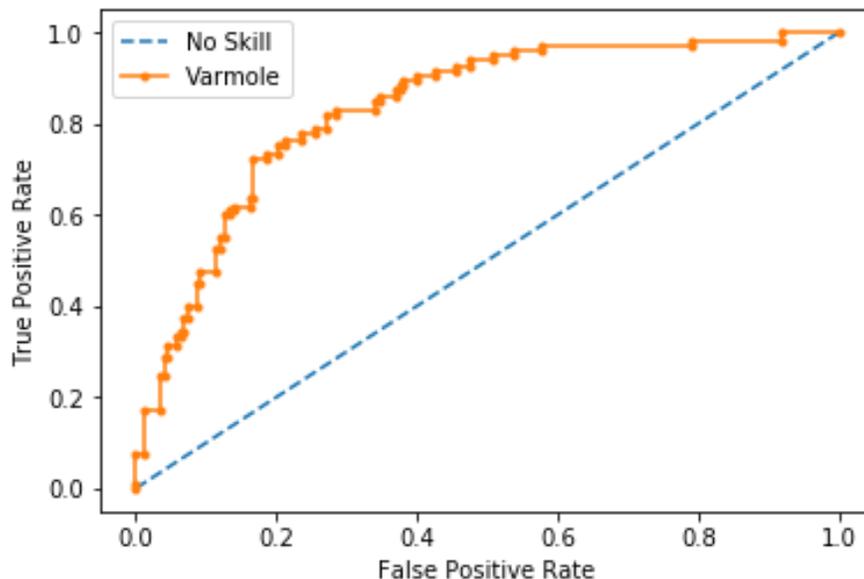


S1. The state-of-the-art classification methods.

We compared Varmole with the following state-of-the-art classifiers: fully connected neural network (NN), k-nearest neighbors (kNN), random forests (RF), Decision Tree (DT), Gaussian Process (GP), AdaBoost, restricted Boltzmann machine (RBM), Gaussian naive Bayes (GNB), linear support vector machine (SVM), SVM with RBF kernel, SVM with L1 penalty, and logistic regression (LR) with L1 penalty. We implemented these methods using the scikit-learn library in Python at <https://scikit-learn.org/stable/> (Pedregosa *et al.*, 2011).

S2. Training of Varmole for classifying schizophrenia

After hyperparameter tuning, we came up with a deep neural network with ReLU activation functions consisting of: (1) the input layer containing 129902 features (2598 genes + 127304 SNPs); (2) the layers above containing 2598x129902, 1000x2598, 500x1000 and 500x1 connections, respectively; (3) the output layer predicting schizophrenia by a probability (i.e., via a single-unit with a sigmoid activation). Other hyperparameters were optimized as follows: Train batch size = 50; Learning rate = 0.001 and Weight decay = 0.01 (with Adam method (Kingma and Ba, 2014)); -1 penalty parameter: 0.0001; Number of training epoch: 60. The performance metrics are as follows: Balanced Accuracy (BACC) = 0.77, Sensitivity = 0.74, Specificity = 0.80, AUROC score = 0.77, The ROC curve is as the following figure:



The mean BACC of other classification method (S1) was as follows: kNN = 0.49, Linear SVM = 0.57, RBF SVM = 0.5, GP = 0.50, DT = 0.62, RF = 0.50, NN = 0.54, AdaBoost = 0.65, GNB = 0.56, SVM with L1 = 0.55, LR with L1 = 0.56, DSPN (RBM-based) = 0.74 (Wang, *et al.*, 2018)

In addition, we compared Varmole with the polygenic risk score (PRS). A recent population study for four cohorts found that the AUROC values of PRS for schizophrenia varied around 0.6 (Zheutlin *et al.* 2019). We calculated the AUROC value of Varmole which is 0.77, showing that the Varmole outperforms PRS for classifying schizophrenia.

S3. Application of Varmole for classifying lung cancer stages.

In addition to classifying schizophrenia, we also applied Varmole to the lung cancer patients for predicting early and late cancer stages. In particular, we used the TCGA genotype and gene expression data (Cancer Genome Atlas Research Network *et al.*, 2013) for two major lung cancer subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). We also used the eQTLs of lung tissue from the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2013), to link SNPs to genes in the Varmole model. The lung cancer gene regulatory network (GRN) linking TFs to target genes was obtained from a recent large-scale study (Zhang *et al.*, 2018). After filtering out SNPs and genes to match with those in the lung eQTLs and lung cancer GRN, we finally input the genotype data of 332 SNPs and the expression data of 118 genes from 121 LUAD patients and 127 LUSC patients for classifying early and late stages. In detail, we grouped patients based on their TMN stages, with (I+IA+IB) as the early stage (N = 133) and II, III, and IV as the late stages (N = 115).

We found that Varmole has also achieved very high classification accuracy (BACC=0.89). The BACC values of other classification methods are NN=0.79, RF=0.82, SVM=0.5, RBM=0.71, AdaBoost=0.77, NB =0.72, LR=0.72, DT=0.70, GP=0.81.

S4. Enrichment analyses of top prioritized genes by Varmole

We performed the enrichment analysis to reveal the enriched pathways and functions among top prioritized genes by Varmole (Supplemental Table S3). In particular, we inputted the top 250 genes (~top 10% of our input genes, which was used as the background) into G:Profiler, a widely used webapp (Raudvere *et al.*, 2019) and reported the enriched terms with the Benjamini-Hochberg FDR < 0.05.

S5. Comparison with the chromatin interaction data in the human brain.

The recent PsychENCODE project has generated the Hi-C data and identified the enhancer-promoter interactions for the human front cortex (Wang *et al.*, 2018) as follows. First, they identified 149,098 promoter-based interactions (“region” to “promoter” pairs) that significantly higher interaction frequencies than a background Weibull distribution (FDR<0.01). Second, they found the promoter-interacting regions that significantly overlapped with the enhancers for the human brain front cortex (Roadmap Epigenomics Consortium *et al.*, 2015); i.e., “enhancer-overlapped region”. Finally, they obtained 90,015 promoter-based interactions that have the enhancer-overlapped regions as a list of Hi-C derived enhancer-promoter interactions (“enhancer-overlapped region” to “promoter” pairs). Additional details about the Hi-C data processing and analysis are available in the supplemental materials of the paper (Wang *et al.*, 2018).

We further compared the SNP-gene pairs from Varmole for schizophrenia with those Hi-C derived enhancer-promoter interactions. In particular, we overlapped SNPs with the enhancer-overlapped regions from the interactions, and the genes with their promoters. Given that the regions are typically small (many are a few dozen or hundred bases only), we extended them by +/- 20 kb to overlap SNPs. Finally, we found that the overlapped SNP-gene pairs have significantly higher importance scores than the rest of the pairs (t-test $p < 5e-5$).

References

Cancer Genome Atlas Research Network *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Liberzon A. *et al.* (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, 1(6): 417–425.

Pedregosa, F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Raudvere, U. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists. *Nucleic Acids Research*, 47(W1):W191-W198

Roadmap Epigenomics Consortium *et al.*, (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518, pages317–330

Wang, D. *et al.* (2018) Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362, 6420.

Zhang, S. *et al.* (2018) Landscape of transcriptional deregulation in lung cancer. *BMC Genomics*, **19**, 435.

Zheutlin, A. B. *et al.* (2019) “Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems.” *Am J Psychiatry* 176 (10): 846–55.