Ultra-accurate complex disorder prediction: Case study of neurodevelopmental disorders

Linh Huynh[1] and Fereydoun Hormozdiari[1,2,3,*]

[1] Genome Center, UC Davis [2] UC Davis MIND institute

[3] Department of Biochemistry and Molecular Medicine, UC Davis

April 22, 2017

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 have a significant impact on increasing the effectiveness of interventions and treatments in improving the prognosis and, in many cases, enhancing the quality of life in the 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. We provide a novel formulation for the **U**ltra-**A**ccurate **D**isorder **P**rediction (UADP) problem and develop a novel combinatorial framework for solving this problem. The primary goal of this novel framework, denoted as Odin (**O**racle for **DI**sorder predictio**N**), is to make an accurate prediction for a subset of affected cases while having virtually zero false positive predictions for unaffected samples. Note that in the Odin framework we will take 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 known recurrent mutated genes in the training set and with a less than 0.5% false positive prediction based on our analysis of unaffected controls. Furthermore, Odin predicted a set of 391 genes that severe variants in these genes can cause autism or other developmental delay disorders. Odin is publicly available at https://github.com/HormozdiariLab/Odin [†]

1 Introduction

The start of the genomics era and sequencing of the first human genome over a decade ago promised significant benefits to public health [1]. These include the potential capability of early detection, pinpointing the causes, and developing novel treatments and therapeutics for most diseases. The sequencing of the human genome has dramatically accelerated biomedical research; however, even a decade after publication, 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 [2–4].

One of the first questions that we need to answer regarding any disease of interest is to calculate the contribution of genetics to the etiology of the disease. The primary metric used for calculating the contribution of genetics to any trait, including diseases, is denoted as genetic heritability ($0 \leq h^2 \leq 1$) [5,6]. However, in many complex diseases, the calculated genetic heritability is by far higher than the fraction of cases that can be explained or predicted by the observed genetic variants. This gap is known as the "missing

[*] contact: fhormozd@ucdavis.edu

[†] A preliminary version of this paper was accepted for presentation in RECOMB-2017 conference.

heritability" problem and is one of the main hindrances in not only building early prediction models for complex disorders but also developing novel treatments [7,8].

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) [9]. ASD is typically diagnosed around the age of two 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 ($0.5 \leq h^2 \leq 0.8$) [10,11].

It is becoming apparent that early treatment and intervention can significantly improve the IQ, language skills, and social interactions in children affected with ASD [12–14]. 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 [15]. However, it is theoretically possible to make an accurate diagnosis of ASD or other neurodevelopmental disorders in some children before any symptoms appear (or even before the child is born) using (perinatal) genetic testing and genome sequencing [16]. Thus, building methods for *early prediction of ASD using genetic variation and other biomarkers* is extremely important and will have significant direct positive effects on public health.

The recent advances in high-throughput sequencing (HTS) technologies have given us the capability to sequence the whole genomes or exomes of many samples. For instance, the consortia focused on complex disorders, such as autism, ID, schizophrenia, and diabetes sequenced tens of thousands of cases [17–20]. Sequencing of samples with these complex disorders had produced genetic maps with tens of thousands of variants with most of them being rare (i.e., minor allele frequency - MAF < 0.05). Unfortunately, in most cases the effect of the variants found is not known (i.e., a variant of unknown significance) and for rare variants it is extremely hard to assign significance solely considering their frequency. In extreme cases of *de novo* variants (i.e., novel variants not inherited from the parents), the same exact variant will likely never be seen in any other sample. Thus, building models for early prediction of complex disorders need to be more sophisticated than just considering the frequency of variants in cases and controls.

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) [21]. Furthermore, there are known rare, large recurrent copy number variations, such as the 16p11.2 deletion or Prader-Willi syndrome, which are known to cause ASD. However, in most cases of ASD, the exact cause of the disorder is not known and no accurate method or model for early prediction of ASD using genetic variants exists. Recently, consortia focused on sporadic ASD have performed whole-exome sequencing on thousands of autism families (affected proband, unaffected sibling and parents) with the hope of finding causative variants in these samples. The enrichment of *de novo* variants in affected probands versus unaffected siblings has indicated that a significant fraction of ASD is the result of *de novo* and rare (MAF < 0.05) variants [17,22]. However, in many cases, it is not clear which *de novo* or rare variants are the real culprit(s) of the phenotype. As we do not expect to see the same *de novo* variant to appear in two different samples, it has proven useful to summarize the observed variants on the genes being affected. This simple approach has provided researchers with enough statistical power to accurately ($p < 0.01$) predict tens of novel ASD genes with high penetrance of likely gene disruptive (LGD) and missense variants. However, these statistically significant genes only cover a fraction of ASD cases estimated to be caused by *de novo* or rare variants. Based on the twin studies, it is estimated that ASD and ID have a genetic heritability of over 0.5, while we can optimistically explain less than 0.2 fraction of the affected children [17, 22] based on observed common or rare genetic variants (including structural variation and copy number variation [23–25]).

It is estimated that hundreds of genes are involved in neurodevelopment and disruption in them can cause ASD or ID [22, 26]. The primary justification for such a high number of genes (high genetic heterogeneity) contributing to a similar disorder (i.e., ASD) is that most of these genes are members of only a few functional "modules" or pathways [27, 28]. Thus, disruption of any of these functional modules results in a similar disorder mainly through interruption of normal neurodevelopment. It is being hypothesized that by using the functional relationship between these genes it is possible to find the causative variants.

Complex disorder prediction using (rare) coding variants: As mentioned above, early prediction of complex disorders using genetic variation is one of the fundamental goals of personalized medicine. Currently, thousands of cases with neurodevelopmental disorders (e.g., ASD) have been studied using WES or targeted sequencing. Thus, we have a very rich set of rare and common coding variants found in samples with ASD and other neurodevelopmental disorders, which can be used for building early prediction models and methods. However, it is also important to realize the intrinsic limitations of rare coding variants in predicting ASD or other complex disorders. Notably, (i) most complex disorders have genetic heritability of significantly less than 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 of a complex disorder (such as autism) with very high accuracy. Therefore, achieving accurate prediction for *all (or even most) affected cases* using solely the coding variants is theoretically not achievable. On the flip side, this also means, theoretically, that we cannot confidently predict a sample as an unaffected control solely based on the observed (coding) variants, as other factors (e.g., environment, epigenetic) can contribute to the disorder. Thus, instead of trying to predict the status of every input sample as affected case or unaffected control, which theoretically is not possible, *we propose to only predict a subset of samples as affected cases with very low false prediction/discovery.*

Ultra-Accurate Disorder Prediction (UADP) problem: A positive diagnosis/prediction of a complex disorder (e.g., ASD) can have a severe negative psychological and economical 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. *Thus, one of the main practical constraints in developing models and methods for prediction of a severe complex disorder is to guarantee a false positive prediction/discovery rate (FDR) of virtually zero.* In other words, it is highly desirable not to have a false prediction of an input unaffected control as an affected case. Note, 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 this paper, we study the UADP problem and provide a framework for solving this problem using rare coding variants. We aim to develop a computational method for positive prediction of a significant fraction of affected cases (due to the prediction limitation from coding variants) with virtually zero false positive prediction of unaffected controls (due to the negative effects of false positive prediction). We choose ASD as a case study since we can utilize a rich dataset of *de novo* mutations. In addition, we also integrate the functional relationship to increase the prediction capability. Approaches such as the one presented in this paper are needed not only to close the missing heritability in many complex disorders but also to translate the biomedical discoveries into actionable items by clinicians.

2 Methods

2.1 UADP problem definition and notations

In the UADP problem we are trying to maximize the number of samples correctly predicted as affected case, while the number of unaffected controls falsely predicted as affected cases must be extremely small. Another way to look at this problem is that we are trying to select a subset of samples such that the total number of unaffected controls picked is negligible while the number of affected cases selected is maximized. Finally, note that because of low recurrence of the same rare and *de novo* variants, we will be using a summarization of coding rare variants based on their effect on each gene and the biological function disrupted to increase our power for prediction.

Training Data: Let n and m be the number of genes and the total number of samples respectively. The

LGD (likely gene disruptive) 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

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

An assumption here is that an LGD mutation will completely knockout or disrupt the copy of the affected gene in the sample. The *diagnosis result (or class)* of the i^{th} sample is a binary value y_i where

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

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: We will use the functional similarity between genes to increase the statistical power in disorder prediction. The assumption is that disruption of genes with similar functionality will result in similar phenotypes. Thus, we would like to develop a framework that can include the similarities between genes (mutational landscape and function) as an additional signal for disease prediction. We denote such a matrix by $P \in [0,1]^{n \times n}$ where $P_{i,j}$ indicates the similarity between genes i^{th} and j^{th} and potentially how disruption of one gene can affect the other gene. For neurodevelopmental disorders such as ASD our goal is to build the matrix P to reflect functional similarity of genes during brain development. One way to calculate such a score for any pair of genes is based on using the coexpression of genes during brain development. Coexpression between two genes i and j is denoted by $R(i,j)$ and is calculated to represent the expression similarity of these two genes in different conditions and tissues. Similar to previous practices, we are using the Pearson correlation of expression profiles between two genes in different conditions as the coexpression value [27, 29–32]. Coexpression has been shown to be a powerful indicator of functional similarity of two genes for neurodevelopment. We also include the similarity of likelihood of observing LGD mutation (pLI) between the two genes in the population [33] for building this matrix. Assuming we are given multiple matrices capturing the similarity of genes with each matrix using different biological concepts, we will use the minimum of similarity scores of two genes among different matrices to build matrix P. In another words, if we have two matrices of gene similarity P' and P'', the matrix P is built by assigning $P_{i,j} = min(P'_{i,j}, P''_{i,j})$. Of course, the choice of these matrices and the way we combine them together can be changed without any need to change the underlying framework and proposed methods. The details of different datasets used to build the matrix P for this study is provided in Section 3.1. We will 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 $D'_{control} = \{\mathbf{z_i} = \mathbf{x_i} \times P \mid \mathbf{x_i} \in D_{control}\}$ and $D'_{case} = \{\mathbf{z_i} = \mathbf{x_i} \times P \mid \mathbf{x_i} \in D_{case}\}$.

2.2 Odin framework

In this subsection, we will introduce the intuition behind our framework Odin (**O**racle for **DI**sorder predictio**N**) as a practical solution for the UADP problem. To build such a conservative prediction model, Odin will intuitively predict 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 satisfying the first condition, we simply use the nearest neighbor approach using a distance function (e.g., Euclidean distance). The closest neighbor of the input sample among the training data should be an affected case so that input sample passes the first condition.

For satisfying the second condition, we will initially 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 controls. 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 (section 1 in the supplemental material) and can not be solved efficiently. Therefore, we propose a relaxation of UCDR that we denote as Weighted Unicolor Clustering with Dimensionality Reduction (WUCDR). In the remainder of this section, we will first formalize the UCDR and WUCDR problems and then present an iterative algorithm to solve the WUCDR problem.

2.2.1 Unicolor Clustering with Dimensionality Reduction (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 constraint 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 $\mathbf{c}$

As a general rule any metric distance function (e.g., Euclidean distance) can be used for the UCDR problem. However, we are using the ℓ_1 distance since it is concordant with the additive model used in common variant studies. The ℓ_1 distance between two points $(a_1, a_2, ..., a_n)$ and $(b_1, b_2, ..., b_n)$ is defined as $\sum_{i=1}^{n} \mid a_i - b_i \mid$. We will denote the region contained with distance r from center $\mathbf{c} \in \mathbb{R}^{|k|}$ as *area of interest* $\mathcal{A}(\mathbf{c}, r)$. 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 small number of training samples. In practice, we will require that the number of selected dimensions be less than $O(\log_2(m))$ (i.e., $k = O(\log_2(m))$) where m is the total number of training samples (both cases and controls).

2.2.2 Weighted Unicolor Clustering with Dimensionality Reduction (WUCDR) Problem

Since UCDR problem is NP-complete (see Section 1 in the supplemental material), we will define a relaxation, where we assign (continuous value) weights to the dimensions. We denote this problem as the Weighted Unicolor Clustering with Dimensionality Reduction (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** ℓ_1 as the distance metric for clustering (we use the notation $w\ell_1$ to represent weighted ℓ_1). In the rest of the paper, we will define the weighted ℓ_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 k dimensions thus, over $n - k$ other dimensions will have weight zero.

2.2.3 Iterative solution for WUCDR

Here we propose an iterative approach to solve the WUCDR problem. It is consist of two main steps. 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 $\mathcal{A}(\mathbf{c}, r)$ (note that the area of interest is considered using weighted ℓ_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 $\mathbf{w} = (w_1, w_2, ..., w_n)$ (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 ℓ_1 distance greater than r to center $\mathbf{c}$ and

(2) the number of blue points, which have weighted ℓ_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 and can be very time-consuming. Thus, we will relax the problem only to 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 and the weights $\mathbf{w}$ we can easily calculate the affected cases (i.e., blue points) covered by the area of interest. Let set S denote the 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 will calculate new weights $\mathbf{w}$ (for each dimension). The objective is to decrease the weighted ℓ_1 distance of points in the set S to center $\mathbf{c}$, while increasing the weighted ℓ_1 distance of points in the set S to the red points ($D'_{Control}$). We will solve the linear programming (LP) problem below to find these new weights

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

Note that in the above LP problem only $\mathbf{w}$ and ρ are unknown variables, while the set S and center $\mathbf{c}$ are calculated in the first step of the method. The constraints in the above LP problem will 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, these weights try to 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. The objective function of the above 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 ℓ_1 distance of all red points to the blue points in set S. Finally, among the weights produced we will only keep the top k weights and convert all of the remaining weights to 0. Note that because of the condition $\sum_{i=1}^{n} w_i \le k/2$, we are guarantied to be able to keep any dimension with value > 0.5 from the LP solution.

Odin framework using WUCDR. As mentioned in the Section 2.2, two conditions should be satisfied for a sample to be predicted as a potential affected case by Odin. The first condition is that the nearest neighbor to the samples should **not** be an unaffected control. Odin uses the ℓ_1 distance function for finding the nearest neighbors of any test sample. The second condition is that the input sample should fall inside the *area of interest* $\mathcal{A}(\mathbf{c}, r)$ 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).

3 Results

3.1 Data Summary

We tested Odin for accurate prediction of neurodevelopmental disorders using the LGD (likely gene disruptive) *de novo* variants from WES and targeted sequenced samples with ASD or ID. Table 1 shows the total number of samples and LGD variants reported from the union of several publications on over 6,000 ASD/ID probands. The union dataset of *de novo* variants used from these publications can be found in [34].

Class - ASD diagnosis (affected or unaffected)	Study	Number of samples	Number of LGD variants	References
Affected ASD/ID probands	Simons Simplex Collection (SSC)	2,508	492	[17, 22, 35–37]
	Autism Sequencing Consortium (ASC)	2,270	185	[18]
	Other studies	1,329	74	[38–40]
	Total cases	6,107	751	
Unaffected siblings or controls	Simons Simplex Collection (SSC)	1,909	248	[17, 22]
	GoNL	250	7	[41]
	Other studies	208	11	[19]
	Total control	2,367	266	

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

For building the gene similarity matrix P, which is used to convert the input variant vector for every sample (i.e, $\mathbf{z_i} = \mathbf{x_i} \times P$), we have used the combination of coexpression values between two genes during brain development [27] and the difference between the likelihood of observing LGD variants in population [33]. We can trivially extend the matrix P to include additional data such as tissue specific networks [42]. We observed that using such a matrix to map the variant vector for each sample into new space results in significantly reducing the ℓ_1 distance of probands with each other ($p < 1.6e-16$). This indicates that using such a transformation indeed helps in increasing the prediction power.

Considering the samples in Table 1, there are few genes with significant recurrence of *de novo* LGD variants in affected cases while having no *de novo* variant in unaffected controls. Any prediction model/method for ASD can be trivially extended to predict a sample as an affected case if they have an LGD *de novo* variant in any of these genes. Thus, in our test data we will not consider any samples with *de novo* variants in any of these genes that are recurrently mutated in our training data. We will call these samples trivial cases/samples and the remaining samples as nontrivial cases/samples. Note that there are nine genes with four of more LGD variants in union of these ASD/ID samples (Table 1). These nine genes are *ADNP, ANK2, ARID1B, CHD2, CHD8, DSCAM, DYRK1A, SCN2A* and *SYNGAP1* and any sample with an LGD variant in any of these genes is considered a trivial case to predict and it is not considered in our analysis.

3.2 Unicolor clustering with dimension reduction

We will first show that the proposed iterative method in Section 2.2.3 for solving the WUCDR problem (number of dimensions selected < 10) does in fact greatly improve the number of cases covered in comparison to the unweighted result (considering all dimensions with weights $w_i = 1$). As shown in Table 2, the optimal result found using the input was only able to cover 45 cases (only 24 of the cases not having LGD variants in recurrently mutated genes, i.e., the number of nontrivial cases covered). However, the WUCDR approach in less than five iterations was able to cover over 71 cases (40 of the cases not having LGD variants in significantly recurrent mutated genes). Thus, our iterative approach for WUCDR improves the number of affected cases covered by over 60% using less than 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 not only the number of cases covered was improved using the WUCDR approach but also the density was increased (see Table 2) per iteration.

Iteration	Number of cases covered	Number of non-trivial cases covered	Density (case/radius)	Number of dimensions (before - after) rounding
0	45	24	0.11	$\geq$ 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)

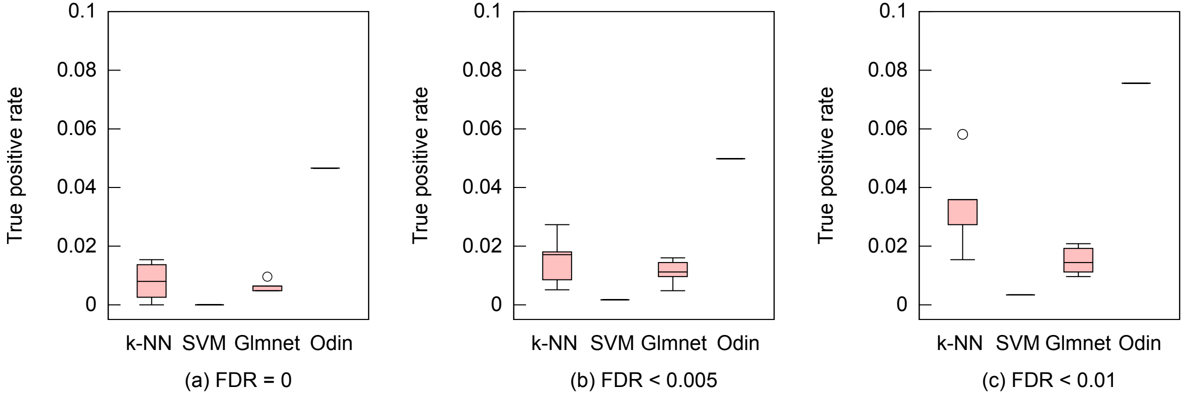
Table 2: Number of ASD/ID affected probands from training dataset covered in each iteration.

3.3 ASD/ID disease prediction results

We will compare the Odin framework in predicting affected ASD cases in comparison to different classification methods. We have used the k-NN classifier (various k-values), support vector machines (SVM) [43], and (lasso and elastic-net) regularization of generalized linear models [44]. We are specifically interested in comparing these methods in predicting ASD in *nontrivial* cases. We will use the leave-one-out (LOO) technique to compare the Odin framework versus prediction power of k-NN, SVM classifiers, (lasso and elastic-net) generalized linear models. As our stated goal is to keep the false positive prediction of unaffected samples as cases close to zero, we will only consider the most conservative results for each method (FDR < 0.01). As can be seen, Odin's true positive rate for predicting ASD is at least two times higher than the best k-NN result (for different values of k) and significantly higher than SVM (Figure 1). It is also significantly higher for different regularized generalized linear models (lasso and elastic net) for different input parameters of α (Glmnet implementation). For each of these tools, we used their intrinsic properties to control/limit the FDR for calculating the TDR. 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 [43] or Glmnet [44]. For Odin the weighted ℓ_1 distance of the sample to the selected center was used. Using these values we could calculate the highest true positive rate for each method given the FDR value using LOO cross-validation. Note that in Odin the full set of samples predicted as affected cases will have an FDR of less than 0.01.

3.4 Developmental delay disorder (DDD) predictions

In addition to the samples reported in Table 1, a set of over 4000 trios with developmental delay disorder (DDD) were whole-exome sequenced [45]. Since linear model (i.e. SVM or Glmnet) was not suitable for this prediction problem (as seen in Figure 1), we only compared Odin against the k-NN approach ($1 \leq k \leq 10$) using the dataset in Table 1 as a training set and the (nontrivial) DDD affected cases [45] for testing (Figure 2). Note that we used the parameters from the LOO cross-validation experiment from the previous section (Section 3.3) to control the FDR. The Odin method was able to accurately predict a higher fraction of nontrivial DDD probands in comparison to k-NN approach (Figure 2a) using the ASD/ID samples as training. We further investigated the overlap between nontrivial DDD affected samples which were correctly



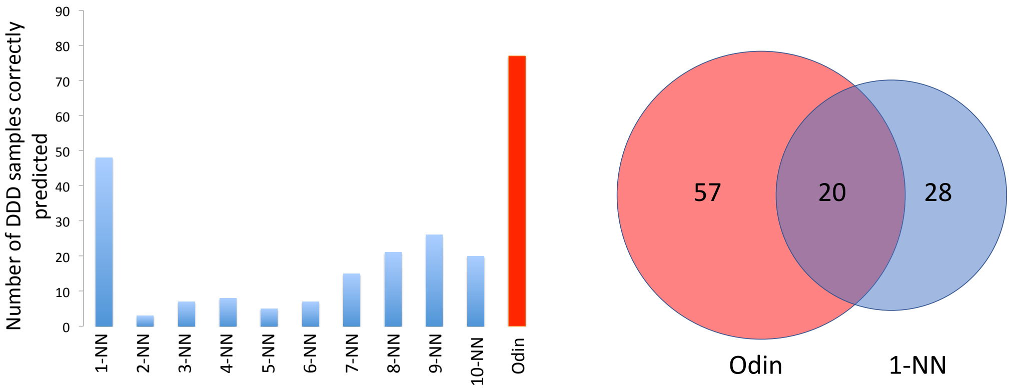
(a) FDR = 0

(b) FDR < 0.005

(c) FDR < 0.01

Figure 1: Prediction results on ASD/ID data set. For k-NN the boxplot shows the result of different values of k tested and for GLmnet the boxplot shows the results for different values of $0 \leq \alpha \leq 1$.

predicted by Odin and 1-NN (nearest neighbor) approach (Figures 2b). Interestingly, there are significant number of samples which were correctly predicted only by one of the methods, which indicates an approach which combines different methods can even outperform Odin. Similar results was observed for FDR < 0.01 (Supplementary Figure 1).



(a) $FDR = 0$

(b) $FDR = 0$

Figure 2: Prediction results on DDD data set

3.5 Autism and developmental delay gene prediction and ranking

Odin is a framework to predict with ultra-accuracy a subset of samples that will develop ASD given the *de novo* variation; however, it can also be used to predict some novel ASD genes. We have utilized Odin to rank all genes for the potential impact of a *de novo* LGD variant disrupting them. Note that similar to the predictions of ASD/ID for subsets of samples made by Odin, if a gene is not selected does not mean it is not an ASD/ID gene. We used the ASD and Siblings variants (Table 1) as the training data and calculated the weighted ℓ_1 distance to the selected center. Ranking all the genes based of the calculated distance, we clearly see an enrichment of known ASD genes closer to the center (Figure 3a). For the set of known ASD genes we used the union of SFARI high-confidence genes and known syndromic genes.

Our analysis also indicated that there are 391 genes where an LGD variant on them will result in a sample falling inside the predicted area of interest (i.e., the inner most sphere/circle in Figure 3a). These 391 genes indicate a set of genes in which Odin has predicted with high probability that their disruption will cause significant (neuro)developmental disorder. Furthermore, there was significant enrichment of LGD variants in these 391 genes in the DDD set (which was not used in the training) versus the ASD set (which was used in the training) as shown in Figure 3b. Interestingly, this clearly indicates that even after normalizing based on expected LGD variants for each disease group the more severe samples tend to be more enriched in LGD variants than their less severe autism samples disrupting these selected 391 genes (Figure 3b). There is also an enrichment of severe *de novo* missense variants (i.e., with CADD score $>$ 25) disrupting these genes (Figure 3c) in affected autism/ID/DDD probands while no such enrichment is seen for control/sibling samples (Figure 3c).

In the Simons Simplex Collection (SSC) we also observed not only that probands with LGD variant tend to have a lower IQ than probands without *de novo* LGD variants, but also the probands with *de novo* LGD variant disrupting one of the genes in the most inner sphere (the 391 genes) have lower IQ than other probands with *de novo* LGD variants (Figure 4a). It is been known that their is a large male to female bias in autism (estimate to be over 4 to 1). Note that in the Simons Simplex Collection (SSC) there are a total of 2478 male probands and 396 female probands (over 6:1 ratio). However, the difference between the number of samples with LGD *de novo* variants in the selected 391 genes in the inner most sphere is 31 to 16 (around 2:1 ratio). This indicates that there is much smaller gap for sex difference for ASD samples with *de novo* LGD variants in the predicted ASD genes by Odin (Figure 4b). We were also interested to see if there are any specific enrichment of expression of these top 391 genes selected in human brain. We used the online tool CSEA (http://genetics.wustl.edu/jdlab/csea-tool-2/) to study the expression profile of these 391 genes. Interestingly, the only significant expression we observed was on the early fetal development and mid-early fetal development of brain (Figure 4c). No significant expression of these genes in any tissues in adult human or mouse brain was observed. Finally, we utilized the predicted probability of observing a missense or LGD *de novo* variant per sample for each gene [35] to calculate the p-values. We could group these genes based on observing significant *de novo* LGD and/or missense variants in affected probands (Figure 4d). The set of genes with only significant missense *de novo* variants observed in cases potentially indicates genes in which an LGD variant will be incompatible with life (i.e., essential genes). However, a missense mutation can result in a severe (neuro)developmental disorder. These genes include *CSNK2A1, SMARCA4, TRRAP, MORC2, PRPF8, TAF1, CNOT1, SF3B1, SMAD4, UBR5, CLASP1, KDM2B,* and *U2AF2*.

3.6 Pathways

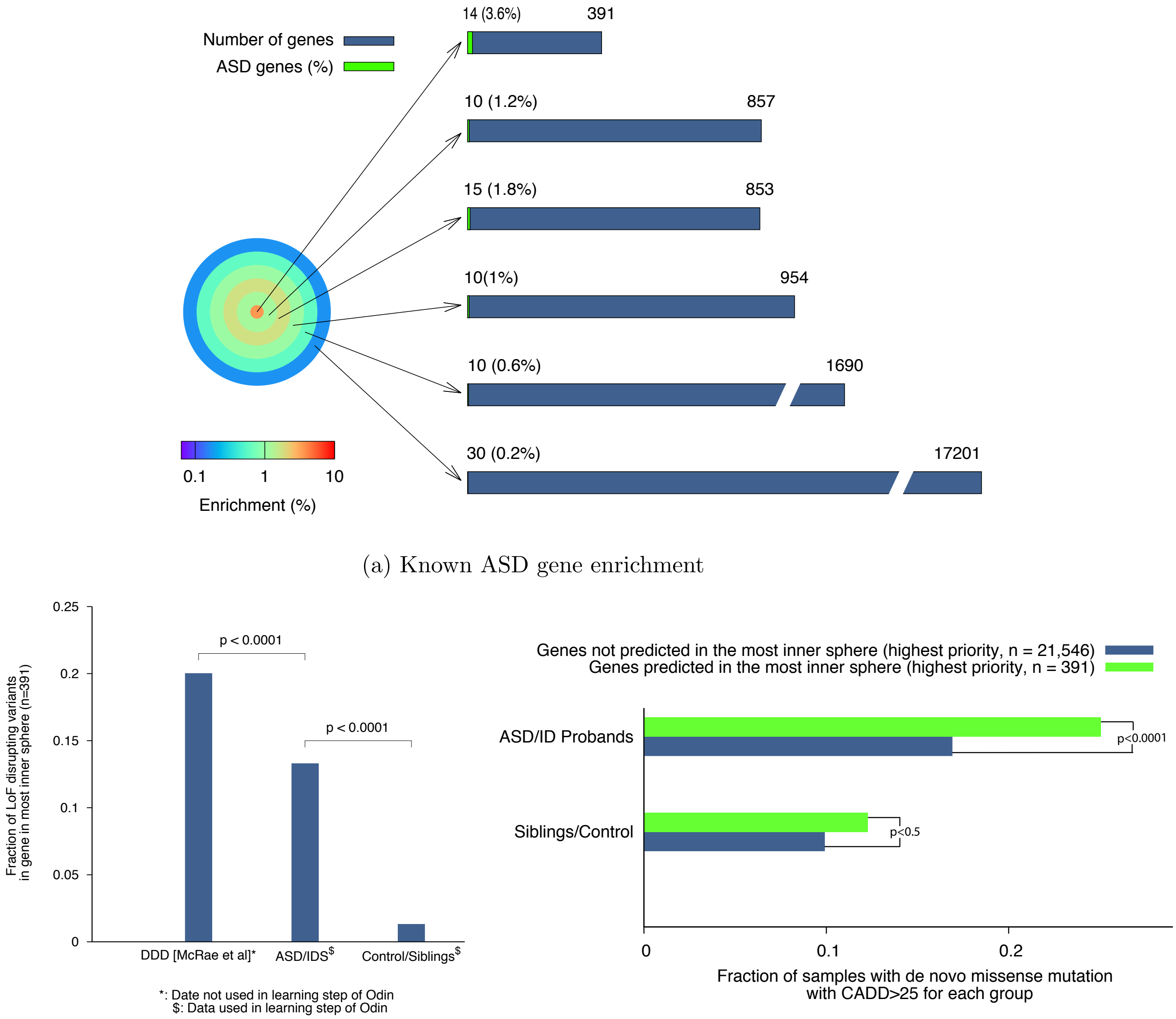
We were interested in studying the properties of the samples that Odin correctly predicted as an affected case. We used the tool David (v.6.7) [46] for the discovery of enriched GO-terms and KEGG pathways for the genes mutated in these samples. For the ASD/ID samples in Table 1, Odin was able to correctly predict ASD status of samples that have *de novo* variants in genes in Wnt pathways [27, 47, 48] and in chromatin regulation [18, 35] (Figure 5a). Similarly, correctly predicted DDD study samples [45] had mutations in chromatin modification and transcription regulation genes (Figure 5b).

4 Conclusion and future works

In this paper, we introduce Odin, a framework for an early prediction of ASD and related disorders from rare genetic variants. Our initial evaluation of the experimental data shows a clear power of this approach in ultra-accurate prediction of ASD using rare genetic variants. The proposed framework can be extended to take into account not only LGD mutations but also *missense mutations* to increase the power of the model in predicting a higher percentage of affected cases. 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 [49] or polyphen-2 score [50]) 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). In addition, we can integrate other information, such as protein interaction [51, 52], tissue-specific networks [42] or the regulation of specifically related pathways such as Wnt [48] or mTOR [53], to increase the prediction capability. For the algorithm, we can improve the first guessed solution (in the first step) of WUCDR (see Section 2.2.3) by utilizing algorithmic techniques in geometry. Finally, our proposed framework here can be extended for predicting the risk of other neurological disorders, such as schizophrenia, epilepsy or Alzheimer.

5 Acknowledgment

We would like to thank Evan E. Eichler, Tychele N. Turner, Phuong Dao, Farhad Hormozdiari, Madeleine Geisheker and Tonia Brown for reading the paper and providing comments.



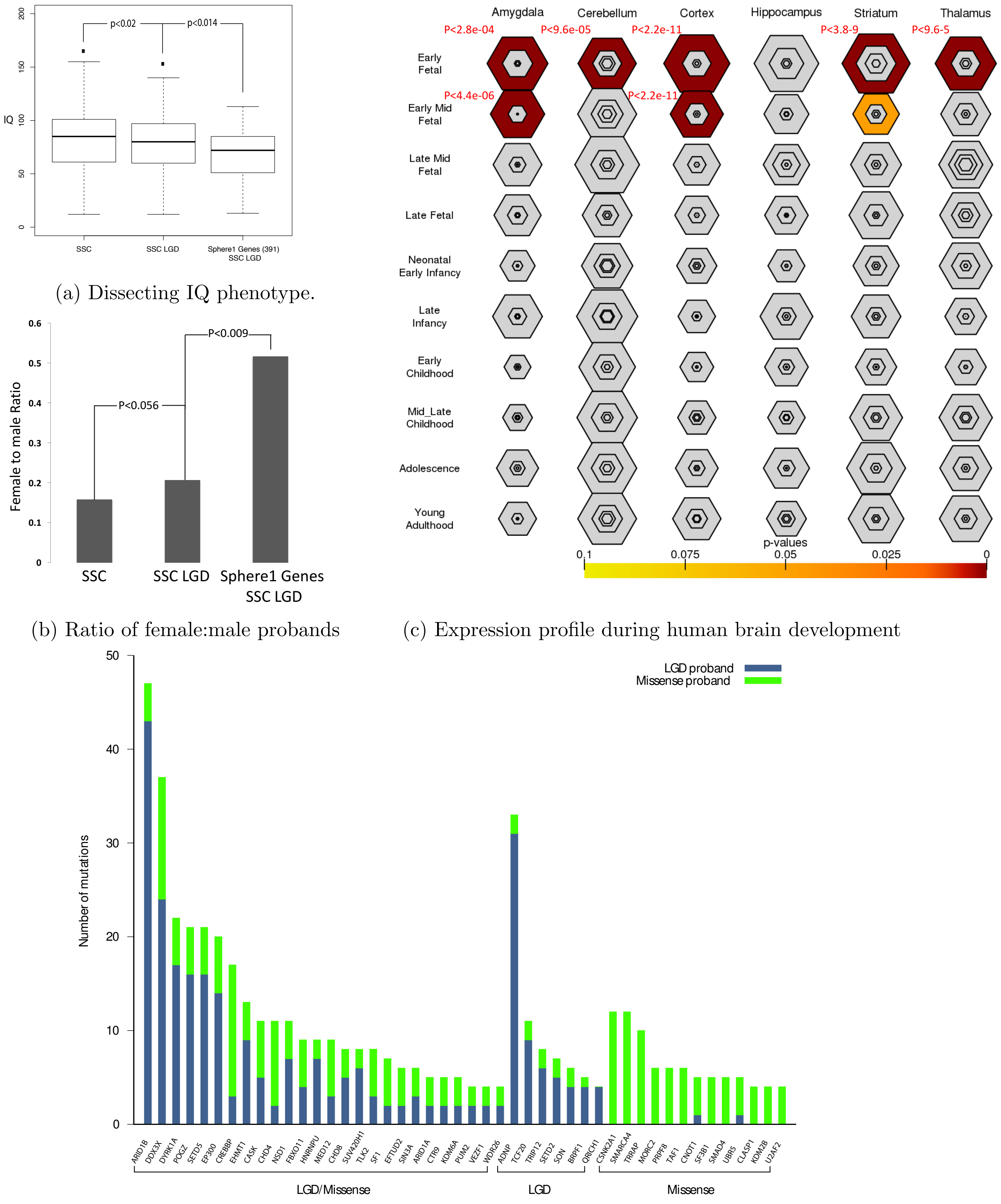
*: Date not used in learning step of Odin
$: Data used in learning step of Odin

(a) Known ASD gene enrichment

(b) LGD variant enrichment

(c) Missense variants (with CADD > 25) enrichment

Figure 3: a) The set of genes which are closer - based on the weighted ℓ_1 distance - to the center selected by Odin are more enrichment in known ASD genes. b) Significant enrichment of LGD variants in ASD/ID/DDD disrupting the genes in inner most sphere (391 total genes). c) Significant enrichment of severe missense variants (CADD$>$ 25) disrupting these 391 genes in ASD/ID/DDD probands.



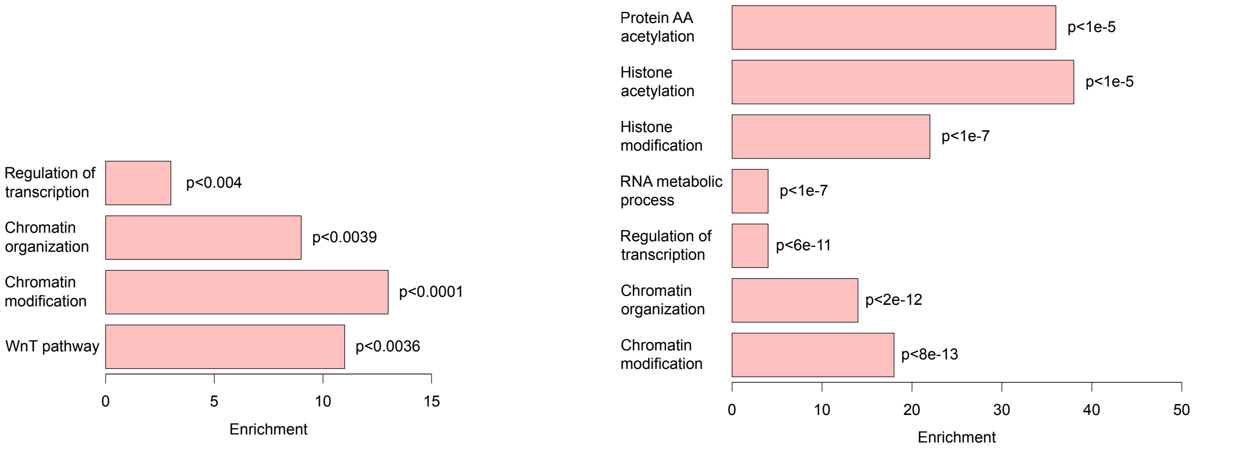
(a) Dissecting IQ phenotype.

(b) Ratio of female:male probands

(c) Expression profile during human brain development

(d) Genes with significant *de novo* variants (with minimum of 4 *de novo* variants).

Figure 4: Properties of genes selected by Odin in the inner most sphere (391 total genes). a) The ASD probands with LGD variant in these genes have significant lower IQ than other ASD probands with LGD variant. b) The gap between female:male ratio is significantly smaller for probands which have LGD variants in these 391 genes. c) These genes are expressed during early fetal and early mid-fetal brain development. d) The subset of genes in inner most sphere which are significantly enriched in variants in probands.



(a) Samples from ASD (Table 1) correctly predicted

(b) Samples from DDD [45] correctly predicted

Figure 5: Pathways and Gene Ontology (GO) enrichment.

References

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, *et al.*, "Exome sequencing identifies the cause of a mendelian disorder," *Nature genetics*, vol. 42, no. 1, pp. 30–35, 2010.

[3] Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, *et al.*, "Clinical whole-exome sequencing for the diagnosis of mendelian disorders," *New England Journal of Medicine*, vol. 369, no. 16, pp. 1502–1511, 2013.

[4] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[5] N. Wray and P. Visscher, "Estimating trait heritability," *Nature Education*, vol. 1, no. 1, p. 29, 2008.

[6] D. S. Falconer, *Introduction to quantitative genetics.* Pearson Education India, 1975.

[7] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[8] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[9] A. P. Association *et al.*, "Diagnostic and statistical manual of mental disorders 5 2013," *American Psychiatric Association.*

[10] S. Sandin, P. Lichtenstein, R. Kuja-Halkola, H. Larsson, C. M. Hultman, and A. Reichenberg, "The familial risk of autism," *JAMA*, vol. 311, no. 17, pp. 1770–1777, 2014.

[11] B. Tick, P. Bolton, F. Happé, M. Rutter, and F. Rijsdijk, "Heritability of autism spectrum disorders: a meta-analysis of twin studies," *Journal of Child Psychology and Psychiatry*, 2015.

[12] L. A. Vismara and S. J. Rogers, "The early start denver model: A case study of an innovative practice," *Journal of Early Intervention*, 2008.

[13] B. A. Boyd, S. L. Odom, B. P. Humphreys, and A. M. Sam, "Infants and toddlers with autism spectrum disorder: Early identification and early intervention," *Journal of Early Intervention*, vol. 32, no. 2, pp. 75–98, 2010.

[14] P. Howlin, I. Magiati, T. Charman, and W. E. MacLean, Jr, "Systematic review of early intensive behavioral interventions for children with autism," *American journal on intellectual and developmental disabilities*, vol. 114, no. 1, pp. 23–41, 2009.

[15] S. H. Kim, S. Macari, J. Koller, and K. Chawarska, "Examining the phenotypic heterogeneity of early autism spectrum disorder: subtypes and short-term outcomes," *Journal of Child Psychology and Psychiatry*, vol. 57, no. 1, pp. 93–102, 2016.

[16] J. O. Kitzman, M. W. Snyder, M. Ventura, A. P. Lewis, R. Qiu, L. E. Simmons, H. S. Gammill, C. E. Rubens, D. A. Santillan, J. C. Murray, *et al.*, “Noninvasive whole-genome sequencing of a human fetus,” *Science translational medicine*, vol. 4, no. 137, pp. 137ra76–137ra76, 2012.

[17] I. Iossifov, B. J. ORoak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, *et al.*, “The contribution of de novo coding mutations to autism spectrum disorder,” *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[18] S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, K. Samocha, A. E. Cicek, Y. Kou, L. Liu, M. Fromer, S. Walker, T. Singh, L. Klei, J. Kosmicki, F. Shih-Chen, B. Aleksic, M. Biscaldi, P. F. Bolton, J. M. Brownfeld, J. Cai, N. G. Campbell, A. Carracedo, M. H. Chahrour, A. G. Chiocchetti, H. Coon, E. L. Crawford, S. R. Curran, G. Dawson, E. Duketis, B. A. Fernandez, L. Gallagher, E. Geller, S. J. Guter, R. S. Hill, J. Ionita-Laza, P. Jimenz Gonzalez, H. Kilpinen, S. M. Klauck, A. Kolevzon, I. Lee, I. Lei, J. Lei, T. Lehtimki, C.-F. Lin, A. Ma’ayan, C. R. Marshall, A. L. McInnes, B. Neale, M. J. Owen, N. Ozaki, M. Parellada, J. R. Parr, S. Purcell, K. Puura, D. Rajagopalan, K. Rehnstrm, A. Reichenberg, A. Sabo, M. Sachse, S. J. Sanders, C. Schafer, M. Schulte-Rther, D. Skuse, C. Stevens, P. Szatmari, K. Tammimies, O. Valladares, A. Voran, W. Li-San, L. A. Weiss, A. J. Willsey, T. W. Yu, R. K. C. Yuen, D. Study, H. M. C. for Autism, U. Consortium, E. H. Cook, C. M. Freitag, M. Gill, C. M. Hultman, T. Lehner, A. Palotie, G. D. Schellenberg, P. Sklar, M. W. State, J. S. Sutcliffe, C. A. Walsh, S. W. Scherer, M. E. Zwick, J. C. Barett, D. J. Cutler, K. Roeder, B. Devlin, M. J. Daly, and J. D. Buxbaum, “Synaptic, transcriptional and chromatin genes disrupted in autism.,” *Nature*, vol. 515, pp. 209–215, Nov 2014.

[19] S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton, S. Casadei, C. Rippey, H. Shahin, C. on the Genetics of Schizophrenia (COGS), P. S. Group, V. L. Nimgaonkar, R. C. P. Go, R. M. Savage, N. R. Swerdlow, R. E. Gur, D. L. Braff, M.-C. King, and J. M. McClellan, “Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network.,” *Cell*, vol. 154, pp. 518–529, Aug 2013.

[20] B. Xu, J. L. Roos, P. Dexheimer, B. Boone, B. Plummer, S. Levy, J. A. Gogos, and M. Karayiorgou, “Exome sequencing supports a de novo mutational paradigm for schizophrenia,” *Nature genetics*, vol. 43, no. 9, pp. 864–868, 2011.

[21] A. O. Caglayan, “Genetic causes of syndromic and non-syndromic autism,” *Developmental Medicine & Child Neurology*, vol. 52, no. 2, pp. 130–138, 2010.

[22] N. Krumm, T. N. Turner, C. Baker, L. Vives, K. Mohajeri, K. Witherspoon, A. Raja, B. P. Coe, H. A. Stessman, Z.-X. He, S. M. Leal, R. Bernier, and E. E. Eichler, “Excess of rare, inherited truncating mutations in autism.,” *Nature genetics*, vol. 47, pp. 582–588, Jun 2015.

[23] G. M. Cooper, B. P. Coe, S. Girirajan, J. A. Rosenfeld, T. H. Vu, C. Baker, C. Williams, H. Stalker, R. Hamid, V. Hannig, *et al.*, “A copy number variation morbidity map of developmental delay,” *Nature genetics*, vol. 43, no. 9, pp. 838–846, 2011.

[24] B. P. Coe, K. Witherspoon, J. A. Rosenfeld, B. W. Van Bon, A. T. Vulto-van Silfhout, P. Bosco, K. L. Friend, C. Baker, S. Buono, L. E. Vissers, *et al.*, “Refining analyses of copy number variation identifies specific genes associated with developmental delay,” *Nature genetics*, vol. 46, no. 10, pp. 1063–1071, 2014.

[25] D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, *et al.*, “Functional impact of global rare copy number variation in autism spectrum disorders,” *Nature*, vol. 466, no. 7304, pp. 368–372, 2010.

[26] M. Ronemus, I. Iossifov, D. Levy, and M. Wigler, "The role of de novo mutations in the genetics of autism spectrum disorders," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 133–141, 2014.

[27] F. Hormozdiari, O. Penn, E. Borenstein, and E. E. Eichler, "The discovery of integrated gene networks for autism and related disorders.," *Genome research*, vol. 25, pp. 142–154, Jan 2015.

[28] A. Krishnan, R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, A. Tadych, N. Volfovsky, A. Packer, A. Lash, and O. G. Troyanskaya, "Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder," *Nature Neuroscience*, 2016.

[29] A. J. Willsey, S. J. Sanders, M. Li, S. Dong, A. T. Tebbenkamp, R. A. Muhle, S. K. Reilly, L. Lin, S. Fertuzinhos, J. A. Miller, *et al.*, "Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism," *Cell*, vol. 155, no. 5, pp. 997–1007, 2013.

[30] N. N. Parikshak, R. Luo, A. Zhang, H. Won, J. K. Lowe, V. Chandran, S. Horvath, and D. H. Geschwind, "Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism," *Cell*, vol. 155, no. 5, pp. 1008–1021, 2013.

[31] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *science*, vol. 302, no. 5643, pp. 249–255, 2003.

[32] I. Ulitsky and R. Shamir, "Identifying functional modules using expression profiles and confidence-scored protein interactions," *Bioinformatics*, vol. 25, no. 9, pp. 1158–1164, 2009.

[33] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, A. O'Donnell-Luria, J. Ware, A. Hill, B. Cummings, *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *BioRxiv*, p. 030338, 2016.

[34] T. N. Turner, Q. Yi, N. Krumm, J. Huddleston, K. Hoekzema, H. A. Stessman, A.-L. Doebley, R. A. Bernier, D. A. Nickerson, and E. E. Eichler, "denovo-db: a compendium of human de novo variants," *Nucleic Acids Research*, p. gkw865, 2016.

[35] B. J. O'Roak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler, "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations.," *Nature*, vol. 485, pp. 246–250, Apr 2012.

[36] T. N. Turner, F. Hormozdiari, M. H. Duyzend, S. A. McClymont, P. W. Hook, I. Iossifov, A. Raja, C. Baker, K. Hoekzema, H. A. Stessman, M. C. Zody, B. J. Nelson, J. Huddleston, R. Sandstrom, J. D. Smith, D. Hanna, J. M. Swanson, E. M. Faustman, M. J. Bamshad, J. Stamatoyannopoulos, D. A. Nickerson, A. S. McCallion, R. Darnell, and E. E. Eichler, "Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory dna.," *American journal of human genetics*, vol. 98, pp. 58–74, Jan 2016.

[37] B. O'roak, H. Stessman, E. Boyle, K. Witherspoon, B. Martin, C. Lee, L. Vives, C. Baker, J. Hiatt, D. A. Nickerson, *et al.*, "Recurrent de novo mutations implicate novel genes underlying simplex autism risk," *Nature communications*, vol. 5, 2014.

[38] J. J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, M. Jian, G. Liu, D. Greer, A. Bhandari, W. Wu, R. Corominas, A. Peoples, A. Koren, A. Gore, S. Kang, G. N. Lin, J. Estabillo, T. Gadomski, B. Singh, K. Zhang, N. Akshoomoff, C. Corsello, S. McCarroll, L. M. Iakoucheva, Y. Li, J. Wang, and J. Sebat, "Whole-genome sequencing in autism identifies hot spots for de novo germline mutation.," *Cell*, vol. 151, pp. 1431–1442, Dec 2012.

[39] R. Hashimoto, T. Nakazawa, Y. Tsurusaki, Y. Yasuda, K. Nagayasu, K. Matsumura, H. Kawashima, H. Yamamori, M. Fujimoto, K. Ohi, S. Umeda-Yano, M. Fukunaga, H. Fujino, A. Kasai, A. Hayata-Takano, N. Shintani, M. Takeda, N. Matsumoto, and H. Hashimoto, "Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder.," *Journal of human genetics*, vol. 61, pp. 199–206, Mar 2016.

[40] A. Rauch, D. Wieczorek, E. Graf, T. Wieland, S. Endele, T. Schwarzmayr, B. Albrecht, D. Bartholdi, J. Beygo, N. Di Donato, A. Dufke, K. Cremer, M. Hempel, D. Horn, J. Hoyer, P. Joset, A. Rpke, U. Moog, A. Riess, C. T. Thiel, A. Tzschach, A. Wiesener, E. Wohlleber, C. Zweier, A. B. Ekici, A. M. Zink, A. Rump, C. Meisinger, H. Grallert, H. Sticht, A. Schenck, H. Engels, G. Rappold, E. Schrck, P. Wieacker, O. Riess, T. Meitinger, A. Reis, and T. M. Strom, "Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study.," *Lancet (London, England)*, vol. 380, pp. 1674–1682, Nov 2012.

[41] G. of the Netherlands Consortium *et al.*, "Whole-genome sequence variation, population structure and demographic history of the dutch population," *Nature Genetics*, vol. 46, no. 8, pp. 818–825, 2014.

[42] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, *et al.*, "Understanding multicellular function and disease with human tissue-specific networks," *Nature genetics*, vol. 47, no. 6, pp. 569–576, 2015.

[43] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[44] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[45] J. F. McRae, S. Clayton, T. W. Fitzgerald, J. Kaplanis, E. Prigmore, D. Rajan, A. Sifrim, S. Aitken, N. Akawi, M. Alvi, *et al.*, "Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation," *bioRxiv*, p. 049056, 2016.

[46] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[47] S. R. Gilman, I. Iossifov, D. Levy, M. Ronemus, M. Wigler, and D. Vitkup, "Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses," *Neuron*, vol. 70, no. 5, pp. 898–907, 2011.

[48] H. O. Kalkman, "A review of the evidence for the canonical wnt pathway in autism spectrum disorders," *Molecular autism*, vol. 3, no. 1, p. 1, 2012.

[49] M. Kircher, D. M. Witten, P. Jain, B. J. ORoak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature genetics*, vol. 46, no. 3, p. 310, 2014.

[50] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature methods*, vol. 7, no. 4, pp. 248–249, 2010.

[51] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, no. 1, p. 88, 2007.

[52] Y.-A. Kim, S. Wuchty, and T. M. Przytycka, "Identifying causal genes and dysregulated pathways in complex diseases," *PLoS Comput Biol*, vol. 7, no. 3, p. e1001095, 2011.

[53] G. Tang, K. Gudsnuk, S.-H. Kuo, M. L. Cotrina, G. Rosoklija, A. Sosunov, M. S. Sonders, E. Kanter, C. Castagna, A. Yamamoto, *et al.*, "Loss of mtor-dependent macroautophagy causes autistic-like synaptic pruning deficits," *Neuron*, vol. 83, no. 5, pp. 1131–1143, 2014.

1 Supplementary materials

1.1 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, G. J., & Yu, Z. (1992). On the equal-subset-sum problem. Information Processing Letters, 42(6), 299-302].*

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 non-empty 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 "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 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 re-write the equation as

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

where $I_1 \cap I_2 = \emptyset$ and $I_1 \cup I_2 = I$. Since $B'_1 = (a_{i_1}, a_{i_2}, \ldots, a_{i_d}) \in F$ then

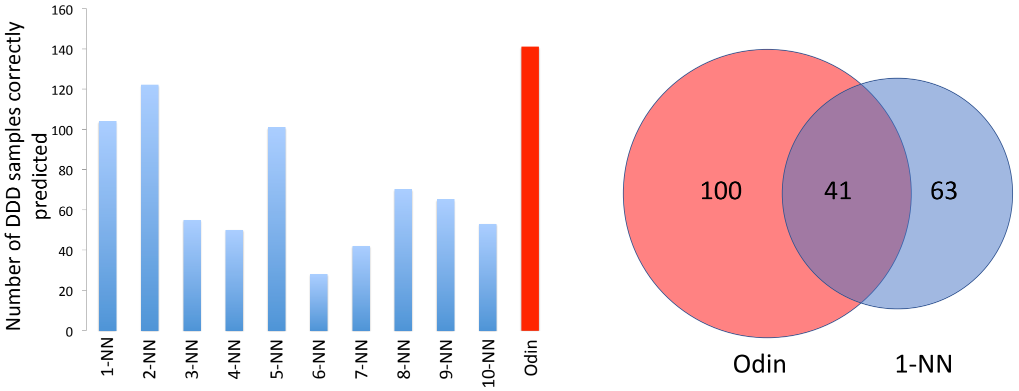
$$\sum_{i_j \in I_1} a_{i_j} - \sum_{i_k \in I_2} a_{i_k} = 0$$

but both a_{i_j} and a_{i_k} are in A that contains positive numbers only so $I_1 \neq \emptyset$ and $I_2 \neq \emptyset$. Therefore, the pair of two sets $A_1 = \{a_{i_j} \mid i_j \in I_1\}$ and $A_2 = \{a_{i_k} \mid i_k \in I_2\}$ 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. □

1.2 Developmental delay disorder (DDD) prediction using Odin

We further analysis Odin's capability in accurate prediction of DDD probands with FDR < 0.01 while using the ASD/ID data (samples in Table 1) as training.



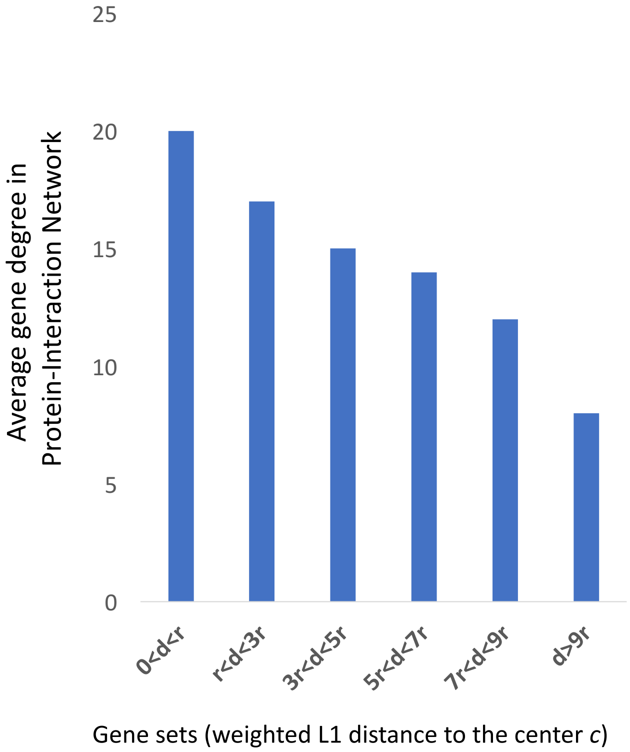
(a) $FDR < 0.01$

(b) $FDR < 0.01$

Supplementary Figure 1: Prediction results on DDD data set

1.3 Protein interaction enrichment

We investigated the changes in genes degree in protein-interaction networks based on their weighted ℓ_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 genes in protein-interaction networks (Supplementary Figure 2).



Supplementary Figure 2: The average degree of genes is higher for set of genes which are closer to the center. The center and the weighted ℓ_1 distance is learned by Odin.