### Title

Combinatorial Approach for Complex Disorder Prediction: Case Study of Neurodevelopmental Disorders

### Permalink

https://escholarship.org/uc/item/9z9219bt

### Journal

Genetics, 210(4)

### ISSN

0016-6731

### Authors

Huynh, Linh
Hormozdiari, Fereydoun

### Publication Date

2018-12-01

### DOI

10.1534/genetics.118.301280

Peer reviewed

# Combinatorial Approach for Complex Disorder Prediction: Case Study of Neurodevelopmental Disorders

**Linh Huynh*** **and Fereydoun Hormozdiari***,†,‡,[1]
*Genome Center, †MIND Institute, and ‡Department of Biochemistry and Molecular Medicine, University of California, Davis, California 95817

**ABSTRACT** Early prediction of complex disorders (*e.g.*, autism and other neurodevelopmental disorders) is one of the fundamental goals of precision medicine and personalized genomics. An early prediction of complex disorders can improve the prognosis, increase the effectiveness of interventions and treatments, and enhance the life quality of affected patients. Considering the genetic heritability of neurodevelopmental disorders, we are proposing a novel framework for utilizing rare coding variation for early prediction of these disorders in subset of affected samples. We provide a combinatorial framework for addressing this problem, denoted as Odin (Oracle for DIsorder predictioN), to make a prediction for a small, yet significant, subset of affected cases while having very low false positive rate (FPR) prediction for unaffected samples. Odin also takes advantage of the available functional information (*e.g.*, pairwise coexpression of genes during brain development) to increase the prediction power beyond genes with recurrent variants. Application of our method accurately recovers an additional 8% of autism cases without any severe variant in known recurrent mutated genes with a <1% FPR. Furthermore, Odin predicted a set of 391 genes that severe variants in these genes can cause autism or other developmental delay disorders. Approaches such as the one presented in this paper are needed to translate the biomedical discoveries into actionable items by clinicians. Odin is publicly available at https://github.com/HormozdiariLab/Odin.

**KEYWORDS** Autism; early disease prediction; complex disorder; neurodevelopmental disorder; *de novo* mutation; rare coding variant

THE start of the genomics era and sequencing of the first human genome over a decade ago promised significant benefits to public health (Lander *et al.* 2001). These include the potential capability of early detection, pinpointing the causes, and developing novel treatments and therapeutics for most diseases. Although sequencing of the human genome has dramatically accelerated biomedical research, progress has been slow in truly unlocking the promise of genetics and genomics in direct application to human health and disease. Notably, the translation of genetic discoveries into actionable items in medicine has not achieved the promised potential. One of the main challenges lies in the fact that discovering the exhaustive set of causative variants for most diseases, except some monogenic Mendelian disorders, has proven to be an elusive and unmet objective (Ng *et al.* 2010; Bamshad *et al.* 2011; Yang *et al.* 2013).

Autism spectrum disorder (ASD) is an umbrella term used to describe a set of neurodevelopmental disorders having a wide range of symptoms, from lack of social interaction, difficulty in communication/language, repetitive behavior, and, in many cases, intellectual disability (ID) (*i.e.*, having an IQ < 70) (American Psychiatric Association 2013). ASD is typically diagnosed around the age of 2 years and is estimated to affect over 1 in 68 children (1.5% of all children). There is a well-known sex bias in ASD as there are four times more male children affected with ASD than female children. Twin study comparisons have shown that genetics play a major role in ASD, and researchers have estimated the heritability of ASD to be one of the highest among complex diseases ($0.5 \le h^2 \le 0.8$) (Sandin *et al.* 2014; Tick *et al.* 2015).

There are some known syndromic subtypes of ASD with known genetic causes, such as Fragile X or Rett syndromes, which are the result of single-gene mutations [*FMR1* or *MECP2*, respectively (Caglayan 2010)]. Furthermore, there are known rare, large recurrent copy number variations

doi: https://doi.org/10.1534/genetics.118.301280
Manuscript received June 20, 2018; accepted for publication September 18, 2018; published Early Online October 8, 2018.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7011308.
[1] Corresponding author: UC Davis Genome Center, 451 Health Science Dr., Davis , CA 95616. E-mail: fhormozd@ucdavis.edu

(CNVs), such as the 16p11.2 deletion or Prader-Willi syndrome, which are known to cause ASD (Sanders *et al.* 2011; Girirajan *et al.* 2012). Recently, several autism and ID sequencing consortia (De Rubeis *et al.* 2014; Iossifov *et al.* 2014) performed whole-exome sequencing (WES) on thousands of autism families (affected proband, unaffected sibling, and parents) with the hope of finding causative variants in these samples. These studies indicated that a significant fraction of ASD was the result of *de novo* and rare (minor allele frequency $< 0.05$) variants (Iossifov *et al.* 2014; Geschwind and State 2015; Krumm *et al.* 2015). However, in many cases, it was not clear which *de novo* or rare variants were the real culprit(s) underlying the phenotype.

It is becoming apparent that early treatment and intervention can significantly improve the IQ, language skills, and social interactions in children affected with ASD (Vismara and Rogers 2008; Howlin *et al.* 2009; Boyd *et al.* 2010). Early diagnosis of ASD in young infants is challenging, mainly due to the fact that most symptoms are not reliably detectable at a very young age and children tend to manifest a heterogeneous set of phenotypes with a diverse range of severity (Kim *et al.* 2016). However, it is theoretically possible to make an accurate diagnosis of ASD or other neurodevelopmental disorders in a subset of children before any symptoms appear (or even before the child is born) using (perinatal) genetic testing and genome sequencing (Kitzman *et al.* 2012).

Although rare coding variants are enriched in the WES data of large ASD/ID proband cohorts (De Rubeis *et al.* 2014; Iossifov *et al.* 2014), it is also important to realize the intrinsic limitations of using these rare coding variants to predict ASD or other complex disorders. Notably, (i) most complex disorders have genetic heritability of significantly $<1$ (*e.g.*, $0.5 < h^2 < 0.8$ for autism), (ii) noncoding variants, which significantly contribute to these disorders, are not found using WES, and (iii) (coding) variants alone do not have the power to *rule out* the possibility of being diagnosed with a complex disorder (such as autism) with very high accuracy. Therefore, achieving accurate prediction for all, or even most, samples (both affected and nonaffected) using solely coding variants is theoretically not achievable. Moreover, a positive diagnosis/prediction of a complex disorder (*e.g.*, ASD) can have a severe negative psychological and economic impact on affected individuals and their family. For instance, a positive prediction of severe developmental disability during prenatal testing can result in a termination of pregnancy. Therefore, it is highly desirable to not have a false positive prediction (*i.e.*, an unaffected sample is predicted as an affected case) in an early disorder prediction method.

We denote the problem of predicting complex disorders that aims to cover a significant fraction of affected cases while having very low FPR prediction for unaffected samples as the Ultra-Accurate Disorder Prediction (UADP) problem. Note that the UADP problem is different from traditional binary classification problems where *each sample* is assigned to one of the two classes (*i.e.*, affected case or unaffected control). In the UADP problem, the goal is to predict a subset of samples as affected cases, while all other samples are not assigned to any class.

In practice, this UADP problem has been addressed successfully in handful of cases. For example, screening for severe variants in few well-known genes with high penetrance for neurodevelopmental disorders (*e.g.*, screen of FMR1 or MECP2 for fragile-X or Rett syndrome). Another example is the screening of CNVs such as 16p11.2 deletion and duplications for autism. Although, these tests give a low FPR, they can only discover a very small fraction of affected cases.

In this paper, we study the UADP problem in ASD. We develop a combinatorial method that utilizes rare coding variants and the available functional information (*e.g.*, pairwise coexpression of genes during brain development) to predict this disorder. By that, our method accurately recovers an additional 8% of autism cases without any severe variant in known recurrent mutated genes with a $<1\%$ FPR. Furthermore, Odin predicted a set of 391 genes that severe variants in these genes can cause autism or other developmental delay disorders.

## Materials and Methods

### Definition and notations

As we do not expect to see the same rare or *de novo* variant to appear in two different samples, it has proven useful to summarize the observed variants on the genes being affected. Here, we assume that an likely gene disruptive (LGD) mutation will completely knockout or disrupt the copy of the affected gene in the sample.

***Training data:*** Let $n$ and $m$ be the number of genes and the total number of samples respectively. The *LGD mutation profile* of the $i$th sample is a binary row vector $\mathbf{x_i} = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ where

$$x_{i_j} = \begin{cases} 1 & \text{if the } i\text{th sample has a LGD mutation at the } j\text{th gene} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The *diagnosis result (or class)* of the $i$th sample is a binary value $y_i$ where

$$y_i = \begin{cases} 1 & \text{if the } i\text{th sample is an affected case} \\ 0 & \text{if the } i\text{th sample is an unaffected control} \end{cases} \tag{2}$$

A dataset $D$ of $m$ input samples is a set of $m$ pairs $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x_m}, y_m)\}$, where each pair $(\mathbf{x_i}, y_i)$ represents the LGD mutation profile and the diagnosis result, respectively, of the $i$th sample. We define the unaffected control set and the affected case set as $D_{control} = \{\mathbf{x_i} | (\mathbf{x_i}, y_i) \in D, y_i = 0\}$ and $D_{case} = \{\mathbf{x_i} | (\mathbf{x_i}, y_i) \in D, y_i = 1\}$, respectively.

***Gene similarity score:*** An assumption here is that disruption of genes with functional relationship/similarity will result in similar phenotypes. Thus, our method not only uses the variant frequency of each gene in cases and controls but also utilizes the similarities between genes as an additional signal for predicting the disorder. We denote a gene similarity matrix $P \in [0,1]^{n \times n}$, where $P_{i,j}$ indicates the similarity between genes $i$th and $j$th. We build this matrix from two auxiliary matrices, $P'$ and $P''$, that are constructed from the functional data and the mutational landscape data, respectively, of each gene. More specifically, we construct the auxiliary matrix $P'$ by using the functional similarity of genes during brain development. We set $P'_{i,j}$ as the Pearson correlation of expression profiles between two genes in different conditions and tissues obtained from the Brainspan dataset (http://www.brainspan.org) as in the previous study (Hormozdiari *et al.* 2015). Then, we construct the auxiliary matrix $P''$ by using the similarity of likelihood of observing LGD mutation (pLI) between the two genes in the population (Lek *et al.* 2016). We set $P''_{i,j} = |pLI(i) - pLI(j)|$, where $pLI(i)$ and $pLI(j)$ are the pLI score of genes $i$th and $j$th, respectively. Finally, we use the minimum similarity scores of two genes to build matrix $P$ as $P_{i,j} = \min(P'_{i,j}, P''_{i,j})$. Of course, the construction of matrix $P$ can be changed without any need to change the underlying framework or the methods proposed in the next section. For example, we can change the construction of auxiliary matrices $P'$ or $P''$, construct more auxiliary matrices, or change the way we combine these auxiliary matrices rather than taking the minimum value.

***Training data transformation:*** We convert every sample by multiplying the vector $\mathbf{x_i}$ by matrix $P$ to produce new vectors $\mathbf{z_i} = \mathbf{x_i} \times P$. We will denote the set of samples $D_{control}$ and $D_{cases}$ converted by the gene similarity matrix $P$ as $D'_{control} = \{\mathbf{z_i} = \mathbf{x_i} \times P | \mathbf{x_i} \in D_{control}\}$ and $D'_{case} = \{\mathbf{z_i} = \mathbf{x_i} \times P | \mathbf{x_i} \in D_{case}\}$

### Odin framework

We propose a framework, denoted as Odin (Oracle for DIsorder predictioN), for solving the UADP problem. Intuitively, Odin predicts an input/test sample to be an affected case if, and only if, it satisfies two conditions:

1. The input sample is "close" to many affected case samples.
2. The input sample is "far" from any unaffected control sample.

For testing the first condition, we simply use the nearest neighbor approach with a distance function (*e.g.*, Euclidean distance). As such, an input sample passes this first condition if its closest neighbor (among the training data) is an affected case.

For testing the second condition, we develop a novel algorithm that first finds a region (after dimension reduction) containing a significant number of affected cases and does not contain any unaffected control. This cluster is denoted as *unicolor cluster*, as it only includes the affected cases. The input sample passes the second condition if it falls inside of this unicolor cluster. We denote the problem of finding such a cluster as unicolor clustering with dimensionality reduction (UCDR). We prove that this problem is a "NP-complete" problem using a reduction from *equal subset sum problem* (see the Appendix section for the NP-completeness proof of UCDR problem). Therefore, we propose a relaxation of UCDR that we denote as weighted unicolor clustering with dimensionality reduction (WUCDR). In the remaining of this section, we first formalize the UCDR and WUCDR problems, and then present an iterative algorithm to solve the WUCDR problem.

***UCDR problem:*** In the UCDR problem, we have a set of *red* and *blue* points in $n$-dimension space $\mathbb{R}^n$, representing unaffected controls (*i.e.*, $D'_{control}$) and affected cases (*i.e.*, $D'_{case}$), respectively. Furthermore, we have an upper bound on the number of dimensions to consider (dimension reduction/feature selection) denoted by $k$. The goal of the UCDR problem is to discover a subset of dimensions with cardinality $k$ ($k \ll n$), a center point $\mathbf{c} \in \mathbb{R}^{|k|}$, and a constant $r$, such that after mapping all the blue and red points to the reduced $k$ dimensions, the following objective and constraints hold:

Objective: *maximize* the total number of blue points with "distance" less than $r$ to center $\mathbf{c}$.
Constraint: there is no red point with "distance" less than $r$ to center $\mathbf{c}$.

Any metric distance function (*e.g.*, Euclidean distance) can be used for the UCDR problem. However, we use the $\ell_1$ distance since it is concordant with the additive model used in common variant studies. The $\ell_1$ distance between two points $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ is defined as $\sum_{i=1}^{n} |a_i - b_i|$. We denote the region limited by a distance $r$ from center $\mathbf{c} \in \mathbb{R}^{|k|}$ as *area of interest* $\mathscr{A}(\mathbf{c}, r)$. Furthermore, any affected case $\mathbf{z_i} \in D'_{case}$ inside the area of interest (*i.e.*, $\ell_1(\mathbf{c}, \mathbf{z_i}) \leq r$) is considered covered by this area. Note that the intuition behind the dimension reduction is to avoid the overfitting issue raised as a result of a large number of dimensions ($> 20,000$ genes) and a smaller number of training samples. We used fivefold cross-validation, and picked the $k$ that had the best true positive rate (TPR) for FPR $< 0.01$ (the desired limit on false positive prediction rate). By that, the optimal value $k = 10$ was selected for all experiments in the paper.

***WUCDR problem:*** Since the UCDR problem is NP-complete (see *Appendix*), we define a relaxation, where we assign (continuous value) weights to the dimensions. We denote this problem as the WUCDR problem. More formally, in addition to selecting $k$ genes/dimensions, we also have to assign weights $0 \leq w_i \leq 1$ to each gene/dimension $i$ and use the *weighted* $\ell_1$ as the distance metric for clustering. In the rest of the paper, we define the weighted $\ell_1$ distance function between two input points $\mathbf{a}$ and $\mathbf{b}$ with weights $\mathbf{w}$ (in $n$ dimensions) as $w\ell_1(\mathbf{a}, \mathbf{b}, \mathbf{w}) = \sum_{i=1}^{n} w_i |a_i - b_i|$. Note that, as we are only allowed to select, at most, $k$ dimensions, thus $n - k$ other dimensions have weight zero.

**Table 1 The total number of ASD/ID-affected probands (cases) and unaffected siblings (controls) used in this study**

| Class-ASD diagnosis (affected or unaffected) | Study | Number of samples | Number of LGD variants | References |
|---|---|---|---|---|
| Affected ASD/ID probands | Simons simplex collection (SSC) | 2508 | 492 | O'Roak *et al.* (2012, 2014), Iossifov *et al.* (2014), Krumm *et al.* (2015), Turner *et al.* (2016) |
| | Autism sequencing consortium (ASC) | 2270 | 185 | De Rubeis *et al.* (2014) |
| | Other studies | 1329 | 74 | Michaelson *et al.* (2012), Rauch *et al.* (2012), Hashimoto *et al.* (2016) |
| | Total (case) | 6107 | 751 | |
| Unaffected siblings or controls | Simons simplex collection (SSC) | 1909 | 248 | Iossifov *et al.* (2014), Krumm *et al.* (2015) |
| | GoNL | 250 | 7 | Francioli *et al.* (2014) |
| | Other studies | 208 | 11 | Gulsuner *et al.* (2013) |
| | Total (control) | 2367 | 266 | |

***Iterative solution for WUCDR:*** Here, we propose an iterative method consisting of two main steps to solve the WUCDR problem. In the first step, given a set of weights $\mathbf{w}$, we find the optimal center $\mathbf{c}$ and radius $r$ to cover a maximum number of affected cases (blue points) in the area of interest $\mathscr{A}(\mathbf{c}, r)$ (note that the area of interest is considered using weighted $\ell_1$ distance). In the second step, we try to find a new set of weights $\mathbf{w}$ given the center $\mathbf{c}$ and the radius $r$.

***First step:*** Given the weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ (all the weights are assigned to 1 at the first iteration), find a center $\mathbf{c}$ and constant $r$ such that

1. all red points have a weighted $\ell_1$ distance greater than $r$ to center $\mathbf{c}$ and
2. the number of blue points, which have weighted $\ell_1$ distance less than $r$ to center $\mathbf{c}$, is maximized.

In general, finding such a center is a hard problem in $n$ dimensional space. Thus, we relax the problem to only consider the blue points as a potential center $\mathbf{c}$. This can be done trivially in polynomial time by considering every blue point as potential center and picking the optimal one. Given a center $\mathbf{c}$, radius $r$ (easily determined from $\mathbf{c}$) and the weights $\mathbf{w}$ we can easily calculate the affected cases (*i.e.*, blue points) covered by the area of interest. Let $S$ denote the set of covered (blue) points (*i.e.*, affected cases), which will be used in the next step for updating the weights.

***Second step:*** Given a center $\mathbf{c}$ and the set of blue points $S$ covered by the area of interest found in the first step, we calculate new weights $\mathbf{w}$ (for each dimension). The objective is to decrease the weighted $\ell_1$ distance of points in the set $S$ to center $\mathbf{c}$, while increasing the weighted $\ell_1$ distance of points in the set $S$ to the red points $(D'_{Control})$. We solve the linear programming (LP) problem (see below) to find these new weights.

Note that, in the LP problem, only $\mathbf{w}$ and $\rho$ are unknown variables, while the set S and center $\mathbf{c}$ are calculated in the first step of the method. The constraints in the LP problem help us find a set of weights that are guaranteed to have all of the points in set S closer to the selected center $\mathbf{c}$ than any red point. Furthermore, the objective function helps us find the weights that squeeze the (blue) points in S further closer to the center $\mathbf{c}$, while increasing the distance of red points to the (blue) points in the set S. More specifically, the objective function of the LP problem has two main terms. The first term aims to reduce the average distance between points in the set S and the center $\mathbf{c}$. Simply stated, the new weights $\mathbf{w}$ would try to make blue points covered in first step (*i.e.*, point in set S) get closer to the center $\mathbf{c}$ (note that both $\mathbf{c}$ and S are from the previous step, not variables in this LP problem). The second term aims to increase the average weighted $\ell_1$ distance of all red points to the blue points in set S. Finally, among the weights produced we will keep only the top k weights and convert all of the remaining weights to 0. Note that, because of the condition $\sum_{i=1}^{n} w_i \leq k/2$, we are guaranteed to be able to keep any dimension with value $> 0.5$ from the LP solution.

***Odin framework using WUCDR:*** As mentioned above, two conditions need to be satisfied for a sample to be predicted as a potential affected case by Odin. The first condition is that the nearest neighbor of this sample must not be an unaffected control. Odin uses the $\ell_1$ distance function for calculating the nearest neighbors of any test sample. The second condition is that the input sample needs to fall inside the *area of interest* $\mathscr{A}$($\mathbf{c}$,r) after performing the same dimension reduction mapping using weights $\mathbf{w}$ (note that $\mathbf{c}$, $r$ and $\mathbf{w}$ are found by the iterative solution of WUCDR).

### Data availability

The code of Odin and related data are publicly available at https://github.com/HormozdiariLab/Odin. The authors state that all data necessary for confirming the conclusions presented in the manuscript are presented fully within the manuscript and the supplemental information (at https://

| Linear programming problem | | |
|---|---|---|
| $\underset{\mathbf{w},\rho}{\text{Minimize}}$ | $\frac{1}{\lvert S\rvert}\sum_{\mathbf{z_i}\in S} w\ell_1(\mathbf{z_i},\mathbf{c},\mathbf{w}) - \frac{1}{\lvert D'_{Control}\rvert \times \lvert S\rvert}\sum_{\mathbf{d_i}\in D'_{Control}}\sum_{\mathbf{z_i}\in S} w\ell_1(\mathbf{z_i},\mathbf{d_i},\mathbf{w})$ | |
| subject to | $w\ell_1(\mathbf{z_i},\mathbf{c},\mathbf{w}) \le \rho$ | $\forall \mathbf{z_i}\in S$ |
| | $w\ell_1(\mathbf{d_i},\mathbf{c},\mathbf{w}) \ge \rho$ | $\forall \mathbf{d_i}\in D'_{Control}$ |
| | $0 \le w_i \le 1$ | $\forall i$ |
| | $\sum_{i=1}^{n} w_i \le k/2$ | |
| | $\rho > 0$ | |

figshare.com/s/ce4e4dc7210e4e4034b3). Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7011308.

## Results

### Data summary

We curated a dataset of *de novo* LGD variants from WES and targeted sequenced samples with ASD or ID to evaluate the performance of Odin in predicting neurodevelopmental disorders. Table 1 shows the total number of samples and LGD variants reported from the union of several publications on over 6000 ASD/ID probands.

We transformed the mutation profile of every sample into a vector in the new space (*i.e.*, $\mathbf{z_i} = \mathbf{x_i} \times P$) by a gene similarity matrix $P$ constructed as in *Materials and Methods*. We observed that such a transformation resulted in significantly reducing the $\ell_1$ distance of probands with each other ($P < 1.6e-16$). This implied that this transformation indeed helped in increasing the prediction power.

In our dataset (Table 1), there are nine genes (*ADNP*, *ANK2*, *ARID1B*, *CHD2*, *CHD8*, *DSCAM*, *DYRK1A*, *SCN2A*, and *SYNGAP1*) with four or more LGD variants in the union of all ASD/ID samples and with no LGD variant in unaffected siblings and controls. A sample with an LGD variant in any of these genes is called a trivial case, and the remaining samples are called nontrivial cases/samples. We only consider nontrivial cases in our analysis, since any prediction model/method for ASD can be extended easily to predict trivial cases.

### Naïve approach for solving UADP problem by utilizing predicted ASD/ID genes

We evaluated the ASD/ID prediction performance (of samples in Table 1) of naïve approaches that used top ranking ASD/ID genes or used genes based on intolerance to LGD variant and expression in the fetal cortex. We realized that none of these approaches provided an acceptable solution to the UADP problem. For example, considering any sample with LGD variant in the top 100 genes from the recently published gene ranking (Krishnan *et al.* 2016) as an affected case resulted a FPR of $>1\%$ and TPR of $<2.5\%$. Similarly, predicting any sample with LGD variant in top genes based on the intolerance to LGD variant [ExAC data (Lek *et al.* 2016)] and high expression in cortex region during early fetal development (from CSEA tool http://genetics.wustl.edu/jdlab/csea-tool-2/) as an affected case resulted a FPR of $>1\%$ and TPR of $<2.8\%$.

### Unicolor clustering with dimension reduction

We first verified if the proposed iterative method for solving the WUCDR problem actually improved the number of cases covered in comparison to the unweighted setting (*i.e.*, considering all dimensions with weights $w_i = 1$). As shown in Table 2, the optimal result found using the initial setting (*i.e.*, unweighted) was able to cover 45 cases (only 24 nontrivial cases, *i.e.*, the ones with no LGD variants in recurrently mutated genes in ASD/ID samples). In contrast, our iterative method converged in less then five iterations and was able to cover over 71 cases (40 nontrivial cases). Thus, our method improved the number of affected cases covered by over 60% using <10 dimensions. We also investigated the "density" of cases inside each selected region. The density was defined as the ratio between the number of affected cases covered and the radius $r$. We observed that our iterative method not only improved the number of cases covered but also increased the density per each iteration (see Table 2).

### ASD/ID disorder prediction results

We compared Odin with different classification methods in predicting nontrivial ASD/ID cases. These methods included the k-NN classifier ($1 \le k \le 20$), support vector machines (SVM; Chang and Lin 2011), glmnet (Lasso and elastic-net regularization of generalized linear models; Friedman *et al.* 2010), and random forest (Liaw *et al.* 2002). For each of these methods, their optimal parameter values were picked and their intrinsic properties were used to control/limit the FPR for calculating the TPR (see *Appendix*). Odin had only one parameter $k$, and we set $k = 10$ (see *Materials and Methods*). Note that, in Odin, the full set of samples predicted as affected cases had a FPR of $<0.01$. We used the leave-one-out (LOO) cross-validation technique to determine the highest TPR of each method given the upper bound on the FPR value. As our stated goal was to keep the false positive prediction of

**Table 2 Number of covered ASD/ID cases (from training dataset), density, and the number of dimensions in each iteration**

| Iteration | Number of cases covered | Number of nontrivial cases covered | Density (case/radius) | Number of dimensions (before-after) rounding |
|---|---|---|---|---|
| 0 | 45 | 24 | 0.11 | ≥20,000 |
| 1 | 53 | 32 | 138.41 | (15–9) |
| 2 | 66 | 39 | 176.64 | (7–7) |
| 3 | 70 | 39 | 179.92 | (10–9) |
| 4 | 71 | 40 | 185.49 | (10–9) |

unaffected samples as cases close to zero, we considered only the most conservative results for each method (*i.e.*, FPR $< 0.01$). Odin's true positive rate for predicting ASD/ID was at least two times higher than the best k-NN result (for different values of k) and significantly higher than SVM and random forest for FPR $< 0.01$ (Figure 1). It was also significantly higher for different regularized generalized linear models (Lasso and elastic net) for different parameter values of $\alpha$ of the glmnet implementation (Figure 1).

We also compared Odin and SVM for a very low FPR by using fivefold cross validation. We selected the best set of parameter values for each method and repeated $>10$ (independent) times to evaluate the performance of two methods. With FPR $<0.01$, Odin was able to achieve a TPR (for nontrivial cases) over 0.06, while SVM achieved a TPR of only 0.04 (for nontrivial cases).

### *Gene prediction and ranking in autism and related disorders*

Odin is a framework that predicts if a sample will develop ASD with an extremely low false positive rate. However, it can also be used to predict some novel ASD genes. Here, we utilized Odin to rank all genes for the potential impact of their LGD variant on ASD. More specifically, we trained Odin with the ASD and siblings variants (Table 1), then each gene was ranked by the distance from the center to a sample that had a LGD variant of this gene. We referred genes in categories "syndromic" and "high confidence" of the SFARI gene collection (Abrahams *et al.* 2013) as known ASD genes. We observed that genes closer to the center were enriched with known ASD genes (Figure 2A). Note that, similar to the conservative way that Odin predicted the ASD/ID risk of a sample, if a gene was not selected it did not mean this gene was not an ASD/ID gene.

Our analysis also indicated 391 genes (Supplemental Material, Table S1) for which an LGD variant resulted in a sample falling inside the predicted area of interest (illustrated as the most inner circle in Figure 2A). In other words, Odin predicted with high probability that the disruption of each of these 391 genes would cause significant (neuro)developmental disorder.

We observed a significant enrichment of LGD variants in these 391 genes in the developmental delay disorder (DDD) sample set (which was not used in training) *vs.* the ASD sample set (which was used in training) as shown in Figure 2B. Interestingly, this clearly indicated that, even after normalizing based on expected LGD variants for each disease group, the more severe samples tended to be more enriched in LGD variants disrupting these selected 391 genes than did their less severe autism samples (Figure 2B). In addition, there was also an significant increase in ratio of *de novo* missense variants with CADD score $>25$ to *de novo* missense variants with CADD score $<25$ disrupting these 391 genes in comparison to the remaining genes for ASD/ID probands (Figure 2C). Interestingly, no such enrichment was observed for control/sibling samples (Figure 2C). Furthermore, we used the *de novo* variants reported in 520 whole-genome sequenced (WGS) samples (Turner *et al.* 2017) that were void of LGD variants to investigate the *de novo* variants disrupting the noncoding regulatory regions of these genes (see *Appendix*). We observed that the noncoding regulatory elements of these 391 genes were significantly disrupted by *de novo* variants in probands *vs.* siblings ($P < 0.004$ : Figure S1, A and B in *Appendix*). Moreover, we also observed that genes that were closer to the center also had more protein–protein interaction than genes that were far from the center (Figure S1C).

We utilized the predicted probability of observing a *de novo* missense/LGD variant per sample for each gene (O'Roak *et al.* 2012) to calculate the *P*-values of observed *de novo* variants in the affected samples. We could group these 391 genes based on observing significant *de novo* LGD and/or missense variants in affected probands (Figure 2D). The set of genes with significant *de novo* missense variants observed only in cases potentially indicated ŁŁan LGD ŁŁin one of these genes ŁŁwould be incompatible with life (i.e., essential genes). However, a missense mutationŁ ŁŁin one of these genes could result in a severe (neuro)developmental disorder. These genes include *CSNK2A1*, *SMARCA4*, *TRRAP*, *MORC2*, *PRPF8*, *TAF1*, *CNOT1*, *SF3B1*, *SMAD4*, *UBR5*, *CLASP1*, *KDM2B*, and *U2AF2*.

Next, we analyzed if there was any specific enrichment of expression of these 391 genes in the human brain. We used the online tool CSEA (Dougherty *et al.* 2010) (http://genetics.wustl.edu/jdlab/csea-tool-2/) to study the expression profile of these genes. Interestingly, the only significant expression we observed was in early fetal development and mid-early fetal brain development (Figure 3A). No significant expression

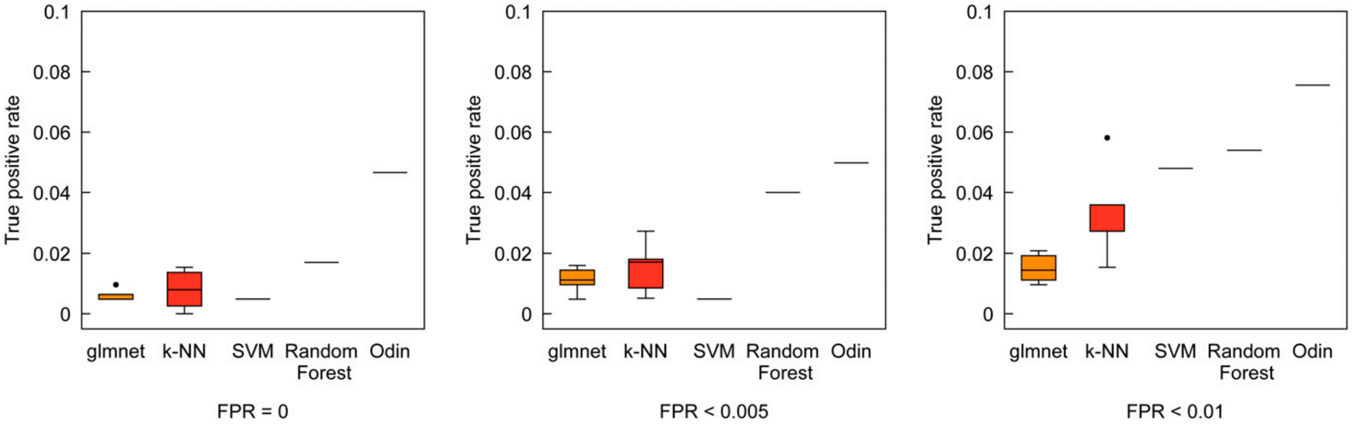

**Figure 1** A comparison between different methods on the ASD/ID prediction. The boxplot shows the results of different values of $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ (for glmnet) and of different values of $k \in [1, 20]$ (for k-NN).

of these genes in any tissues in adult human or mouse brain was observed.

In the Simons Simplex Collection (SSC), we also observed not only that probands with *de novo* LGD variants tended to have a lower IQ than probands without *de novo* LGD variants, but also that probands with a *de novo* LGD variant disrupting one of these 391 genes had lower IQ than probands with other *de novo* LGD variants (Figure 3B). It is known that there is a large male to female bias in autism (estimated to be over 4:1). In the SSC, there was a total of 2478 male probands and 396 female probands (over 6:1 ratio). However, the difference between the number of samples with *de novo* LGD variants in the selected 391 genes was 31:16 (~2:1 ratio). This indicated that there was a much smaller gap in sex difference for ASD samples with *de novo* LGD variants in these 391 genes (Figure 3C).

Finally, we analyzed the biological function of the top ASD/ID genes predicted by Odin. We used the tool David (Huang *et al.* 2009, version 6.8) for discovery of enriched gene ontology (GO)-terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for top 391 genes. As we expected, these genes were enriched in transcription, spliceosome, chromatin modification, and Wnt pathways, which have been indicated previously to be major contributing factors in neurodevelopmental disorders (Gilman *et al.* 2011; Kalkman 2012; Ben-David and Shifman 2013; De Rubeis *et al.* 2014; Hormozdiari *et al.* 2015; Lelieveld *et al.* 2016).

## Discussion

In this paper, we formalized the problem of predicting a complex disorder while enforcing a virtually zero false positive prediction to make it directly applicable in clinical diagnosis. We denoted this specific problem as the Ultra-Accurate Disorder Prediction (UADP) problem. We showed that simple approaches of utilizing the predicted ASD/ID gene rankings were not a viable solution for the UADP problem. Then, we introduced the framework Odin for solving the UADP problem in autism and related disorders using *de novo* LGD variants. Our evaluation of experimental data showed that Odin outperformed other approaches in this prediction task.

Note that Odin is not meant to replace other approaches for disorder gene discovery and ranking, but rather for accurate prediction of the disorder in a subset of cases given the genetic variation. One of the drawbacks of this approach is that it can only work if the penetrance of the genetic variation to cause the disorder is very high. It is not applicable in cases where the goal is to only predict if the probability of disorder is significantly higher than the general population.

Although Odin indicated some well-known pathways (*e.g.*, Wnt and chromatin remodelers) related to autism (Ben-David and Shifman 2013; Hormozdiari *et al.* 2015), it did not indicate other important pathways (*e.g.*, long-term potentiation and synaptic function), which were also related to this disease. The main reason is that the current formulation of Odin only considers one center that may represent only one module/pathway and thus other modules are not being covered. Thus, an extension of Odin that considers multiple centers may significantly improve the complex disease prediction and pathways covered.

In addition, other potential extensions also need to be investigated. First, the proposed framework can be extended to take into account not only LGD mutations but also missense mutations to increase the prediction power. As we have shown, there is clear enrichment of severe missense mutations (CADD score >25) to genes closer to the predicted center. We can adapt evolutionary based scores [*e.g.*, CADD score (Kircher *et al.* 2014) or polyphen-2 score (Adzhubei *et al.* 2010)] to define an additive summarization function to assign a disruption score for each gene (*i.e.*, a continuous value in comparison to a binary value as done in this paper). Second, we can integrate other information such as protein–protein interaction (Sharan *et al.* 2007; Kim *et al.* 2011), tissue-specific networks (Greene *et al.* 2015) or the regulation of specifically related pathways [*e.g.*, Wnt (Kalkman 2012) or mTOR (Tang *et al.* 2014)] to increase prediction capability. Furthermore, for the algorithm, we can improve the first guessed solution of WUCDR (the first step of the iterative

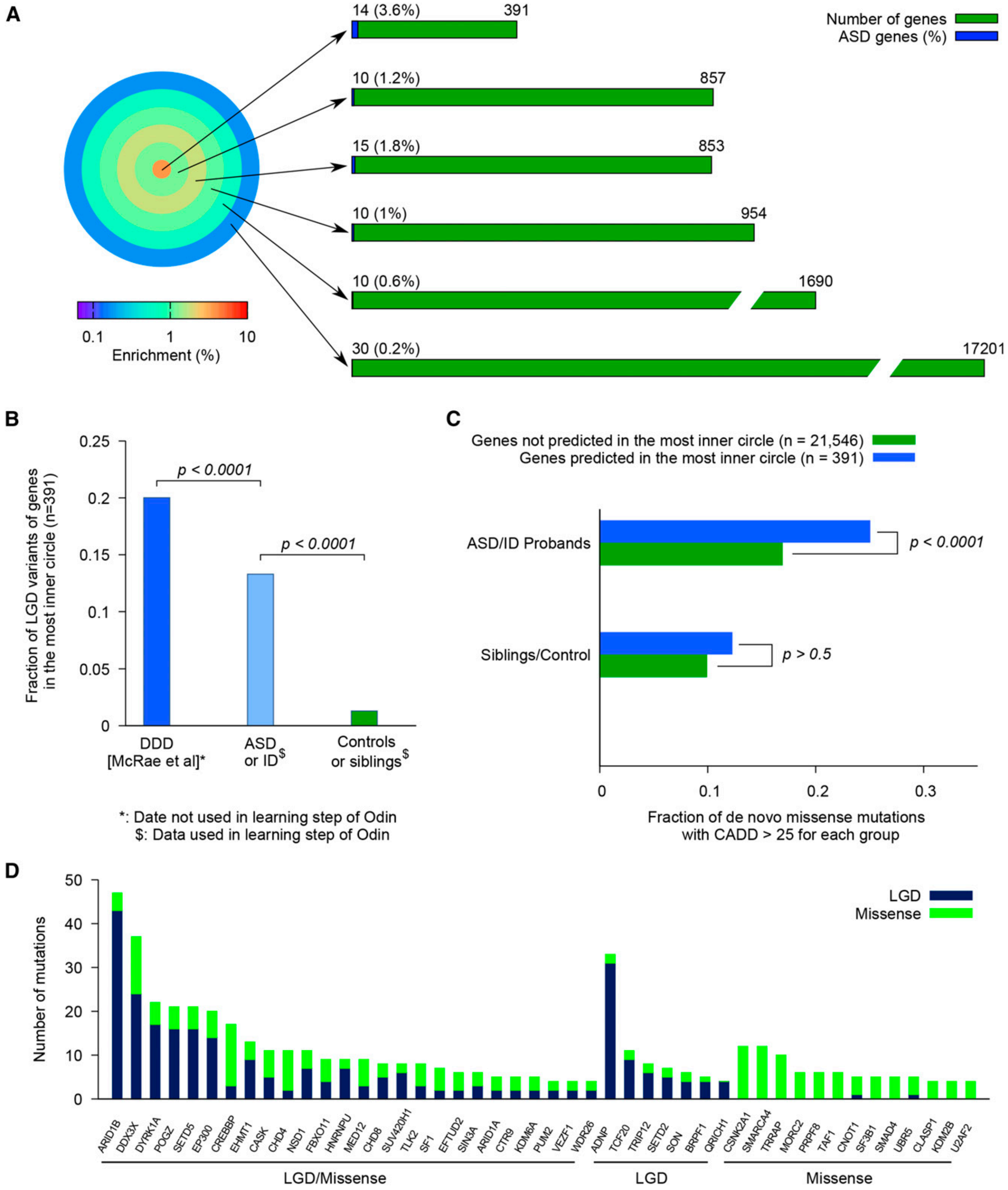

**Figure 2** ASD gene ranking by Odin. (A) The circle on the left illustrates six gene groups ranked by Odin (closer to the center means higher ASD rank) and the enrichment of each group with known ASD gene (from SFARI gene collection), the bar chart on the right provides the number of genes, the number of known ASD genes, and the enrichment (*i.e.*, their ratio), respectively, of each group. (B) Enrichment of LGD variants disrupting one of top 391 genes in ASD/ID/DDD. (C) Enrichment of severe missense variants (CADD > 25) of top 391 genes in ASD/ID probands. (D) Genes in the most inner circle with at least four significant *de novo* variants.

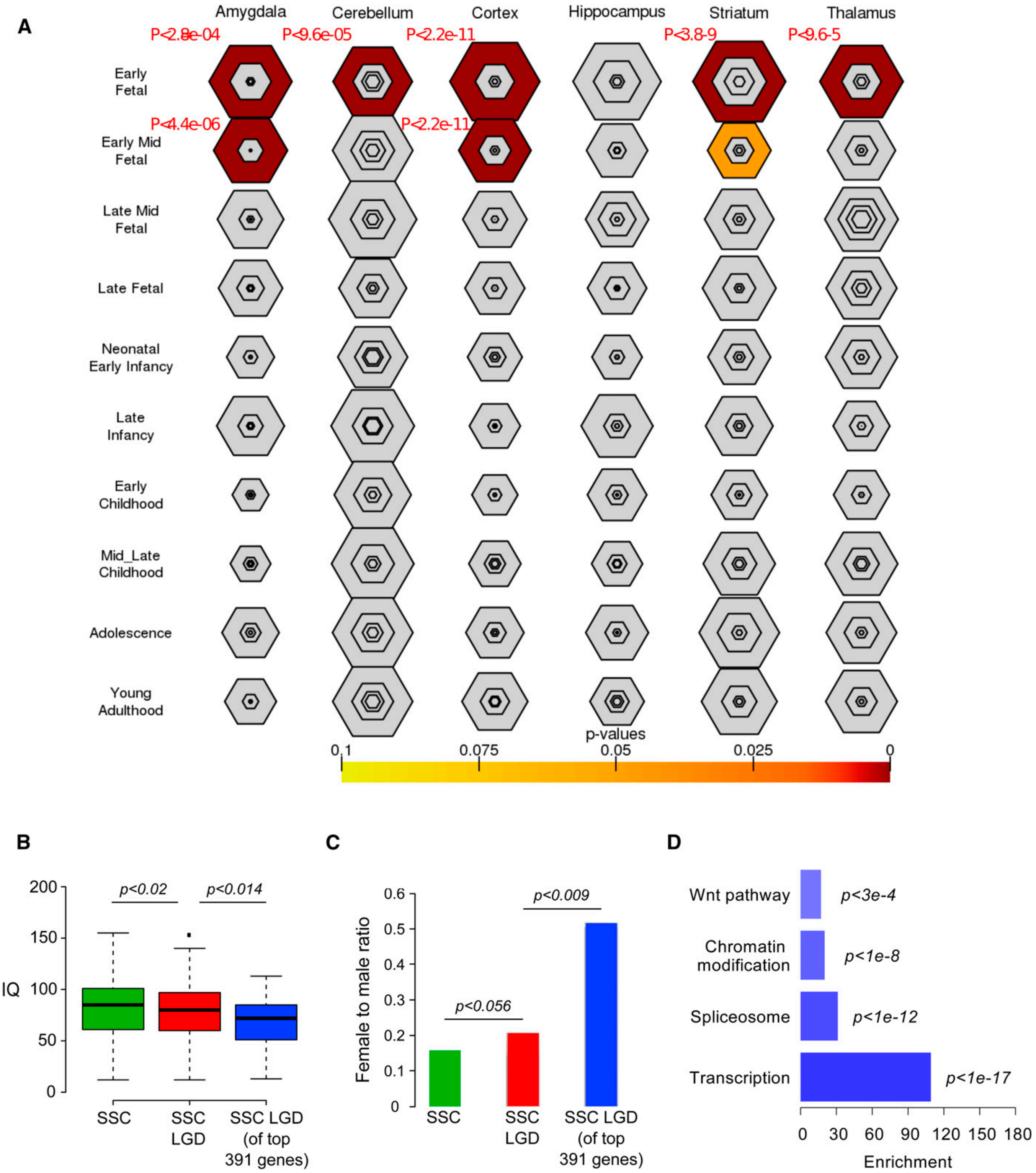

**Figure 3** Further analysis of top 391 ASD genes selected by Odin. (A) Enrichment of expression of our selected genes in human brain development. The reported *P*-values are calculated by the CSEA tool (http://genetics.wustl.edu/jdlab/csea-tool-2/) after Benjamini-Hochberg statistical correction. (B and C) IQ and female:male ratio of three groups including all SSC probands, SSC probands with *de novo* LGD variants, and SSC probands with *de novo* LGD variants in our selected 391 genes. (D) Pathway and GO term enrichment of top 391 ASD genes.

method in *Materials and Methods*) by utilizing algorithmic techniques in geometry. Finally, the framework proposed here can be extended to predicting the risk of other neurological disorders, such as schizophrenia, epilepsy, or Alzheimer's disease.

## Acknowledgments

We would like to thank Evan E. Eichler, Tychele N. Turner, Phuong Dao, Farhad Hormozdiari, Madeleine Geisheker, A. Ercument Cicek, and Tonia Brown for reading the paper and providing comments. None of the authors have any competing interests. This work is supported in part by the Sloan Research Fellowship number G-2017-9159 (to FH).

## Literature Cited

Abrahams, B. S., D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow *et al.*, 2013 Sfari gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism 4: 36. https://doi.org/10.1186/2040-2392-4-36

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.*, 2010 A method and server for predicting damaging missense mutations. Nat. Methods 7: 248–249. https://doi.org/10.1038/nmeth0410-248

American Psychiatric Association, 2013 *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, Arlington, TX. https://doi.org/10.1176/appi.books.9780890425596

Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond *et al.*, 2011 Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 12: 745–755. https://doi.org/10.1038/nrg3031

Ben-David, E., and S. Shifman, 2013 Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. Mol. Psychiatry 18: 1054–1056. https://doi.org/10.1038/mp.2012.148

Boyd, B. A., S. L. Odom, B. P. Humphreys, and A. M. Sam, 2010 Infants and toddlers with autism spectrum disorder: early identification and early intervention. J. Early Interv. 32: 75–98. https://doi.org/10.1177/1053815110362690

Caglayan, A. O., 2010 Genetic causes of syndromic and non-syndromic autism. Dev. Med. Child Neurol. 52: 130–138. https://doi.org/10.1111/j.1469-8749.2009.03523.x

Chang, C.-C., and C.-J. Lin, 2011 LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2: 27.

De Rubeis, S., X. He, A. P. Goldberg, C. S. Poultney, K. Samocha *et al.*, 2014 Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515: 209–215. https://doi.org/10.1038/nature13772

Dougherty, J. D., E. F. Schmidt, M. Nakajima, and N. Heintz, 2010 Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. Nucleic Acids Res. 38: 4218–4230. https://doi.org/10.1093/nar/gkq130

Francioli, L. C., A. Menelaou, S. L. Pulit, F. Van Dijk, P. F. Palamara *et al.*, 2014 Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46: 818–825. https://doi.org/10.1038/ng.3021

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33: 1. https://doi.org/10.18637/jss.v033.i01

Geschwind, D. H., and M. W. State, 2015 Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 14: 1109–1120. https://doi.org/10.1016/S1474-4422(15)00044-7

Gilman, S. R., I. Iossifov, D. Levy, M. Ronemus, M. Wigler *et al.*, 2011 Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron 70: 898–907. https://doi.org/10.1016/j.neuron.2011.05.021

Girirajan, S., J. A. Rosenfeld, B. P. Coe, S. Parikh, N. Friedman *et al.*, 2012 Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N. Engl. J. Med. 367: 1321–1331. https://doi.org/10.1056/NEJMoa1200395

Greene, C. S., A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya *et al.*, 2015 Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. 47: 569–576. https://doi.org/10.1038/ng.3259

Gulsuner, S., T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton *et al.*, 2013 Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell 154: 518–529. https://doi.org/10.1016/j.cell.2013.06.049

Hashimoto, R., T. Nakazawa, Y. Tsurusaki, Y. Yasuda, K. Nagayasu *et al.*, 2016 Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. J. Hum. Genet. 61: 199–206. https://doi.org/10.1038/jhg.2015.141

Hormozdiari, F., O. Penn, E. Borenstein, and E. E. Eichler, 2015 The discovery of integrated gene networks for autism and related disorders. Genome Res. 25: 142–154. https://doi.org/10.1101/gr.178855.114

Howlin, P., I. Magiati, T. Charman, and W. E. MacLean, Jr., 2009 Systematic review of early intensive behavioral interventions for children with autism. Am. J. Intellect. Dev. Disabil. 114: 23–41. https://doi.org/10.1352/2009.114:23-41

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4: 44–57. https://doi.org/10.1038/nprot.2008.211

Iossifov, I., B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm *et al.*, 2014 The contribution of de novo coding mutations to autism spectrum disorder. Nature 515: 216–221. https://doi.org/10.1038/nature13908

Kalkman, H. O., 2012 A review of the evidence for the canonical Wnt pathway in autism spectrum disorders. Mol. Autism 3: 10. https://doi.org/10.1186/2040-2392-3-10

Kim, S. H., S. Macari, J. Koller, and K. Chawarska, 2016 Examining the phenotypic heterogeneity of early autism spectrum disorder: subtypes and short-term outcomes. J. Child Psychol. Psychiatry 57: 93–102. https://doi.org/10.1111/jcpp.12448

Kim, Y.-A., S. Wuchty, and T. M. Przytycka, 2011 Identifying causal genes and dysregulated pathways in complex diseases. PLOS Comput. Biol. 7: e1001095. https://doi.org/10.1371/journal.pcbi.1001095

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper *et al.*, 2014 A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46: 310–315. https://doi.org/10.1038/ng.2892

Kitzman, J. O., M. W. Snyder, M. Ventura, A. P. Lewis, R. Qiu *et al.*, 2012 Noninvasive whole-genome sequencing of a human fetus. Sci. Transl. Med. 4: 137ra76. https://doi.org/10.1126/scitranslmed.3004323

Krishnan, A., R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong *et al.*, 2016 Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. 19: 1454–1462. https://doi.org/10.1038/nn.4353

Krumm, N., T. N. Turner, C. Baker, L. Vives, K. Mohajeri *et al.*, 2015 Excess of rare, inherited truncating mutations in autism. Nat. Genet. 47: 582–588. https://doi.org/10.1038/ng.3303

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921. https://doi.org/10.1038/35057062

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks *et al.*, 2016 Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285–291. https://doi.org/10.1038/nature19057

Lelieveld, S. H., M. R. Reijnders, R. Pfundt, H. G. Yntema, E.-J. Kamsteeg *et al.*, 2016 Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat. Neurosci. 19: 1194–1196. https://doi.org/10.1038/nn.4352

Liaw, A., and M. Wiener, 2002 Classification and regression by randomforest. R News 2: 18–22.

Michaelson, J. J., Y. Shi, M. Gujral, H. Zheng, D. Malhotra *et al.*, 2012 Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151: 1431–1442. https://doi.org/10.1016/j.cell.2012.11.019

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor *et al.*, 2010 Exome sequencing identifies the cause of a Mendelian disorder. Nat. Genet. 42: 30–35. https://doi.org/10.1038/ng.499

O'Roak, B. J., L. Vives, S. Girirajan, E. Karakoc, N. Krumm *et al.*, 2012 Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485: 246–250. https://doi.org/10.1038/nature10989

O'Roak, B. J., H. Stessman, E. Boyle, K. Witherspoon, B. Martin *et al.*, 2014 Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat. Commun. 5: 5595. https://doi.org/10.1038/ncomms6595

Rauch, A., D. Wieczorek, E. Graf, T. Wieland, S. Endele *et al.*, 2012 Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet (London, England) 380: 1674–1682. https://doi.org/10.1016/S0140-6736(12)61480-9

Sanders, S. J., A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha *et al.*, 2011 Multiple recurrent de novo CNVs, including duplications of the 7q11. 23 Williams syndrome region, are strongly associated with autism. Neuron 70: 863–885. https://doi.org/10.1016/j.neuron.2011.05.002

Sandin, S., P. Lichtenstein, R. Kuja-Halkola, H. Larsson, C. M. Hultman *et al.*, 2014 The familial risk of autism. JAMA 311: 1770–1777. https://doi.org/10.1001/jama.2014.4144

Sharan, R., I. Ulitsky, and R. Shamir, 2007 Network-based prediction of protein function. Mol. Syst. Biol. 3: 88. https://doi.org/10.1038/msb4100129

Tang, G., K. Gudsnuk, S.-H. Kuo, M. L. Cotrina, G. Rosoklija *et al.*, 2014 Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits. Neuron 83: 1131–1143 (erratum: Neuron 83: 1482). https://doi.org/10.1016/j.neuron.2014.07.040

Tick, B., P. Bolton, F. Happé, M. Rutter, and F. Rijsdijk, 2015 Heritability of autism spectrum disorders: a meta-analysis of twin studies. J. Child Psychol. Psychiatry 57: 585–595. https://doi.org/10.1111/jcpp.12499

Turner, T. N., F. Hormozdiari, M. H. Duyzend, S. A. McClymont, P. W. Hook *et al.*, 2016 Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am. J. Hum. Genet. 98: 58–74. https://doi.org/10.1016/j.ajhg.2015.11.023

Turner, T. N., B. P. Coe, D. E. Dickel, K. Hoekzema, B. J. Nelson *et al.*, 2017 Genomic patterns of de novo mutation in simplex autism. Cell 171: 710–722.e12. https://doi.org/10.1016/j.cell.2017.08.047

Woeginger, G. J., and Yu, Z, 1992 On the equal-subset-sum problem. Inform. Process. Lett. 42: 299–302. https://doi.org/10.1016/j.cell.2017.08.047

Vismara, L. A., and S. J. Rogers, 2008 The early start Denver model: a case study of an innovative practice. J. Early Interv. 31: 91–108. https://doi.org/10.1177/1053815108325578

Yang, Y., D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis *et al.*, 2013 Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. 369: 1502–1511. https://doi.org/10.1056/NEJMoa1306555

*Communicating editor: M. Sillanpää*

## Appendix

## Complexity of the UCDR Problem

We show that an instance of the decision version of the UCDR problem is NP-complete.

**Remark 1.** Given a set of positive (rational) numbers. The problem of determining if there exists two disjoint nonempty subsets whose elements sum up to the same value is NP-complete (Woeginger and Yu, 1992).

The problem in Remark 1 was called "*equal subset sum problem*." Notice that the pair of two subsets in the solution is not necessary a partition (*i.e.* there may be some elements that are in the original set but are not in either of these two sub-sets).

**Theorem 2.** Given a set of points in a $n$-dimension space where each point was assigned a color either blue or red. The problem of determining if there exists a nonempty dimension subset and a center point such that all blue points are not farther to that center point in comparison to red points (by the $L_1$ norm in the reduced dimension space) is NP-complete. We call the problem, the "UCDR decision problem."

*Proof*. We will reduce the equal subset sum problem (Remark 1) to a special instance of the UCDR decision problem.

Assume we are given a set of positive rational numbers $A = \{a_1, a_2, \ldots, a_n\}$ We create two blue points $B_1 = (a_1, a_2, \ldots, a_n)$, $B_2 = (-a_1, -a_2, \ldots, -a_n)$ and one red point $R = (0, 0, \ldots, 0)$. We consider the UCDR decision problem of three points $B_1, B_2$, and $R$. Suppose that this UCDR decision problem has a solution that includes a dimension subset $I = \{i_1, i_2, \ldots, i_d\} \subseteq \{1, 2, \ldots, n\}$ and a center $C$.

Now, we only consider the reduced space with $d$ dimensions from $I$. We denote $B_1', B_2'$, and $R'$ as the corresponding points of $B_1$, $B_2$, and $R$, respectively, in the reduced space.

Let $H$ be the smallest (by volume) $L_1$ norm ball that has the center $C$ and contains both $B_1'$ and $B_2'$. Thus, since $B_1'$ or $B_2'$ (or both) must be on a facet of $H$, we can assume $B_1'$ is on a facet of $H$ without losing generality. Since $H$ is convex and $R' = (B_1' + B_2')/2$, $H$ also contains $R'$. But if $B_2'$ is not on the same facet of $B_1'$, then $R'$ will be inside $H$ and thus $d(C, R') < d(C, B_1')$. Therefore, both $B_1', B_2'$ and $R'$ must be on the same facet of $H$. Let $F$ be that facet, since $H$ is a $L_1$ norm ball then any point $(x_{i_1}, x_{i_2}, \ldots, x_{i_d}) \in F$ must satisfy an equation that has the form

$$\pm x_{i_1} \pm x_{i_2} \pm \ldots \pm x_{i_d} = s$$

Since $R' = (0, 0, \ldots, 0) \in F$, so $s$ must be 0. Thus we can rewrite the equation as

$$\sum_{i_j \in I_1} x_{i_j} - \sum_{i_k \in I_2} x_{i_k} = 0$$

where $I_1 \cap I_2 = \varnothing$ and $I_1 \cup I_2 = I$. Since $B_1' = (a_{i_1}, a_{i_2}, \ldots, a_{i_d}) \in F$ then

$$\sum_{i_j \in I_1} a_{i_j} - \sum_{i_k \in I_2} a_{i_k} = 0$$

but both $a_{i_j}$ and $a_{i_k}$ are in $A$ that contains positive numbers only so $I_1 \neq \varnothing$ and $I_2 \neq \varnothing$. Therefore, the pair of two sets $A_1 = \{a_{i_j} | i_j \in I_1\}$ and $A_2 = \{a_{i_k} | i_k \in I_2\}$ is a solution of the equal subset sum problem of the set $A$.

Thus, a solution of the UCDR decision problem is also a solution of the equal subset sum problem. Conversely, we can also easily verify that a solution of the equal subset sum problem is also a solution of the UCDR decision problem. Therefore, if we can solve the decision version of UCDR then we can solve the equal subset sum problem which is NP-complete (Remark 1). Since it is easy to verify this problem is in NP, it is also NP-complete.

## Further Analysis

### *Enrichment of non-LGD variants disrupting the selected genes and their regulatory elements in ASD probands*

The whole genome of a total of 516 ASD simplex families from SSC was recently sequenced and *de novo* variants in the affected probands and unaffected sibling were predicted and validated (Turner *et al.* 2017). Note that these families were selected to be void of LGD variants based on WES. Thus, they were not part of the samples that contributed to Odin training. However, we did observe a significant number of affected probands in comparison of unaffected siblings had non-LGD coding and noncoding *de novo* variants disrupting the coding or the regulatory elements of the genes in the most inner circle (Figure S1, A and B). The subset of genes in the selected 391 genes in the most inner circle, which had a *de novo* variant disrupting their coding or regulatory elements in probands or siblings, is depicted in Figure S1, A and B. Furthermore, we also observed the significant enrichment after removing the known SFARI high confidence and syndromic autism genes from the set of 391 genes considered.

### *Protein interaction enrichment*

We investigated the degree of change in genes in protein-interaction networks based on their weighted $\ell_1$ distance to the center found using Odin. There is an interesting correlation between distance calculated by Odin for each gene and the average degree of that gene in the protein-interaction network (Figure S1C).

## Experiments Details and Commands

In the union of the ASD/ID datasets considered in this study (Table 1) there are a total of 684 affected ASD/ID cases/probands with LGD variants, and 245 control and unaffected siblings with LGD variants. We compared the results of Odin against k-NN, SVM, glmnet (Lasso and elastic-net), and random forest for predicting ASD/ID with low FPR $( < 1\%)$. We used a LOO approach to compare these methods. We used the scores/confidence/probability outputted by each method for each prediction to control the number of unaffected samples predicted as case by mistake (denoted as FPR). More specifically, in k-NN, we used the difference of number of affected cases and unaffected controls in the *k* closest neighbor; for SVM and generalized linear models, we used the predicted probability (or distance) given by the libSVM (Chang and Lin 2011) or glmnet (Friedman *et al.* 2010); for random forest, we changed the probability cut-off value for determining cases. For both glmnet and libSVM, the optimal set of training parameters was first picked considering all the data as input. As such, parameters “gamma” and “cost” of libSVM and the parameter *s* of glmnet were set to optimal values learned using the full dataset.

The exact commands used for each program is as follows:

### *SVM experiments*

The command for training and testing used for SVM is based on libSVM version 3.21 implementation (Chang and Lin 2011). Using the full dataset, we first found the optimal parameters for “gamma” and “cost” and set these to 0.25 and 0.03125, respectively, for the libSVM classifier. Then, for the LOO experiment, we used the following commands in training dataset: svm-train -b 1 -w0 5 -w1 1 -c 0.03125 -g 0.25 training-data, and, in the case of test data, we used the following command: svm-predict -b 1 testing-data training-data.model output.

### *Lasso and elastic-net (glmnet) experiments*

The commands used for glmnet (Lasso and elastic-net; Friedman *et al.* 2010). In training dataset, we use the following command: fit=glmnet(training-data.features, training-data.class, alpha=a (we ran with parameters $a \in \{0, 0.25, 0.5, 0.75, 1\}$, and, in the case of test data, we used predict(fit, testing-data, s=0.042645) (the value s was calculated as *lambda.min* as instructed in https://web.stanford.edu/hastie/glmnet/glmnet_alpha.html).

### *K-NN experiments*

We implemented the k-NN classifier and tested and reported the results for *k* ranging from 1 to 20.

### *Random forest experiments*

We used the package randomForest in R with the following command for training

```
rf <- randomForest (training-data.features, training-data.class, ntree = 1000)
```

and the following command for testing

```
pred <- predict(rf, testing-data, type=”vote”)
```