Received: 27 October 2017 | Revised: 28 February 2018 | Accepted: 30 March 2018

DOI: 10.1002/ajmg.b.32638

RESEARCH ARTICLE

WILEY American Journal of medical genetics Part B Neuropsychiatric Genetics

Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes

Yannis J. Trakadis | Sameer Sardaar | Anthony Chen | Vanessa Fulginiti | Ankur Krishnan

Department of Human Genetics, McGill University, Montreal, Québec, Canada

Correspondence
Yannis J. Trakadis, Department of Human Genetics, McGill University, Room A04.3140, 1001 Boul. Decarie, Montreal, Quebec H4A 3J1, Canada
Email: yannis.trakadis@mcgill.ca

Our hypothesis is that machine learning (ML) analysis of whole exome sequencing (WES) data can be used to identify individuals at high risk for schizophrenia (SCZ). This study applies ML to WES data from 2,545 individuals with SCZ and 2,545 unaffected individuals, accessed via the database of genotypes and phenotypes (dbGaP). Single nucleotide variants and small insertions and deletions were annotated by ANNOVAR using the reference genome hg19/GRCh37. Rare (predicted functional) variants with a minor allele frequency $\leq$1% and genotype quality $\geq$90 including missense, frameshift, stop gain, stop loss, intronic, and exonic splicing variants were selected. A file containing all cases and controls, the names of genes with variants meeting our criteria, and the number of variants per gene for each individual, was used for ML analysis. The supervised machine-learning algorithm used the patterns of variants observed in the different genes to determine which subset of genes can best predict that an individual is affected. Seventy percent of the data was used to train the algorithm and the remaining 30% of data ($n = 1{,}526$) was used to evaluate its efficiency. The supervised ML algorithm, gradient boosted trees with regularization (eXtreme Gradient Boosting implementation) was the best performing algorithm yielding promising results (accuracy: 85.7%, specificity: 86.6%, sensitivity: 84.9%, area under the receiver-operator characteristic curve: 0.95). The top 50 features (genes) of the algorithm were analyzed using bioinformatics resources for new insights about the pathophysiology of SCZ. This manuscript presents a novel predictor which could potentially enable studies exploring disease-modifying intervention in the early stages of the disease.

KEYWORDS
artificial intelligence, diagnostic, genomic, prediction, psychosis

1 | INTRODUCTION

Schizophrenia (SCZ) is a severe mental illness with an estimated incidence of ~1% and a heritability of ~80% (Bassett et al., 1995). However, classic linkage and association studies have not been very successful in identifying major susceptibility loci (Altmüller et al., 2001; Cardno & Gottesman, 2000; DeLisi et al., 2002; Risch, 2000; Sullivan, Kendler, & Neale, 2003). With the development of new 'omics technologies', there are new opportunities to better understand the pathophysiology of SCZ. Recently, several whole exome sequencing (WES) studies have been performed in patients with SCZ, allowing simultaneous screening for variants in the coding portion of all genes in a patient's genome.

One interesting finding highlighted by several WES studies is the importance of de novo mutation (DNM) in individuals with SCZ and other neuropsychiatric diseases (Girard et al., 2011; Li et al., 2016; Xu, Ionita-Laza, et al., 2012). For instance, Li et al. (2016) reported a higher prevalence of DNMs in probands across four neuropsychiatric disorders: SCZ, autism spectrum disorder (ASD), encephalopathy (EE), and intellectual disability (ID). This study retrieved rare DNMs from 3,555 trios across the four neuropsychiatric disorders targeted, in addition to unaffected siblings (control), from 36 studies using WES or whole genome sequencing. The 3,334 DNMs selected had a minor allele frequency <0.1% and were exonic, loss of function (LOF), variants. These variants were analyzed for an association with each of the four targeted diseases and a higher prevalence of DNMs was noted in the

probands of all four disorders compared to controls. After transmission and de novo association analysis, a total of 764 potential candidate genes with *P*-value generated by the TADA program (PTADA) $\leq$0.05 were identified in the four disorders: 277 for SCZ, 330 genes for ASD, 109 for EE, and 106 for ID. Fifty-three candidate genes were shown to be associated with more than one disorder, suggesting a possibly shared genetic etiology underlying these four disorders. In addition, among the 764 candidate genes, 8 genes harbored recurrent DNMs in 1,024 SCZ trios, 12 from 1,038 ASD trios, 8 from 291 EE trios, and 15 from 220 ID trios. No single gene was found to harbor recurrent extreme DNMs in 982 controls.

On top of the role of DNMs, WES studies have emphasized the importance of *inherited* functional variants in SCZ. Takata et al. (2014) focused on LOF variants in WES data from patients with SCZ ($n = 231$) and control ($n = 34$) trios. Both inherited and de novo variants were included in this analysis. Two LOF variants were identified in *SETD1A*, a subunit of the histone methyltransferase protein complex. This is of particular interest because there is evidence that disruption of chromatin modification, specifically histone H3 methylation, is linked with the pathogenesis of SCZ. Transmission pattern analyses in this study revealed that LOF variants are more likely to be transmitted to affected individuals than controls, especially for private LOF variants in genes intolerant to functional genetic variation. Similarly, Singh et al. (2016) analyzed the burden of rare LOF variants (with minor allele frequency $<$0.1%) on WES data of 4,264 SCZ cases, 9,343 controls, and 1,077 published trios. *SETD1A* was again found to be associated with SCZ risk (Singh et al., 2016). Another WES study, performed by Genovese et al. (2016), used data from 4,946 Swedish patients with SCZ and identified rare protein damaging variants that were present in single individuals. These variants were not present in the Exome Aggregation Consortium (consisting of 45,376 individuals, after excluding the subjects from this study and other subjects ascertained for psychiatric disorders). In addition, the frequency of these variants in the SCZ cases was significantly increased relative to 6,242 unrelated Swedish controls. Furthermore, the elevated rate of these rare variants in individuals affected with SCZ was found to be several times greater than the rate of DNMs. This suggests that the observed significant excess of variants in individuals with SCZ is mostly inherited. At the pathophysiology level, given these variants were concentrated in neuronally expressed genes, whose ribonucleic acids (RNAs) interact with synaptically localized proteins, this study provided additional evidence for the role of synaptic dysfunction in the pathogenesis of SCZ.

In conclusion, these studies have advanced our understanding of the molecular etiology of SCZ and emphasized the role of both DNM and inherited variants in its pathogenesis. However, their focus has not been on identifying a predictor for individuals at high risk for SCZ and testing its accuracy using a different test sample. The quantity of WES data is amenable to machine learning (ML), a method of data analysis that automates analytical model building and constructs algorithms that iteratively learn from data to optimize data-based predictions. Our hypothesis is that ML analysis of WES data (both inherited and DNM functional variants) can be used to identify individuals at high risk for SCZ. If true, this could potentially enable disease-modifying clinical trials in the early stages of the disease.

2 | METHODS

2.1 | Exome data source and annotation

This study was approved by the research ethics board of the McGill University Health Centre. WES data for 2,545 individuals with SCZ and 2,545 unaffected individuals were accessed via the database of genotypes and phenotypes (dbGaP), study phs000473.v1.p1. The downloaded dataset includes sample information, including a numerical identifier of each individual, affection status, and vcf files indicating the variants in all genes of the exome for each individual in the dataset. The median age of cases in this study was 54 (interquartile range = 45 to 62 years old), while the median age of controls was 57 (interquartile range = 48 to 65 years old). Older controls were selected to ensure they were properly classified because they had greater time at risk for psychiatric hospitalizations. Cases were selected if they were (1) hospitalized two or more times with a discharge diagnosis of SCZ, (2) 18 years old or older, and (3) both parents were born in Scandinavia. Controls were randomly selected from the Swedish population registry and were not reported to be matched to cases based on age or sex. They were included in the study if they were (1) never hospitalized for SCZ or bipolar disorder, (2) 18 years or older and (3) both parents were born in Scandinavia.

Single nucleotide variants and small insertions and deletions were annotated by ANNOVAR using the reference genome hg19/GRCh37. The variants were segregated using an in-house program; affected and unaffected individuals were processed separately. Rare (predicted functional) variants with a minor allele frequency $\leq$1% and genotype quality $\geq$90 including missense, frameshift, stop gain, stop loss, intronic, and exonic splicing variants in the whole exome were selected. The output data consisted of a list of variants, the gene where the variant occurs, and the individuals with the variant. If a position included a variant that did not meet our filtering criteria, the position was considered wildtype. The format of the list of selected variants was rearranged in a tabular format indicating the sample ID, the gender and the total number of variants per gene meeting our criteria for each individual. If an individual did not have a single variant meeting our criteria in a specific gene, the gene was assigned a zero (i.e., no variants meeting our criteria) for that individual. A file containing all individuals (cases and controls), the names of genes with variants meeting our criteria, and the number of variants per gene for each individual, was used for ML analysis.

2.2 | Machine learning

The supervised machine-learning algorithm used the patterns of variants observed in the different genes to determine which subset of genes can best predict that an individual is affected. Only genes with at least one variant in five or more individuals (either affected or unaffected) were included in the analysis. As a data preprocessing step, feature values were standardized to have mean value of zero and

TABLE 1 Algorithms comparison table

Algorithm	Accuracy (%)	Sn (%)	Sp (%)	Precision (%)	Recall (%)	F1	Accuracy *p* value
XGBoost	85.7	84.9	86.6	86.9	84.9	85.9	9.11E-179
L1.Logistic	74.6	72.0	77.3	76.0	72.0	73.9	2.87E-86
SVM	70.7	70.8	70.6	70.5	70.8	70.7	1.18E-59
Random. Forest	81.7	82.0	81.3	81.1	82.0	81.6	2.57E-141

Sn = sensitivity; Sp = specificity. Summary of the performance of the predictors used. Across these methods, XGBoost clearly outperformed the other methods in all categories.

standard deviation 1. In addition, we filtered out features in our data showing more than 90% Pearson correlation. Features were then selected in combination of lasso regularized (L1) logistic regression, random forest, and extreme gradient boosting algorithms' implicit feature selection. As a result, the number of relevant features for prediction purposes came out to 1,155 down from 17,138 features. To prevent overfitting, a regularized implementation of the gradient boosting algorithm called eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) was used. The implementation of the XGBoost algorithm was done in the R programing language (Team RC, 2016) in XGBoost library (Chen & He, 2015). The dataset was split into 70% for training and 30% testing purposes. Since the distribution of case and control in our dataset is balanced, accuracy (i.e., number of correct predictions/total number of predictions) is a well-defined performance indicator. The performance of each algorithm was documented and the genes of highest importance across different ML methods compared. (For more details regarding the ML approach, please refer to "Supporting Information Methods 1.")

2.3 | Literature review and pathway analysis

The top 50 genes (by relative importance values of features defining the best performing algorithm) were identified and analyzed further. Database for annotation, visualization, and integration discovery (DAVID) (Dennis et al., 2003; Huang, Sherman, & Lempicki, 2008, 2009) was used to search for pathways overrepresentation among these genes. (For more details regarding the analysis with DAVID, please refer to "Supporting Information Methods 2.")

3 | RESULTS

The supervised ML algorithm, gradient boosted trees (GBTs) with regularization, (XGBoost implementation) yielded the highest accuracy of 85.7% in predicting the occurrence of SCZ, outperforming in all categories other traditional classifiers such as SVMs, Random Forests, and Logistic Regression (Table 1). The *p* value of 9.11×10^{-179} indicates that XGBoost algorithm is performing better than a random predictor simply predicting the majority class. XGBoost achieved the highest specificity (86.6%), sensitivity (84.9%), precision (86.9%), and Recall (84.9%). We also used the F1 score to compare the predictor performance as it considers both the precision and recall to construct the metric. The XGBoost predictor achieved the highest F1 score (85.9%). The performance of the algorithm on the test data ($n = 1{,}526$) can be summarized in Table 1 and the confusion matrix depicted in Table 2. A receiver-operator characteristic (ROC) curve was plotted (Figure 1) and the area under the ROC curve (AUC) was 0.95, indicating a high accuracy of the test to correctly distinguish patients with SCZ from controls (of note, AUC = 1 indicates a perfectly discriminating test). The top 25 features (genes) according to each method are summarized in Figure 2. It is clear from the graph that the majority of the top 25 genes are shared across the different feature selection methods.

The relative importance of the 50 most important genes for the best performing algorithm, XGBoost, is graphically depicted in Figure 3. In addition, the relative importance, OMIM number, and function of each of these genes are summarized in Supporting Information Table 1. Analysis of the top 50 genes using DAVID, showed an overrepresentation of the KEGG pathway termed "MAPK signaling pathway" (hsa04010) (Kanehisa et al., 2017) based on genes *RASGRP2*, *RASGRP3*, *RASGRP4*, and *MYC*, which were present in our list. The JAK-STAT signaling pathway (*IL21R*, *IL21*, and *MYC* genes) was also overrepresented. A subsequent search was performed using only genes with relative importance higher than 0.4, and the cGMP-PKG signaling pathway, calcium ion signaling pathway (*PLN* and *SLC25A4* genes) were highlighted. Certain gene entries were not identified in DAVID as their role and functional annotations in the curated pathways in DAVID were incomplete (Huang et al., 2008, 2009).

Of note, some of the important genes from Supporting Information Table 1 exhibited mutations only in controls of our dataset (see Supporting Information Table 2).

TABLE 2 The performance of the best performing algorithm on the test data

n = 1526	Predicted: unaffected	Predicted: affected
Actual: control (unaffected)	TN = 645	FP = 100
Actual: case (affected)	FN = 118	TP = 663

FN = false negative; FP = false positive; TN = true negative; TP, true positive. The performance of the best performing algorithm on the test data can be summarized in the confusion matrix depicted above. The best supervised machine-learning algorithm was gradient boosted trees (GBTs) with regularization, which had 85.7% accuracy, 84.5% sensitivity, and 86.9% specificity.

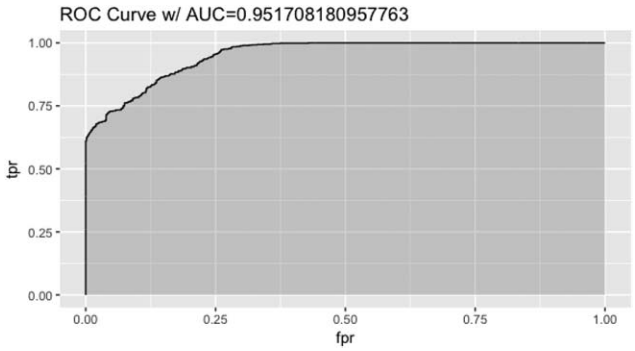

FIGURE 1 ROC curve for the best performing algorithm, XGBoost. A ROC curve is plotted with sensitivity values on the *Y*-axis and the corresponding false-positive ratio (1−specificity) on the *X*-axis. The AUC indicates the accuracy of a test for correctly distinguishing patients with psychosis from controls, where AUC = 1 indicates a perfectly discriminating test

The complete list of genes defining the best performing algorithm along with their relative importance to making correct prediction is listed in Supporting Information Table 3. Of note, Gain implies improvement in accuracy brought by a particular feature in splits of the trees in the model. Cover measures the relative number of observations concerning a feature, and frequency measures the relative times a feature was used in tree splits of the model. The Gain metric is the best indicator of variable importance in prediction accuracy, and so this table is sorted in decreasing Gain (Importance) value.

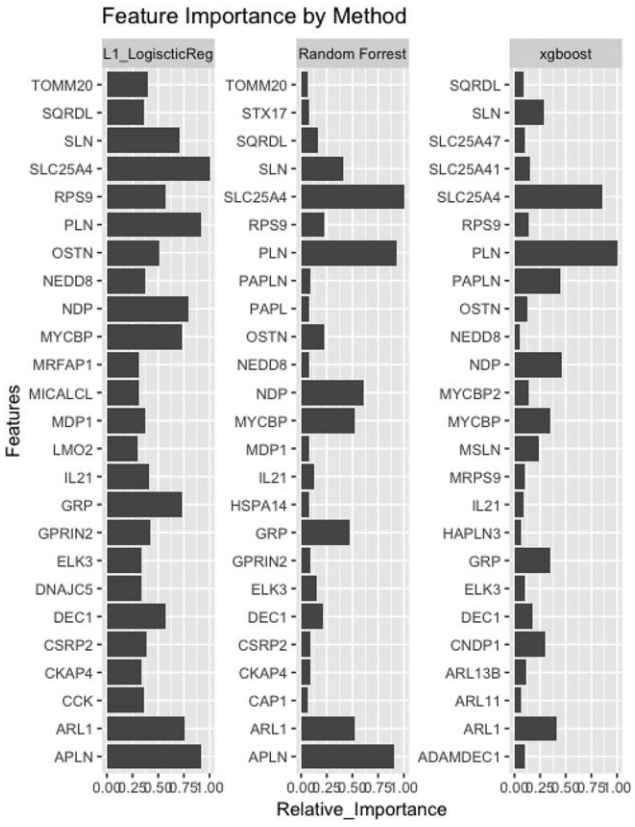

FIGURE 2 The top 25 features (genes) according to each ML method used. The top 25 features (genes) according to each method are summarized in above. This graph illustrates that the majority of top 25 genes are shared across the different feature selection methods

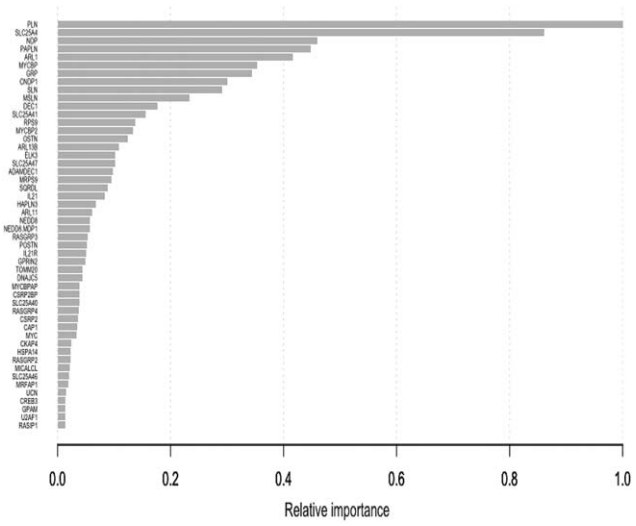

FIGURE 3 The relative importance of top 50 features (genes) for XGBoost algorithm. The relative importance of top 50 features (genes) in determining the predictions for the best supervised machine-learning algorithm, gradient boosted trees (GBTs) with regularization can be seen in the plot above

4 | DISCUSSION

We have presented a novel way for predicting individuals at high risk for SCZ based on exome data. The supervised ML algorithm, GBTs with regularization, yielded the highest accuracy of 85.7% in predicting the occurrence of SCZ, outperforming other traditional classifiers such as SVMs, Random Forests, and Logistic Regression. This algorithm was shown to provide advantages as demonstrated by higher F1 score (85.9%) when compared with the other widely used prediction algorithms. In brief, GBT is a tree-based ensemble method, which builds powerful predictive algorithms by training a series of tree-based classifiers, each attempting to correct the mistakes of the previous model, thus, iteratively improving the prediction performance of its weak (base) learners. It is important to emphasize that the evaluation of the prediction algorithms, employed a testing sample which was never used in training or feature selection steps. This method of evaluation resembles the expected outcomes in an application (clinical and research) environment, which can be considered encouraging given the relatively good performance of the algorithm (specificity: 86.6%, sensitivity: 84.9%, *AUC*: 0.95) for the testing set.

Looking inside the XGBoost algorithm, it should be highlighted that it only uses 372 features (genes) out of all the available features to make predictions. If we look back, we started with ~18,000 genes having at least one mutation in either cases or controls, and based on our approach only 372 of them are relevant in making ~86% accurate prediction. The complete list of these 372 genes with their relative

importance to making correct prediction is listed in Supporting Information Table 3. It is clear from Figure 2 that the majority of top 25 genes are shared across the different feature selection methods so looking a bit closer at the genes and pathways involved may give us new clues about the pathophysiology of SCZ.

When focusing on the top 50 genes of the XGBoost algorithm (Supporting Information Table 1), several were found to have evidence in the literature for a link with SCZ or play a role in neuropsychiatric diseases. For example, *GRP* encodes gastrin-releasing peptide (GRP), which is a mammalian neuropeptide and a homolog of the amphibian peptide bombesin. GRP regulates neurogenesis and neuronal development. In humans, GRP and bombesin like peptides (BLPs) bind preferentially to the GRP receptor (GRPR). This facilitates GRPR mediated signal transduction in the central nervous system, which plays a role in synaptic plasticity, memory, social interactions, and feeding behavior (Roesler & Schwartsmann, 2012). Studies using animal models have reported potential links between disrupted GRPR signaling pathways and psychosis (Kauer-Sant'Anna et al., 2007; Meller, Henriques, Schwartsmann, & Roesler, 2004). GRPR is believed to interact with other signaling pathways such as mitogen-activated protein kinase (MAPK) signaling and cAMP/PKA signaling and play a role in hippocampal memory formation (Roesler, Luft, et al., 2006). Mouse models of SCZ show elevated hippocampal neurogenesis with reduced neuronal maturity, which were reversed via GRP treatment, confirming GRP's role in these processes (Walton et al., 2014). In addition, SCZ patient have been reported to have lower BLP levels (Gerner, van Kammen, & Ninan, 1985; Meller et al., 2004). Hence, there are different lines of evidence supporting the role of *GRP* in SCZ.

MYCBP2 is another gene used by our predictor. It encodes MYC-binding protein 2, also known as protein associated with MYC (PAM), one of the PHR proteins involved in axonal development, as well as, in intracellular signaling pathways (Grill, Murphey, & Borgen, 2016). The PHR proteins (including MYCBP2) may act as cellular signaling hubs and their network dysregulation can lead to neurodevelopmental disorders such as SCZ (Grill et al., 2016; Wojda, Salinska, & Kuznicki, 2003). In addition, Wojda et al. (2003) reported an increased prevalence of autoantibodies that potentially target PAM and HSP60 in subjects with SCZ.

CNDP1 is also a gene with high relative importance in our predictor and substantial evidence for a link with SCZ in the literature. It encodes for carnosine dipeptidase 1 enzyme, which degrades carnosine and homocarnosine, both of which are believed to have neuroprotective and neurotransmitter functions in the brain (Teufel et al., 2003). Maccarrone et al. (2013) performed a proteomics study on the CSF of patients with MDD, BD, and SCZ compared to controls and found CNDP1 to be one of the reliable biomarkers to distinguish between all three neuropsychiatric disorders and controls.

Another gene that needs to be underlined is *ELK3*. Viana et al. (2016) found a differentially methylated position in *ELK3* to be associated with SCZ. A transcriptomic analysis study on gene networks and blood biomarkers using methamphetamine-associated psychosis (a model of SCZ), found *ELK3* along with *SINA3* to be top-scoring biomarkers. *SINA3* is known to be involved in the circadian clock function and *ELK3* was found to be co-expressed with SINA3, therefore suggesting its involvement in the circadian clock function too (Breen et al., 2016). Of note, disrupted circadian clock functions and sleep-wake cycles have been previously implicated in SCZ (Monti et al., 2013; Niculescu et al., 2000). Finally, other genes included in the predictor described in our study, which have been reported to be significantly differentially expressed in human subjects with SCZ or relevant animal models, include *ARL1* (Ibáñez et al., 2014), *CAP1* (Wong et al., 2005), *GPRIN2* (Narayan et al., 2008), *RASGRP3* (Arion, Horváth, Lewis, & Mirnics, 2010), and *CHI3L1* (Wojda et al., 2003; Zheng, Fu, Shen, & Xu, 2007).

Interestingly, some genes listed as highly important with regard to the XGBoost predictive algorithm showed variants, meeting our criteria, only in the controls of our dataset (Supporting Information Table 2). Several of these genes appear to have published evidence for a link with SCZ or other neuropsychiatric disease. For example, mutations in *NDP* can cause Norrie disease (OMIM #310600), a genetic condition associated with psychosis. In addition, polymorphisms in *NDP* may be risk factors for hallucinations and delusions in SCZ (Sun, Jayathilake, Zhao, & Meltzer, 2012). Finally, norrin, a cysteine rich protein encoded by the *NDP* gene, is known to activate the Wnt/beta-catenin pathway (Xu, Li, et al., 2012), which has been found to be associated with SCZ in genetic and postmortem studies (Cotter et al., 1998; Freyberg, Ferrando, & Javitch, 2009; Lovestone, Killick, Di Forti, & Murray, 2007; Miyaoka, Seno, & Ishino, 1999; Okerlund & Cheyette 2011; Yang et al., 2003; Zandi et al., 2008). Another gene in this group, *NEDD8*, is believed to be involved in proteolysis by the Ub-proteasome system (Mori et al., 2005), which has been implicated in the pathophysiology of SCZ (Rubio, Wood, Haroutunian, & Meador-Woodruff, 2013). Earlier studies showed that disruption to the ubiquitin proteasome pathway is one of the top pathways disrupted in SCZ (Bousman et al., 2010). In additon, a gene expression study of the superior temporal gyrus of subjects with SCZ found significantly altered expression of *NEDD8* in SCZ subjects (Bowden, Scott, & Tooney, 2008). Another gene of the predictor, which only had variants in the control subjects of our study and previous evidence of differential expression in patients with SCZ, is *CCK*. A polymorphism in *CCK* has been associated with the presence of hallucinations (Lenka, Arumugham, Christopher, & Pal, 2016) and the receptor of CCK has been found to be associated with positive symptoms in SCZ (Zheng et al., 2012). There is also evidence suggesting that CCK has a role in mediating neuronal gene expression and SCZ (Hansen et al., 2008) via changes in signaling (Curley & Lewis, 2012; Hashimoto et al., 2008) and the expression of transcription factors related to circadian rhythm control, the misregulation of which has been implicated in SCZ, as discussed above (Hansen et al., 2008; Monti et al., 2013).

Analysis of the top 50 genes used by the best performing algorithm, XGBoost, using DAVID software for pathway prediction, showed that the MAPK signaling pathway (hsa04010) and JAK-STAT signaling pathway (hsa04630) were overrepresented. MAPK-associated pathways are coupled with many neurotransmitter receptors and integrate extracellular and intracellular stimuli including peptide growth factors, cytokines, hormones, and various cellular stressors such as oxidative

stress and endoplasmic reticulum stress stimuli. These signaling pathways regulate a variety of cellular activities including proliferation, differentiation, survival, and death (Balu & Coyle, 2011; Crisafulli et al., 2015; Funk et al., 2012; Kim & Choi, 2010; Kyosseva et al., 1999; Perkins et al., 2007; Reichenberg, 2010; Sweatt, 2001; Walsh et al., 2008). Disturbances in the MAPK signaling are believed to affect Ca^{2+} homeostasis, neurotransmitter receptors, transcription factors, interactions between different signaling pathways and other neuroplasticity-related biological functions (Funk et al., 2012; Reichenberg, 2010; Sweatt, 2001). Studies on rare structural variants in genes involved in neurodevelopmental pathways (Walsh et al., 2008), microRNA expression (Perkins et al., 2007), and expression of proteins associated with MAPK signaling (Funk et al., 2012) have shown that disrupted MAPK signaling pathway is implicated in the pathophysiology of SCZ.

A search using only genes with relative importance higher than 0.4, showed that the *cGMP-PKG signaling pathway* (hsa04022) and *calcium ion signaling pathway* (hsa04020) were overrepresented. cGMP (cyclic guanosine monophosphate) is an intracellular, nucleotide second messenger that facilitates the action of natriuretic peptides (NPs) and nitric oxide (NO). *cGMP signaling cascade* is believed to play a role in synaptic functional plasticity and neurogenesis leading to an increased interest in cGMP associated pathways to find newer and better drug targets to treat neuropsychiatric diseases, like SCZ, where compromised neuroplasticity is believed to be important in their pathophysiology (Halene & Siegel, 2007; Kleppisch & Feil, 2009; Menniti et al., 2007; Schmidt et al., 2008; Shim et al., 2016; Reierson et al., 2011). *Calcium signaling* plays an important role in regulating a variety of neuronal processes such as neurotransmitter release, neuron excitability, synaptic plasticity which are further responsible for learning and memory (Wojda et al., 2008). Intracellular Ca^{2+} dyshomeostasis and disrupted signaling appears to be a common theme in cellular processes affected in SCZ (Ripke et al., 2013), such as dopamine, glutamate, γ-aminobutyric acid, and serotonin neurotransmission, oxidative stress, alterations in mitochondrial and cytosolic metabolism, dysregulation of myelination, and deficiencies in growth factors (Gunduz-Bruce, 2009; Howes & Kapur, 2009; Lewis & Moghaddam, 2006). A growing body of literature links disturbances and abnormalities in *intracellular signaling pathways* involving molecules such as *MAPK* (Balu & Coyle, 2011; Crisafulli et al., 2015; Funk et al., 2012; Kim & Choi, 2010; Kyosseva et al., 1999; Perkins et al., 2007; Reichenberg, 2010; Sweatt, 2001; Walsh et al., 2008), *cGMP* (Halene & Siegel, 2007; Kleppisch & Feil, 2009; Menniti et al., 2007; Schmidt et al., 2008; Shim et al., 2016; Reierson et al., 2011), and *Ca^{2+} signaling and homeostasis* (Berridge, 2012, 2013, 2014; Bojarski et al., 2010; Forstner et al., 2017; Gunduz-Bruce, 2009; Harrison, 2015; Howes & Kapur, 2009; Jimerson et al., 1979; Lewis & Moghaddam, 2006; Mäki-Marttunen et al., 2016; Purcell et al., 2014; Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Wojda et al., 2008) to neuropsychiatric disorders such as SCZ.

It should be highlighted that the great majority of the important genes had mutations in both cases and controls. Specifically, for each of the 346 out of the 372 (93%) relatively more important genes, at least 30% of individuals with mutations were cases and 30% were controls. This is consistent with the literature supporting a polygenic paradigm for SCZ, involving different susceptibility genes. Variants in susceptibility genes can modify the risk for SCZ but no single variant is sufficient to independently cause the disease.

4.1 | Possible interpretation in the context of the threshold hypothesis

As discussed above, Genovese et al. (2016) showed that the frequency of rare protein damaging variants was significantly increased in SCZ cases relative to controls. The observed significant excess of variants in individuals with SCZ were mostly inherited from unaffected parents and LOF variants were more likely to be transmitted to affected individuals than controls, especially private LOF variants in genes intolerant to functional genetic variation. The "threshold hypothesis" posits that SCZ is manifested when an individual's combined liability crosses over a certain fixed threshold value (McGue, Gottesman, & Rao, 1983). We postulate that affected individuals have randomly accumulated (mostly inherited) genetic variants in a specific combination of SCZ susceptibility genes, which cooperatively increase one's predisposition to SCZ. This could be one possible explanation for the constant prevalence of SCZ despite the reduced reproductive fitness of affected individuals, but also for the observed high heritability of SCZ despite the absence of major SCZ susceptibility loci. Published literature suggests that SCZ is not manifested by mutations in a fixed network of susceptibility genes but rather by mutations in networks of genes that can vary across affected individuals in different populations (International Schizophrenia Consortium, 2008, 2009; Ripke et al., 2013). It is thus possible that the predisposition of a child to be affected is mainly influenced by the mutations that his/her parents, *as a couple rather than individually*, carry on the corresponding susceptibility genes. An offspring who inherits sufficient variants in one of the SCZ networks of genes will be predisposed to SCZ. This mechanism of inheritance would lead to a genetically heterogeneous population of patients with SCZ. However, it would also imply that ML can be used to identify individuals at high risk to develop the disease, as demonstrated by our study. In summary, we do not believe that SCZ genes in our predictor can be separated in susceptibility versus protective genes but rather hypothesize that the role of each gene depends on the environmental and genomic background of the corresponding patient (e.g., the presence of copy number variants [CNVs] or point mutations in other genes).

4.2 | Limitations

It is important to mention some limitations of the present study. Given, we do not have access to the clinical characteristics of the patients with SCZ in our analysis, we are not sure if this is a representative group of patients in the general population and thus how generalizable our findings are in other patients. At the same time, studies often use performance measures (AUC, accuracy, sensitivity, and specificity) based only on cross validation. In our case, by evaluating the performance of the prediction model using an independent testing set, albeit from the same population, we can estimate the

true expected performance of the model when used in a clinical or research environment for early identification of individuals at high risk for SCZ. However, it is important to emphasize that the population included in this analysis represents a select group of individuals motivated to participate in research and may not represent the general population.

5 | FUTURE DIRECTIONS AND CONCLUSION

A genomic study by Pocklington et al. (2015), which focused on CNVs in 11,355 cases and 16,416 controls, found that large, rare CNVs (>100 kb, frequency <1%) in affected individuals are enriched for genes involved in both GABAergic and glutamatergic neurotransmission. Likewise, a study focusing on genes differentially expressed in patients with SCZ relative to controls highlighted different pathways, including transmission across chemical synapses, postsynaptic membrane, voltage-gated potassium channel complexes, and axon guidance (Fromer et al., 2016). This study used sequenced RNA from dorsolateral prefrontal cortex of subjects with SCZ ($N = 258$) and control subjects ($N = 279$), available through the CommonMind Consortium dataset. A total of 693 genes differentially expressed in cases versus controls were reported and the differential expression of these genes was further validated using an independent sample from the human brain collection core, consisting of 131 SCZ subjects and 176 controls. Another transcriptome study, performed by Gulsuner et al. (2013), found that DNMs in SCZ have a spatial and temporal mapping to a fetal prefrontal cortical network. DNMs were identified in persons with SCZ and the genes harboring them were mapped onto transcriptome profiles of normal human brain tissues from age 13 weeks gestation to adulthood. In the dorsolateral and ventrolateral prefrontal cortex during fetal development, these genes formed a network significantly enriched for transcriptional coexpression and protein interaction. The 50 genes in this network function in neuronal migration, synaptic transmission, signaling, and transcriptional regulation. This study not only underlined the role that disruptions of fetal prefrontal cortical neurogenesis have in SCZ, but also exemplified how genomic and transcriptome data can be combined to better understand the pathophysiology of SCZ.

Finally, it is important to highlight that environment plays an important role in SCZ risk. For example, epidemiological studies have found evidence that difficult social environments such as severe bullying could increase one's risk for psychosis twofold (Schreier et al., 2009). In addition, several in utero and postnatal environmental threats are thought to be interacting with genetics to increase one's predisposition to disease (Brown et al., 2004; Schreier et al., 2009). A nested case-control study studying the effect of maternal infection (in this case influenza) on SCZ risk investigated birth records in California between 1959 and 1966 and determined the diagnosis of children 40 years later. After confirming the presence of influenza anti-body in maternal serum, individuals exposed to influenza in the first trimester were found to have a sevenfold increase risk for SCZ (Brown et al., 2004). Although, viruses do not usually penetrate the placenta, it is believed that maternal immune response indirectly impacts the fetus through exposure to proinflammatory cytokines. The investigation of 17 adult SCZ cases that were exposed to increased level of interluken-8 in utero had changes to the structure of their brain recorded by magnetic resonance imaging that were consistent with SCZ (Ellman et al., 2010).

Our manuscript presents an algorithm to predict patients at high risk for SCZ based on rare, predicted functional, variants with ~86% accuracy (*AUC*: 0.95) and a novel approach for exploring the genetics of different (neuropsychiatric) conditions with complex inheritance. It would be important to replicate our findings, and the overall approach, not only in large independent SCZ studies, including cases with matched controls, but also in family trios (Singh et al., 2016; Takata et al., 2014). Taking the studies mentioned above into consideration, the generalizability of our model can potentially be improved by training the predictor on different levels of data acquired from multiple cohorts. Given the role of environment, it is unlikely to find a genetic test that predicts with 100% certainty the individuals who will develop SCZ solely based on genomic data. If, however, such a predictive algorithm is trained using genomic data (including both exonic mutations and CNVs), transcriptome data, and ML approaches using other phenotypic information, such as speech (Bedi et al., 2015) or neuroimaging data (Gheiratmand et al., 2017), it might be possible to effectively prioritize high risk individuals meriting early clinical evaluation and/or participation in research studies. This would provide an opportunity to prospectively explore the impact of different genetic and environmental factors on an individual's risk for SCZ and explore the role of early interventions, such as stress management counseling, in decreasing this risk.

ACKNOWLEDGMENTS

We thank the Montreal Children's Hospital Foundation for financially supporting Dr. Trakadis' research. Data used in the preparation of this manuscript were obtained from the Database of genotypes and phenotypes (dbGaP) after McGill IRB approval. Raw data used are available in study phs000473.v1.p1. We thank the authors and dbGaP for access to the WES data. We thank Bill Qi and Alexandre Dionne Laporte for bioinformatics support. The variants were segregated using an in-house program from Dr. Rouleau's team. Dr. Trakadis would like to acknowledge, and thank, Dr. Rouleau for his continuous support in his research.

AUTHORS' CONTRIBUTIONS

YT designed the study and put the paper together. VF, AC, and SS developed the scripts used for variant prioritization under the supervision of YT. SS performed the machine-learning analysis. AC and AK performed the literature review for the genes identified. All authors reviewed and provided feedback on the manuscript.

CONFLICT OF INTEREST

The authors declared that they have no conflict of interest.

ORCID

Yannis J. Trakadis http://orcid.org/0000-0002-8740-4416

REFERENCES

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H., & Wjst, M. (2001). Genomewide scans of complex human diseases: True linkage is hard to find. *American Journal of Human Genetics*, *69*(5), 936–950.

Arion, D., Horváth, S., Lewis, D. A., & Mirnics, K. (2010). Infragranular gene expression disturbances in the prefrontal cortex in schizophrenia: Signature of altered neural development? *Neurobiology of Disease*, *37*(3), 738–746.

Balu, D. T., & Coyle, J. T. (2011). Neuroplasticity signaling pathways linked to the pathophysiology of schizophrenia. *Neuroscience and Biobehavioral Reviews*, *35*(3), 848–870.

Bassett, A. S., McAlduff, J., Bury, A., Bindseil, K., Hodgkinson, K. A., & Honer, W. G. (1995). Reproductive fitness in familial schizophrenia. *Schizophrenia Research*, *15*(1–2), 35–36.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., … Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophrenia*, *1*(1), 15030.

Berridge, M. J. (2012). Calcium signalling remodelling and disease. *Biochemical Society Transactions*, *40*(2), 297–309.

Berridge, M. J. (2013). Dysregulation of neural calcium signaling in Alzheimer disease, bipolar disorder and schizophrenia. *Prion*, *7*(1), 2–13.

Berridge, M. J. (2014). Calcium signalling and psychiatric disease: Bipolar disorder and schizophrenia. *Cell and Tissue Research*, *357*(2), 477–492.

Bojarski, L., Debowska, K., & Wojda, U. (2010). In vitro findings of alterations in intracellular calcium homeostasis in schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *34*(8), 1367–1374.

Bousman, C. A., Chana, G., Glatt, S. J., Chandler, S. D., May, T., Lohr, J., … Everall, I. P. (2010). Positive symptoms of psychosis correlate with expression of ubiquitin proteasome genes in peripheral blood. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153B*(7), 1336–1341.

Bowden, N. A., Scott, R. J., & Tooney, P. A. (2008). Altered gene expression in the superior temporal gyrus in schizophrenia. *BMC Genomics*, *9*(1), 199.

Breen, M. S., Uhlmann, A., Nday, C. M., Glatt, S. J., Mitt, M., Metsalpu, A., … Illing, N. (2016). Candidate gene networks and blood biomarkers of methamphetamine-associated psychosis: An integrative RNA-sequencing report. *Translational Psychiatry*, *6*(5), e802.

Brown, A. S., Begg, M. D., Gravenstein, S., Schaefer, C. A., Wyatt, R. J., Bresnahan, M., … Susser, E. S. (2004). Serologic evidence of prenatal influenza in the etiology of schizophrenia. *Archives of General Psychiatry*, *61*(8), 774–780.

Cardno, A. G., & Gottesman, I. I. (2000). Twin studies of schizophrenia: From bow-and-arrow concordances to star wars Mx and functional genomics. *American Journal of Medical Genetics Part Genetics*, *97*(1), 12–17.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining, San Francisco, CA. pp. 785–794.

Chen, T., & He, T. (2015). *Xgboost: Extreme gradient boosting*. https://github.com/dmlc/xgboost, R package version 0.4-2.

Cotter, D., Kerwin, R., Al-Sarraji, S., Brion, J. P., Chadwich, A., Lovestone, S., … Everall, I. (1998). Abnormalities of Wnt signalling in schizophrenia–Evidence for neurodevelopmental abnormality. *Neuroreport*, *9*(7), 1379–1383.

Crisafulli, C., Drago, A., Calabrò, M., Spina, E., & Serretti, A. (2015). A molecular pathway analysis informs the genetic background at risk for schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *59*, 21–30.

Curley, A. A., & Lewis, D. A. (2012). Cortical basket cell dysfunction in schizophrenia. *The Journal of Physiology*, *590*(4), 715–724.

DeLisi, L. E., Shaw, S. H., Crow, T. J., Shields, G., Smith, A. B., Larach, V. W., … Stewart, J. (2002). A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. *The American Journal of Psychiatry*, *159*(5), 803–812.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, *4*(9), R60.

Ellman, L. M., Deicken, R. F., Vinogradov, S., Kremen, W. S., Poole, J. H., Kern, D. M., … Brown, A. S. (2010). Structural brain alterations in schizophrenia following fetal exposure to the inflammatory cytokine interleukin-8. *Schizophrenia Research*, *121*(1–3), 46–54.

Forstner, A. J., Hecker, J., Hofmann, A., Maaser, A., Reinbold, C. S., Mühleisen, T. W., … Mattheisen, M. (2017). Identification of shared risk loci and pathways for bipolar disorder and schizophrenia. *PloS One*, *12*(2), e0171595.

Freyberg, Z., Ferrando, S. J., & Javitch, J. A. (2009). Roles of the Akt/GSK-3 and Wnt signaling pathways in schizophrenia and antipsychotic drug action. *The American Journal of Psychiatry*, *167*(4), 388–396.

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., … Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442–1453.

Funk, A. J., McCullumsmith, R. E., Haroutunian, V., & Meador-Woodruff, J. H. (2012). Abnormal activity of the MAPK-and cAMP-associated signaling pathways in frontal cortical areas in postmortem brain in schizophrenia. *Neuropsychopharmacology*, *37*(4), 896–905.

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landén, M., … McCarroll, S. A. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience*, *19*(11), 1433–1441.

Gerner, R. H., van Kammen, D. P., & Ninan, P. T. (1985). Cerebrospinal fluid cholecystokinin, bombesin and somatostatin in schizophrenia and normals. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *9*(1), 73–82.

Gheiratmand, M., Rish, I., Cecchi, G. A., Brown, M. R. G., Greiner, R., Polosecki, P. I., … Dursun, S. M. (2017). Learning stable and predictive network-based patterns of schizophrenia and its clinical symptoms. *NPJ Schizophrenia*, *3*(1), 22. 16

Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., … Thibodeau, P. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genetics*, *43*(9), 860–863.

Grill, B., Murphey, R. K., & Borgen, M. A. (2016). The PHR proteins: Intracellular signaling hubs in neuronal development and axon degeneration. *Neural Development*, *11*(1), 8.

Gulsuner, S., Walsh, T., Watts, A. C., Lee, M. K., Thornton, A. M., Casadei, S., … McClellan, J. M. (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, *154*(3), 518–529.

Gunduz-Bruce, H. (2009). The acute effects of NMDA antagonism: From the rodent to the human brain. *Brain Research Reviews*, *60*(2), 279–286.

Halene, T. B., & Siegel, S. J. (2007). PDE inhibitors in psychiatry–future options for dementia, depression and schizophrenia? *Drug Discovery Today*, *12*(19–20), 870–878.

Hansen, T. V. O., Borup, R., Marstrand, T., Rehfeld, J. F., & Nielsen, F. C. (2008). Cholecystokinin-2 receptor mediated gene expression in neuronal PC12 cells. *Journal of Neurochemistry*, *104*(6), 1450–1465.

Harrison, P. J. (2015). Recent genetic findings in schizophrenia and their therapeutic relevance. *Journal of Psychopharmacology*, *29*(2), 85–96.

Hashimoto, T., Arion, D., Unger, T., Maldonado-Aviles, J. G., Morris, H. M., Volk, D. W., … Lewis, D. A. (2008). Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. *Molecular Psychiatry*, *13*(2), 147–161.

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: Version III—The final common pathway. *Schizophrenia Bulletin*, *35*(3), 549–562.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57.

Ibáñez, K., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A., & Valencia, A. (2014). Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. *PLoS Genetics*, *10*(2), e1004173.

International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, *455* (7210), 237.

International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748.

Jimerson, D. C., Post, R. M., Carman, J. S., Van Kammen, D. P., Wood, J. H., Goodwin, F. K., & Bunney, W. E. (1979). CSF calcium: Clinical correlates in affective illness and schizophrenia. *Biological Psychiatry*, *14*(1), 37–51.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361.

Kauer-Sant'anna, M., Andreazza, A. C., Valvassori, S. S., Martins, M. R., Barbosa, L. M., Schwartsmann, G., … Kapczinski, F. (2007). A gastrin-releasing peptide receptor antagonist blocks D-amphetamine-induced hyperlocomotion and increases hippocampal NGF and BDNF levels in rats. *Peptides*, *28*(7), 1447–1452.

Kim, E. K., & Choi, E. J. (2010). Pathological roles of MAPK signaling pathways in human diseases. *Biochimica Et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1802*(4), 396–405.

Kleppisch, T., & Feil, R. (2009). cGMP signalling in the mammalian brain: Role in synaptic plasticity and behaviour. *Handbook of Experimental Pharmacology*, *191*, 549–579.

Kyosseva, S. V., Elbein, A. D., Griffin, W. S. T., Mrak, R. E., Lyon, M., & Karson, C. N. (1999). Mitogen-activated protein kinases in schizophrenia. *Biological Psychiatry*, *46*(5), 689–696.

Lenka, A., Arumugham, S. S., Christopher, R., & Pal, P. K. (2016). Genetic substrates of psychosis in patients with Parkinson's disease: A critical review. *Journal of the Neurological Sciences*, *364*, 33–41.

Lewis, D. A., & Moghaddam, B. (2006). Cognitive dysfunction in schizophrenia: Convergence of γ-aminobutyric acid and glutamate alterations. *Archives of Neurology*, *63*(10), 1372–1376.

Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., … Xia, K. (2016). Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NP denovo database. *Molecular Psychiatry*, *21*(2), 298–297.

Lovestone, S., Killick, R., Di Forti, M., & Murray, R. (2007). Schizophrenia as a GSK-3 dysregulation disorder. *Trends in Neurosciences*, *30*(4), 142–149.

Maccarrone, G., Ditzen, C., Yassouridis, A., Rewerts, C., Uhr, M., Uhlen, M., … Turck, C. W. (2013). Psychiatric patient stratification using biosignatures based on cerebrospinal fluid protein expression clusters. *Journal of Psychiatric Research*, *47*(11), 1572–1580.

Mäki-Marttunen, T., Halnes, G., Devor, A., Witoelar, A., Bettella, F., Djurovic, S., … Dale, A. M. (2016). Functional effects of schizophrenia-linked genetic variants on intrinsic single-neuron excitability: A modeling study. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(1), 49–59.

McGue, M., Gottesman, I. I., & Rao, D. C. (1983). The transmission of schizophrenia under a multifactorial threshold model. *American Journal of Human Genetics*, *35*(6), 1161–1178.

Meller, C. A., Henriques, J. A. P., Schwartsmann, G., & Roesler, R. (2004). The bombesin/gastrin releasing peptide receptor antagonist RC-3095 blocks apomorphine but not MK-801-induced stereotypy in mice. *Peptides*, *25*(4), 585–588.

Menniti, F. S., Chappie, T. A., Humphrey, J. M., & Schmidt, C. J. (2007). Phosphodiesterase 10A inhibitors: A novel approach to the treatment of the symptoms of schizophrenia. *Current Opinion in Investigational Drugs*, *8*(1), 54–59.

Miyaoka, T., Seno, H., & Ishino, H. (1999). Increased expression of Wnt-1 in schizophrenic brains. *Schizophrenia Research*, *38*(1), 1–6.

Monti, J. M., BaHammam, A. S., Pandi-Perumal, S. R., Bromundt, V., Spence, D. W., Cardinali, D. P., & Brown, G. M. (2013). Sleep and circadian rhythm dysregulation in schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *43*, 209–216.

Mori, F., Nishie, M., Piao, Y. S., Kito, K., Kamitani, T., Takahashi, H., & Wakabayashi, K. (2005). Accumulation of NEDD8 in neuronal and glial inclusions of neurodegenerative disorders. *Neuropathology and Applied Neurobiology*, *31*(1), 53–61.

Narayan, S., Tang, B., Head, S. R., Gilmartin, T. J., Sutcliffe, J. G., Dean, B., & Thomas, E. A. (2008). Molecular profiles of schizophrenia in the CNS at different stages of illness. *Brain Research*, *1239*, 235–248.

Niculescu, A. B., Segal, D. S., Kuczenski, R., Barrett, T., Hauger, R. L., & Kelsoe, J. R. (2000). Identifying a series of candidate genes for mania and psychosis: A convergent functional genomics approach. *Physiological Genomics*, *4*(1), 83–91.

Okerlund, N. D., & Cheyette, B. N. (2011). Synaptic Wnt signaling—A contributor to major psychiatric disorders? *Journal of Neurodevelopmental Disorders*, *3*(2), 162–174.

Perkins, D. O., Jeffries, C. D., Jarskog, L. F., Thomson, J. M., Woods, K., Newman, M. A., … Hammond, S. M. (2007). microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biology*, *8*(2), R27.

Pocklington, A. J., Rees, E., Walters, J. T., Han, J., Kavanagh, D. H., Chambert, K. D., … Owen, M. J. (2015). Novel findings from CNVs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron*, *86*(5), 1203–1214.

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., … Duncan, L. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, *506*(7487), 185–190.

Reichenberg, A. A. (2010). The assessment of neuropsychological functioning in schizophrenia. *Dialogues in Clinical Neuroscience*, *12*(3), 383.

Reierson, G. W., Guo, S., Mastronardi, C., Licinio, J., & Wong, M. L. (2011). cGMP signaling, phosphodiesterases and major depressive disorder. *Current Neuropharmacology*, *9*(4), 715–727.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45* (10), 1150–1159.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, *405*(6788), 847–856.

Roesler, R., Luft, T., Oliveira, S. H., Farias, C. B., Almeida, V. R., Quevedo, J., … Schwartsmann, G. (2006). Molecular mechanisms mediating gastrin-releasing peptide receptor modulation of memory consolidation in the hippocampus. *Neuropharmacology*, *51*(2), 350–357.

Roesler, R., & Schwartsmann, G. (2012). Gastrin-releasing peptide receptors in the central nervous system: Role in brain function and as a drug target. *Frontiers in Endocrinology*, *3*, 159. 17

Rubio, M. D., Wood, K., Haroutunian, V., & Meador-Woodruff, J. H. (2013). Dysfunction of the ubiquitin proteasome and ubiquitin-like systems in schizophrenia. *Neuropsychopharmacology*, *38*(10), 1910–1920.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427.

Schmidt, C. J., Chapin, D. S., Cianfrogna, J., Corman, M. L., Hajos, M., Harms, J. F., … Proulx-LaFrance, C. (2008). Preclinical characterization of selective phosphodiesterase 10A inhibitors: A new therapeutic approach to the treatment of schizophrenia. *The Journal of Pharmacology and Experimental Therapeutics*, *325*(2), 681–690.

Schreier, A., Wolke, D., Thomas, K., Horwood, J., Hollis, C., Gunnell, D., … Salvi, G. (2009). Prospective study of peer victimization in childhood and psychotic symptoms in a nonclinical population at age 12 years. *Archives of General Psychiatry*, *66*(5), 527–536.

Shim, S., Shuman, M., & Duncan, E. (2016). An emerging role of cGMP in the treatment of schizophrenia: A review. *Schizophrenia Research*, *170*(1), 226–231.

Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., … Barrett, J. C. (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nature Neuroscience*, *19*(4), 571–577.

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: Evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*, *60*(12), 1187–1192.

Sun, J., Jayathilake, K., Zhao, Z., & Meltzer, H. Y. (2012). Investigating association of four gene regions (GABRB3, MAOB, PAH, and SLC6A4) with five symptoms in schizophrenia. *Psychiatry Research*, *198*(2), 202–206.

Sweatt, J. D. (2001). The neuronal MAP kinase cascade: A biochemical signal integration system subserving synaptic plasticity and memory. *Journal of Neurochemistry*, *76*(1), 1–10.

Takata, A., Xu, B., Ionita-Laza, I., Roos, J. L., Gogos, J. A., & Karayiorgou, M. (2014). Loss-of-function variants in schizophrenia risk and SETD1A as a candidate susceptibility gene. *Neuron*, *82*(4), 773–780.

Team RC. (2016). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Teufel, M., Saudek, V., Ledig, J. P., Bernhardt, A., Boularand, S., Carreau, A., … Smirnova, T. (2003). Sequence identification and characterization of human carnosinase and a closely related non-specific dipeptidase. *Journal of Biological Chemistry*, *278*(8), 6521–6531.

Viana, J., Hannon, E., Dempster, E., Pidsley, R., Macdonald, R., Knox, O., … Mill, J. (2016). Schizophrenia-associated methylomic variation: Molecular signatures of disease and polygenic risk burden across multiple brain regions. *Human Molecular Genetics*, *26*(1), 210–225.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., … Stray, S. M. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, *320*(5875), 539–543.

Walton, N. M., de Koning, A., Xie, X., Shin, R., Chen, Q., Miyake, S., … Tamura, K. (2014). Gastrin-releasing peptide contributes to the regulation of adult hippocampal neurogenesis and neuronal development. *Stem Cells*, *32*(9), 2454–2466.

Wang, X. F., Wang, D., Zhu, W., Delrahim, K. K., Dolnak, D., & Rapaport, M. H. (2003). Studies characterizing 60 kda autoantibodies in subjects with schizophrenia. *Biological Psychiatry*, *53*(5), 361–375.

Wojda, U., Salinska, E., & Kuznicki, J. (2008). Calcium ions in neuronal degeneration. *IUBMB Life*, *60*(9), 575–590.

Wong, A. H., Lipska, B. K., Likhodi, O., Boffa, E., Weinberger, D. R., Kennedy, J. L., & Van Tol, H. H. (2005). Cortical gene expression in the neonatal ventral-hippocampal lesion rat model. *Schizophrenia Research*, *77*(2–3), 261–270.

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., … Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*, *44*(12), 1365–1369.

Xu, L. M., Li, J. R., Huang, Y., Zhao, M., Tang, X., & Wei, L. (2012). AutismKB: An evidence-based knowledgebase of autism genetics. *Nucleic Acids Research*, *40*(D1), D1016–D1022. https://doi.org/10.1093/nar/gkr1145

Yang, J., Si, T., Ling, Y., Ruan, Y., Han, Y., Wang, X., … Zhang, D. (2003). Association study of the human FZD3 locus with schizophrenia. *Biological Psychiatry*, *54*(11), 1298–1301.

Zandi, P. P., Belmonte, P. L., Willour, V. L., Goes, F. S., Badner, J. A., Simpson, S. G., … Potash, J. B. (2008). Association study of Wnt signaling pathway genes in bipolar disorder. *Archives of General Psychiatry*, *65*(7), 785–793.

Zhao, X., Tang, R., Gao, B., Shi, Y., Zhou, J., Guo, S., … Li, S. (2007). Functional variants in the promoter region of Chitinase 3–Like 1 (CHI3L1) and susceptibility to schizophrenia. *American Journal of Human Genetics*, *80*(1), 12–18.

Zheng, C., Fu, Q., Shen, Y., & Xu, Q. (2012). Investigation of allelic heterogeneity of the CCK-A receptor gene in paranoid schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *159B*(6), 741–747.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Trakadis YJ, Sardaar S, Chen A, Fulginiti V, Krishnan A. Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. *Am J Med Genet Part B*. 2019;180B:103–112. https://doi.org/10.1002/ajmg.b.32638