sample presenting alternative variants or with all samples presenting only NA genotypes. Finally, after QC filtering, we had 292 OCD cases, 601 controls, and a total of 624,516 unique SNVs and indels. The samples of the control group belonged to the following projects: controls, centenarians, healthy parents

of intellectual disability probands, and chronic lymphocytic leukemia, cystic fibrosis and fibromyalgia, and stroke patients.

S1.2. Quality control of samples captured with NimbleGen v3

From our in-house WES dataset, we selected only those samples that we were going to use in subsequent analysis: 266 OCD cases and 206 unrelated Spanish samples that did not present any neuropsychiatric disorder or related, which we used as controls in the RVAS. All these samples had been whole-exome captured with NimbleGen v3

We excluded samples with 0 variants and variants that had more then 20 % of NA calls in the samples, which made a total of 838,503 excluded variants (41.2 %). We then filtered by a minimum number of variants per sample of 47,500 (Supplementary Figure S4), excluding 5 samples. We did not filter by transition to transversion (Ti/Tv) ratio per sample because we considered that all samples met the standards (Supplementary Figure S5). Then, we performed a PCA and filtered out outliers based on the first two principal components, PC1 and PC2 (Supplementary Figure S6). After selecting a minimum threshold of -5 for PC1 and of -10 for PC2 and a maximum threshold of 10 for PC2, we excluded 15 samples. Next, we excluded 16,815 positions with any sample presenting alternative variants or with all samples presenting only NA genotypes. Finally, after QC filtering, we had 253 OCD cases, 188 controls, and a total of 490,150 unique SNVs and indels. The samples of the control group belonged to the following projects: controls, healthy parents of intellectual disability probands, and fibromyalgia, and stroke patients.

S1.3. Quality control of samples from targeted sequencing

From the targeted sequencing samples, we excluded those with less than 150 variants (Supplementary Figure S7), and those variants that had more then 20 % of NA calls in the samples. We also filtered out samples with a Ti/Tv ratio

lower than 1.8 and higher than 4, in order to delete outliers (Supplementary Figure S7). We did not apply hard filters by number of mutations per sample or Ti/Tv ratio because we included these variables later in the RVAS model, as covariates. Next, we performed a PCA but we did not filter by PC, as there were no outliers (Supplementary Figure S8). After QC filtering, we had 427 OCD cases, 1474 controls, and a total of 13,751 unique SNVs and indels.

S2. Development of a drd4 knockout zebrafish model (ZeClinics methodology)

S2.1. Zebrafish maintenance

Adults wild-type zebrafish (*Danio rerio*), strain AB, purchased from KIT-European Zebrafish Resource Center (EZRC) were maintained at 28–29ºC on a light cycle of 14h light: 10h dark (lights on at 7am; lights off at 9 pm).

S2.2. CRISPR/Cas9 design for gene Knock Out

Gene sequences were retrieved using http://www.ncbi.nlm.nih.gov/gene and http://www.ensembl.org/Danio_rerio/Info/Index. sgRNAs were designed using the online tool http://crispor.tefor.net/, based on exon site and high efficacy and not off-target published algorithms.

S2.3. Zebrafish embryo preparation and sgRNA's microinjection

Fertilized zebrafish embryos were collected in E3 medium in Petri dishes. At 1-cell stage (0-0.5 hours post fertilization (hpf)), >20 embryos (to identify the best sgRNA candidates for KO efficiency) per gene were injected. At 3 hpf, deformed or not fertilized embryos were discarded. 20-25 embryos of 48 hpf per injected pool were selected for Double Strand Break (DSB) efficacy analysis. To do so different steps were followed:

i) Genomic DNA extraction

ii) PCR with HIFI-Taq from genomic DNA using Diagnosis primers (XX_Fw/XX_Rv). Fw and Rv primers should be ~100-150 bp from DNA break. Hence, PCR size should be ~300 bp. PCR in a 10 uL volume.
iii) Clean with PCR cleaning columns (Qiagen better). Elute in 20 uL dH2O.
iv) Take 1 uL of clean PCR for T7 endonuclease reaction (when KO).

Injected embryos were grown to adulthood to identify F0 mutant founders, which were outcrossed to wt animals to generate F1 heterozygous mutants. F1 embryo pools were selected to grow to adulthood through the above mentioned protocol.

When F1 animals reached adulthood, their fins were individually clipped and used to extract genomic DNA. The region of interest was amplified and analysed by Sanger sequencing to identify isogenic mutant animals promoting the appearance of early stop codons in *drd4a* and *drd4rs*.

After selecting knockout alleles for each gene (F1 generation), we crossed single heterozygous between each other in order to generate double heterozygous animals (F2 generation). Through the cross of these double het animals we obtained single and double homozygous larvae for performing the proposed phenotypic analysis.

S2.3. Behavioural protocol

Putative behavioural alterations were assessed by comparing locomotion differences among the different genotypes: *wild-type*, *ra* or *rs* single *knock-out*, *ra* and *rs* double *knock-out*. Larvae were analysed at 120 hpf by locomotion assessment using the EthoVision XT 12 software and the DanioVision device from Noldus Information Technologies, Wageningen, The Netherlands. This closed system consists of a camera placed above a chamber with circulating water and a temperature sensor set at 28 ◦C. Individualized larvae in a 48-wells plate are placed in the chamber, which can provide different stimuli (light/dark environment, tapping, sound) controlled by the software. Prior to each experiment, larvae are left for 10 minutes in dark for acclimation, then

predetermined series of alternating dark and light environment and external stimuli are presented to the larvae. The final experimental protocol is divided in two main parts: first a 50 minutes of dark/light alternating environments phase (10 min each), then a series repeated tapping for 30 seconds (1 tapping/sec).

Different sets of information can be extrapolated from the different phases: the first phase is useful to detect anomalies in larval movement and deviations from the stereotyped behaviour (natural locomotor behaviour of zebrafish is active in dark and immobile in light) and changes in the larval total locomotion. Moreover, measuring the time spent by the larvae in the center or in the periphery of the well allows us to extrapolate the anxiety state of the individual (thigmotaxis). Finally, in the second part of the trial, defects in short memory and learning could be evaluated by testing the capacity to gradually reduce response to the external stimuli, a process known as “habituation”.

S2.4. Statistical analysis

Data was analysed using the IBM SPSS Statistics version 20.0 software (Armonk, NY, USA). Data are presented as mean ± standard error (SE). Prior to the analyses, the Shapiro-Wilk test was used to assess the normality of the distribution of the dependent variables. Statistical analysis of the data for the locomotive parameters was performed using One-way ANOVA followed by the Dunnett test. Results were statistically compared between genetic groups and wt (negative control) group. Differences were considered statistically significant when $p<0.05$.

S3. Diversity measures

We estimated α- and β-diversity measures within samples. α-diversity refers to species richness (number of taxa) within a single sample, while β-diversity refers to dissimilarity in taxonomic abundance profiles from different samples.

We estimated α-diversity within samples as measured by Observed diversity, Chao1 index, Abundance-based Coverage Estimator (ACE), Shannon, Simpson, Inverse Simpson, and Fisher Diversity indices using the estimate_richness function from the Phyloseq package. We also calculated Faith's phylogenetic diversity and species richness using the pd function from the picante package (version 1.6.2)[1]. Boxplots were generated using ggplot2 (version 2.2.1)[2]. Statistical significance of α-diversity differences between groups was evaluated with Mann–Whitney U test when samples were independent, and with Wilcoxon rank-sum test when samples were paired.

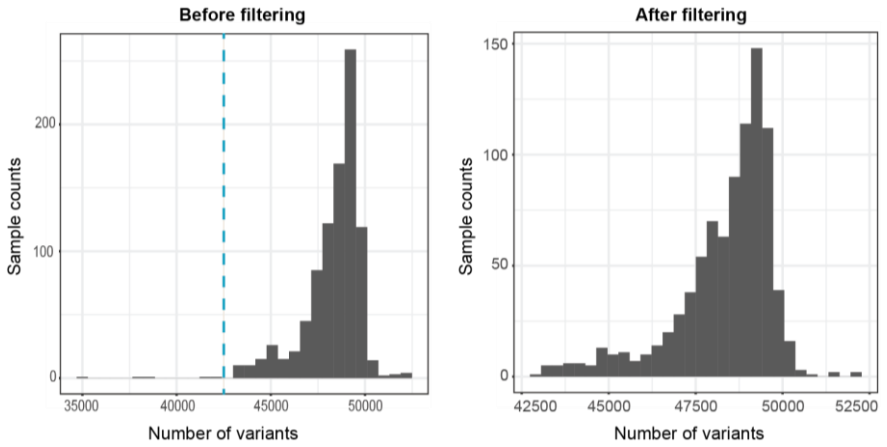
The observed diversity index measures the number of different species per sample, which is defined as "richness". It does not considers the abundances of the species or their relative abundance distributions. The Chao1 index is also a qualitatively measure of alpha diversity which, beside species richness, considers the ratio of singletons (n = 1) to doubletons (n = 2) giving more weight to rare species. The ACE incorporates data from all species with fewer than 10 individuals, rather than just singletons and doubletons. The Shannon diversity index relates taxa richness and evenness, which is defined as the relative abundances of the different species making up the samples' richness. The Simpson diversity index considers the number of species present, as well as the abundance of each species, but it has a strong dependency on the few most common species. The inverse of Simpson index refers to the effective number of taxa types that is obtained when the weighted arithmetic mean is used to quantify average proportional abundance of taxa types in the dataset of interest. Fisher is an alpha diversity measure with an inherent assumption of a logarithmic series-type rank abundance structure of communities. Finally, Faith's Phylogenetic Diversity (PD) is the phylogenetic analogue of taxon richness and is expressed as the number of tree units which are found in a sample.

We estimated β-diversity as the weighted and unweighted UniFrac distance between samples with the Unifrac function, as well as the Jensen-Shannon Divergence (JSD) with the JSD function, both from the Phyloseq package, and we also calculated the Bray-Curtis dissimilarity and Canberra index using the

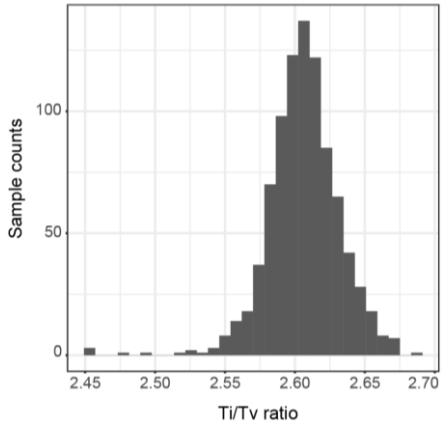
vegdist function in the vegan package (version 2.4.6)[3]. Furthermore, the adonis function in the vegan package was used to perform a PERMANOVA test on β-diversity with 999 permutations considering even dependence of samples (paired OCD samples after and before treatment) using the “strata” argument within the adonis function. We used a Principal Coordinate Analysis (PCoA) to visualize the clustering of the samples.

Bray-Curtis dissimilarity is a statistic used to quantify the compositional dissimilarity between two samples, based on abundance or read count data. It is dominated by the abundant species so that rare species add very little to the value of the coefficient. The Canberra metric is not affected as much by the more abundant species in the community, and thus differs from the Bray-Curtis measure. The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. UniFrac is a distance metric used for comparing biological communities. It differs from dissimilarity measures in that it incorporates information on the relative relatedness of community members by incorporating phylogenetic distances between observed organisms in the computation. Unweighted uniFrac metric is purely based on sequence distances (does not include abundance information), while in weighted UniFrac metric branch lengths are weighted by relative abundances (includes both sequence and abundance information).

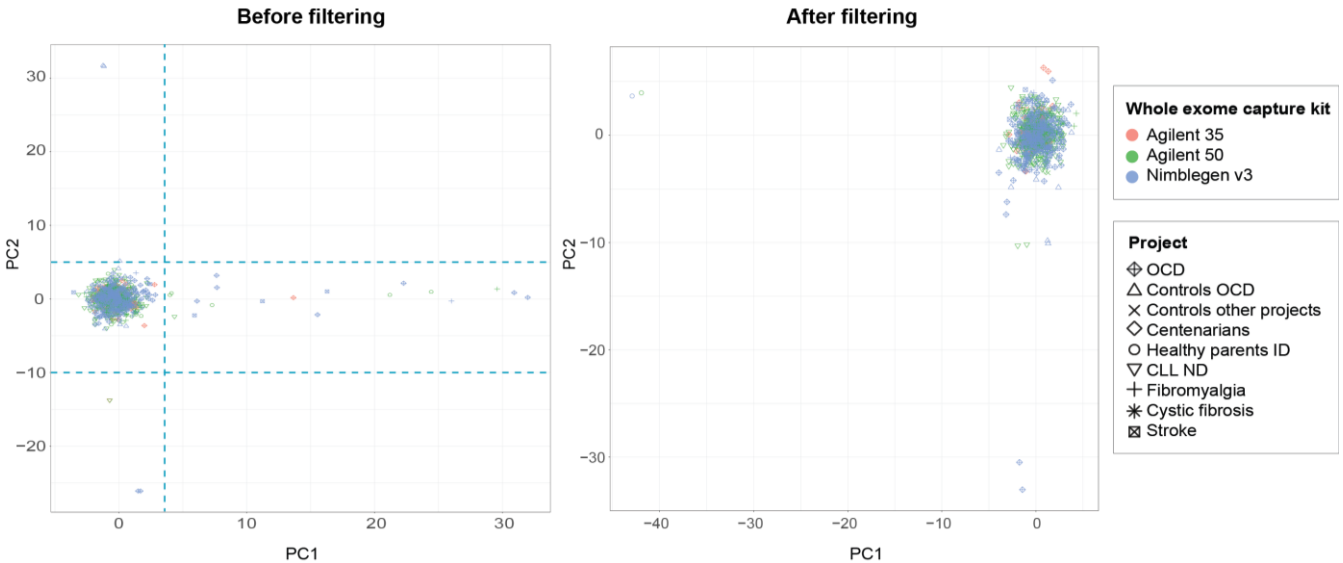
SUPPLEMENTARY FIGURES



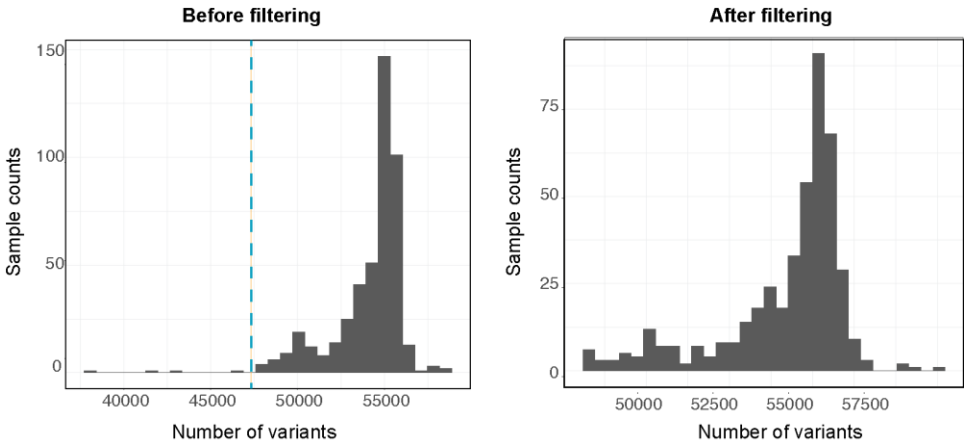
Supplementary Figure S1. Number of variants per sample before and after filtering. We removed those samples wit less than 42,500 variants, as they represented outliers of our sample dataset.



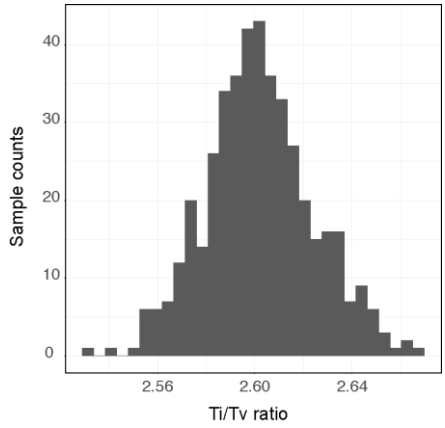
Supplementary Figure S2. Transition/transversion (Ti/Tv) ratio per sample.



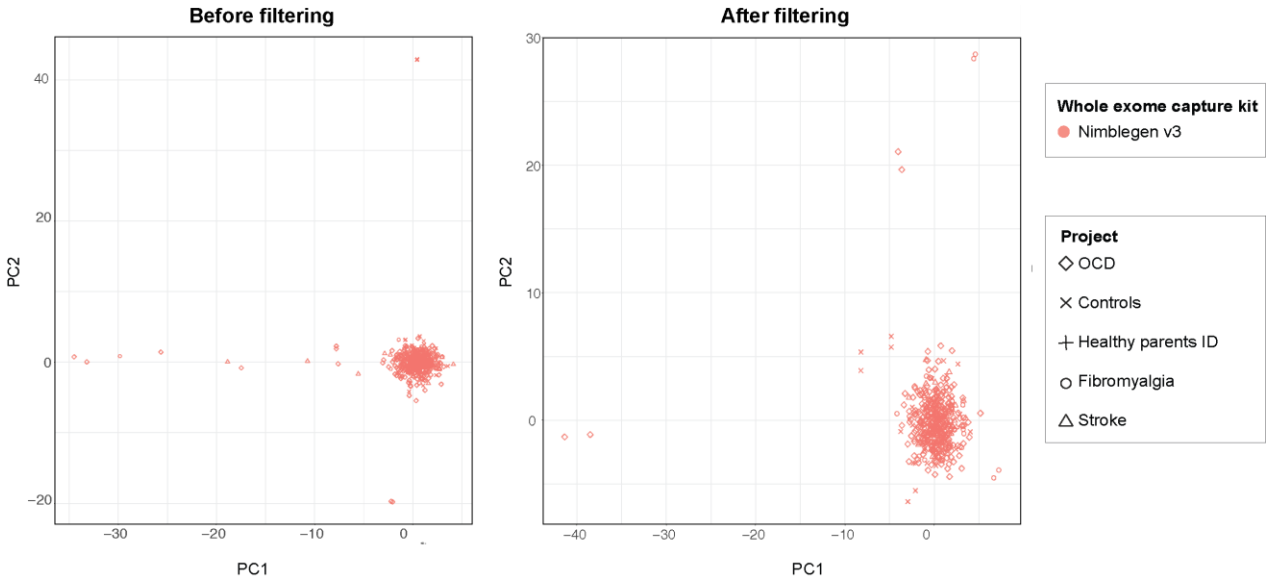
Supplementary Figure S3. PCA of all samples involved in posterior RVAS before and after filtering. We performed the PCA analysis selecting all synonymous SNVs without linkage disequilibrium of all samples. Agilent 35: Agilent SureSelect Human All Exon 35Mb Kit; Agilent 50: Agilent SureSelect Human All Exon 50Mb Kit; NimbleGen v3: NimbleGen SeqCap EZ Library v3.0; OCD: obsessive compulsive disorders; ID: intellectual disability; CLL: chronic lymphocytic leukemia.



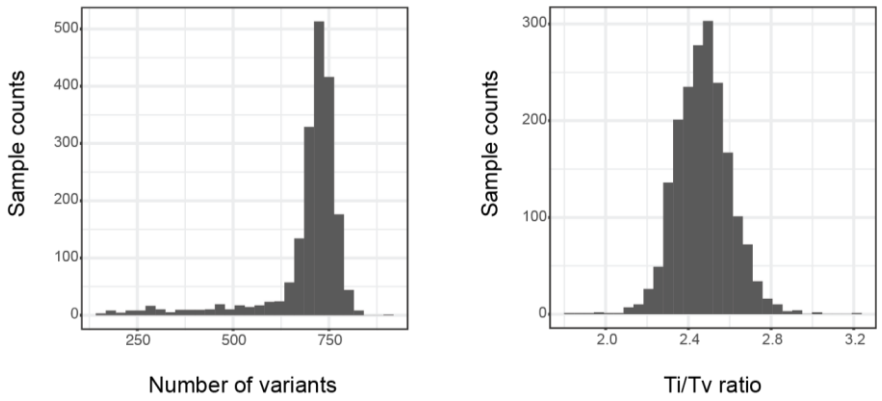
Supplementary Figure S4. Number of variants per sample before and after filtering. We removed those samples wit less than 47,500 variants, as they represented outliers of our sample dataset.



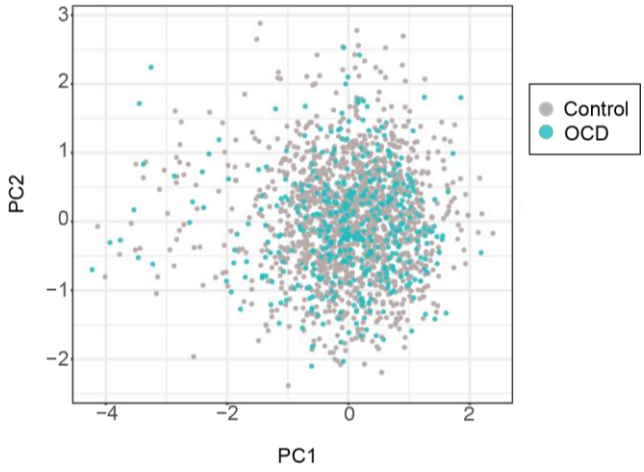
Supplementary Figure S5. Transition/transversion (Ti/Tv) ratio per sample.



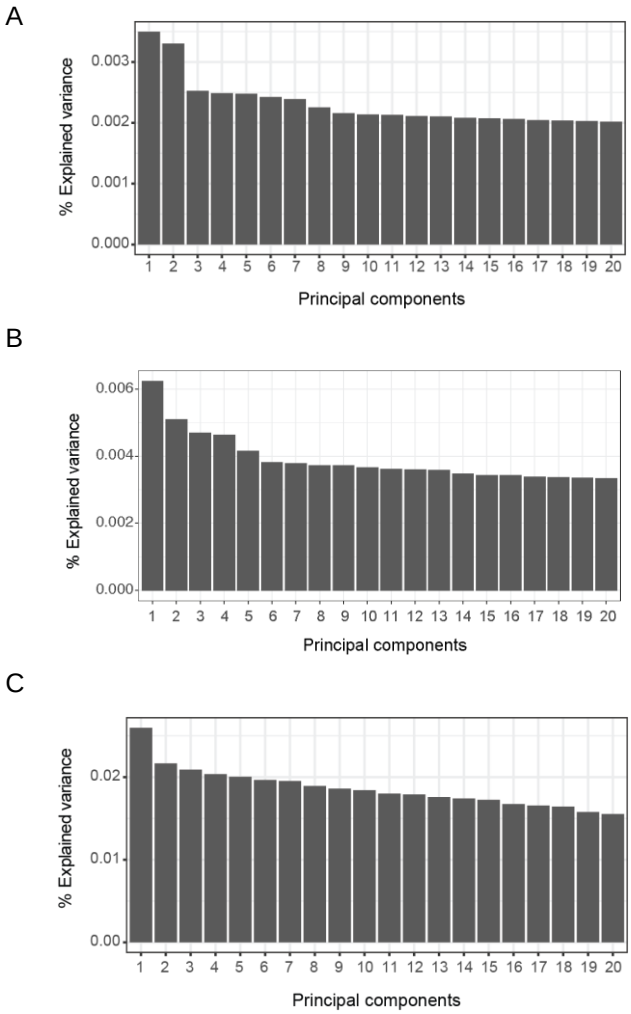
Supplementary Figure S6. PCA of all samples involved in posterior RVAS before and after filtering. We performed the PCA analysis selecting all synonymous SNVs without linkage disequilibrium of all samples. NimbleGen v3: NimbleGen SeqCap EZ Library v3.0; OCD: obsessive compulsive disorders; ID: intellectual disability.



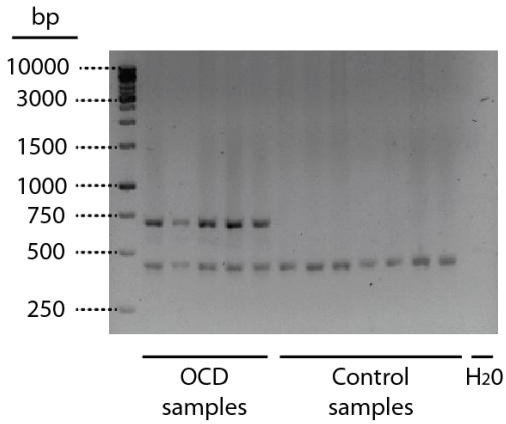
Supplementary Figure S7. Targeted resequencing quality control. Number of variants (left) and Ti/Tv ratio (right) per sample after filtering.



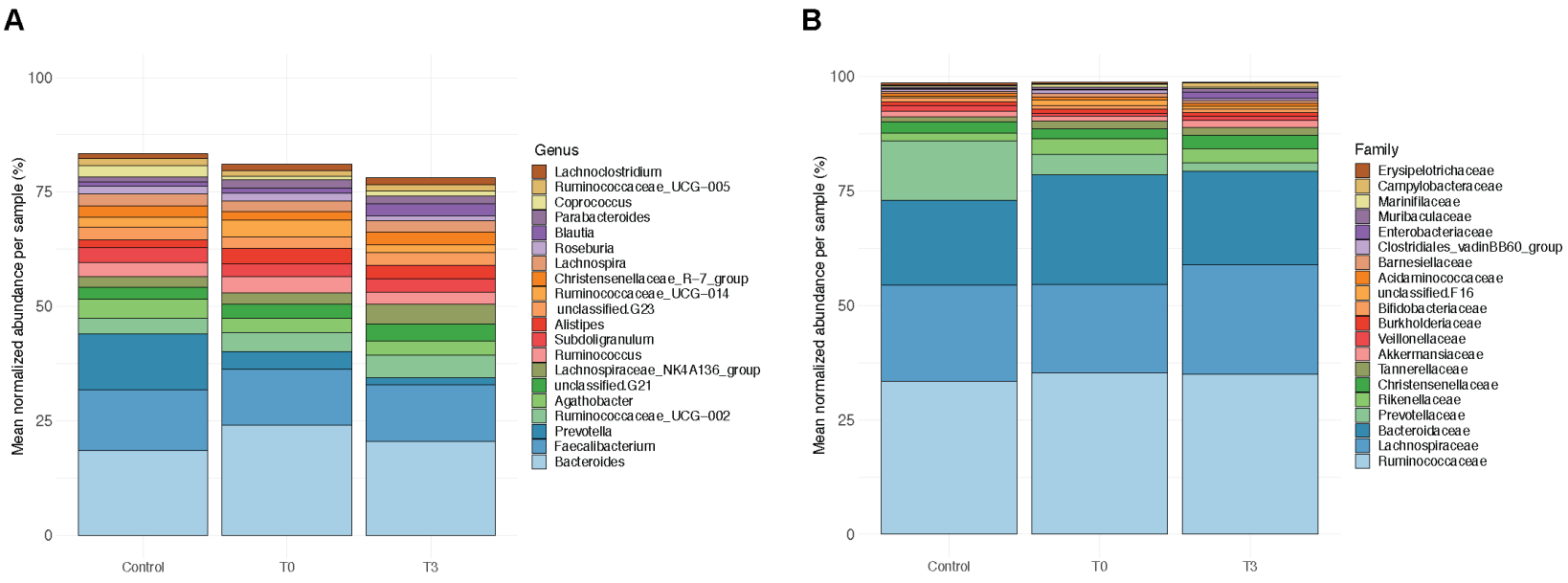
Supplementary Figure S8. Targeted resequencing quality control. PCA of all samples involved in the targeted resequencing study. We performed the PCA analysis selecting all synonymous SNVs without linkage disequilibrium of all samples.



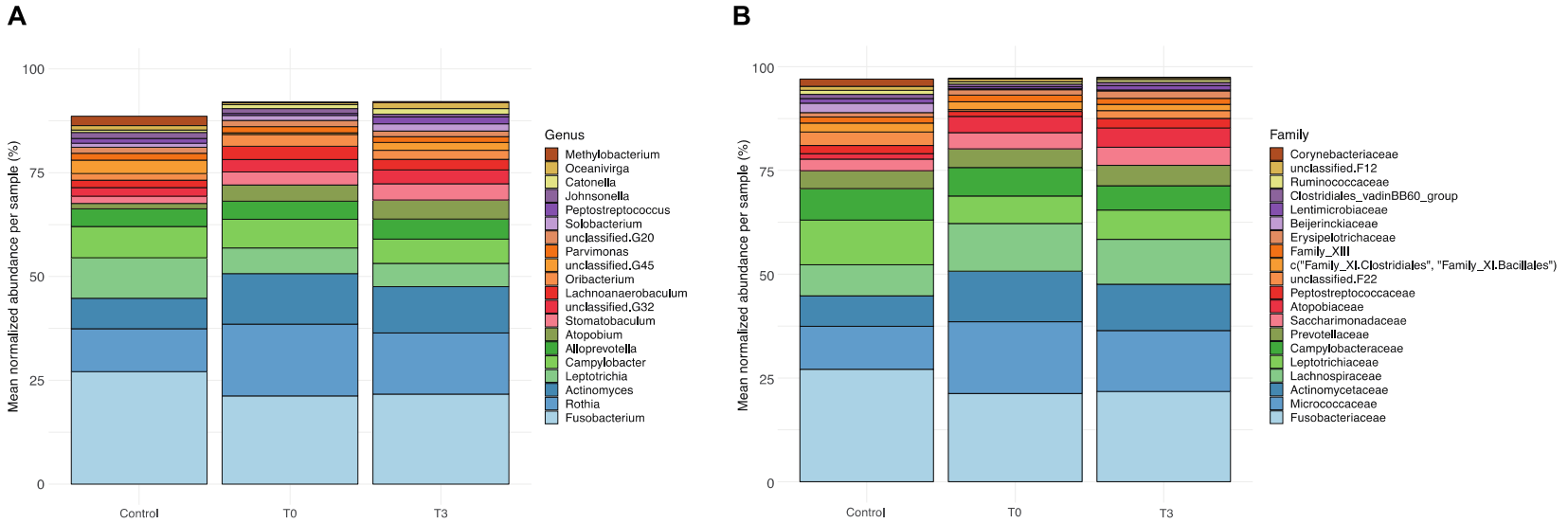
Supplementary Figure S9. Percentage of explained variance by the first 20 principal components derived from (A) PCA of samples whole-exome sequenced with Agilent 35, Agilent 50 and NimbleGen v3; (B) PCA of samples whole-exome sequenced with NimbleGen v3; and (C) PCA of the targeted resequencing samples. We selected the 10 first principal components to be included as covariates in subsequent RVAS.



Supplementary Figure S9. Gel electrophoresis of the multiplex PCR designed to detect the *DRD4* 13-bp frameshift deletion in B-lymphoblastoid cell lines of OCD patients carrying the deletion and controls. OCD samples (the first 5 lanes) presented the 674 bp, band corresponding to the deletion. Control samples (the next 7 lanes) only presented the positive control band of 429 bp, which meant that they did not have the deletion.



Supplementary Figure S11. Gut bacterial abundances for OCD T0, OCD T3 and controls at the genus (A) and family (B) level.



Supplementary Figure S12. Oro-pharyn1geal bacterial abundances for OCD T0, OCD T3 and controls at the genus (A) and family (B) level.

SUPPLEMENTARY TABLES

Supplementary tables are provided in electronic format.

Supplementary Table S1: Clinical diet questionnaire

Supplementary Table S2: Diet questionnaire answers from OCD patients

Supplementary Table S3: Diet questionnaire answers from controls

Supplementary Table S4: OCD patient data according to age and gender

Supplementary Table S5: Clinical data from OCD patients included in the metagenomics study

Supplementary Table S6: List of primers sequences used for Sanger sequencing validation

Supplementary Table S7: Primers used in the multiplex PCR for DRD4 validation

Supplementary Table S8: PCR amplification conditions for DRD4 validation

Supplementary Table S9: Targeted sequencing design

Supplementary Table S10: Gene list from Venn Diagram analyses comparing different RVAS methods

Supplementary Table S11: Gene list from Venn Diagram analyses comparing RVAS performed with whole-exome samples captured with one or three kits

Supplementary Table S12: Enriched gene ontology-based sets from RVAS results

Supplementary Table S13: ExAC data of the common variants of the study (1) associated to OCD with statistical significance (adjusted p-value < 0.005**Supplementary Table S14:** DE genes with FC >1.5 or FC <0.665 and nominal p-value <0.05

Supplementary Table S15: Enriched gene ontology-based sets from DE analyses OCD T0 versus controls results

SUPPLEMENTARY BIBLIOGRAPHY

1. Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26,** 1463–4 (2010).
2. Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).
3. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, E. S. and H. W. CRAN - Package vegan. (2018).

ANNEX

List of publications

Costas J, Carrera N, Alonso P, Gurriarán X, Segalàs C, Real E, López-Solà C, Mas S, Gassó P[9], Domènech L, Morell M, Quintela I, Lázaro L, Menchón JM, Estivill X, Carracedo Á.

Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. Transl Psychiatry. 2016 Mar 29;6:e768. doi: 10.1038/tp.2016.34.

Kumar R, Gardner A, Homan CC, Douglas E, Mefford H, Wieczorek D, Lüdecke HJ, Stark Z, Sadedin S, Broad CMG, Nowak CB, Douglas J, Parsons G, Mark P, Loidi L, Herman GE, Mihalic Mosher T, Gillespie MK, Brady L, Tarnopolsky M, Madrigal I, Eiris J, Domènech Salgado L, Rabionet R, Strom TM, Ishihara N, Inagaki H, Kurahashi H, Dudding-Byth T, Palmer EE, Field M, Gecz J.

Severe neurocognitive and growth disorders due to variation in THOC2, an essential component of nuclear mRNA export machinery. Hum Mutat. 2018 Aug;39(8):1126-1138. doi: 10.1002/humu.23557. Epub 2018 Jun 14.

Articles in revision

Hana Susak, Serra-Saurina L., Mattia Bosio, Raquel Rabionet Janssen, Laura Domènech Salgado, Xavier Estivill, Georgia Escaramís* and Stephan Ossowski*.

Bayesian Rare Variant Association Test using Integrated Nested Laplace Approximation

Francesc Muyas, Mattia Bosio, Anna Puig, Hana Susak, Laura Domènech-Salgado, Georgia Escaramis, Luis Zapata, Xavier Estivill, Raquel Rabionet Janssen, Stephan Ossowski

Allele Balance Bias Identifies Systematic Genotyping Errors and False Disease Associations

Mattia Bosio, Oliver Drechsel, Rubayte Rahman, Francesc Muyas, Raquel Rabionet, Daniela Bezdan, Laura Domènech Salgado, Hyun-Gyu Hor, Jean-Jacques Schott, Francina Munell, Roger Colobran, Alfons Macaya, Xavier Estivill, Stephan Ossowski.

eDiVA – Classification and Prioritization of Pathogenic Variants for Clinical Diagnostics

María Alemany Navarro, Javier Costas, Eva Real, Cinto Segalàs, Sara Bertolín, Laura Domènech, Raquel Rabionet, Ángel Carracedo, Jose Manuel Menchón, Pino Alonso.

Do Polygenic Risk and Stressful Life Events Predict Pharmacological Treatment Response in Obsessive Compulsive Disorder?