Computational genomics and genetics of developmental disorders

Hongjian Qi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Computational genomics and genetics of developmental disorders

Hongjian Qi

Computational genomics is at the intersection of computational applied physics, math, statistics, computer science and biology. With the advances in sequencing technology, large amounts of comprehensive genomic data are generated every year. However, the nature of genomic data is messy, complex and unstructured; it becomes extremely challenging to explore, analyze and understand the data based on traditional methods. The needs to develop new quantitative methods to analyze large-scale genomics datasets are urgent. By collecting, processing and organizing clean genomics datasets and using these datasets to extract insights and relevant information, we are able to develop novel methods and strategies to address specific genetics questions using the tools of applied mathematics, statistics, and human genetics.

This thesis describes genetic and bioinformatics studies focused on utilizing and developing state-of-the-art computational methods and strategies in order to identify and interpret *de novo* mutations that are likely causing developmental disorders. We performed whole exome sequencing as well as whole genome sequencing on congenital diaphragmatic hernia parents-child trios and identified a new candidate risk gene *MYRF*. Additionally, we found male and female patients carry a different burden of likely-gene-disrupting mutations, and isolated and complex patients carry different gene expression levels in early development of diaphragm tissues for likely-gene-disrupting mutations.

To increase the power to detect risk genes and risk variants, we developed a deep neural network classifier called MVP to accurately predict the pathogenicity of missense variants. MVP implemented an advanced structure of ResNet model and based on two independent data sets,

MVP achieved clearly better results in prioritizing pathogenic variants than other methods. Additionally, we studied the genetic connection between developmental disorders and cancer. We found that in developmental disorder patients predicted deleterious *de novo* mutations are more enriched in cancer driver genes than non cancer driver genes. A Hidden Markov Model was implemented to discover cancer somatic missense mutation hotspots and we demonstrated many cancer driver genes shared a similar mode of action in developmental disorders and caner. By improving ability to interpret missense mutations and leveraging cancer genomics data, we can improve risk gene inference in developmental disorders.

Contents

List of Figures

List of Tables

Acknowledgements

First, and foremost, I would like to express my deepest gratitude to my thesis advisor, Dr.Yufeng Shen, who has provided invaluable mentorship and guidance through my Ph.D life. Four years ago, I approached Yufeng to be my research advisor with limited background in biology and I am so thankful that he was willing to take the risk and gave me the opportunity to explore the area of human genetics. It is a great journey to do research in this exciting field as a medical detective and I really appreciated Yufeng's immense knowledge and tremendous support for me.

I am very lucky to have many collaborators and my sincere thanks to Dr. Wendy Chung, who is like a co-mentor to me. Wendy has great taste for research problem, extraordinary scientific acumen and is a visionary leader in clinical genetics. I truly cherish the opportunities to conduct research with her group.

I am grateful to my thesis committee – Dr. Chris Wiggins, Dr. Aron Pinczuk, Dr. Kyle Mandli and Dr. Raul Rabadan for their insights and suggestions during committee meetings and their willingness to discuss my research and provide feedbacks.

I also would like to thank the patients and their families for their generous contribution to the study and I am grateful for the experimental research collaborators for their hard work to generate the data used in analysis.

I could ever thank my family enough, especially my mother, for her never-ending support, selfless love, sacrifice and dedication.

Finally, thank you, Ga-in, Gakki, all lab mates from Shen lab, and so many friends who have been with me and supported me along this journey.

To my parents

Chapter 1

Introduction

1.1 Overview

The richness of data science lies in its connection with real-world problems: its goal is to solve real problems, and to do so we have to develop new tools with models rooted in these domains. Computational genomic and genetics is at the intersection of computational applied physics, applied math, statistics, computer science and biology, it is also big data science as tens of Zeta bytes of genomics data have been generated every year trying to solve domain specific genetic problems[1]. However, the genomics data is messy, unstructured and heterogeneous; it has become extremely challenging to understand, analyze, and interpret multi-dimensional genomics datasets with traditional methods. New technologies are needed to address the computational challenges and integrate various types of data from next generation sequencing experiments.

Over the past few decades, technological and methodological advances in human genetics and genomics have allowed accurate identification of genetic mutations in patients that are involved in rare diseases[2-7]. One of the fundamental problems of medical genetics is to associate the patients' genotype to the phenotype[8], which is the clinical feature of the rare diseases. This problem is complicated by various factors, such as the large mount of mutations observed in each individual and diverse type of human genetic variations from large chromosome abnormality to a single nucleotide change. Additionally, the genetic architecture varies across different diseases; some diseases are monogenic and caused by high penetrance variants[9] while some diseases are polygenic and involved combination actions of many genes with small effect size mutations[10,11].

Currently more than 1,500 genes have been found to cause a broad range of developmental disorder diseases[12,13], while many more disease associated risk genes still remained to be discovered. Unfortunately, for patient carries a rare developmental disorder, unless the patient does have a well-established pathogenic variant, it still remains a considerable

challenge to make accurate diagnosis and establish the connection between the disease and the tens of thousands of variants identified in the individual. In these situations, mutations must be prioritized to make further investigation[12].

The primary focus of this thesis is to develop novel computational methods to aid prioritization of mutations that are likely to be contributing to the diseases and identify risk genes that are associated with the disease.

1.2 Next generation sequencing

Next generation sequencing (NGS) technologies, also known as high-throughput sequencing, have revolutionized human genetics through massively parallel sequencing of multiple genes simultaneously[12,14,15]. NGS technologies are particularly useful to identify small genetic variations such as single nucleotide variants (SNVs) or small insertions and deletions (indels) that might contribute to diseases[15,16]. The sensitivity to detect SNVs and indels is very high for NGS technologies and has been proved with superior quality to detect mutations[17]. With the advancement of parallel sequencing and continuous decreasing in cost per genome, it is feasible to perform large-scale sequencing on collected parents data, a typical research strategy is illustrated in Fig 1.1.

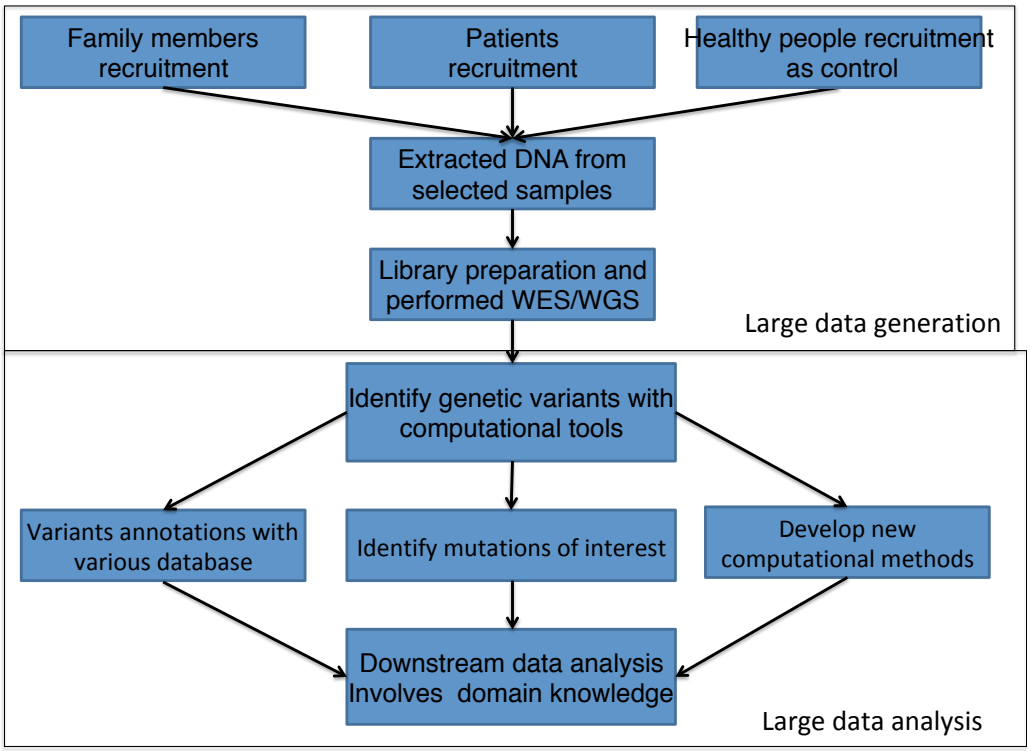


Fig1.1 Strategy and key steps to apply NGS to human genetics

NGS can be used to sequence entire genomes. Whole genome sequencing (WGS) studies have proved fruitful in uncovering risk candidate genes and disease associated mutations.[18] The strategy to sequence family members including the child and both of their unaffected parents offers the ability to identify *de novo* mutations that are only occurred in the child and filter out rare benign inherited variants from parents. Those *de novo* mutations are sufficiently rare and multiple mutations hits in a gene from unrelated patients provide strong evidence for a causal link to the diseases. Consequently, for families where the disease affects neither parent, sequencing of the parents-child trio rather than proband alone can dramatically increase the clinic diagnostics in individuals with potential genetic disorders. NGS can also sequence specific areas of interest of human genome, currently the most widely used targeted sequencing region is

whole exome sequencing (WES)[19]. The creation of exome-capture kits allowed researchers to sequence only coding regions, such sequencing experiments is faster and cheaper than sequence the whole genome, therefore accelerating the discovery of protein-coding mutations that are associated with disease. NGS can also be used to detect large genetic variations including copy number variation (CNV)[20], which includes amplification or deletions of segments with more than thousands of base pairs of DNA. Recently WGS and WES have become primary choices for CNV detection and for studying of human diseases. Researchers have demonstrated the large impact of CNVs on a wide collection of pediatric conditions[21], including congenital heart disease and other various developmental disorders[22].

Successful application of sequencing technology studies can be used to diagnose rare, severe, and likely monogenic disorders, such as Kabuki syndrome where the missense mutations in *KMT2B* that were considered causal, usually occurred as *de novo* mutations in the affected individual[23]. Another example is the application of WES/WGS to epileptic encephalopathy, a severe brain disorder that can be caused by multiple variants in multiple genes, where researchers discovered 31 novel genes using WES technologies recently[24]. These studies proved that sequencing technology is especially useful and critical for identifying disease-associated genes and mutations.

1.3 Interpretation of sequence variants

Identification and interpretation of the genetic variants responsible for causing diseases can be very challenging in clinical genetic testing since many variants, even in well-established risk genes, are classified as variants of uncertain significance (VUS)[25], unless they are highly recurrent in patients. Our current categorization of genetic mutation falls in a range given the

understanding from clinical significance. A variant is almost certainly pathogenic for a disorder disease if it is a mutation directly contributes to the development of disease and is well established as disease causing in the literature and databases with a wide consensus[9,25], while a benign variant is considered not to be the cause of the disease[26] and mostly likely having no effects on patients.

The genetic interpretation of missense variants is particularly challenging because missense variants are the most abundant type of coding mutations and play important roles in a wide range of human genetic diseases, there was often insufficient patients sample size to determine whether the amino acid change is either detrimental or neutral as variants that were likely to cause the tested disease are usually deleterious and had a low population frequency under severe selection[27]. In order to improve the power to identify damaging missense variants given the same sample size, many *in silico* methods such as CADD[28], VEST3[29], metaSVM[30], M-CAP[31], and REVEL[32] have been developed to utilize information from allele frequency in population, protein structures, conservation and advanced machine learning model such as gradient boosted decision trees. Those methods facilitated the interpretation and predictions by defining damaging missense variants with prediction sore suppress certain threshold. Unfortunately, those methods sometimes yield discordant predictions and have limited performance in recent large scale sequencing data. In Chapter 5, we proposed a new prediction method, MVP, which uses a deep learning approach to leverage large training data sets and achieved better performance in prioritizing pathogenic missense variants than previous methods.

1.4 Genetic basis of congenital diaphragmatic hernia

Congenital diaphragmatic hernia (CDH) is a severe birth defect of the diaphragm and lungs affecting about 1 per 3,000 live births[33,34]. In infants with CDH, malformation of the diaphragm creates a hole in the corner that could allow the abdominal organs to push into the chest cavity and thus disrupt the normal lungs development. About half of CDH patients are syndromic cases with associated anomalies including pulmonary hypoplasia, pulmonary hypertension and heart failure[35,36] while the remaining are isolated cases. Despite advances in prenatal and postnatal care, CDH is still a life-threatening pathology in infants with high mortality and morbidity.

The etiology of CDH is largely unknown in most cases, but there is strong evidence that genetic factors play an important role in the development of CDH[37]. The genetic contribution can be established by familial aggregation[37], rare disease associated with CDH[38], chromosome abnormalities[9]. In about one third of CDH patients, potential genetic causes can be identified in a wide range of genetic defects, such as small genetic variations such as snps or indels, or large genetic variants like chromosomal anomalies or copy number variations. Trisomy 13, 18, 21 are the most frequent CDH-associated aneuploidies[38]. Most single genes identified in CDH through the analysis of recurrent chromosomal anomalies[38]. Individual genes implicated in CDH including *GATA4, ZFPM2, NR2F2* and *WT1*, many of them encode transcription factors and are pleiotropic genes that effect diaphragm development and have also been associated with other congenital anomalies in heart, brain, and genitalia.

CDH is usually a sporadic condition, which refers to that CDH occurs infrequently within families; the low reproductive fitness of CDH patients can led to the hypothesis that *de novo* mutations with large effect sizes may explain a significant fraction of CDH patients. In 2013, the

DHREAMS study recruited 39 trios of unaffected parents and CDH children and performed WES. They observed an excess burden of likely to be deleterious *de novo* mutations among genes highly expressed during diaphragm development[7]. No recurrent gene is identified with more than two damaging *de novo* mutations, indicating the genetic heterogeneity and potentially big number of candidate genes that could cause CDH. In chapter 4, we replicated the results of an excess of *de novo* mutation burden with a cohort of 357 CDH trios and identified a new CDH risk gene *MYRF*.

1.5 Thesis outline

The remaining of this thesis will be organized into five chapters.

In Chapter 2, we described the detail procedures and pipelines used to detect and characterize *de novo* mutation in parents-child trio studies.

In Chapter 3, we described the genetic analysis of congenital diaphragmatic hernia (CDH). We complied genetic data from whole exome and genome sequencing of a cohort of 357 child-parent trios and identified *MYRF* as a new candidate risk gene for CDH. *MYRF* harbored four deleterious *de novo* mutations in four unrelated CDH patients, which is more than expected significantly (p-value < 10^{-9}). We had also demonstrated that there are different genetic architectures for female and male CDH patients without additional anomalies: female isolated cases carry a substantial contribution from *de novo* mutations in whereas male isolated cases carry little contribution from *de novo* mutations. A manuscript, *Genetic analysis of de novo variants reveals sex differences in complex and isolated congenital diaphragmatic hernia and indicates MYRF as a candidate gene*, is currently under review.

In Chapter 4, we examined the link between the genetic component of developmental disorders and cancers through analyzing whether genes associated with cancer could affect genes contribute developmental disorders. We observed a significant enrichment of loss of function and predicted damaging missense variants in cancer driver genes among cases with developmental disorders; then we proposed a Hidden Markov Model to demonstrate that predicted damaging missense *de novo* mutations are enriched in cancer mutation hotspots, suggesting a similar mode of dysregulation of the mutated proteins. Results have been published in the paper: *Deep genetic connection between cancer and developmental disorders.*

In Chapter 5 we developed a deep learning method called MVP to better predict and identify missense variants pathogenicity. MVP used advanced ResNet structure to train on large mounts of putative pathogenic variants aggregated from several curated clinical databases. To explicitly consider the difference in mode of action and genetic effect size of pathogenic missense variants, we trained MVP model for constrained genes and non constrained genes separately. We assessed the performance of MVP along with other methods using two independent datasets, *de novo* germline mutations from recent large-scale genetic studies and cancer somatic mutation hotspots, and showed that MVP achieved much better precision under similar sensitivity than previously published methods, especially in genes that are not severely constrained. The manuscript *MVP: predicting pathogenicity of missense variants by deep learning* is currently under review.

Chapter 6 will present overall conclusions and discuss future research directions to better understand the genetics of developmental disorders.

Chapter 2

Identification and characterization of *de novo* mutations

2.1 Introduction

Next-generation sequencing (NGS) has become more affordable and accessible to researchers and clinical geneticists. As a result, our understanding of the genetics of developmental disorders has rapidly advanced over the past few years. A recent highlight is that *de novo* mutations play an important role in sporadic diseases such as autism spectrum disorder[39] and congenital heart disease[5], those mutations are either newly formed during gamete formation or occur very early in embryonic development and, thus, are unique to the child when compared to the parent. New disease associated risk genes are discovered by the recurrent *de novo* mutations within same gene and excess of *de novo* mutations when compared to background led to the discovery of new pathways and prioritization of genes that relevant to the disease[40].

In principle, detecting *de novo* mutations is straightforward. We search for Mendelian error in the child genotype, which is an allele in the individual that could not have been inherited from either of its biological parents. One major issue in *de novo* mutation identification is to distinguish true positives from false positives, it is complicated by sequencing errors, sequence reads mapping errors as well as the rare occurrences of *de novo* mutations. Therefore, we needed to establish a pipeline with rigorous thresholds to determine high quality candidate *de novo* mutations.

2.2 Genetic variation data generation from child-parents family

Identification of *de novo* mutations in one family requires the genetic information of the child and both parents. The study designs of finding *de novo* mutations that only exist in affected children but not in healthy parents using trio based sequencing data are similar for patients with severe early onset diseases. In the work of congenital diaphragmatic hernia study, we performed whole exome sequencing (WES) at the University of Washington of 79 trios. We collected genomic DNA from whole blood or saliva and/or skin/diaphragm tissue samples from the affected patients and both parents. Genomic DNA (~3 μg) was extracted and sheared to 200-300 base pairs using covaris acoustic adaptor, the fragments were end-repaired, adenylated, and sequencing adaptor oligonucleotides ligated for sequencing preparation. Libraries were barcoded using the Illumina index read strategy and was subsequently enriched for sequences with 5′ and 3′ adapters by PCR amplification with primers complementary to the adapter sequences. Exon was captured with the Nimblegen SeqCap EZ Exome V2 exome capture reagent (Roche). Samples were multiplexed and sequenced with paired-end 75bp reads on Illumina HiSeq 2500 platform according to the manufacturer's instructions. The sequencer outputted the paired end raw sequencing data into two matching FASTQ format files.

The exome sequencing variant calling pipeline was based on the Broad Institute's best practices. Briefly, we aligned FASTQ files of raw sequence reads to the reference genome (build GRCh37) using BWA-mem software[41], and then we used Picard (v1.67) software to mark PCR duplicates. Variants were jointly called using the Genome Analysis Toolkit (GATK) HaplotypeCaller in all WES samples[42], and GATK generated a standard Variant Call Format (VCF) file with genetic variation information for the sequenced samples. We annotated the VCF

file using ANNOVAR software[43] to have complete gene annotation as well as function annotation. Further downstream analyses such as identification of *de novo* mutation can be performed by in house scripts. Figure 2.1 provides an overview of the exome sequencing data processing pipeline.

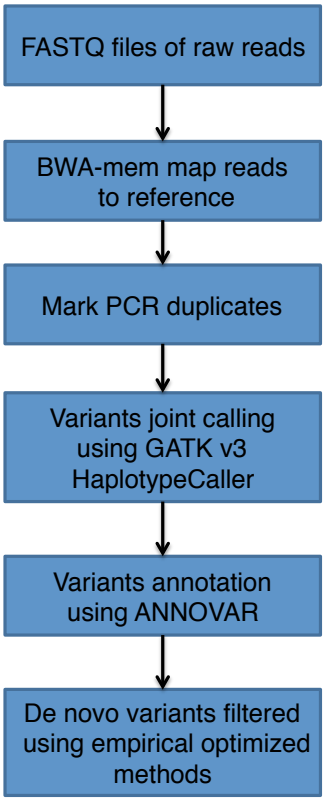


Figure 2.1 Exome sequencing variant calling pipeline

2.3 Pipeline to identify *de novo* mutations

All candidate *de novo* mutations were called from child-parents sequencing data using customized scripts written in python. The scripts took two inputs, one GATK generated VCF file with genotype information for each individuals and one pedigree file, which described the family relationship. To identify potential *de novo* mutations, we first selected the sites where the child had a heterozygous or homozygous alternative genotype and both parents had homozygous reference, then we filtered for high quality variants based on genotype information and various site-level annotations, which summarize context information from the samples as well as information from other databases.

The filters we used to identify high confident *de novo* mutations are empirical optimized. In order to remove miscalled genotype in the proband, we first set a threshold of alternate allele balance (minimum 20% if alternate read depth greater than or equal to 10 or minimum 28% if alternate read depth less than 10). Ideally the child with a heterozygous mutation should have 50% of the sequencing reads carrying the alternative allele, however, there is sequencing error and a slight bias towards reference since it is easier to capture sequences with the reference allele than the alternative allele, so we allowed the minimum allele balance of the child to be 20%. To avoid miscalling a homozygous genotype in the parents, we required the alternate allele balance to be less than 3.5% in the parents. We further filtered based on depth (minimum 10 reads total and 5 alternate allele reads) in proband and minimum depth of 10 reference read in parents, failing such requirements indicates the sites are poorly sequenced and we were under power to determine true genotype in the samples. We also filtered mutation based on PL, which is normalized Phred-scaled likelihoods of the possible genotypes considered in the variant record

for each sample and GQ, which is assigned genotype quality. The basic formula for calculating PL is:

$$PL = -10 * \log\ P(Genotype|Data)$$

where P(Genotype | Data) is the conditional probability of the Genotype given the sequence data that we have observed. For the typical case of a monomorphic site (one single alternative allele) in a human diploid cell, the PL field will contain three numbers, corresponding to the three possible genotypes (0/0, 0/1, and 1/1). The PL values are "normalized" so that the PL of the most likely genotype is 0 and all others are scaled relative to the most likely genotype. Therefore, we set a minimum PL of 60 in the proband which corresponds to the genotype in question being a million times less likely to be the true genotype than the reported most likely genotype. The value of GQ is simply the difference between the second lowest PL and the lowest PL (which is always 0), since we only want to keep the reference genotype in parents; we set the minimum GQ to be 30 in the parents.

For site-level annotation filtering, we set the max Fisher Strand to 25 to reduce sequence bias and minimum Quality by Depth to 2 to reach a high variant confidence. We also required the max population frequency to be 0.1% from ExAC database[44] and cohort allele count to be 6 since given the size of the dataset, any mutation was seen multiple times in other individuals in population or the same data set was not likely to be a true *de novo* mutation. Additionally, variants located in segmental duplication regions (maximum score 0.98) were excluded.

In the end, all candidate *de novo* variants were manually inspected in the Integrated Genomics Viewer[45] (IGV, http://software.broadinstitute.org/software/igv/) to future remove false positive.

2.4 Quality control of *de novo* mutations

The Sanger sequencing method was developed by Frederick Sanger and colleagues in 1977[46], it was the most widely used sequencing method to sequence single strands of DNA with a high degree of certainty and had become the gold standard for DNA sequencing. It is widely accepted that variants found using NGS should be validated with Sanger confirmation. To support our choice of filters used to find *de novo* mutations identification, we submitted some of the identified *de novo* variants to Sanger confirmation. In a recent study of congenital heart disease, we submitted 409 *de novo* mutations for validation by Sanger sequencing, 394 of them were confirmed (specificity 96.3%)[47]. In the study of congenital diaphragmatic hernia described in Chapter 3, we validated all the *de novo* likely gene disrupting (including frameshift, nonsense and splicing site) mutations by Sanger sequencing, all 40 were confirmed (specificity 100%). Overall, more than 95% of variants were confirmed to be *de novo*, indicating high precision and confirming the overall robustness of the approach.

The average *de novo* mutation rate is estimated to be 1.2×10^{-8} per nucleotide per generation[48], so we can estimated that about 1 *de novo* per trio will be observed in coding region. The overall *de novo* mutation rate in the 1213 probands with congenital heart disease (1.05 events per proband on average) was consistent with a background *de novo* mutation rate[47]. Additionally, the average number of synonymous *de novo* variants per trio is comparable between cases and controls samples that are identified by similar pipelines.

Given a particular *de novo* mutation rate, due to random variation, the frequency of *de novo* mutations detected per proband is expected to follow a Poisson distribution[49]. In the congenital heart disease study, the observed data in both case and controls is consistent with the Poisson process model (Fig. 2.2).

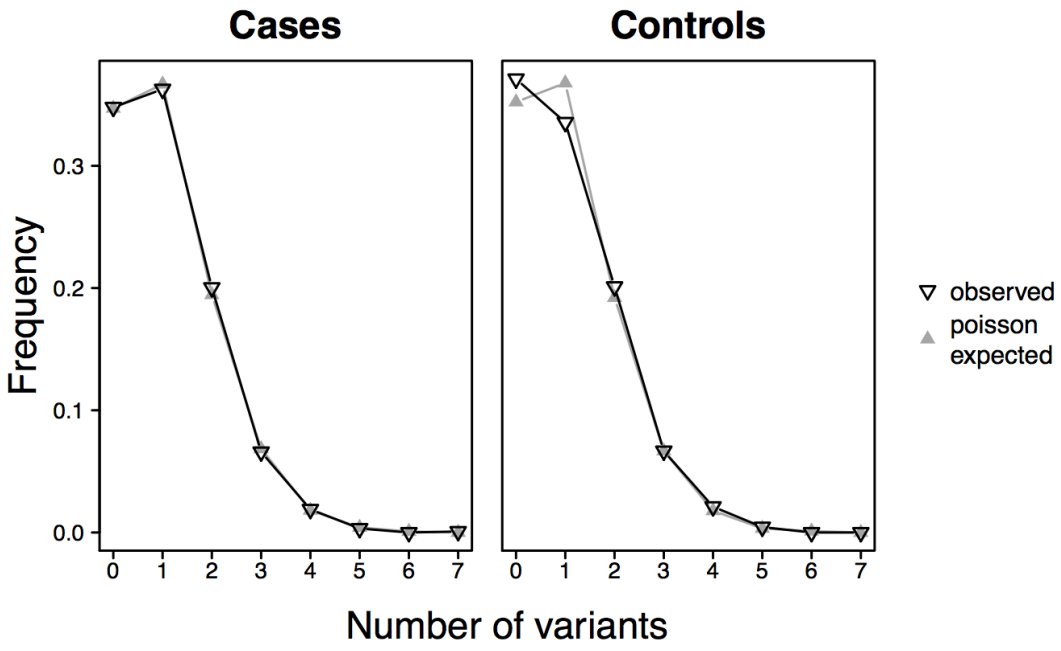


Figure 2.2 Distribution of *de novo* mutations per person follows an expected Poisson distribution from Homsy et al.

Chapter 3

De novo mutations reveal sex differences in complex and isolated congenital diaphragmatic hernia and indicate *MYRF* as a candidate gene

3.1 Introduction

Congenital diaphragmatic hernia (CDH) is an anatomical defect of the diaphragm that leads to the protrusion of abdominal viscera into the thoracic cavity, compressing the lungs in utero and resulting in lung hypoplasia. CDH affects approximately 1 in 3000 live births and is often lethal[33-35]. It can be isolated (50-60%) or associated with other birth defects and neurodevelopmental disorders[35,36]. Among these, cardiovascular malformations are the most common (~35% of CDH patients)[50]. Lung hypoplasia and the associated pulmonary hypertension are the main cause of the mortality and morbidity of CDH. Despite greatly improved survival rate with neonatal and surgical interventions, the overall mortality remains at ~30%[51-53].

The diaphragm develops between the fourth and tenth weeks of human gestation and in mice between embryonic day (E) 10.5 and E15.5[54]. Both environmental and genetic factors have been implicated. The mesenchymal-derived pleuroperitoneal folds (PPFs) play a key role in diaphragm development, and mutations in PPF-derived muscle connective tissue fibroblasts can result in CDH[55]. Most genes implicated in CDH have been identified through recurrent chromosomal anomalies and mutant mice[54,56-61]. The etiology is unclear for most CDH patients. The historical low reproductive fitness of CDH has limited the number of familial cases for genetic analysis. Others and we have reported an enrichment of *de novo* deleterious genetic events in sporadic CDH patients[7,62,63], especially LGD (likely gene disrupting) variants in complex cases.

To identify novel risk genes and compare the genetic architecture of complex and isolated cases, we performed whole exome sequencing (WES) in 79 proband-parent trios and whole genome sequencing (WGS) in 192 trios. Combined with previously published cases[7,62], we analyzed a total of 357 trios (Table 3.1), including 148 complex and 209 isolated cases.

Table 3.1 Summary of CDH WES/WGS data sets.

Batch	Phenotype	Number of trios	Total	Previously published
DHREAM_WES	Complex	39	39	Yu et al 2015
BOSTON_WES	Complex	29	74	Longoni et al 2017
	Isolated	45		
DHREAM_WES II	Complex	65	79	
	Isolated	14		
DHREAM_WGS	Complex	42 (27 with WES negative)	192	
	Isolated	150		
Total	Complex	148	357	
	Isolated	209		

We observed that there is different contribution from *de novo* variants in female and male CDH cases, and genes implicated by LGD variants in complex and isolated CDH cases have distinct expression patterns in early diaphragm development. Finally, we identified *MYRF* as a new candidate risk gene with *de novo* variants in four complex CDH patients.

3.2 Results

3.2.1 Clinical data of the cohort

Patients were recruited from the multicenter, longitudinal DHREAMS study[64] and from the Boston Children's Hospital/Massachusetts General Hospital. In the combined cohort, there were 210 (59%) male and 147 (41%) female CDH patients. The gender distribution with increase male prevalence (1.4:1) is consistent with published retrospective and prospective studies[53,65]. Among the 148 complex cases, the most frequent anomalies were congenital heart disease (41%), but neurodevelopmental delay, gastrointestinal, and other malformations were common (Table 3.2). A total of 209 (59%) patients had isolated CDH without additional anomalies at last contact[63]. In the DHREAMS cohort of 283 patients, 229 were part of the neonatal cohort (with 56% males), of which 152 had formal neurodevelopmental assessments at 2 years and/or 5 years.

Nine (5.9%) patients evaluated had neurodevelopmental delay (NDD) with scores greater than 2 standard deviations below the mean.

Table 3.2 Clinical and phenotypic summary of CDH patients (n=357)

Characteristics	**Number**	**Percentage (%)**
Male/Female	210/147	59/41
Left/Right/Other CDH location	269/56/32	75/16/9
White/Asian/Black/Other or unknown	240/13/10/94	67/4/3/26
Isolated cases	209	59
Complex cases	148	41
congenital heart disease	60	41
gastrointestinal anomaly	14	10
structural brain anomaly	15	10
other congenital malformations	67	45
neurodevelopmental delay	14	10

3.2.2 Significant enrichment of coding *de novo* variants in both complex and isolated CDH

We identified 461 protein-coding *de novo* variants (~1.29 per patient), including 190 damaging *de novo* variants in LGD and predicted deleterious missense variants ("D-mis" defined as CADD score ≥ 25). The overall *de novo* frequency in cases was 1.33 (255/192) in WGS and 1.25 (206/165) in WES. 41.2% (147/357) of probands carried at least one damaging *de novo* variant, including one *de novo* LGD in 8.4% (30/357), one *de novo* D-mis in 22.7% (81/357), and two or more damaging *de novos* in 10.1% (36/357).

We observed an overall enrichment of damaging *de novo* variants (fold enrichment (FE)=1.7, P-value=$4.2x10^{-4}$ for LGD, and FE=1.5, P-value=$3.2x10^{-6}$ for D-mis, respectively) in all CDH patients based on the expected mutation rate calibrated by the method described in Samocha et al.[40,66](Table 3.3). The positive predictive value (PPV) estimated from the enrichment rate for LGD and D-mis variants is 35%, which indicates about 67 damaging *de novo* variants contribute to CDH. The enrichment is still significant when stratifying complex and isolated CDH or by sex (Table 3.3). 22% of complex and 16% of isolated cases are explained by damaging *de novo* variants.

Table 3.3 Enrichment of *de novo* variants in CDH cases. ^LGD: likely-gene-disrupting, including frameshift, stopgain, stoploss, and splicing variants; *D-mis: missense predicted to be damaging by CADD phred score >= 25; ~Background expectation calibrated based on Samocha et al 2014 and Ware et al 2015[40,66].

Case groups	Variant type	Number of variants	Background expectation~	Fold enrichment	P-value
All (n=357)	silent	108	109	0.99	5.37E-01
	missense	290	240	**1.21**	**9.60E-04**
	D-mis*	136	90	**1.52**	**3.21E-06**
	LGD^	54	33	**1.65**	**4.24E-04**
	D-mis and LGD	190	123	**1.55**	**9.81E-09**
Complex (n=148)	D-mis*	61	37	**1.64**	**2.08E-04**
	LGD^	23	13	**1.69**	**1.23E-02**
	D-mis and LGD	84	51	**1.66**	**1.22E-05**
Isolated (n=209)	D-mis*	75	53	**1.43**	**2.02E-03**
	LGD^	31	19	**1.61**	**8.03E-03**
	D-mis and LGD	106	72	**1.48**	**9.04E-05**
Female (n=147)	D-mis*	64	37	**1.71**	**4.84E-05**
	LGD^	26	13	**1.89**	**2.02E-03**
	D-mis and LGD	90	51	**1.76**	**5.74E-07**
Male (n=210)	D-mis*	72	52	**1.38**	**5.53E-03**
	LGD^	28	19	**1.47**	**3.25E-02**
	D-mis and LGD	100	71	**1.40**	**7.78E-04**

We then tested whether the burden of damaging *de novo* variants were concentrated in constrained genes (defined as ExAC pLI≥0.5)[67] across variant types and sub-phenotypes. Overall, the burden of LGD variants was concentrated in constrained genes for both complex and

isolated cases. The burden of D-mis variants was concentrated in constrained genes for complex cases, whereas for isolated cases, the burden of D-mis variants was concentrated in other genes (pLI<0.5 or not available) (Table 3.4). This suggests that *de novo* pathogenic variants in constrained genes are more likely to cause syndromic abnormalities while such variants in other genes are more likely to cause isolated cases. Since other genes are generally not dosage sensitive, the observed burden of D-mis in these genes suggests a role of dominant negative or gain of function in isolated CDH.

Table 3.4 *De novo* enrichment based on geneset and sub-phenotype distribution. * Geneset was grouped as constrained gene (ExAC pLI>=0.5) and Other gene (ExAC pLI<0.5)

Variant type	Phenotype	Geneset*	Number of variants	Observation rate	Background expectation	Expectation rate	Fold enrichment	p-value
LGD	Complex (n=148)	Constrained	12	0.08	4.9	0.03	**2.43**	**0.005**
		Other	11	0.07	8.7	0.06	1.27	0.255
	Isolated (n=209)	Constrained	15	0.07	7	0.03	**2.15**	**0.005**
		Other	16	0.08	12.2	0.06	1.31	0.173
D-mis	Complex (n=148)	Constrained	31	0.21	15.2	0.10	**2.04**	**0.0002**
		Other	30	0.20	22	0.15	1.36	0.06
	Isolated (n=209)	Constrained	28	0.13	21.4	0.10	1.31	0.098
		Other	47	0.23	31.1	0.15	**1.51**	**0.005**

3.2.3 Different contribution of *de novo* variants to male and female CDH cases

Although CDH is more common in males, the enrichment of damaging *de novo* variants is higher in females than in males (FE=1.8 in female, FE=1.4 in male) (Table 3.3). We estimated that 27% of females could be explained by LGD or D-mis variants compared to 14% of males. In

female cases, the enrichment rate of LGD or D-mis is comparable between complex and isolated cases (Table 3.5).

Table 3.5 *De novo* enrichment based on sex and sub-phenotype distribution

Variant type	Phenotype	Gender	Number of variants	Observation rate	Background expectation	Expectation rate	Fold enrichment	p-value
LGD	Complex	Male (n=88)	14	0.16	8	0.09	1.75	0.0339
		Female (n=60)	9	0.15	5.6	0.09	1.61	0.115
	Isolated	Male (n=122)	14	0.12	11.1	0.09	1.26	0.226
		Female (n=87)	17	0.20	8.1	0.09	**2.09**	**0.00431**
D-mis	Complex	Male (n=88)	36	0.41	21.9	0.25	1.64	0.0035
		Female (n=60)	25	0.42	15.3	0.26	1.64	0.0136
	Isolated	Male (n=122)	36	0.30	30.4	0.25	1.19	0.174
		Female (n=87)	39	0.45	22.1	0.26	**1.76**	**0.00075**

In contrast, in male cases, the enrichment rate is much higher in complex cases than isolated cases. In fact, there is essentially no enrichment of LGD or D-mis variants in male isolated cases (Figure 3.1a and Table 3.5). Furthermore, in isolated female cases, LGD variants are mainly enriched in constrained genes (FE=3.3, P=0.001, Figure 3.1a), and D-mis variants were mainly in other genes (FE=2.2, P=0.0002) (Supplementary Table 3.6, Figure 3.1a). In complex CDH, the difference in enrichment rate of LGD and D-mis *de novo* variants in constrained genes between female and male cases is much smaller; and there is no significant enrichment of D-mis in other genes in either female or male cases (Table 3.6, Figure 3.1b).

Table 3.6 *De novo* enrichment based on geneset in sex and sub-phenotype distribution.
* Geneset was grouped as constrained gene (ExAC pLI>=0.5) and Other gene (ExAC pLI<0.5)

Variant type	Phenotype	Geneset*	Number of variants	Observation rate	Background expectation	Expectation rate	Fold enrichment	p-value
LGD	Complex male (n=88)	Constrained	6	0.07	2.9	0.03	2.09	0.071
		Other	8	0.09	5.1	0.06	1.56	0.147
	Complex female (n=60)	Constrained	6	0.10	2.1	0.03	**2.91**	**0.019**
		Other	3	0.05	3.5	0.06	0.85	0.686
D-mis	Complex male (n=88)	Constrained	17	0.19	8.9	0.10	**1.92**	**0.00964**
		Other	19	0.22	13	0.15	1.46	0.0715
	Complex female (n=60)	Constrained	14	0.23	6.3	0.11	**2.22**	**0.00554**
		Other	11	0.18	9	0.15	1.23	0.29
LGD	Isolated male (n=122)	Constrained	5	0.04	4	0.03	1.26	0.366
		Other	9	0.07	7.1	0.06	1.27	0.285
	Isolated female (n=87)	Constrained	10	0.12	3	0.03	**3.34**	**0.00108**
		Other	7	0.08	5.1	0.06	1.36	0.257
D-mis	Isolated male (n=122)	Constrained	17	0.14	12.3	0.10	1.38	0.117
		Other	19	0.16	18.1	0.15	1.05	0.445
	Isolated female (n=87)	Constrained	11	0.13	9.1	0.11	1.20	0.311
		Other	28	0.32	13	0.15	**2.15**	**0.000204**

3.2.4 Genes implicated by *de novo* LGD variants in complex and isolated CDH cases have distinct expression patterns in early diaphragm development

Genes associated with CDH are often expressed in pleuroperitoneal folds (PPF), an early structure critical in the developing diaphragm[68,55]. We analyzed the expression patterns of genes with LGD and D-mis variants using a mouse E11.5 PPF data set[69]. Isolated and complex cases have different patterns of LGD and missense variant burden. In complex cases, LGD *de novo* variants are dramatically enriched in genes in the top quartile of expression in developing

diaphragm (E11.5) (FE=4.7, P-value=$7x10^{-7}$) (Table 3.7, Fig. 3.2). By contrast, in isolated cases, the burden of LGD *de novo* variants is distributed across genes with a broad range of expression in PPF (Table 3.7, Fig. 3.2).

Table 3.7 The expression pattern of genes with damaging *de novo* variants in mouse embryonic day (E) 11.5 * Total of 18,000 genes were included for ranking in quartilation.

Variant type	Phenotype	Quartile *	Number of variants	Observation rate	Background expectation	Expectation rate	Fold enrichment	p-value
LGD	Complex	Q1	16	0.11	3.4	0.02	**4.67**	**0.000000698**
		Q2	3	0.02	3.5	0.02	0.87	0.67
		Q3	2	0.01	3.1	0.02	0.65	0.81
		Q4	2	0.01	2.6	0.02	0.78	0.73
	Isolated	Q1	9	0.04	4.8	0.02	1.86	0.06
		Q2	11	0.05	4.9	0.02	2.25	0.01
		Q3	8	0.04	4.3	0.02	1.84	0.07
		Q4	2	0.01	3.6	0.02	0.55	0.88
D-mis	Complex	Q1	22	0.15	9.2	0.06	**2.39**	**0.000232**
		Q2	17	0.12	10.5	0.07	1.62	0.04
		Q3	14	0.10	9.7	0.07	1.45	0.11
		Q4	7	0.05	6.3	0.04	1.12	0.43
	Isolated	Q1	20	0.10	13	0.06	1.54	0.04
		Q2	21	0.10	14.8	0.07	1.42	0.08
		Q3	16	0.08	13.6	0.07	1.17	0.30
		Q4	16	0.08	8.8	0.04	1.81	0.02

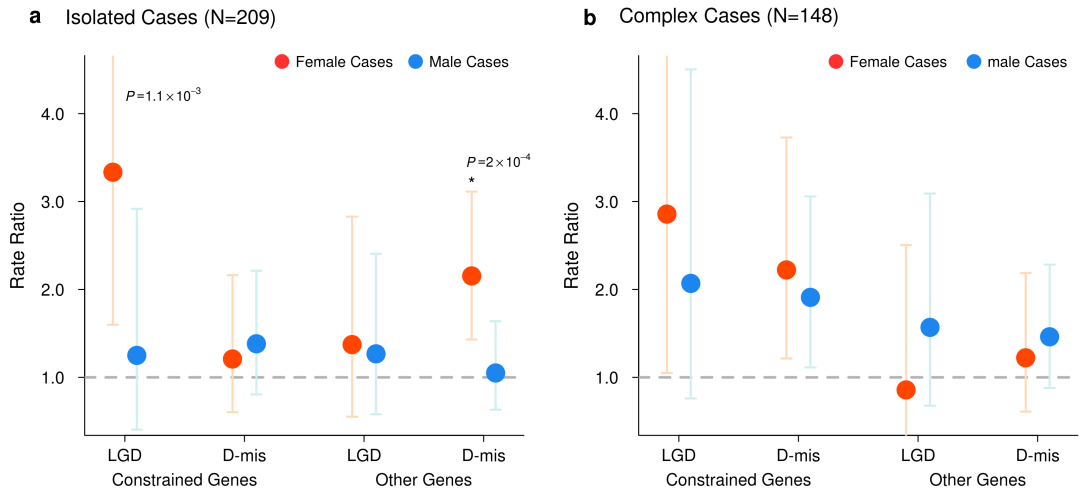


Figure 3.1. Female and male CDH cases have different enrichment rate of damaging *de novo* variants. (a) Enrichment of LGD variants and D-mis in constrained or other genes in isolated female and male cases. Constrained genes with LGD variants and other genes with D-mis variants are mainly enriched in female isolated cases. There is no enrichment of damaging *de novo* variants in isolated male cases. (b) Enrichment of LGD and D-mis variants in constrained or other genes in complex female and male cases. Both LGD and D-mis *de novo* variants were mainly enriched in constrained genes in complex cases. P-values shown are from tests of enrichment analysis. Red dots represent female cases, blue dots represent male cases. Bars represent the 95% confidence intervals (CIs) of the point estimates. Constrained genes: genes with ExAC pLI≥0.5. Other genes: genes with pLI<0.5 or no pLI estimate from ExAC; D-mis are missense variants with CADD Phred score≥25.

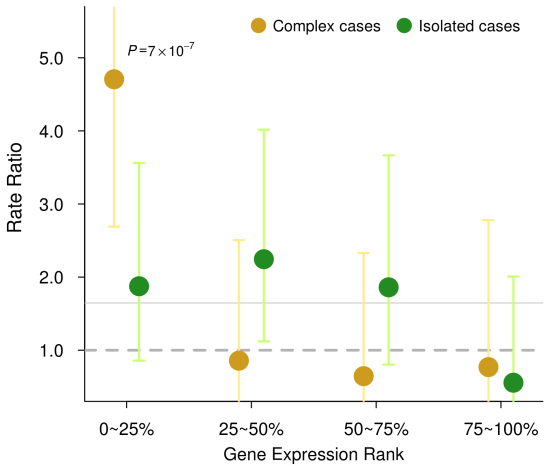


Figure 3.2. Isolated and complex cases have different enrichment patterns of LGD *de novo* variants. Enrichment rate of LGD *de novo* variants are shown in gene sets grouped by expression rank in E11.5 pleuroperitoneal folds (PPFs). In complex CDH cases, LGD *de novo* variants are dramatically enriched in the genes within the top quartile (0-25%) of expression in developing diaphragm (E11.5), and show no trend of enrichment in other quartiles. In isolated cases, LGD *de novo* variants have similar enrichment (~2x) across the 0-75% range of PPF gene expression. *P* values shown are from a test of enrichment. Bars represent the 95% CIs of the point estimates.

3.2.5 *MYRF* is a novel candidate risk gene of CDH

Two genes are observed with multiple damaging *de novo* variants. Wilms tumor 1 (*WT1)* has been previously implicated in CDH[70] and has two D-mis variants. Myelin Regulatory Factor (*MYRF)*, a transcription factor, has one *de novo* LGD and three D-mis variants (Fig. 3) in four complex CDH patients ($p=2x10^{-10}$, based on comparison to expectation from background mutations [40,66]) (Table 3.8). A recent study of congenital heart disease (CHD) [47,71] reported three additional damaging *de novo* missense variants (p.F387S, p.Q403H and p.L479V) in *MYRF* (Table 3.8, Fig 3.3). All four CDH patients had CHD (Table 3.8). The CHD patient with the *MYRF* p.Q403H variant had hemidiaphragm eventration. Genitourinary anomalies were present in six of the seven patients, a female had a blind-ending vagina with no internal sex organs and five males had ambiguous genitalia or undescended testes. *MYRF* is a constrained gene intolerant of loss of function variants in the general populations (ExAC[67] pLI=1). Although it has not previously been implicated in CDH or CHD, it is highly expressed in developing diaphragm and heart (ranked top 21% and 14% in mice E11.5 PPF[69] and E14.5 heart [5], respectively). Genital malformation may share developmental processes[72] because PPF is physically connected dorsally to urogenital ridge.

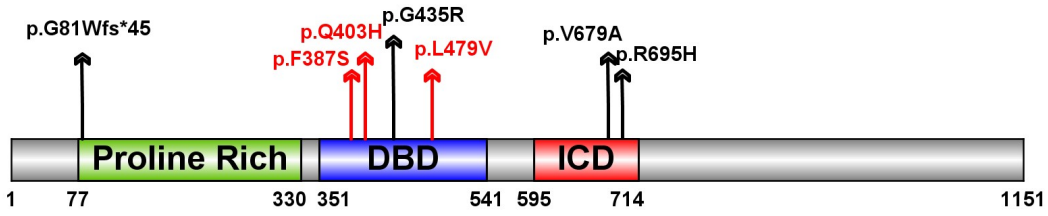
Table 3.8. *De novo* variants of *MYRF* identified in CDH and CHD patients. Abbreviation: CDH(Congenital diaphragmatic hernia) ; CHD (congenital heart disease); ASD (Atrial Septal Defect); VSD (Ventricular septal defect); TOF (Tetralogy of Fallot). The last three patients are ascertained by CHD as describe in Jin et al 2017 [71]

Sample ID	Sex	Diaphragm defect	Heart defect	Genital defect	Protein	CADD
01-1008	Male	CDH	ASD,VSD,TOF	bilateral undescended testes	p.G81Wfs *45	27.3
01-0429	Female	CDH	VSD	no internal genital organs, external blind-ending vagina	p.G435R	32
04-0042	Male	CDH	ASD,VSD	NA	p.V679A	25.9
05-0050	Male	CDH	hypoplastic left heart syndrome	ambiguous genitalia	p.R695H	34
1-02264	Male	NA	abnormal aorta	ambiguous genitalia, hypospadias, undescended testis	p.F387S	27.9

1-03160	Male	right hemi-diaphragm eventration	abnormal atrial septum, pulmonary vein and arteries, systemic vein, aorta, aortic valve, mitral valve, ventricular septum	undescended testis	p.Q403H	27.6
1-07403	Female	NA	abnormal aorta and aortic valve	NA	p.L479V	23.9

The three variants identified in CHD patients and p.G435R are located in the conserved DNA binding domain (DBD) of *MYRF* (Fig. 3.3), and could alter DNA binding[73]. The other two D-mis variants (p.V679R and p.R695H) are located in the intramolecular chaperone auto-processing domain (ICD) in a leucine zipper[41]. Mutations in the leucine zipper of the ICD domain may inhibit the trimerization of *MYRF*, resulting in the failure of formation of the N-terminal trimer[41] which is important for the transcription factor function[74]. *MYRF* is thought to be an essential transcription factor for oligodendrocyte differentiation and myelination[75]. Conditional deletion of *Myrf* impaired motor learning[76,77] and the individual with the p.V679A variant we assessed at two years old had intellectual disability.

Figure 3.3. *De novo* variants identified in *MYRF*. Schematic of the MYRF protein with predicted sequence features, including N-terminal Proline Rich region, DNA-binding domain (DBD) and intramolecular chaperone domain (ICD). Variants identified in CDH indicated as black arrow, variants identified in congenital heart disease cases indicated with red arrows.



3.3 Discussions

CDH is slightly more common in males. For the first time, our study suggests that male and female CDH may have a different genetic architecture, especially among isolated CDH cases. Damaging *de novo* variants with large effect have a substantial contribution to isolated female cases but little contribution to isolated male cases. Given the higher frequency of males among isolated cases, a plausible explanation is that polygenic risk from inherited variants alone can cause isolated CDH in males, but due to a female protective effect[78], additional highly penetrant *de novo* variants are often required to cause CDH in females to pass the threshold of liability. This is similar to what has been observed in autism which is also more common among males [39]. Since there is a similar male/female ratio in overall cohort and our neonatal cohort (1.4:1), this difference is unlikely due to ascertainment or survival bias. The parental ages for male and female probands were similar and cannot account for the differences we observed in *de novo* variants.

Additionally, we found genes implicated in isolated and complex cases have distinct expression patterns in early development. In complex CDH, the burden of LGD and D-mis variants are concentrated in genes highly expressed in the FFP, an early embryonic diaphragm precursor, consistent with the pleiotropic effects of these genes on diaphragm and other organogenesis. By contrast, the burden of LGD variants in isolated cases is distributed across genes with a broader range of expression in PPF. Since the bulk expression data from PPFs is the sum of different cell types[68], the lack of correlation of LGD enrichment and expression level in PPF suggests the possibility that a substantial portion of the implicated genes in isolated cases could be expressed only in sub populations of cells in the PPF that are not relevant to organogenesis in other parts of the body. Single-cell mRNA-sequencing will be necessary to

analyze gene expression pattern in specific cell types and further assess the cell type(s) responsible for isolated CDH.

Finally, *MYRF* is a novel candidate risk gene of CDH. The four CDH patients carrying damaging *de novo* variants in *MYRF* all have congenital heart defects, genitourinary anomalies including ambiguous genital, and this likely represents a novel syndrome. As we identify larger numbers of patients with mutations in genes associated with CDH, we will be able to better describe the spectrum of disease associated with the gene as well as the clinical outcomes including risk of pulmonary hypertension and respiratory complications which are life threatening concerns for CDH patients. Identification of additional high risk CDH genes should elucidate the developmental biology and provide targets for treatment and prevention.

3.4 Material and methods

3.4.1 Patient cohorts

A total of 357 CDH patients and their unaffected parents were recruited for analysis in this study, including 74 trios from Boston Children's Hospital (BCH) and Massachusetts General Hospital (MGH)[62] (Boston Cohort) and 39 trios from a previous study [7] (Table 4.1). Two hundred and eighty-three trios were recruited as part of the DHREAMS (Diaphragmatic Hernia Research & Exploration; Advancing Molecular Science) study (http://www.cdhgenetics.com/)[63]. Neonates, children and fetal cases with a diagnosis of diaphragm defects were eligible for DHREAMS. Clinical data were abstracted from the medical chart by study personnel at each of 16 clinical sites. Data on prenatal history, neonatal outcome, and longitudinal follow-up data including Bayley III and Vineland II developmental assessments and a parent interview about the patient's health since discharge at 2 years of age and/or 5 years of age were gathered in our birth

cohort. A complete family history of diaphragm defects and major malformations was collected on all patients by a single genetic counsellor, and no patients had a family history of CDH.

Patients without additional birth defects or neurodevelopmental disorder (NDD) at last contact were classified as isolated, and patients with the additional birth defects or NDD were classified as non-isolated (Details previously published[7,63]). The diaphragm lesion was classified as left, right, bilateral or central. Pulmonary hypoplasia, cardiac displacement and intestinal herniation were considered to be part of the diaphragm defect sequence and were not considered to be an additional malformation. Subjects from BCH and MGH were described previously[62]. A blood, saliva, and/or skin/diaphragm tissue sample was collected from the affected patient and both parents. All participants provided informed consent/assent for participation in this study, which was approved by the institutional review boards of each participate study site.

3.4.2 Whole exome and whole genome sequencing of case trios

We included previously two sets of WES data for analysis[7,62]. We performed whole exome sequencing (WES) at the University of Washington in 79 additional trios using genomic DNA largely from whole blood (73 trios, 93.4%), with a minority from saliva or tissues. DNA was processed with the Nimblegen SeqCap EZ Exome V2 exome capture reagent (Roche) and TruSeq DNA Sample Prep Kits (Illumina). Samples were multiplexed and sequenced with paired-end 75bp reads on Illumina HiSeq 2500 platform according to the manufacturer's instructions (Illumina, Inc, San Diego, California, USA).

We sequenced another 192 trios at Baylor College of Medicine using whole genome sequencing (WGS) as part of NIH Gabriella Miller Kids First Pediatric Research Program.

Among these, 27 trios that had no damaging *de novo* variants in previously published WES data were selected as "WES-negative" cases for WGS[7]. Genomic libraries were prepared by the Illumina TruSeq DNA PCR-Free Library Prep Kit. DNA was sheared into 350-bp average length using sonication on a Covaris LE220 instrument. The fragmented DNA was end-repaired, A-tailed and indexed using TruSeq Illumina adapters with overhang-T added to the DNA. The libraries were validated on a Bioanalyzer DNA High Sensitivity chip by size and quality, then pooled in equal quantities and sequenced as paired-end reads of 150-bp lengths on an Illumina HiSeq X platform.

3.4.3 Alignment and quality controls

Mapping, alignment, and variant calling were done according to the Broad Institute's best practices using Burrows-Wheeler Aligner (bwa-mem, version 0.7.10)[79] and Genome Analysis Toolkit (GATK; version 3.3) (https://software.broadinstitute.org/gatk/best-practices/). Briefly, we mapped WES or reads to the reference genome (build GRCh37) using BWA-mem [79], mark PCR duplicates using Picard (v1.67), performed local realignment and quality recalibration using GATK[42].We jointly called variants in all WES samples using the GATK HaplotypeCaller. The output file was generated in the universal variant call format (VCF). We used the same procedure to analyze WGS samples.

Among new samples sequenced by WES, the mean depth of coverage is 59± 21 with 93±2.5% bases read with at least 15x (DP15) in target regions. Among new samples sequenced

by WGS, the mean depth of coverage is 39±2, with 99±0.25% bases read at least 15x (Fig. 3.4).

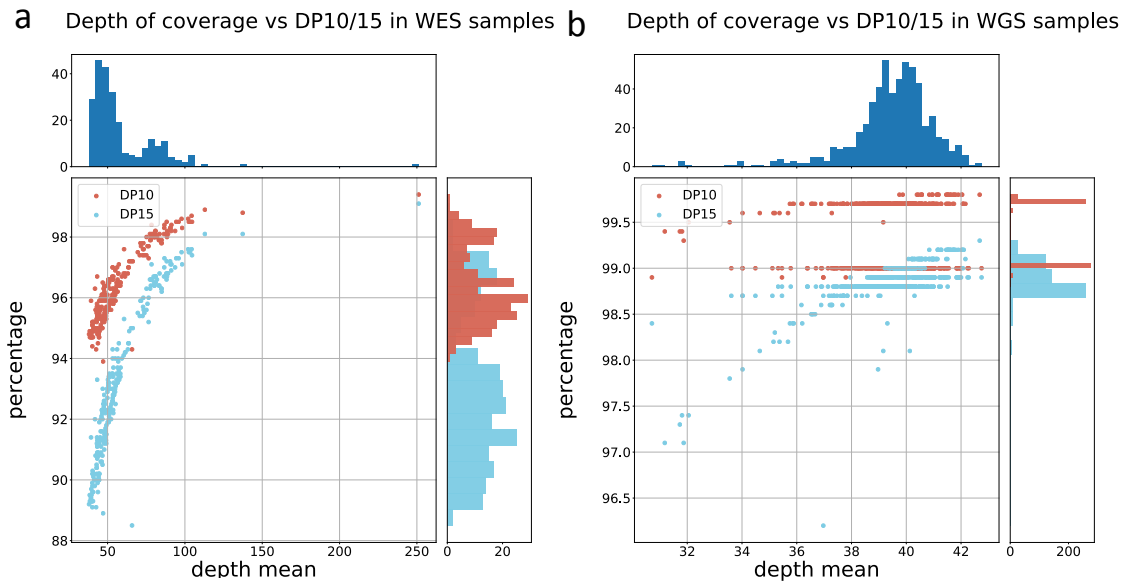


Figure 3.4 Depth of coverage quality for case cohorts. (a) Mean depth and DP10/DP15 percentage in WES case samples. (b) Mean depth and DP10/DP15 percentage in WGS case samples. DP10 is percentage of targeted bases that are covered by at least 10 reads. DP15 is percentage of targeted bases that are covered by at least 15 reads. Top panel is histogram of mean depth coverage. Right panel is DP10/DP15 histogram.

We performed principal component analysis of common variants (allele frequency >5%) using Eigenstrat[80] to determine the population structure and ancestry of both cases and controls, with HapMap 3 sample collection data[81] as a reference.

3.4.4 Detection of *de novo* snps and indels

We used Plink[82] (http://pngu.mgh.harvard.edu/purcell/plink/) to estimate Identity by Descent (IBD)[83] to confirm the relatedness among familial trios. All trios were matched to parents-offspring with relatedness.

A variant that presents as a heterozygous genotype in the offspring and homozygous reference genotypes in both parents was considered to be a potential *de novo* variant. We used an established stringent filtering method to identify *de novo* variants as described previously [7,40,47].

Briefly, we required the candidate variants have depth (minimum 5 alternate allele reads), alternate allele fraction (minimum 20%), Fisher Strand (FS) (maximum 25), Quality by depth (QD) (minimum 2), Phread-scaled genotype likelihood (PL) (minimum 60), population allele frequency(maximum 0.1% in ExAC), and parental read characteristics (minimum depth of 10 reference reads; alternate allele fraction less than 5%, minimum GQ of 30) . Additionally, variants located in segmental duplication regions (maximum score 0.98) were excluded. All candidate *de novo* variants were manually inspected in the Integrated Genomics Viewer (IGV, http://software.broadinstitute.org/software/igv/). In addition, we validated all the *de novo* likely gene disrupting (LGD) (including frameshift, nonsense and splicing site) variants by dideoxynucleotide sequencing. Of 40 case variants that were submitted for validation by Sanger sequencing, all 40 were confirmed (precision =100%).

Among the 27 "WES-negative" cases, there were 12 *de novo* variants identified by WGS that were not detected by WES[7].

3.4.5 Annotation of variants

We used ANNOVAR[43] to annotate variants and aggregate allele frequency and *in silico* functional predictions, then used average allele frequency in Exome Aggregation Consortium (ExAC) data to define rare variants (frequency < 1e-4). Rare *de novo* variants were classified as silent, missense, and likely-gene-disrupting ("LGD", which includes stopgain, stoploss, canonical splicing site, or frameshift variants). In-frame insertions or deletions were not considered in the genetic analysis. We defined deleterious missense variants ("D-mis") by CADD[28] phred-scale score ≥25.

3.4.6 Global or gene set burden between case and mutation background rate

Baseline rate for different classes of *de novo* variants in each GENCODE coding gene were calculated for the longest transcript using a previously described mutation model. The expected number of variants in different gene sets were based on the 3-nucleotide context-specific mutation rate estimated by Samocha et al.[40,66] and calculated by summing up the class-specific variant rate in each gene in the gene set multiplied by twice the number of patients. The observed number of variants in each gene set and case group was then compared with the baseline expectation using Poisson test.

We used Poisson test to assess the significance of excess of observed *de novo* variants over expectation which was defined as enrichment rate (r). The positive predictive value (PPV) for *de novo* variants in each class was calculated as $(r-1)/r$. The Estimated number of true risk variants in each class is the number of observed variants (m) in cases multiplied by PPV: m * (r-1)/ r. The most severe predicted functional effect variants (LGD and D-mis) were used in further burden analyses based on the different phenotype, gender, gene set, and expression data.

3.4.7 Percent of CDH attributable to *de novo* variants

We calculated the percent of CDH patients with pathogenic variants in isolated and complex CDH groups, in male and female case groups, respectively. The fraction of individuals carrying at least one damaging *de novo* variant was determined, by subtracting the expected rate of damaging *de novo* variants per individual.

The formula is as follows:

$$\frac{(n1 - r * s1)}{s1} * 100\%$$

where n1 is the total number of sub-group CDH patients with at least one *de novo* deleterious variant, r is the expected rate per healthy individual with at least one *de novo* deleterious variant, where the rate was estimated by 10,000 simulations of Poisson distribution of variants per person, and s1 is the total number of sub-group CDH patients.

3.4.8 Expression profile during diaphragm development

Mouse developing diaphragm (MDD) gene expression datasets from the pleuroperitoneal folds (PPFs)[69] at embryonic day 11.5 (E11.5) were used in this study. High diaphragm expression is defined as the top quartile of probe sets based on RMA (Robust Multi-Array Average)-normalized expression levels of microarray data[7].

3.4.9 Single genes with multiple *de novo* mutations

For *MYRF*, the number of observed deleterious *de novo* mutations was compared to the expected deleterious mutation background using a Poisson test. The p-value passed Bonferroni correction with all protein-coding genes annotated in CCDS[84].

Chapter 4

Genetic connections between developmental disorders and cancer

4.1 Introduction

De novo or rare functional variants with large effect sizes have large contributions[49] to developmental disorders (DDs), such as developmental delay, autism, intellectual disability, and epilepsy[6,24,39,85,86]. However, most of DDs risk genes are still unknown: there are about 100 to 200 known candidate risk genes[39,86] yet the estimated number of risk genes that contribute to DDs is about 1,000[39]. Additionally, it is challenging to clinically interpret *de novo* or rare variants, especially for missense variants, even in known risk genes.

Cancer and DDs have common dysregulated cellular processes, such as proliferation, growth, and differentiation[87-89]. There are well-known genes and pathways implicated in both, with recurrent somatic mutations in cancer and highly penetrant germline *de novo* variants in DDs. Classic examples include *PTEN* [90 601728], a negative regulator in ALK pathway implicated in autism[91] and many types of cancers[92], and *PTPN11* [90 176876], a phosphatase in RAS/MAPK signal- ing pathway implicated in both Noonan syndrome and leukemia[93]. Recent large-scale genomic studies of cancer[94,95] and DDs[6,24,39,85,86,96-98] revealed a substantial number of genes implicated in both classes of diseases. There was reported increased burden of rare nonsynonymous variants in proto-oncogenes in autism patients[99]. Some of these genes share similar modes of action through cancer somatic mutations and DDs germline *de novo* variants[100]. For example, *PTPN11* is known to harbor gain of function mutations that make it constitutively active in both cancer and Noonan syndrome patients; *EP300* [90 602700], a tumor suppressor, has a large fraction of likely-gene-disrupting (LGD) mutations that mostly likely lead to loss-of-function in both diseases[101].

In this study, we aim to quantify the genetic connection between cancer and DDs, and investigate the feasibility of utilizing cancer genomics data to help improve risk gene and variant

discovery in genetic studies of DDs. Driver genes are much more frequently mutated in cancer, and with ongoing international efforts in cancer precision medicine, there is an accelerated accumulation of cancer somatic mutation data. Such data will provide an unprecedented opportunity to study empirical functional consequences of mutations at virtually every base in cancer driver genes. Elucidating such a connection could lead to a better understanding of molecular mechanisms of both cancer and DDs.

We compiled data sets of *de novo* variants from recently published studies on DDs[6,24,39,85,86,94,102], including autism, intellectual disabilities, epilepsy, and developmental delays. We also assembled a large number of candidate cancer driver genes from various sources, including Cancer Census, The Cancer Genome Atlas (TCGA), and The Candidate Cancer Gene Database (CCGD)[94,103-105]. We compared the burden of *de novo* variants in candidate cancer driver genes and non-driver genes among DD cases. We then estimated the fraction of DD risk genes that are also cancer drivers. Finally, we investigated whether germline *de novo* variants and cancer somatic mutations in this set of overlapping genes have similar modes of action.

4.2 Results

4.2.1 Burden of germline *de novo* variants in DD patients among candidate cancer driver genes

To investigate the contribution of cancer driver genes to DDs, we compiled a large dataset of 6,294 germline *de novo* coding variants from 5,542 DD cases drawn from recent published studies, including 3,953 cases with autism spectrum disorder[6,39], 1,133 cases with various DDs from Deciphering Developmental Disorders study[86], 192 cases with epileptic

encephalopathies[24], and 264 cases with intellectual disability[85,96-98] (Table 4.1). We used 1,911 parents-unaffected sibling trios from the Simons Simplex Collection (SSC) as controls[39].

Table 4.1 Dataset of developmental disorders cases and parents-unaffected sibling trios from the Simons Simplex Collection.

Disease		Samples	Reference
Developmental disorders (DD) cases Total: 5542	Autism spectrum disorder (ASD)	3953	De Rubeis et al 2014 Iossifov et al 2014
	Deciphering Developmental Disorders (DDD)	1133	DDD 2014
	Epileptic encephalopathies (EE)	264	Epi4K Consortium 2013
	Intellectual disability (ID)	192	Hamdan et al 2014 Rauch et al 2012 de Ligt et al 2012 Gilissen et al 2014
Control	Simons Simplex Collection	1911	Iossifov et al 2014

Aggregating various DDs with shared but distinct genetic risk architectures can yield additional findings in risk genes[47,86,106]. We reannotated these variants using ANNOVAR software[43], and predicted the functional consequences of missense variants in silico using meta-SVM[30]. The following analyses are focused on LGD (which includes stopgain, stoploss, frameshifting, and splicing variants) or predicted-damaging missense (D-mis, predicted by meta-SVM) variants. The overall rate of silent *de novo* variants is similar between cases and controls (0.25 per subject).

To include a broad set of cancer driver genes[107], we obtained 568 cancer census genes from COSMIC database[105], 773 genes with MutSigCV[108] q-value less than 0.1 from individual The Cancer Genome Atlas (TCGA) studies curated by cBioPortal[94], and 325 candidate driver genes from forward genetic screens in mice by The Candidate Cancer Gene Database (CCGD)[104]

(see Materials and Methods section). In total, we compiled a list of 1,481 candidate cancer driver genes (Fig. 4.1), all other genes not classified as cancer driver genes were considered as non-cancer driver genes.

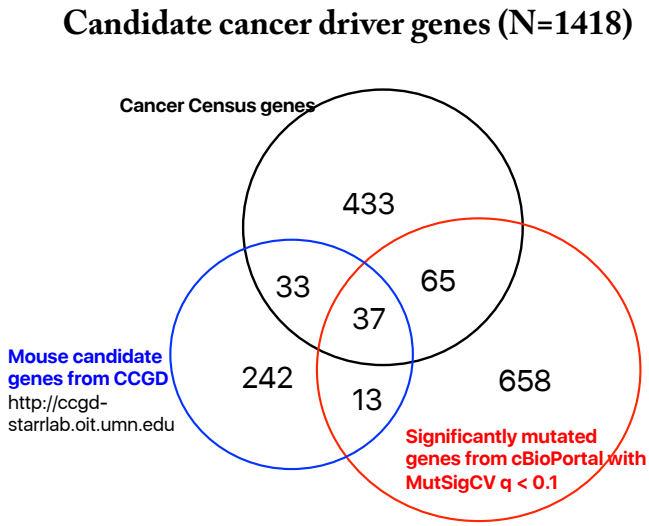


Figure 4.1 Venn diagram of cancer driver genes from three sources: cancer census genes from COSMIC, significantly mutated genes from individual TCGA studies curated by cBioPortal and candidate driver genes from forward genetic screens in mice by The Candidate Cancer Gene Database (CCGD)

Among all candidate cancer driver genes, there is significant enrichment of LGD or D-mis germline *de novo* variants in DD cases compared with controls (Table 4.2 (b)), and such enrichment is significant in both autism and other types of DD (Table 4.2 (c)). Moreover, among DD cases, candidate cancer driver genes show significantly higher enrichment of germline *de novo* variants than non-drivers (Table 4.2 (a); odds ratio = 2.0, P value = 4.5e–6). Such enrichment cannot simply be explained by known constrained genes in cancer driver genes, as we observed that among constrained genes[67], there is still a significantly greater burden in drivers than in non-drivers (odds ratio = 2.1; Table 4.2 (d)). Based on fold enrichment of *de novo*

variants in cases compared to controls, we estimate that there are about 391 causative LGD variants in total, of which 163 are in candidate cancer driver genes (42%), and that there are about 327 causative D-mis variants in total, of which 114 are located in cancer driver genes (35%). Therefore, about 38% (95% confidence interval (CI): [29%, 51%]) of all potentially causative damaging (LGD or D-mis) *de novo* variants observed in these DD cases are located in candidate cancer drivers.

Table 4.2 Burden of *de novo* germline mutations in candidate cancer driver genes, non-cancer driver gene and all genes. Damage missense mutations are predicted by meta-SVM. Burden tests between case and control were performed using Binomial tests; Burden comparisons between cancer and non-cancer drivers were performed using Fisher's exact test.

(a) Burden of germline *de novo* variants in candidate cancer driver genes (N=1481) comparing to non-driver genes (N=17396). The null hypothesis in Fisher's exact test is that the fraction of germline *de novo* variants of each type located in cancer driver genes is the same between DD cases and controls.

Type of *de novo* variants	Number of *de novo* variants in 5542 DD cases		Number of *de novo* variants in 1911 Controls		Fisher's exact	
	Cancer driver genes	Non-cancer driver genes	Cancer driver genes	Non-cancer driver genes	Odds ratio	P-value
LGD	233	671	24	153	2.2	3.3E-04
missense	518	3412	132	1002	1.2	0.19
D-mis	190	766	26	191	1.8	0.006
LGD/D-mis	423	1437	50	344	2	4.5E-06
silent	146	1238	52	427	1	0.86

(b) Comparison between cases and controls among cancer driver genes, non-cancer driver genes, and all genes.

Gene group	Type of *de novo* variant	Cases (N=5542)	Controls (N=1911)	Fold enrich-ment	P-value
Cancer driver genes	LGD	233	24	3.3	6.30E-11
	Missense	518	132	1.4	0.0016

(N=1481)	Dmis	190	26	2.5	1.10E-06
	LGD/Dmis	423	50	2.9	3.20E-16
	silent	146	52	1	0.87
Non-cancer driver genes (N=17370)	LGD	671	153	1.5	1.90E-06
	Missense	3412	1002	1.2	6.30E-06
	Dmis	766	191	1.4	3.90E-05
	LGD/Dmis	1437	344	1.4	3.40E-10
	silent	1238	427	1	1
All genes (N=18851)	LGD	904	177	1.8	3.40E-13
	Missense	3930	1134	1.2	8.30E-08
	Dmis	956	217	1.5	8.10E-09
	LGD/Dmis	1860	394	1.6	2.90E-20
	silent	1384	479	1	0.93

(c) Enrichment of LGD or D-mis *de novo* variants in various types of DD among candidate cancer driver genes.

Type of NDD	Type of variant	Number of variants in cases	Number of variants in controls (N=1911)	Fold enrichment	p-value
Autism (N=3953)	LGD	121	24	2.437	1.75E-05
	Missense	318	132	1.165	0.145
	Dmis	97	26	1.804	0.00681
	LGD/Dmis	218	50	2.108	4.27E-07
	silent	108	52	1.004	1
Epileptic encephalopathies (N=264)	LGD	1	24	0.302	0.355
	Missense	28	132	1.535	0.0513
	Dmis	13	26	3.619	0.000464
	LGD/Dmis	14	50	2.027	0.032
	silent	8	52	1.114	0.695
Developmental delay or intellectual disability (N=1325)	LGD	111	24	6.67	9.19E-23
	Missense	172	132	1.879	4.62E-08
	Dmis	80	26	4.438	6.23E-13
	LGD/Dmis	191	50	5.509	7.52E-34
	silent	30	52	0.832	0.435

(d) Among constrained genes, cancer driver genes harbor greater burden of *de novo* variants than non-cancer driver genes in case-control comparison. Constrained genes are defined as genes with pLI score >= 0.9 from ExAC database. The null hypothesis of the Fisher's exact test is that among constrained genes, the burden of *de novo* variants in cancer drivers comparing to non-drivers is independent of case/control status.

	Cases (N=5542)		Controls (N=1911)		Fisher's exact test	
	Constrained gene in driver N=507	Constrained gene in non-driver N=2619	Constrained gene in driver N=507	Constrained gene in non-driver N=2619	Odds ratio	P-value
LGD	188	297	9	32	2.25	0.043

Missense	302	1004	59	255	1.3	0.11
D-mis	133	300	13	53	1.8	0.08
LGD/D-mis	321	597	22	85	2.1	0.002
silent	82	314	21	98	1.2	0.515

Among the candidate cancer drivers that harbor damaging *de novo* variants in DD cases, several pathways are enriched (Table 4.3), including transcriptional regulation (e.g., lysine degradation), core developmental pathways (e.g., Wnt and Hippo signaling), pathways related to cell junctions and adhesion, and ubiquitin mediated proteolysis.

Table 4.3 Functional term enrichment analysis of all cancer driver genes with damaging (LGD or Dmis) *de novo* mutations in all DD cases. Adjusted p-values were calculated by Enrichr.

Groups	KEGG term	Enrich-ment rate	Adjusted P-value	Genes
Tran-scription regulation	Lysine degradation (hsa00310)	16.06	5.44E-05	*KMT2E, KMT2D, SETD2, KMT2A, NSD1, KMT2C, ASH1L, WHSC1, WHSC1L1*
	Transcriptional misregulation in cancer (hsa05202)	6.70	8.40E-04	*DDX5, KMT2A, PAX3, PAX5, PBX1, MYCN, SIN3A, RARA, TCF3, WHSC1, MET, BIRC3, KDM6A*
Core develop-mental pathways	Wnt signaling pathway (hsa04310)	5.23	4.27E-02	*TCF7L2, CREBBP, SMAD4, TBL1XR1, CUL1, CTNNB1, EP300, RAC1*
	Hippo signaling pathway (hsa04390)	5.46	2.75E-02	*SMAD2, SOX2, TCF7L2, CRB1, SMAD4, WWC1, CTNNB1, ACTB, LLGL2*
	Signaling pathways regulating pluripotency of stem cells (hsa04550)	5.23	4.27E-02	*SMAD2, SOX2, SMAD4, MAP2K1, KAT6A, CTNNB1, TCF3, FGFR2*
Cell adhesion and junctions	Adherens junction (hsa04520)	12.54	6.44E-05	*SMAD2, TCF7L2, PTPRB, CREBBP, SMAD4, CTNNB1, EP300, RAC1, MET, ACTB*
	Focal adhesion (hsa04510)	5.51	6.96E-03	*MAP2K1, HGF, KDR, PTEN, CTNNB1, BRAF, COL6A6, RAC1, ARHGAP5, MET, ACTB, BIRC3*
	Rap1 signaling pathway (hsa04015)	4.84	2.45E-02	*GRIN2A, MAP2K1, HGF, KDR, GNAS, CTNNB1, BRAF, RAC1, MET, ACTB, FGFR2*
	Tight junction (hsa04530)	6.01	1.68E-02	*MYH2, PTEN, MYH9, CTNNB1, MYH11, ASH1L, SPTAN1, ACTB, LLGL2*
	Bacterial invasion of	7.14	3.53E-02	*CTNNB1, CBLB, RAC1, CBL, MET,*

	epithelial cells (hsa05100)			*ACTB*
Signaling	cAMP signaling pathway (hsa04024)	5.60	6.73E-03	*CREBBP, GRIN2A, MAP2K1, CREB1, PTCH1, FSHR, GNAS, EP300, BRAF, CACNA1D, RAC1, ATP1A1*
	Thyroid hormone signaling pathway (hsa04919)	9.44	7.94E-05	*MED12, CREBBP, MAP2K1, NOTCH1, MED13, SIN3A, TSC2, CTNNB1, EP300, ATP1A1, ACTB, MTOR*
Ubiquitin	Ubiquitin mediated proteolysis (hsa04120)	6.10	1.64E-02	*MAP3K1, CUL3, UBR5, CUL1, CBLB, BIRC6, BRCA1, CBL, BIRC3*
Cancer	Pathways in cancer (hsa05200)	5.84	1.37E-05	*PTEN, CBLB, CBL, SUFU, EP300, RAC1, SMAD2, TCF7L2, CREBBP, MAP2K1, HSP90AA1, SMAD4, PTCH1, HGF, BRAF, MLH1, MTOR, MSH6, MSH2, RARA, GNAS, CTNNB1, MET, FGFR2, BIRC3*
	Colorectal cancer (hsa05210)	14.97	4.49E-05	*MSH6, SMAD2, TCF7L2, SMAD4, MAP2K1, MSH2, CTNNB1, BRAF, RAC1, MLH1*
	Proteoglycans in cancer (hsa05205)	7.32	5.44E-05	*SMAD2, MAP2K1, DDX5, PTCH1, HGF, CBLB, BRAF, PTPN11, ANK3, CBL, ACTB, MTOR, KDR, CTNNB1, RAC1, MET*
	Prostate cancer (hsa05215)	11.47	5.44E-05	*TCF7L2, HSP90AA1, CREBBP, MAP2K1, CREB1, PTEN, CTNNB1, EP300, BRAF, MTOR, FGFR2*
	Renal cell carcinoma (hsa05211)	11.25	9.63E-04	*CREBBP, MAP2K1, HGF, EP300, PTPN11, BRAF, RAC1, MET*
	Endometrial cancer (hsa05213)	10.71	8.52E-03	*TCF7L2, MAP2K1, PTEN, CTNNB1, BRAF, MLH1*
	Melanogenesis (hsa04916)	6.50	2.88E-02	*TCF7L2, CREBBP, MAP2K1, CREB1, GNAS, CTNNB1, EP300*
	Chronic myeloid leukemia (hsa05220)	7.63	2.88E-02	*SMAD4, MAP2K1, PTPN11, BRAF, CBLB, CBL*
	Acute myeloid leukemia (hsa05221)	8.14	4.27E-02	*TCF7L2, MAP2K1, RARA, BRAF, MTOR*
	Thyroid cancer (hsa05216)	12.80	2.88E-02	*TCF7L2, MAP2K1, CTNNB1, BRAF*

4.2.2 Cancer driver genes comprise about a third of DD risk genes

To identify a broad set of candidate risk genes of development disorders, we applied TADA[6,109], a probabilistic method for identifying risk genes based on *de novo* or rare inherited variants. We used only *de novo* LGD and D-mis mutations in this analysis and gene-specific background mutation rate[40,66]. We ranked all genes by FDR, defined DD candidate risk genes using increasing thresholds of 10% bins, and then calculated the percentage of true DD risk genes that overlap with cancer driver genes (Table 4.4). The estimated overlap percentage is 45% at FDR 0.1; the ratio decreases at larger FDR values but is still significantly greater than what is expected by chance at FDR of 0.5. Overall, we estimate that cancer drivers comprise more than a third of risk genes contributing to developmental diseases.

Table 4.4. **Number of developmental disorder (DD) candidate risk genes at different FDR values estimated by TADA, and corresponding overlapping cancer driver genes**. For each FDR threshold, number of true risk genes is estimated by FDR definition.

FDR by TADA	Number of candidate risk genes	Estimated number of true risk genes	Number of candidate risk genes that are cancer drivers	Estimated number of true risk genes that are cancer drivers	Estimated percentage of cancer drivers among true risk genes
≤ 0.1	134	120.6	56	54.6	45%
≤ 0.2	186	148.8	66	62.2	41%
≤ 0.3	269	188.3	74	65.7	34%
≤ 0.4	421	252.6	92	74.6	29%
≤ 0.5	649	324.4	124	90.5	27%

4.2.3 Germline *de novo* variants disrupt DD risk genes through similar modes of action as somatic mutations in cancer drivers

Cancer driver genes are generally categorized as tumor suppressors or oncogenes, with the exception of genes that play either role in different cancer types[110]. The molecular consequence of a driver somatic mutation is usually loss-of-function in a tumor suppressor gene and gain-of-function in an oncogene. There are a number of known DD risk genes disrupted by germline variants via similar modes of action as cancer driver genes disrupted by somatic mutations. For example, gain-of-function germline variants in *SOS1* [90 182530] and *PTPN11* genes are implicated in Noonan syndrome[111]. Both genes are also oncogenes with gain-of-function somatic mutations in leukemia[93]. To quantify the similarity of modes of action between cancer and DDs in individual genes and pathways, we investigated the patterns of cancer somatic mutations and DD germline *de novo* variants. We made two assumptions: (a) loss-of-function mutations include both truncating mutations (LGD, including stop- gain, stoploss, splicing, and frameshifting), and a subset of missense mutations. Tumor suppressors tend to harbor both types of mutations, generally with a large fraction of LGD mutations[112]; (b) gain-of-function mutations are mostly composed of missense mutations. We note that genes with dominant negative mutations are often exceptions.

We reasoned that tumor suppressors are likely haploinsufficient[112] as DD risk genes. To test that, we identified likely tumor suppressor genes and likely non-suppressor genes based on the fraction of LGD mutations among all somatic SNVs and indels in a given gene, across all cancers. Specifically, we grouped the candidate cancer driver genes into four bins using data from COSMIC, with fractions of LGD mutations at 0%–5%, 5%–10%, 10%–15%, and ≥ 15%, respectively. Among these bins, the genes in the 0%–5% bin are likely non-suppressors, and the

ones in the 15% or larger bin are likely tumor suppressor genes. This tier classification is consistent with an independent study of predicted tumor suppressor genes[112] using TCGA data. More than half of the likely tumor suppressor genes overlap with the predicted confident tumor suppressor genes; there is almost no overlap of likely non-suppressor with predicted confident tumor suppressor genes[112] (Fig. 4.2). Compared with the number of germline LGD *de novo* variants expected from background mutation rate[40,66] in DD cases, we observed a 2.4 enrichment in likely non-suppressor genes (Table 4.5), which, as expected, is below the overall enrichment in cancer driver genes (3.3×; Table 4.2). On the contrary, we observed more than 10-fold enrichment of LGD variants in likely-tumor suppressor genes than expected (Fig. 4.3; Table 4.5), representing a 4.2× greater enrichment than in likely non-suppressors. This indicates that tumor suppressor genes implicated in DD patients through germline *de novo* variants often confer disease risk via loss of function.

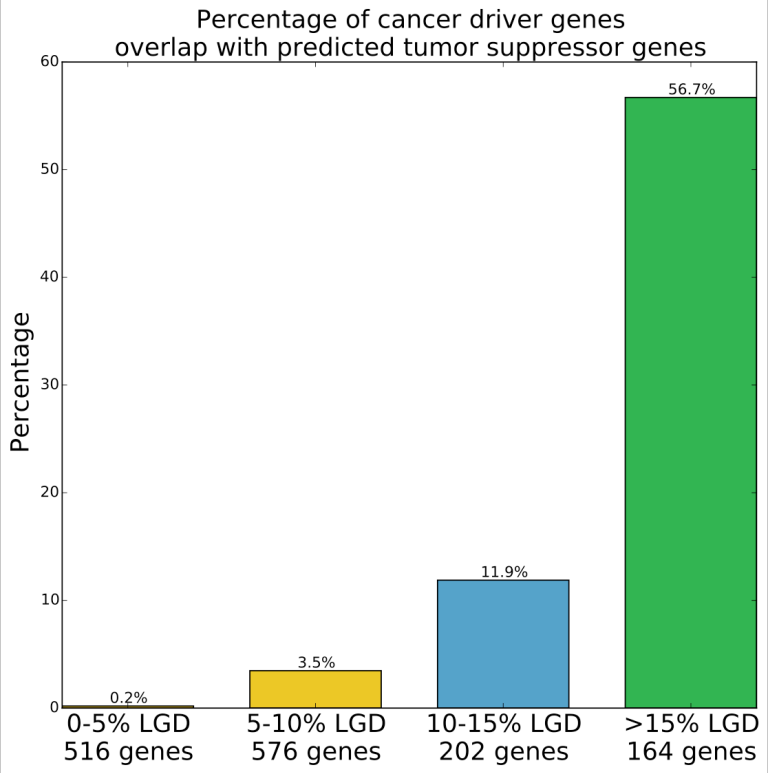


Figure 4.2 Percentage of cancer driver genes overlap with the predicted confident tumor suppressor genes. Green bar is likely tumor suppressor genes with more than 15% LGD variants in COSMIC, red bar (the left-most group) is likely non-suppressors with less than 5% LGD variants in COSMIC

Table 4.5 (a) Enrichment of germline *de novo* variants in cases among cancer driver genes with different LGD% in COSMIC.

LGD % in COSMIC	Observed number of *de novo* LGD variants in NDD cases	Expected number of *de novo* LGD variants in NDD cases	Enrichment	p-value
0-5%	31	13	2.4	1.8e-5
5-10%	96	21.5	4.5	< 1e-20
10-15%	36	8.4	4.3	1.5e-12
>15%	70	6.97	10.0	< 1e-20

(b) Cancer driver genes of different LGD% in COSMIC have different rate of *de novo* LGD variants in NDD cases. The table shows the number of genes that have at least one *de novo* coding variants among cases. Comparison using Fisher's exact test: odds ratio = 4.9, p-value = 1.1e-5.

LGD % in COSMIC	Genes with LGD variants	Genes without LGD variants
0-5%	18	104
>15%	29	34

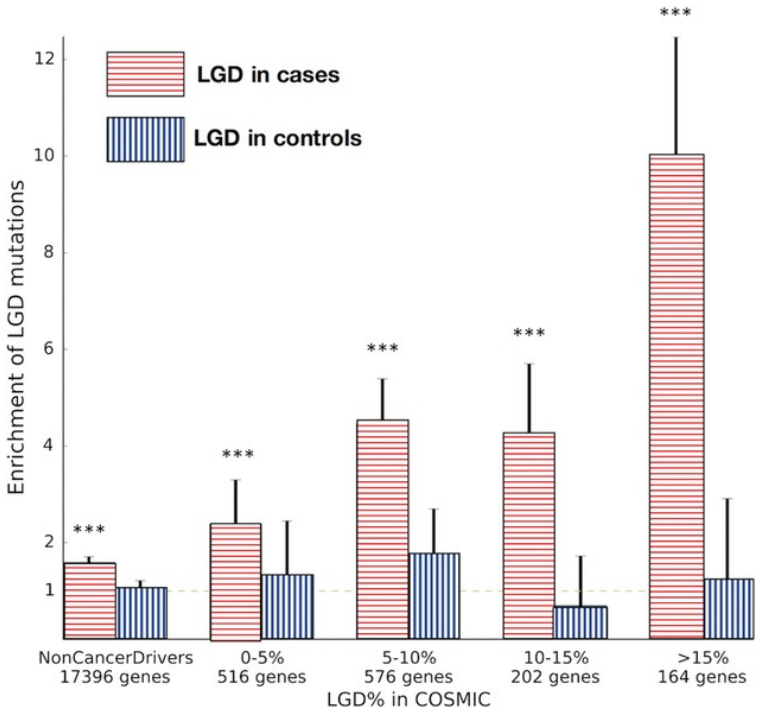


Figure 4.3 Enrichment of germline LGD *de novo* variants in DD patients and controls among candidate cancer driver genes and non-cancer driver genes. Cancer driver genes are grouped based on fraction of LGD somatic mutations among all reported point mutations or small indels in COSMIC. The group with >15% of LGD mutations are likely tumor suppressors. Enrichment values were estimated by comparing observed number of germline *de novo* LGD variants to expectation from background mutation rate in cases or controls. Red bars represent DD cases, blue bars represent controls, error bars represent 95% confident interval. P values (*** indicates P value < 0.001) were calculated using Poisson tests with expected value estimated from background mutation rate.

Functional missense mutations, whether gain or loss of function, disrupt cellular processes in very specific ways. For example, these mutations can cause (gain or loss of) enzymatic activity or (loss of) regulation of protein stability/activity, or affect interaction with other proteins. Therefore, functional missense mutations tend to form clusters in specific regions. We denote these clusters as cancer mutation hotspots. We found that for amino acid positions where there were at least three reported somatic missense mutations in COSMIC, there are 34 *de novo* missense variants in cases and just 1 in controls (fold enrichment = 12, P value = 6.9e–4;

Table 4.6). There was a consistent trend among *de novo* D-mis variants at positions with 1 or 2 reported somatic missense mutations (fold enrichment = 3.1, P value = 0.028; Table 4.6).

Table 4.6 Enrichment of germline missense *de novo* variants in reported somatic missense mutation positions in COSMIC among all candidate cancer driver genes (N=1481).

	Number of somatic missense mutations reported in COSMIC	5542 cases	1911 controls	Enrich-ment	P-value
All missense *de novo* variants	1-2	92	25	1.27	0.34
	>2	34	1	11.7	6.9E-04
D-mis *de novo* variants	1-2	36	4	3.1	0.028
	>2	21	0	NA	0.0041

Several methods have been developed to find mutation hotspots for the purpose of finding cancer driver genes with a high accuracy[113,114]. Among the reported cancer mutation hotspots, we observed a similar trend of enriched *de novo* mutations in DD cases (Table 4.7).

Table 4.7 Enrichment of germline *de novo* missense variants in NDD cases among cancer somatic missense hotspots reported in recent published studies

a) Enrichment in hotspots from Chang et al., 2015

Variant type in hotspots	Case	Control	Fold enrichment	p-value
Missense	5	0	NA	1
D-mis	3	0	NA	1

b) Enrichment in hotspots from Yang et al., 2015

Variant type in hotspots	Case	Control	Fold enrichment	p-value
Missense	4	0	NA	0.211
D-mis	2	0	NA	1

To reach optimal power for this study with a balance of accuracy and sensitivity, we implemented a HMM to predict these hotspots (Materials and Methods section) in genes that are

already implicated as candidate drivers. We collected all somatic missense mutations from COSMIC for each gene and applied our HMM-based methods to detect missense mutation hotspots in all candidate cancer driver genes. Comparing DD cases with controls, we observed a 16× fold enrichment (P value = 1.8e–5) of germline *de novo* D-mis variants in cancer mutation hotspots (Table 4.8), which indicates that almost all such mutations contribute to DDs and corresponds to a class vulnerability value of 90%[39], much greater than D-mis variants (about 25%) in non-cancer drivers (Fig. 4.4). Based on fold enrichment, the estimated number of DD-causative *de novo* missense variants among all candidate cancer drivers is about 135, and the estimated number of such variants in cancer somatic mutation hotspots is 67. This suggests that a large portion (about 50%) of causative *de novo* missense variants in DD cases among cancer driver genes have similar modes of action as cancer somatic mutations.

Table 4.8. **Enrichment of germline *de novo* missense variants in NDD cases located in cancer somatic missense hotspots.**

(a) There are significantly more germline *de novo* missense variants located in hotspots in NDD case comparing to controls.

Variant type in hotspots	Case	Control	Fold enrichment	P-value
Missense	95	11	3	0.00013
D-mis	47	1	16.2	1.8e-5

(b) Among all germline *de novo* missense variants in cancer driver genes, the ones in NDD cases are more likely to be located at cancer somatic missense mutation hotspots than the ones in controls.

Variant type in hotspots	Case		Control		Odds ratio	P-value
	Hotpots	Not hotspots	Hotpots	Not hotspots		
Missense	95	391	11	102	2.3	0.01
D-mis	47	138	1	23	7.8	0.02

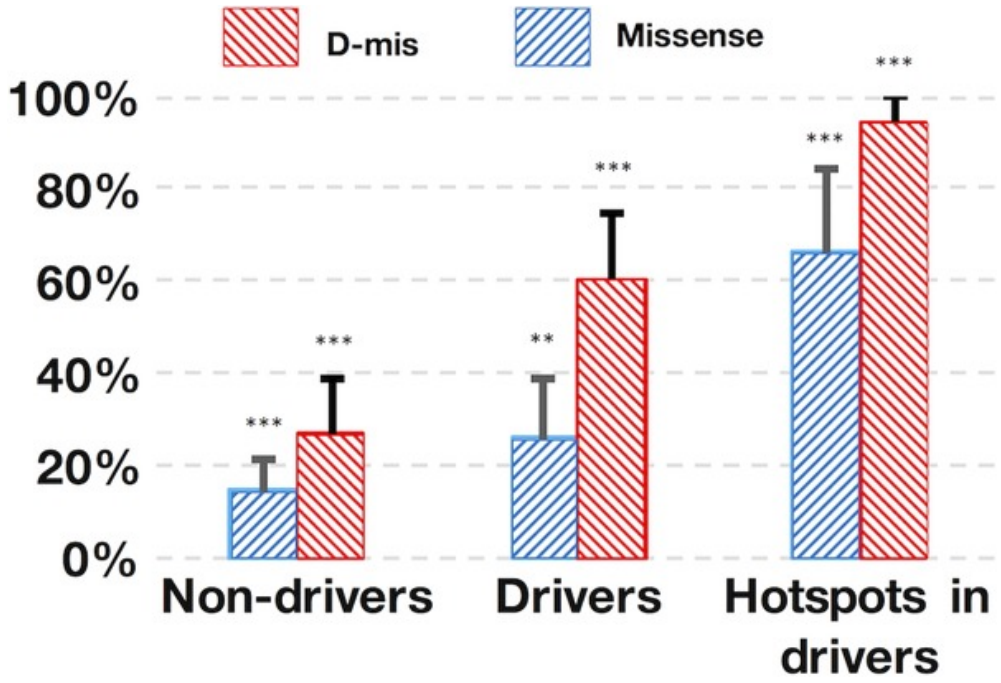


Figure 4.4 Class vulnerability of *de novo* missense variants in different groups of genes. Class vulnerability is defined as the probability of a variant being associated with the disease. D-mis is defined as missense predicted to be damaging by meta-SVM. P values (*** indicates P value < 0.001; ** indicates P value < 0.01) were calculated using Binomial tests.

Figure 4.5 shows a few representative genes. *EP300*, a known tumor suppressor, has one D-mis *de novo* variant at a cancer mutation hotspot in one DD patient (Fig. 4.5A), consistent with its implicated role through loss of function with other five LGD *de novo* variants in DD data sets. *ARID2* [90 609539] is another tumor suppressor and part of SWI/SNF chromatin remodeling complex. One autism patient had a germline *de novo* missense variant in *ARID2* at a somatic mutation hotspot (Fig. 4.5B). *DDX3X* [90 300160], a tumor suppressor implicated in intellectual disability[86], has three missense *de novo* variants in our compiled DD data sets and all located in cancer hotspots (Fig. 4.5C). *MAP2K1* [90 176872], a proto-oncogene, has a missense *de novo* variant located in a cancer hotspot (Fig. 4.5D) in an autism case, suggesting that the variant plays a similar role as gain-of-function mutations implicated in syndromes[115] with ASD features. *PPP2R1A* [90 605983], a recently discovered DD risk gene[86], harbors three missense *de novo* variants

in two cancer hotspots (Fig. 4.5E). *PPP2R1A* is likely an oncogene in ovarian clear cell carcinoma[116], consistent with its gain of function roles in both cancer and DDs. *CNOT3* [90 604910], a tumor suppressor gene[117], has two *de novo* LGD variants and two *de novo* D-mis variants, one of which is located in a mutation hotspot (Fig. 4.5F), indicating it is a potential DD risk gene. Not all driver genes have similar mode of action. *CTNNB1* [90 116806], a central player in Wnt signaling, is a proto-oncogene in various cancers[118]. It has a very small fraction (0.5%) of LGD somatic mutations in COSMIC, and most missense somatic mutations disrupt the phosphorylation sites at the N-terminal end that are required for phosphorylation- dependent degradation. In contrast to somatic mutations in cancer, it is usually haploinsufficient and harbors LGD variants in patients with neurodevelopmental syndromes[119]. In the DD data sets, we compiled there are seven LGD *de novo* variants, consistent with a haploinsufficiency mechanism. In addition, there is a missense variant in an autism case. This missense variant is not located in any somatic mutation hotspot (Fig. 4.5G), and is therefore unlikely to cause gain-of-function in *CTNNB1*. This is consistent with the notion that, this variant is either implicated in autism via loss of function similar to other LGD variants, or not associated with the disease. *SMARCA4* [90 603254], a tumor suppressor gene[120,121], harbors three deleterious missense *de novo* mutations in the DD cases, none of which is located in cancer mutation hotspots (Fig. 4.5H). This is consistent with previous report that SMARCA4 may have gain of function or dominant negative mutations in DDs[122].

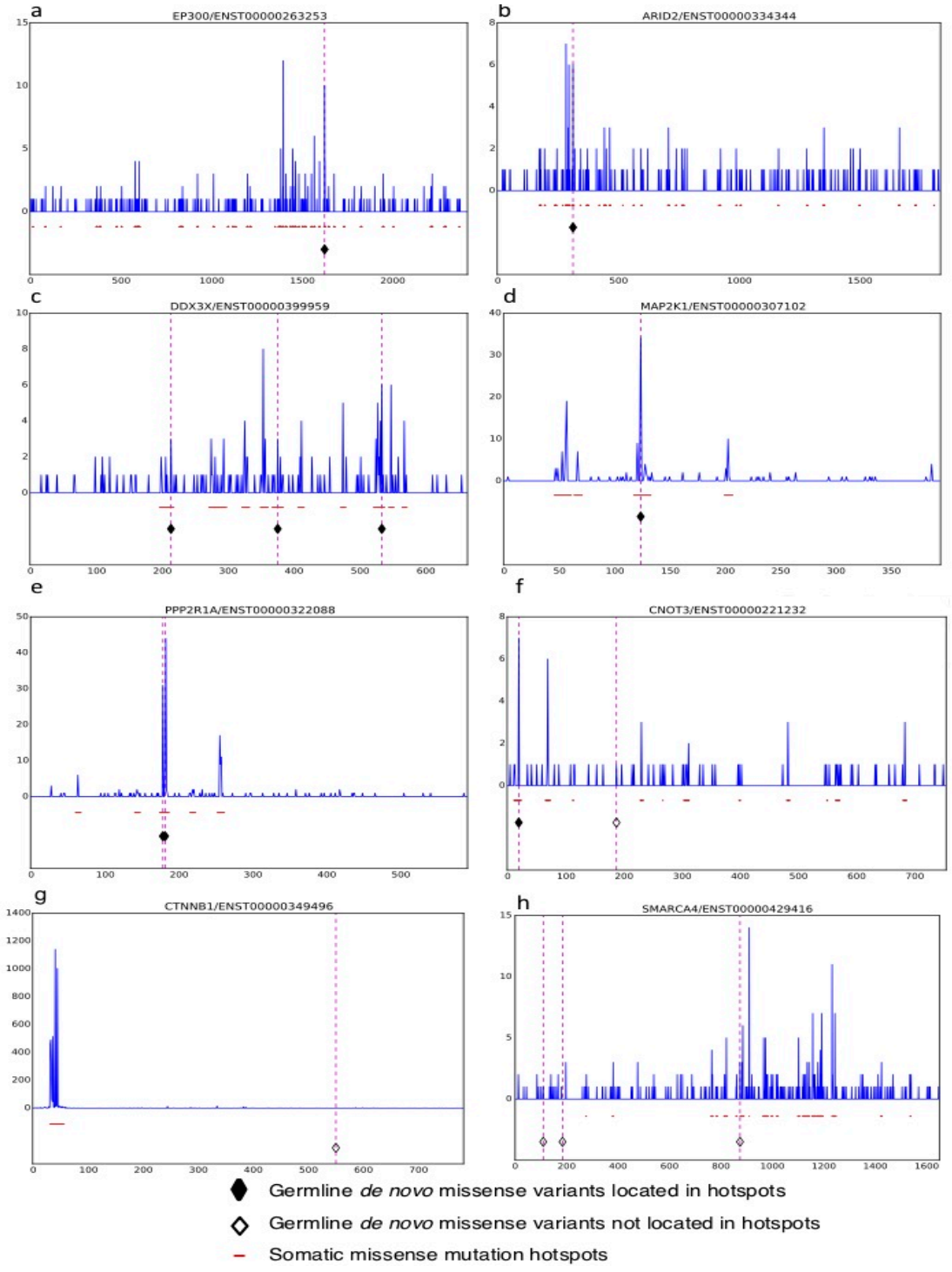


Figure 4.5 Examples of germline *de novo* missense variants in DD patients superimposed with cancer somatic mutation hotspots. Blue spike lines are somatic missense counts at each amino acid position in cancer. Red dashes indicate predicted hotspot positions by the Hidden Markov Model method. Filled diamonds show germline *de novo* variants that are located in somatic hotspots, and hollow diamonds represent germline *de novo* variants that are not located in somatic hotspots. The following genes are shown: A: *EP300*; B: *ARID2*; C: *DDX3X*; D: *MAP2K1*; E: *PPP2R1A*; F: *CNOT3*; G: *CTNNB1*; H: *SMARCA4*.

4.3 Discussion

Recent large-scale exome sequencing studies of DDs uncovered many candidate risk genes and pathways through deleterious germline *de novo* mutations. Many of these genes and pathways have been previously implicated in cancer through somatic mutations. Such genetic connection is reasonable because both classes of diseases involve disruption of similar fundamental cellular processes such as growth, proliferation, and differentiation. In this study, we hypothesize that quantifying such connection between DDs and cancer would lead to better understanding of how genes are disrupted through mutations, and ultimately allow us to leverage the vast amount of caner mutation data to improve genetic discovery in DD studies. Based on data from recently published large-scale DD studies and cancer genomics resources, we found that in DD patients there is a significantly greater burden of functional *de novo* mutations in candidate cancer driver genes than in non-cancer driver genes. And such enrichment trend holds in both candidate tumor suppressors and oncogenes (Table 4.9).

Table 4.9 Enrichment of damaging *de novo* variants in tumor suppressors and oncogenes. Tumor suppressors are defined as the cancer driver genes with fraction of LGD mutation reported in COSMIC > 15% or p-value < 0.001 in Davoli et al 2013. Oncogenes are defined the cancer driver genes with p-value < 0.001 in Davoli et al 2013.

Gene group	Type of *de novo* variant	Cases (N=5542)	Controls (N=1911)	Fold enrich-ment	P-value
Tumor suppressor genes (N=208)	LGD	98	2	6.8	4.7E-8
	Missense	108	27	1.4	0.14
	D-mis	47	4	4.1	0.002
Oncogenes (N=61)	LGD	12	1	4.1	0.21
	Missense	34	8	1.5	0.38

	D-mis	19	2	3.3	0.13

Specifically, about 38% of all potentially causative damaging *de novo* mutations observed in these DD patients are located in cancer drivers, and about 27%–45% of DD risk genes are likely cancer driver genes. This indicates that we can prioritize known cancer driver genes to find candidate risk genes in DD studies.

Additionally, we investigated whether driver somatic mutations in cancer and causative germline *de novo* variants in DDs have similar modes of action. We found that likely tumor suppressor genes, that is, the ones with larger fraction of LGD mutations (≥15%) reported in COSMIC, have a significantly higher burden of germline *de novo* LGD variants than likely non-suppressors (somatic LGD fraction <5%) in DD patients, indicating that tumor suppressor genes often exert DD risk through loss of function germline *de novo* variants that disrupt molecular pathways in DD similar to the ones in cancer. Several well-known oncogenes have gain-of-function germline *de novo* missense variants that cause DDs. However, in general it remains a challenge to infer whether a missense mutation causes gain or loss of function of the gene. We therefore asked whether missense mutations in cancer (somatic) driver genes and DD germline risk genes have similar modes of action. We hypothesized that in both tumor suppressors and oncogenes, functional somatic missense mutations in driver genes occur in "hotspots" in a driver gene. We indeed found a stronger enrichment of damaging *de novo* missense variants located in these hotspots in DD patients than in controls. Specifically, we estimated that about 48% of causative *de novo* missense variants observed in DDs among cancer driver genes are located in hotspots, indicating that missense mutations also often have similar mode of action in cancer and DDs. This ratio is likely an underestimate, since the power of detecting mutation hotspots is limited in many cancer driver genes due to relatively small number of mutations. We observed

this enrichment of hotspot variants in both tumor suppressors and oncogenes. Strikingly, the case control comparison indicates that when located in cancer mutation hotspots, most of the germline *de novo* missense mutations in DD patients are implicated with the disease. This suggests that in addition to using cancer driver genes to prioritize candidate risk gene in DD studies, we can leverage cancer somatic mutation data to improve functional assessment of germline rare or *de novo* variants in these genes observed in DD patients, potentially improving both risk gene discovery in genetic studies and genetic diagnosis in clinical testing.

Based on the evidence of LGD variants in tumor suppressors and D-mis variants located in cancer somatic mutation hotspots, we identified two new candidate risk genes for DDs. The first is *ARID2*, which harbors a *de novo* D-mis variant in an autism patient. The variant is located at the second most recurrently mutated position reported in COSMIC. A potential role of ARID2 in autism is consistent with its recently implicated role in causing intellectual disability with *de novo* LGD variants[123]. The second gene is *CNOT3*, which harbors two *de novo* LGD variants and two *de novo* D-mis variants in four different patients, including one with autism and three with undiagnosed DDs[86]. *CNOT3* is a tumor suppressor[117] with a very large fraction (~24%)[105] of LGD mutations among all reported somatic mutations, indicating that its suppressor role is through haploinsufficiency[112]. One of the *de novo* D-mis variants is located at the most recurrently mutated site reported in COSMIC. *CNOT3* is a component of CCR4–NOT complex, which is one of the major cellular mRNA deadenylases[124] and has a broad role in post-transcriptional regulation of gene expression[125]. Post-transcriptional regulation of gene expression has been implicated as a major pathway with neurodevelopment disorders[6]. This supports *CNOT3* as a candidate risk gene of DDs. Future genetic and functional studies are required to confirm and validate these two candidate risk genes.

In summary, our results suggest that we can view tumors as natural laboratories for assessing the deleterious effects of mutations that are applicable to germline variants, which will enable us to improve identification of causal genes and variants in DDs. Our study is still limited by inadequate number of sequenced cancer genomes in a few ways. First, we have limited power to detect mutation hotspots in a substantial portion of cancer driver genes due to a relatively small number of mutations, especially among the genes that are mutated in a small fraction of cancer patients or cancer types. This lack of power leads to lower sensitivity and specificity. Recent works on clustering of somatic mutations in 3D[126] or pooled homologous domains[127] present promising directions to improve the power. Second, many genes have a diverse set of functions, and clinically distinct types of cancer[128,129] or diseases often involve disruption of different functions of the same gene. Categorizing these disruptions as gain or loss of function is overly simplification. Although our approach of detecting somatic mutation hotspots does not rely on such simplified assumption, the complexity does lead to decreased power in detection of somatic mutation hotspots, and increased difficulty in utilization and interpretation of the somatic mutation hotspots in DDs. Finally, we do not have a complete catalog of cancer driver genes, and our list of candidate cancer driver genes may contain a non-negligible number of false positives. Ongoing international efforts in cancer precision medicine are generating much larger cancer mutation data sets. With prudent data sharing practices, this will improve cancer driver genes and mutation hotspots detection in the future, and make cancer data more valuable to genetic studies and diagnosis of DDs.

4.4 Material and methods

4.4.1 Candidate cancer driver genes

The candidate cancer driver genes list is are comprised of census genes from COSMIC[105], significantly mutated genes from TCGA studies curated by cBioPortal[103] and candidate genes from forward genetic screens in mice in The Candidate Cancer Gene Database (CCGD)[104] (Fig. 4.1). For cBioPortal data, we included genes with MutSigCV q-value less than 0.1 in individual TCGA studies as significantly mutated genes. We excluded results from the Adrenocortical Carcinoma and Pancreatic Adenocarcinoma datasets because while these two cancer datasets have a moderate number of samples, many of these genes have q-values less than 0.1. For CCGD data, we only considered the genes with relative rank A[104]. We further filtered these CCGD genes based on mutation data in TCGA. Specifically, we counted the variants of various functional categories (LGD, missense, silent) reported in TCGA[95], and tested whether there is significant excess ($P < 0.05$) of missense or LGD mutations compared with silent mutations based on germline gene-specific background mutation rates[40]. We note that the background somatic mutation rate is affected by various processes[130,131] that are different to germline mutations. However, the usage of germline background in this study is justified by the observation that there is a very strong correlation between observed ratio of missense/silent (or LGD/silent) somatic mutations and ratio of missense/silent (or LGD/silent) germline background mutation rate among non-candidate cancer driver genes (correlation coefficient = 0.46; Fig. 4.6). All other genes not included in cancer driver genes are classified as non-cancer driver genes.

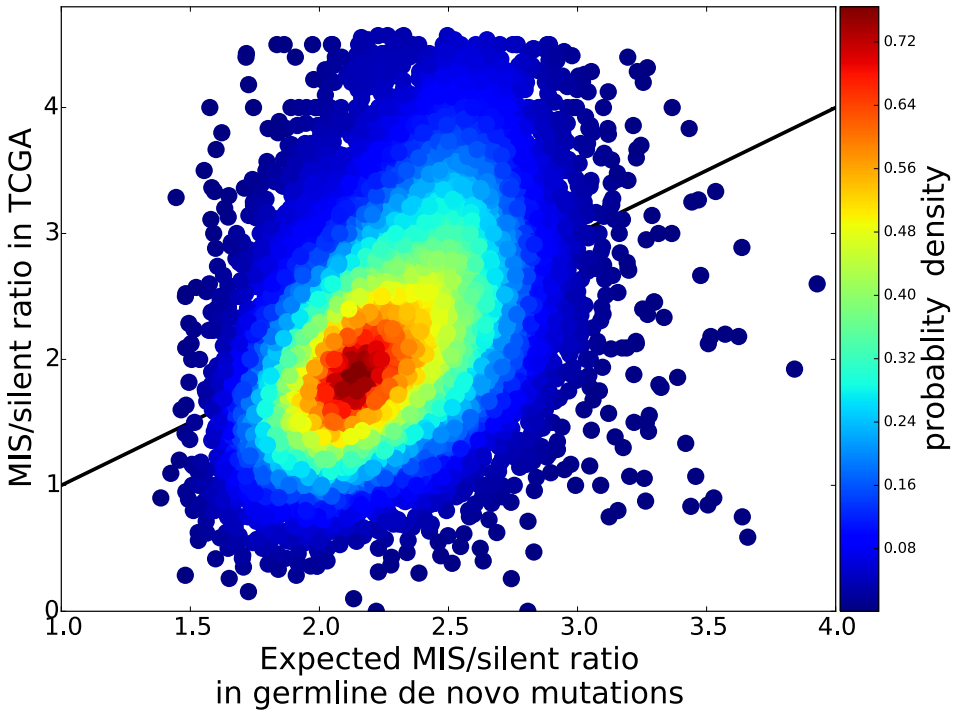


Figure 4.6 Observed cancer somatic missense/silent mutation ratio versus expected ratio using germline background *de novo* missense/silent mutation rate from Samocha et al 2014. We require the silent variants counts to be more than 5 and remove outliers. The black line indicates equal value. Color bar indicates the probability density.

4.4.2 Germline *de novo* mutations of DDs

We compiled germline *de novo* variants from 5,542 DD cases in recent published studies, including 3,962 cases with autism spectrum disorder[6,39], 1,133 cases from Deciphering Developmental Disorders study[86], 191 cases with epileptic encephalopathies[24], and 264 cases with intellectual disability[85,96-98]. We re-annotated these mutations using ANNOVAR[43] software to have complete gene annotation as well as function annotation. The functional consequence of missense mutations is predicted in silico by meta-SVM. In this study, we only consider mutations in the exonic regions.

4.4.3 Burden test and estimation of number of causative damaging *de novo* mutations

Burden test between case and control: Within gene sets, we counted the number of mutations inside the gene set of different functional categories (LGD, missense, silent) for both cases and controls. We assumed that *de novo* variants are sequences of individual Bernoulli trials and we used the portion of case trios as the success probability to calculate the two-side binomial distribution P value as well as fold enrichment.

Burden comparison between cancer and non-cancer drivers: We counted the number of *de novo* variants in candidate cancer driver genes and non-cancer drivers of different functional categories (LGD, missense, silent) for both cases and controls. We used two-side Fisher's exact test to test the null hypothesis that the case/control burden of various categories is the same among cancer driver genes and non-cancer driver genes.

Estimation of number of causative *de novo* mutations and class vulnerability in gene sets: In a group of genes (e.g., cancer drivers), there are L1 LGD (or D-mis) mutations from n_1 cases and L_2 LGD (or D-mis) mutations from n_2 controls, we estimate the number of causative variants C by:

$$C = L_1 - \frac{L_2 * n_1}{n_2}$$

and class vulnerability V by:

$$V = \frac{C}{L_1}$$

Using the L_1 and L_2 as the Poisson distribution rate to simulate 10,000 trials, we can calculate the 95% confident intervals of causative variants and class vulnerability.

Estimation of percentage of causative mutations in cancer driver genes: We first counted the number of *de novo* LGD (or D-mis) variants in all genes and in candidate cancer

driver genes for both cases and controls, then we used the variant counts as the Poisson distribution rate to simulate 10,000 trials. Dividing the number of simulated causative variants in cancer drives genes by the simulated mutations in all genes, we obtain the expectation as well as a 95% confident interval.

4.4.4 Infer candidate risk genes of DDs

TADA (transmission and *de novo* association)[6,109] is a Bayesian method for identification of risk genes using rare or *de novo* variants. We tallied the occurrence of *de novo* variants in two categories: LGD and D-mis. We used gene-specific mutation rate[40] as the parameter for the Poisson distribution and calculated its corresponding false discovery rate (FDR) using other default parameters.

We defined DD candidate risk genes using FDR calculated by TADA. With each FDR threshold, we obtained the number of candidate DD risk genes (N) and the number Nc of such genes that are also candidate cancer drivers. We estimated the number (F) of false positive DD risk genes by FDR definition:

$$F = N * FDR$$

To estimate the fraction (f) of true DD risk genes that overlap with candidate cancer driver genes, we assumed false positive DD risk genes overlap with candidate cancer driver genes just by chance, which is determined by background germline *de novo* mutation rate. In most TADA FDR bins (FDR < 0.5), the false positive risk genes should have at least one damaging *de novo* variant (LGD or D-mis). By calculating the sum of germline damaging mutation rate in cancer driver genes divided by all genes, we determined that the overlap rate by chance is r = 10%.

Finally, for each TADA FDR bin, the fraction of true DD risk genes that are also candidate cancer driver genes was estimated by:

$$f = \frac{N_c - F * r}{N - F}$$

4.4.5 Hidden Markov Model to infer cancer somatic missense mutation hotspots

We implemented a Hidden Markov Model (HMM) to predict somatic missense mutation hotspots in each candidate cancer driver gene. We assume that the background somatic mutation rate is uniform across a given gene. For each transcript in the given gene, we inspected the somatic missense mutations from COSMIC. We counted all missense mutations at each given amino acid site, regardless of actually amino acid changes, to identify mutation hotspots. We defined hotspots in two ways: (1) highly recurrent mutation sites and (2) sites with non-background states prediction by HMM. Recurrent mutated positions were defined as having more than 3.5 median-absolute-deviation number of mutations. After excluding recurrent sites, we took a sliding window of size 8 and summed the number of mutations for each position to reduce the fluctuation of mutations in a neighborhood region. We used the smoothed position-specific mutation counts as the input to a HMM with Poisson emission probability and three hidden states, including: (a) the "background" state, (b) possible mutation hotspot state, and (c) probable mutation hotspot state. We used germline mutation background to estimate the fraction of missense mutations that are drivers in each gene. This is based on the observation that, among non-cancer driver genes, the ratio of reported missense/silent somatic mutations is close to gene-specific background mutation rate estimated by Samocha et al. 2014 (regression slope = 0.97 and intercept close to zero; Fig. 4.7). We simulated the missense mutation counts (S1) and silent mutation counts (S2) in each gene using the corresponding recorded COSMIC data (C1, C2) as

the Poisson distribution rate. With the missense/silent ratio from germline *de novo* background (R1), we estimated the mean and 95% confidence interval of the fraction of missense mutations (f) that are drivers in each gene by:

$$f = \frac{S_1 - S_2 * R_1}{S_1}$$

We used the upper bound of 95% CI as the maximum allowed (M) percentage of missense drivers from HMM. To obtain reasonable initial values for HMM parameters, we then calculated the expected number of driver missense mutations per position (T) by: $T = \frac{C_1 * (1-f)}{L}$ L where L is the total transcript length. We set the lambda (mean of a Poisson distribution) for the background state to be at least T. To restrict the number of transitions between background and hotspot states, we took the average of the diagonal of the transition matrix of the Baum–Welsh result with 0.99 if the corresponding transition matrix elements were smaller than 0.99 in each iteration. After convergence, we used the Viterbi algorithm to find the most probable state path and forward–background algorithm to calculate posterior marginal probabilities of hidden states for each position. To identify the hotspots, we took positions with the non-background states as the hotspots, with exception that if the fraction of somatic missense mutations in those hotspots exceeded M, we ranked those positions by their marginal probability of being background states (increasingly), and included such positions until the fraction of missense mutations in hotspots reached M.

Chapter 5

Predicting pathogenicity of missense variants by deep learning

5.1 Introduction

Missense variants are the most common type of coding genetic variants and are a major class of genetic risk across a broad range of common and rare diseases. Previous studies have estimated that there is a substantial contribution from *de novo* missense mutations to structural birth defects[47,71] and neurodevelopmental disorders[6,13,39]. However, only a small fraction of missense *de novo* mutations are pathogenic[39]. As a result, the statistical power of detecting individual risk genes based on missense variants or mutations is limited[27]. In clinical genetic testing, many of missense variants in well-established risk genes are classified as variants of uncertain significance, unless they are highly recurrent in patients. Previously published *in silico* prediction methods have facilitated the interpretation of missense variants, such as CADD[28], VEST3[29], metaSVM[30], M-CAP[31], and REVEL[32]. However, based on recent *de novo* mutation data, they all have limited performance with low positive predictive value (Table 5.1), especially in non-constrained genes (defined as ExAC[67] pLI<0.5).

Table 5.1 Estimated number of pathogenic missense *de novo* mutations using published methods by recommended thresholds. The table indicates their thresholds, estimated number of risk variants and positive predictive values in Congenital heart disease and Autism spectrum disorder data.

a) Evaluation among all genes

		Congenital heart disease (CHD)		Autism spectrum disorder (ASD)	
	Threshold	Estimated number of risk variants	Estimated Precision	Estimated number of risk variants	Estimated precision
All missense	N/A	264	0.17	264	0.13
M-CAP	> 0.025	219	0.26	202	0.18
Meta-SVM	> 0	115	0.31	105	0.22
MutationTaster	> 0.5	187	0.18	201	0.14
Polyphen	> 0.5	170	0.22	183	0.17
SIFT	< 0.05	151	0.17	183	0.15
VEST3	> 0.8	115	0.28	134	0.24
CADD	> 15	195	0.17	237	0.15
REVEL	> 0.5	133	0.33	162	0.3

b) Evaluation among constrained genes (ExAC pLI≥0.5)

		Congenital heart disease (CHD)		Autism spectrum disorder (ASD)	
	Threshold	Estimated number of risk variants	Estimated precision	Estimated number of risk variants	Estimated precision
All missense	N/A	140	0.29	163	0.25
M-CAP	> 0.025	115	0.37	118	0.29
Meta-SVM	> 0	64	0.44	64	0.34
MutationTaster	> 0.5	102	0.25	134	0.23
Polyphen	> 0.5	85	0.32	113	0.30
SIFT	< 0.05	90	0.29	107	0.24
VEST3	> 0.8	83	0.44	96	0.39
CADD	> 15	106	0.27	147	0.26
REVEL	> 0.5	87	0.49	90	0.40

c) Evaluation among non-constrained genes (ExAC pLI<0.5)

		Congenital heart disease (CHD)		Autism spectrum disorder (ASD)	
	Threshold	Estimated number of risk variants	Estimated precision	Estimated number of risk variants	Estimated precision
All missense	N/A	124	0.12	101	0.07
M-CAP	> 0.025	104	0.20	84	0.11
Meta-SVM	> 0	50	0.23	40	0.14
MutationTaster	> 0.5	85	0.13	66	0.07
Polyphen	> 0.5	84	0.17	70	0.10
SIFT	< 0.05	61	0.11	76	0.09
VEST3	> 0.8	31	0.14	38	0.12
CADD	> 15	89	0.12	90	0.09
REVEL	> 0.5	45	0.2	71	0.21

Here we hypothesize that missense variant pathogenicity prediction can be improved in a few dimensions. First, conventional machine learning approaches have limited capacity to leverage large amount of training data compared to recently developed deep learning methods[132]. Second, databases of pathogenic variants curated from the literature are known to have a substantial frequency of false positives[133], which are likely caused by common issues across databases and therefore introduce inflation of benchmark performance. Developing new

benchmark data and methods can help to assess and improve real performance. Finally, previous methods do not consider gene dosage sensitivity[67,134], which can modulate the pathogenicity of deleterious missense variants, as hypomorphic variants are pathogenic only in dosage sensitive genes[13]. With recently published metrics of mutation intolerance, it is now feasible to consider gene dosage sensitivity in predicting pathogenicity. Based on these ideas, we developed a new method, MVP, to improve missense variant pathogenicity prediction.

5.2 Results

5.2.1 Derivation of the MVP score

MVP uses many correlated predictors, which can be broadly grouped into two categories: (a) "raw" features computed at different scales, per base pair (e.g. amino acid constraint score and conservation), per local context (e.g. protein structure and modification) as well as per gene (e.g. gene mutation intolerance, sub-genic regional depletion of missense variants [135]); (b) deleteriousness scores from selected previous methods. We reason that the variants in constrained genes (ExAC pLI≥0.5) and non-constrained genes (ExAC pLI<0.5) may have different modes of action of pathogenicity, therefore, trained our models for the two gene sets separately. We included 38 features for the constrained gene model, and 21 features for the non-constrained gene where we removed most published prediction methods features due to limited prediction accuracy (Table 5.1).

MVP uses a deep residual neural network (ResNet)[136] model. There are two layers of residual blocks, consisting of convolutional filters and activation layers, and two fully connected layers with sigmoid output (Fig. 5.1).

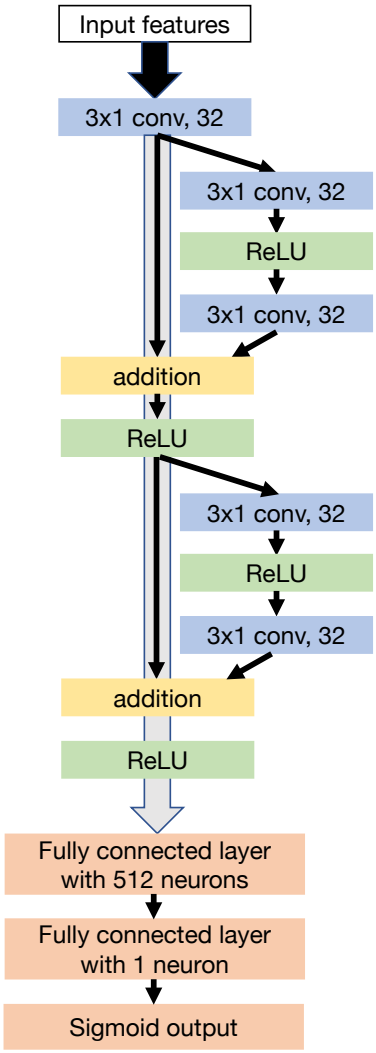


Figure 5.1 The ResNet neural network architecture of MVP. Building blocks are arranged as shown in the figure. Parameters and dimensions of input and output are indicated in the boxes. Blue boxes are convolutional filters, green boxes are ReLU activation, yellow boxes are addition of output from 2 layers, orange boxes are fully connected layers.

The convolutional filters can exploit spatial locality by enforcing a local connectivity pattern between "neurons" of adjacent layers and identify nonlinear interactions at higher levels of the network. To take advantage of this characteristic, we ordered the predictors based on their correlation, as highly correlated predictors are clustered together (Fig. 5.2). Notably, some protein-related predictors are weakly correlated with previous scores, suggesting that they may include additional information and can help improve the overall prediction accuracy. For each

missense variant, we defined MVP score by the rank percentile of the ResNet's raw sigmoid output relative to all 76 million possible missense variants.

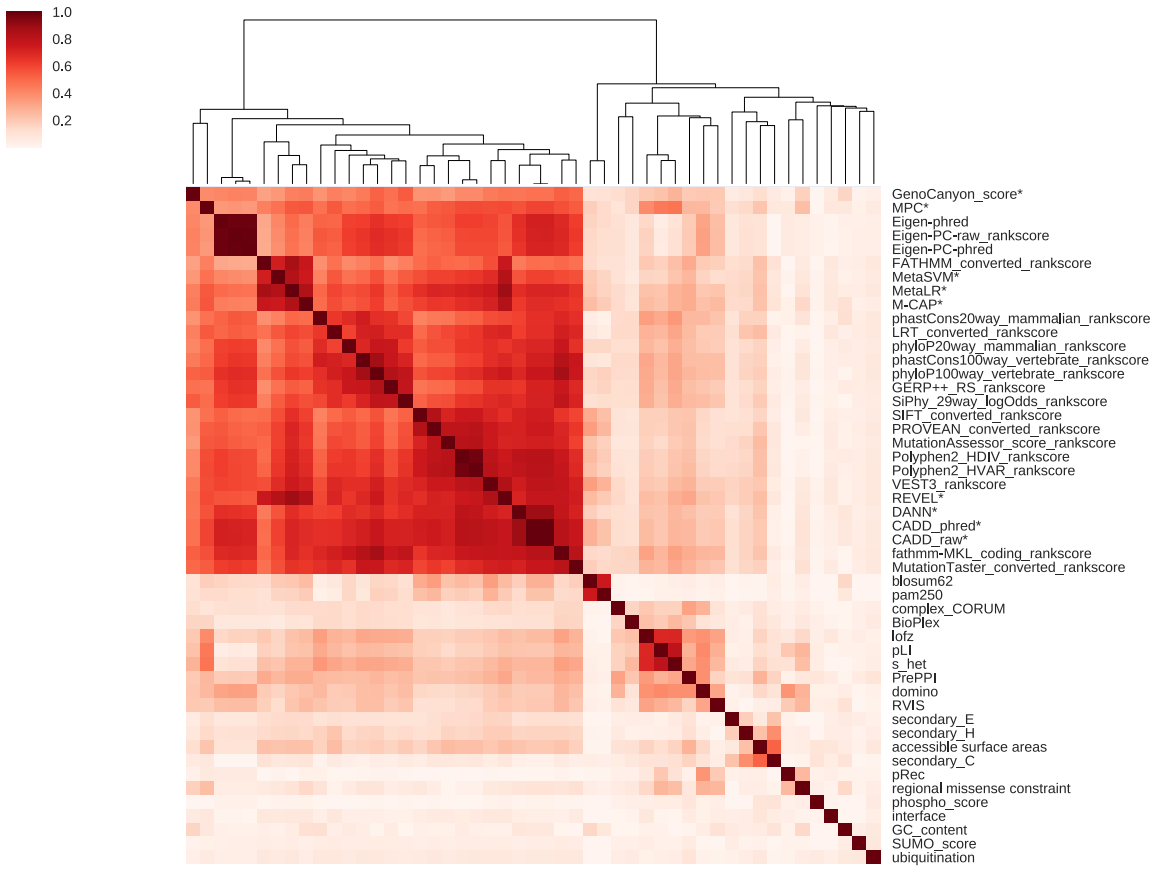


Figure 5.2 Correlation and hierarchical clustering of features and additional published methods. We calculated pairwise Spearman correlation of all features and additional published methods across data points used in the training**.** Color key indicates absolute value of Spearman correlation coefficient among features and predictors. Columns are ordered by hierarchical clustering. Published methods marked with * are not used in training.

We obtained large curated datasets of pathogenic variants as positives and random rare missense variants from population data as negatives for training (Table 5.2). Using 6-fold cross-validation on the training set, MVP achieved mean area under the curve [97] of 0.99 in constrained genes and 0.97 in non-constrained genes.

Table 5.2 Summary statistics of training and testing data sets

Data sets				Total variants	Total genes
Training sets	Positive	HGMD positive variants		22390	1058
		Uniprot positive variants		12875	1070
		Clinvar pathogenetic variants		4424	813
		Total unique positive variants		32074	1914
	Negative	Uniprot negative		5190	3240
		Human-derived changes		39593	11739
		Randomly selected DiscovEHR rare variants		42415	14311
		Total unique negative variants		86620	16786
Testing sets	Curated benchmark datasets	VariBench dataset		3333	459
		DiscovEHR rare variants excluded from training		3486	2960
	Cancer datasets	Cancer hotspot		875	204
		DiscovEHR rare variants excluded from training		8771	4801
	De novo datasets	Cases	Autism spectrum disorder (ASD)	2133	1843
			Congenital heart disease (CHD)	1530	1373
		Controls	Simons Simplex Collection unaffected siblings	869	817

5.2.2 Comparing MVP to different model structures

We first compared MVP with a three layers feed forward network with 256 * 256 * 256 neurons. The feed forward network has 637,534,208 parameters. Given the limited number of training dataset, it quickly go to over-fitting after two iterations and result in large fluctuation in performance. In the CNN framework, there are 12,416 parameters in the residual layers and total 636,161 parameters in constrained model and 357,633 parameters in non-constrained model. In MVP model, we put highly correlated features closely so that first residual layers can capture local context interaction within groups while high order residual layers can capture non-linear interaction between groups. We then tested different model structure of various numbers of residual blocks to assess the performance. With all other parameter fixed, two residual blocks parameters has 12,544 parameters before fully connected layers, the model saturated around 20 iterations. Adding a third residual block increases the number to 18,752 parameters and the model saturated around 8 iterations, which indicated over fitting quickly. The results indicated

that the model is sensitive to the parameters in early layers and larger as well as cleaner data set are needed to train a deeper network to fully utilize the power of deep learning model. We also tested model between 512 neurons and 1024 neurons in fully connected layers, 1024 neurons will double the number of parameters in the fully connected layers and results in over-fitting quickly. Other hyper parameters are kernel size of 3, pooling size of 2, depths with 32 and ReLu as activation functions are commonly used in deep learning model.

5.2.3 Comparing MVP to published methods in synthetic data

To evaluate predictive performance of the MVP and compare it with other methods, we evaluated the performance in an independent curated testing dataset from VariBench[30,137] (Fig. 5.3). MVP outperformed all other scores with an AUC of 0.96 and 0.92 in constrained and non-constrained genes, respectively. A few recently published methods (REVEL, M-CAP, VEST3, and metaSVM) were among the second-best predictors and achieved AUC around 0.9.

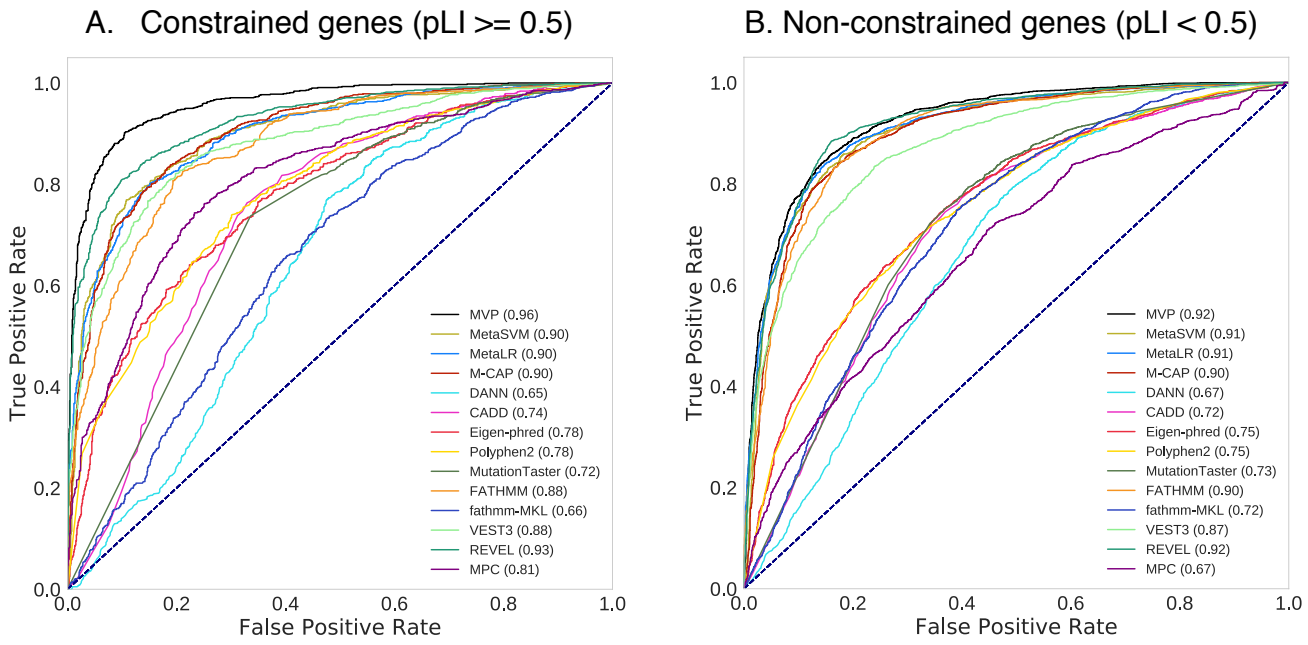


Figure 5.3 Comparing MVP with previous methods by ROC curves using VariBench testing data. (A) Performance evaluation in constrained genes. (B) Performance evaluation in non-constrained genes. The performance of each method is evaluated by the ROC curve and AUC score indicated in parenthesis. Higher AUC score indicates better performance.

Systematic false positives caused by similar factors across training and VariBench data sets could inflate the performance in testing. To address this issue, we obtained two additional types of data for further evaluation. First, we compiled cancer somatic mutation data, including missense mutations located in inferred hotspots based on statistical evidence from a recent study[138] as positives, and randomly selected variants from DiscovEHR[139] database as negatives. In this dataset, the performance of all methods decreased, but MVP still achieved the best performance of AUC of 0.91 and 0.85 in constrained and non-constrained genes, respectively (Fig. 5.4).

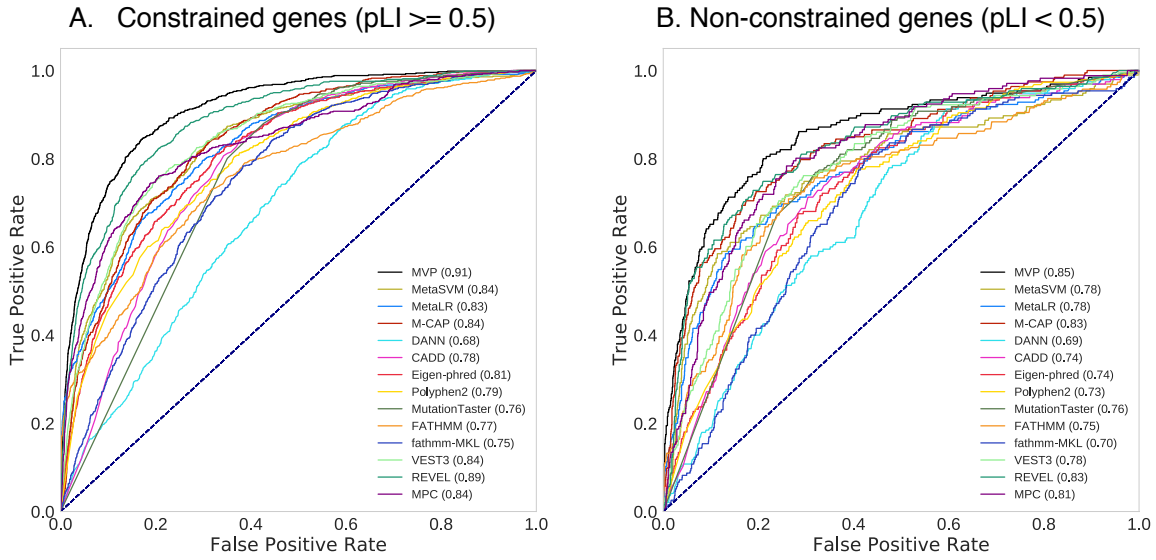


Figure 5.4 ROC curves for existing prediction scores and MVP scores of cancer somatic mutation data sets. (A) Constrained genes: evaluation of 699 cancer mutations located in hotspots from 150 genes, and 6989 randomly selected mutations from DiscovEHR database excluding mutations used in training. (B) Non-constrained genes: evaluation of 177 cancer mutations located in hotspots from 55 genes and 1782 randomly selected mutations from DiscovEHR database excluding mutations used in training. The performance of each method is evaluated by the ROC curve and AUC score indicated in parenthesis. Higher AUC score indicates better performance.

Databases of pathogenic variants curated from the literature are known to have a substantial frequency of false positives. There are likely similar factors causing false positives across different databases. Therefore, dividing the datasets into training and testing data does not create truly independent data for performance assessment, and as a result, the AURC calculated from VariBench data is likely inflated for methods trained on these dataset, including MVP and other methods with best AUROC values. This is supported by results in Figure 5.5: using cancer somatic mutation hotspots as positives, and randomly selected rare variants from DiscovEHR as negatives, the area under receiver operating characteristic curve (AUROC) of all methods trained by HGMD or UniProt is substantially decreased (Figure 5.5). Notably, MPC, which was trained

on a small set of high-confidence ClinVar data, saw increased performance in cancer data, especially in non-constrained genes.

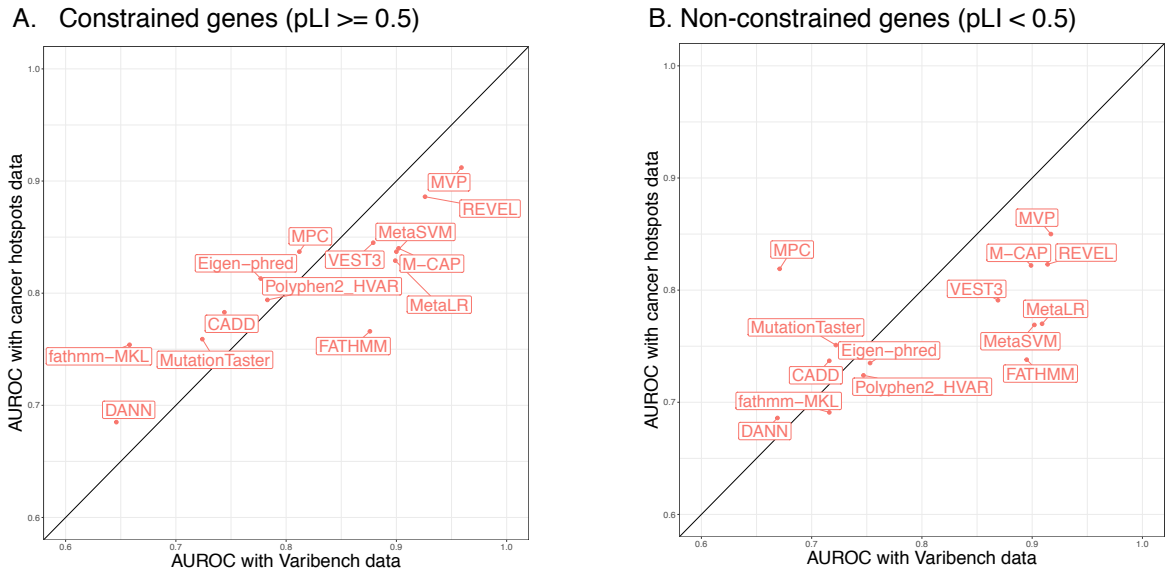


Figure 5.5. **Comparison of AUC using VariBench data versus cancer mutation hotspots data for MVP and previous methods**. X-axis indicates the AUC with VariBench data; y-axis indicates the AUC with cancer hotspots data. (A) Comparison in constrained genes. (B) Comparison in non-constrained genes.

To investigate the contribution of features to MVP predictions, we performed cross-one-group-out experiments and used the differences in AUC as an estimation of feature contribution (Fig. 5.6). We found that in constrained gene, conservation scores and published deleteriousness predictors have relatively large contribution, whereas in non-constrained genes, protein structure and modification features and published predictors are most important.

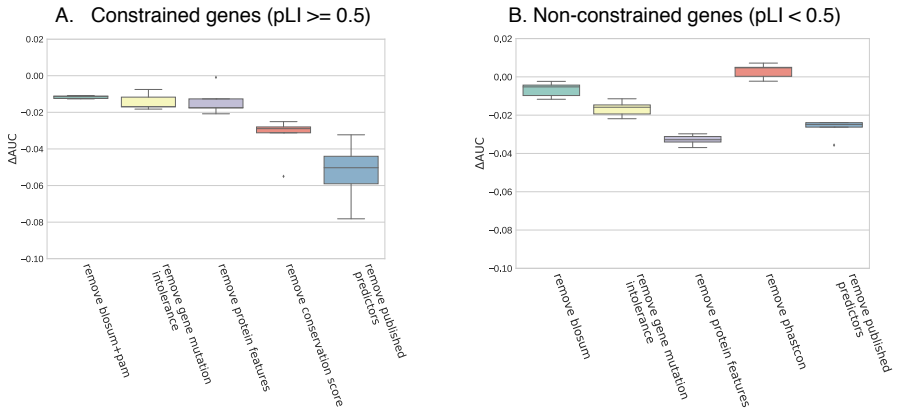


Figure 5.6 Measuring the contribution of features to MVP prediction performance in cancer mutation hotspots data. Performance contribution is measured by AUC reduction (ΔAUC) from excluding a group of features. Since features within a group are often highly correlated, we did measure the contribution of an entire group instead of individual features in the group. (A) Constrained genes; (B) Non-constrained genes. Error bar is estimated by subsampling of large number of negatives.

5.2.4 Comparing MVP to published methods in *de novo* mutation data

To test the utility in real genetic studies, we obtained germline *de novo* missense variants (DNMs) from 2645 cases in a congenital heart disease (CHD) study[71], 3953 cases in autism spectrum disorder (ASD) studies[6,39,71], and DNMs from 1911 controls (unaffected siblings) in Simons Simplex Collection[6,39,71]. Since genes with cancer mutation hotspots are relatively well studied in both constrained and non-constrained gene sets, assessment using *de novo* mutations can provide additional insight with less bias (Table 5.3).

Table 5.3 Number and percentage of genes in testing datasets that are overlapped with genes used in training.

	constrained genes			non-constrained genes		
dataset	number of genes	number of genes overlapped with training genes	percentage of genes in training	number of genes	number of genes overlapped with training genes	percentage of genes in training
varibench positive	149	112	0.75	310	227	0.73
cancer hotspot	149	67	0.45	55	21	0.38

positive						
ASD *de novo*	874	151	0.17	1359	183	0.13
CHD *de novo*	640	103	0.16	1007	139	0.14
SSC control *de novo*	372	59	0.16	641	77	0.12

Because the true pathogenicity of most of the *de novo* mutations is unknown, we cannot directly evaluate the performance of prediction methods. To address this issue, we calculated the enrichment rate of predicted pathogenic DNMs by a method with a certain threshold in the cases compared to the controls, and then estimated precision and the number of true risk variants (Methods), which is a proxy of recall since the total number of true positives in all cases is a (unknown) constant independent of methods. We compared the performance of MVP to other methods by estimated precision and recall-proxy (Fig. 5.7). Based on the optimal thresholds of MVP in cancer hotspot ROC curves, we used a score of 0.7 in constrained genes and 0.75 in non-constrained genes to define pathogenic DNMs. In constrained genes, we observed an enrichment of 2.2 in CHD and an enrichment of 1.9 in ASD, achieving estimated precision of 0.55 and 0.47 (Fig. 5.7A and 5.7D), respectively. This indicates that about 50% of the MVP-predicted pathogenic DNMs contribute to the diseases. In non-constrained genes, we observed an enrichment of 1.9 in CHD and 1.4 in ASD, respectively, and 0.32 and 0.28 in estimated precision (Fig. 5.7B and 5.7E). In all genes combined, MVP achieved an estimated precision of 40% for both CHD and ASD (Fig. 5.7C and 5.7F). The next best methods reached 25% (M-CAP) and 20% (MPC[135] and REVEL) given the same recall-proxy for CHD and ASD, respectively.

Furthermore, the estimated precision of MVP with DNMs at optimal threshold is much closer to the expected precision based on ROC of cancer hotspots data than the value from VariBench data, supporting that there is less performance inflation in testing using cancer data.

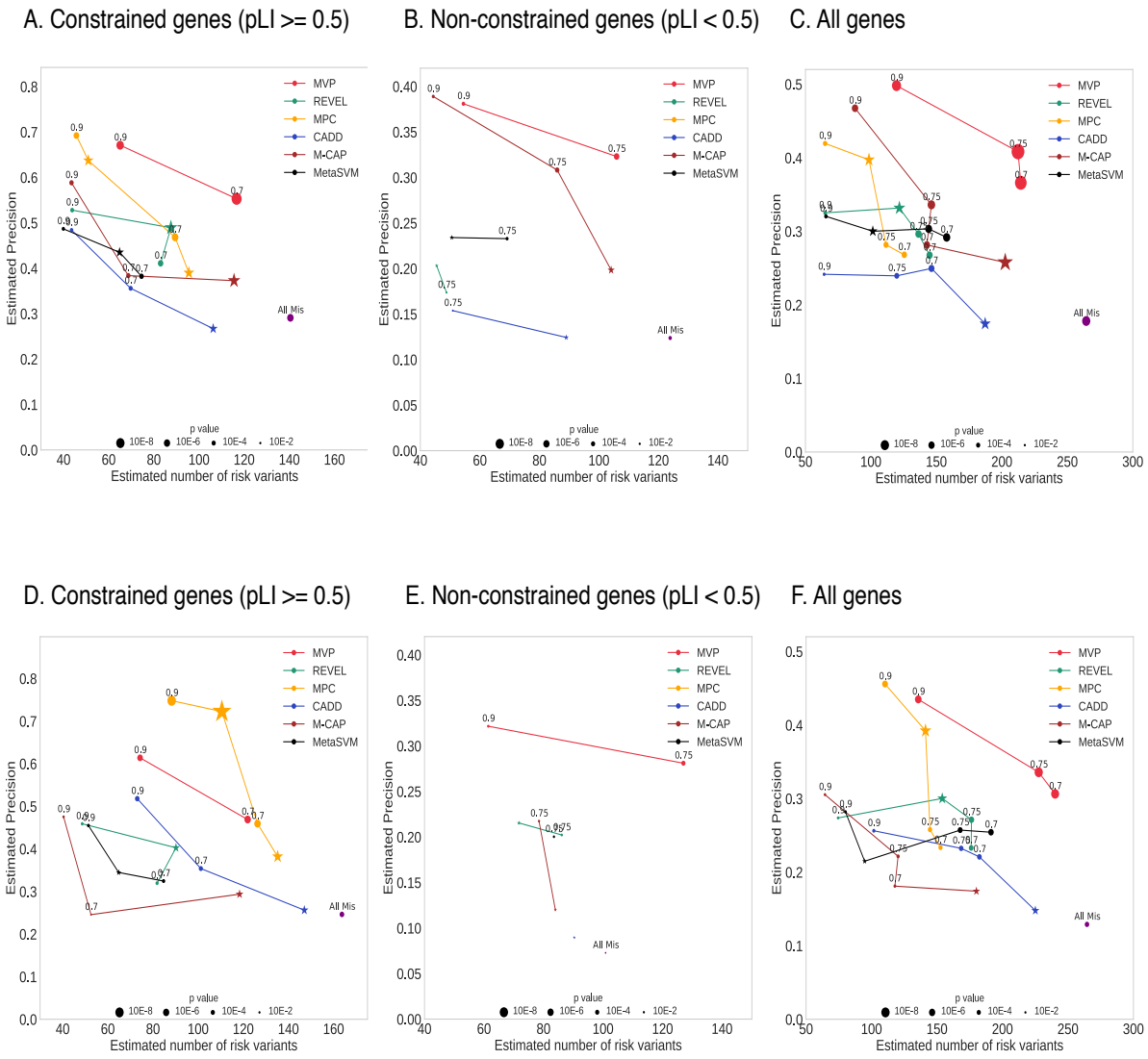


Figure 5.7. **Comparison of MVP and previously published methods using *de novo* missense mutations from CHD and ASD studies by precision-recall-proxy curves**. Numbers on each point indicate rank percentile thresholds, star points indicate thresholds recommended by publications. The positions of "All Mis" points are estimated from all missense variants in the gene set without using any pathogenicity prediction method. The point size is proportional to –log (p-value). P-value is calculated by binomial test, only points with p value less than 0.05 are shown. (A, B, C) Performance in CHD DNMs in constrained genes, non-constrained genes, and all genes, respectively. (D, E, F) Performance in ASD DNMs in constrained genes, non-constrained genes, and all genes, respectively.

Previous studies have estimated that deleterious *de novo* coding mutations, including loss of function variants and damaging missense variants, have a small contribution to isolated CHD[71]. Here, we used MVP to revisit this question. With the definition of damaging DNMs in Jin et al 2017[71] (based on metaSVM[30]), the estimated contribution of deleterious *de novo* coding mutations to isolated CHD is about 4.3%. With MVP score of 0.75, the estimation is 7.8%(95% CI = [6.5%, 9.1%]), nearly doubling the previous estimate.

5.3 Discussion

We developed a new method, MVP, to predict pathogenicity of missense variants. MVP is based on residual neural networks, a supervised deep learning approach, and was trained using a large number of curated pathogenic variants from clinical databases, separately on constrained genes and non- constrained genes. Using cancer mutation hotspots and *de novo* mutations from CHD and ASD, we showed that MVP achieved overall better performance than published methods, especially in non-constrained genes. Nevertheless, the fraction of pathogenic variants among *de novo* missense variants in non-constrained genes is low in both CHD and ASD, leading to relatively poor performance by all methods. MVP achieved substantially better performance than other methods in these genes, partly attributed to inclusion of protein structure-based predictors (Figure 5.6B).

Further improvement in protein structure prediction and the utilization of protein structure in the model[93] would be the key to improve MVP. Finally, all methods are limited by the size and the potentially high false positive rate of the training data. Systematic efforts such as ClinVar[140] will eventually produce better training data to improve prediction performance.

5.4 Material and methods

5.4.1 Training and testing data sets

Training data sets: We compiled 22,390 missense mutations from Human Gene Mutation Database Pro version 2013 (HGMD)[141] database under the disease mutation (DM) category, 12,875 deleterious variants from UniProt[30,142], and 4,424 pathogenic variants from ClinVar database[140] as true positive (TP). In total, there are 32,074 unique positive training variants. The negative training sets include 5,190 neutral variants from Uniprot[30,142], randomly selected 42,415 rare variants from DiscovEHR database[139], and 39,593 observed human-derived variants[28]. In total, there are 86,620 unique negative training variants (Table 5.2).

Testing data sets: We have three categories of testing data sets (Table 5.2). The three categories are: (a) Benchmark data sets from VariBench [30,137] as positives and randomly selected rare variants from DiscovEHR database[139] as negatives; (b) cancer somatic missense mutations located in hotspots from recent study[138] as positives and randomly selected rare variants from DiscovEHR database[139] as negatives; (c) and *de novo* missense mutation data sets from recent published exome-sequencing studies[6,39,71]. All variants in (a) and (b) that overlap with training data sets were excluded from testing. We tested the performance in constrained genes (ExAC pLI $\geq$ 0.5) and non-constrained gene (ExAC pLI < 0.5)[67] separately.

To focus on rare variants with large effect, we selected ultra-rare variants with MAF $<10^{-4}$ based on gnomAD database to filter variants in both training and testing data sets.

We applied additional filter of MAF < 10^{-6} for variants in constrained genes in both cases and controls for comparison based on a recent study [135,143].

5.4.2 Features and architecture used in MVP deep learning model

MVP uses many correlated features as predictors. There are six categories: (1) local context: GC content within 10 flanking bases on the reference genome; (2) amino acid constraint, including blosum62[144] and pam250[145]; (3) conservation scores, including phyloP 20way mammalian and 100way vertebrate[146], GERP++[147], SiPhy 29way[148], and phastCons 20way mammalian and 100way vertebrate[149]; (4) Protein structure, interaction, and modifications, including predicted secondary structures[150], number of protein interactions from the BioPlex 2.0 Network[151], whether the protein is involved in complexes formation from CORUM database[152], number of high-confidence interacting proteins by PrePPI [153], probability of a residue being located the interaction interface by PrePPI (based on PPISP, PINUP, PredU), predicted accessible surface areas were obtained from dbPTM[154], SUMO scores in 7-amino acids neighborhood by GPS-SUMO [155], phosphorylation sites predictions within 7 amino acids neighborhood by GPS3.0[156], and ubiquitination scores within 14-amino acids neighborhood by UbiProber [157]; (5) Gene mutation intolerance, including ExAC metrics[67] (pLI, pRec, lof_z) designed to measure gene dosage sensitivity or haploinsufficiency, RVIS[158], probability of causing diseases under a dominant model "domino"[159], average selection coefficient of loss of function variants in a gene "s_het" [160], and sub-genic regional depletion of missense variants [135]; (6) Selected deleterious or pathogenicity scores by previous published methods obtained through dbNSFPv3.3a[161], including Eigen[162], VEST3[29], MutationTaster[163], PolyPhen2 [164], SIFT [165], PROVEAN[166], fathmm-MKL [167], FATHMM [167], MutationAssessor[168], and LRT[169].

For consistency, we used canonical transcripts to define all possible missense variants[135]. Missing values of protein complex scores are filled with 0 and other features are filled with -1.

Since pathogenic variants in constrained genes and non-constrained genes may have different mode of action, we trained our models on constrained and non-constrained variants separately with different sets of features (38 features used in constrained model, 21 features used in non-constrained model).

MVP is based on a deep residual neural network model (ResNet)[136] for predicting pathogenicity using the predictors described above. To preserve the structured features in training data, we ordered the features according to their correlations (Fig. 5.2). The model (Fig. 5.1) takes a vector of the ordered features as input, followed by a convolutional layer of 32 kernels with size 3 x 1 and stride of 1, then followed by 2 computational residual units, each consisting of 2 convolutional layers of 32 kernels with size 3 x 1 and stride of 1 and a ReLU[170] activation layer in between. The output layer and input layer of the residual unit is summed and passed on to a ReLU activation layer. After the two convolutional layers with residual connections, 2 fully connected layers of 320 x 512 and 512 x1 are used followed by a sigmoid function to generate the final output[171].

$$Sigmod(x) = \frac{1}{1 + e^{-x}}$$

In training, we randomly partitioned the synthetic training data sets into two parts, 80% of the total training sets for training and 20% for validation. We trained the model

with batch size of 64, used adam[172] optimizer to perform stochastic gradient descent[173] with logarithmic loss between the predicted value and true value. After one full training cycle on the training set, we applied the latest model weights on validation data to compute validation loss.

To avoid over fitting, we used early stopping regularization during training. We computed the loss in training data and validation data after each training cycle and stopped the process when validation loss is comparable to training loss and do not decrease after 5 more training cycle, and then we set the model weights using the last set with the lowest validation loss. We applied the same model weights on testing data to obtain MVP scores for further analysis.

5.4.3 Previously published methods for comparison

We compared MVP score to 13 previously published prediction scores, namely, M-CAP[31], DANN[174], Eigen[162], Polyphen2[164], SIFT[165], MutationTaster[163], FATHMM[167], REVEL[32], CADD[28], metaSVM[30], metaLR[30], VEST3[29], and MPC[135].

5.4.4 Normalization of scores using rank percentile

For each method, we first obtained predicted scores of all possible rare missense variants in canonical transcripts, and then sort the scores and converted the scores into rank percentile. Higher rank percentile indicates more damaging, e.g., a rank score of 0.75 indicates the missense variant is more likely to be pathogenic than 75% of all possible missense variants.

5.4.5 ROC curves and optimal points estimation

We plotted Receiver operating characteristic (ROC) curves and calculated Area Under the Curve [97] values in training data with 6-fold cross validation and compared MVP performance with other prediction scores in curated benchmark testing datasets (Fig. 5.3) and cancer hotspot mutation dataset (Fig. 5.4). For each prediction method, we varied the threshold for calling pathogenic mutations in a certain range and computed the corresponding sensitivity and specificity based on true positive, false positive, false negative and true negative predictions. ROC curve was then generated by plotting sensitivity against 1 – specificity at each threshold.

We define the optimal threshold for MVP score as the threshold where the corresponding point in ROC curve has the largest distance to the diagonal line (Fig. 5.4). Based on the true positive rate and false positive rate at the optimal points in ROC curves, we can estimate the precision and recall in *de novo* precision-recall-proxy curves.

5.4.6 Precision-recall-proxy curves

Since *de novo* mutation data do not have ground truth, we used the excess of predicted pathogenic missense *de novo* variants in cases compared to controls to estimate precision and proxy of recall. For various thresholds of different scores, we can calculate the estimated number of risk variants and estimated precision based on enrichment of predicted damaging variants in cases compared to controls. We adjusted the number of missense *de novo* mutation in controls by the synonymous rate ratio in cases verses controls, assuming the average number of synonymous

as the data sets were sequenced and processed separately), which partly reduced the signal but ensures that our results were not inflated by the technical difference in data processing.

Table 5.4. Comparison of cases and controls in rate of synonymous *de novo* variants

	Number of synonymous variants	Rate per cases compared to controls
Autism spectrum disorder (ASD)	1026	1.027
Congenital heart disease (CHD)	701	1.049
Simons Simplex Collection unaffected siblings (controls)	483	N/A

Denote the number of cases and controls as N_1 and N_0, respectively; the number of predicted pathogenic *de novo* missense variants as M_1 and M_0, in cases and controls, respectively; the rate of synonymous *de novo* variants as S_1 and S_0, in cases and controls, respectively; technical adjustment rate as α; and the enrichment rate of variants in cases compared to controls as R.

We first estimate α by:

$$\alpha = \frac{S_1}{S_0}$$

Then assuming the rate of synonymous *de novo* variants in cases and controls should be identical if there is no technical batch effect, we use α to adjust estimated enrichment of pathogenic *de novo* variants in cases compared to the controls by:

$$R = \frac{\frac{M_1}{N_1}}{\frac{M_0}{N_0} \times \alpha}$$

Then we can estimate number of true pathogenic variants (M_1') by:

$$M_1' = \frac{M_1(R-1)}{R}$$

And then precision by:

$$\widehat{Precision} = \frac{M_1'}{M_1}$$

5.4.7 Estimation of precision for a method at a certain threshold based on ROC curves

Denote the number of all true positives (pathogenic variants in cases) in a *de novo* mutation data set as $\boldsymbol{P}$, the estimated number of true positive detected by all methods at any threshold (including estimation from "all missense" without prediction methods) as a set $\boldsymbol{\mathcal{P}}$, the number of all negatives (non-pathogenic variants in cases) in the *de novo* mutation data as $\boldsymbol{N}$, the number of true positives by a method at a threshold as *TP*, the number of false positives by a method at a threshold as *FP*, and the baseline precision as $\boldsymbol{B}$, defined as:

$$B \equiv \frac{P}{P+N}$$

P+*N* is just the total number of *de novo* mutations in cases. We can estimate *B* by:

$$\hat{B} = \frac{\max(\boldsymbol{\mathcal{P}})}{P+N}$$

Therefore, N/P can be estimated as:

$$\frac{N}{P} = \frac{1}{1/\hat{B} - 1}$$

From the ROC curve, denote true positive rate (which is also called *recall* or *sensitivity*) as *TPR*, and false positive rate as *FPR*. We obtain *FPR* and *TPR* for a method at a certain threshold from cancer or VariBench ROC curves, and then use them to estimate number of true and false positives:

$$\widehat{TP} = P \cdot TPR$$

$$\widehat{FP} = N \cdot FPR$$

Therefore, the estimated precision of a method at a threshold based on ROC curve is:

$$\widehat{Precision} = \frac{\widehat{TP}}{\widehat{TP} + \widehat{FP}} = \frac{1}{1 + \frac{\widehat{FP}}{\widehat{TP}}} = \frac{1}{1 + \frac{FPR}{TPR} \cdot \frac{N}{P}} = \frac{1}{1 + \frac{FPR}{TPR} * (\frac{1}{\hat{B}} - 1)}$$

Chapter 6

Conclusions and future work

The primary goal of this thesis was to develop novel computational methods to analyze genomic data and facilitate gene discovery in genetic studies. By establishing new methods of statistical genetics and integrating biological domain knowledge, we can identify a group of genes important in the etiology and improve the understanding of the genetic architecture of developmental disorder. A better understanding of genetic variations, especially the rare inherited or *de novo* coding mutations in patients, can help identify the genetic causes of diseases and guide clinic decisions to provide better treatment.

Advances in sequencing technology have enabled the ability to identify major causes of severe developmental disorders. Large-scale whole exome sequencing as well as whole genome sequencing experiments are performed in congenital diaphragmatic hernia (CDH) patients. In this thesis, we combined samples from Boston Children's Hospital, Massachusetts General Hospital and DHREAMS study to maximize our power to identify the pathogenic mutations in sporadic CDH cases. The combined cohort is one of the largest and most well characterized cohorts of patients with CDH in the world. In previous studies, a significant burden of damaging *de novo* coding variants was identified in CDH patients, we replicated this finding in our cohort and showed that female patients carried almost the entire burden in isolated CDH while females and males carried similar burden in complex CDH. Additionally complex CDH cases carried excess of *de novo* LGD variants mostly in genes highly expressed in developing diaphragm while isolated CDH cases had a broad range of gene expression levels for *de novo* LGD variants distribution. We also identified *MYRF* as a new candidate risk gene for CDH. *MYRF* is a transcription factor with high probability of mutations intolerance. All patients have

additional anomalies including congenital heart defects and genitourinary defects with *MYRF* mutations, this phenomenon is likely representing a novel syndrome; functional genomics studies using cell and animal models of *MYRF* are currently in progress.

International efforts in cancer genomic studies have produced rich cancer somatic mutation data set and a great opportunity to investigate the functionality of cancer driver genes at variant level. Dysregulation of fundamental cellular processes, such as cell generation, cell division and growth, and cell differentiation, can cause cancer and developmental disorders. Integration of data sets from cancer can improve the interpretation of genetic data in developmental disorders studies. In this thesis, we studied the deep genetic relations between cancer and developmental disorders and gave a quantitative assessment of the shared mode of action of mutations and the total number of overlapping risk genes among cancer driver genes. Based on these results, we can leverage massive amounts of somatic mutation data in cancer studies to improve our understanding of variants conferring developmental disorders.

To improve power in genetic studies and accurate interpretation of missense variants in clinical genetic testing, we developed MVP, a deep neural network based method, in order to achieve better prediction of the pathogenicity of missense variants. MVP uses a deep learning approach to leverage large training data sets and many correlated predictors. We compiled cancer somatic mutation hotspots dataset and *de novo* germline mutations from developmental disorders as benchmark data, MVP achieved overall superior performance in identifying and prioritizing pathogenic missense variants than previously published methods.

Recent whole exome sequencing studies have been designed to assess the relative impact of inherited variants on developmental disorders[71,175,176]. In recent studies of congenital heart disease (CHD), researcher identified an enrichment of rare inherited heterozygous loss of functions variants only in isolated CHD patients but not syndromic CHD patients, an aggregation analysis of *de novo* and rare inherited variants can reveal the distinct disease causation and genetic architectures for syndromic and isolated CHD. However, inherited variants are unlikely to be completely penetrant and each individual variant with small effect sizes contributes a little to the phenotype. Prioritization of inherited variants remains a serious challenge and large number of patients collections are required to provide statistical power for detecting risk variants with small effect sizes, continued efforts in large cohorts sequencing will contribute to a more complete picture of the pathogenesis of disease. Additionally, rare inherited dominant and recessive mutations in CDH patients has not been systematically studied due to low prevalence; further discovery of effect inherited mutations with impact requires a more comprehensive analytical strategies and larger cohorts

In this thesis, we only focused on the analysis and interpretation of mutations located in the coding region and previously researcher has estimated that coding pathogenic *de novo* mutations[13] can contribute to about half of the parents with severe developmental disorders. Many studies have been performance to understand the effects of the remaining 98% non coding DNA. Numerous GWAS studies[177,178] have greatly improved our understanding in human diseases mostly from common noncoding variants, ENCODE project[179,180] aimed to produce high-quality data to understand how noncoding DNA contribute to gene expression regulation. Unfortunately it is still largely unknown which rare noncoding variants contribute to disease with large effect, as it is much harder

to assess the effects and interpret the DNA changes in noncoding regions than coding regions currently. Noncoding variants can be functional in many ways, such as disrupting DNA sequence motifs in enhancer or promoter regions, disrupting mRNA binding specificity and changing DNA accessibility[179,181,182]. Whole genome sequencing technology become affordable for large scale studies in recent years and provided an unprecedented opportunity to study the contribution of rare inherited mutations or *de novo* mutations in the noncoding region, it can interrogate more of the noncoding genome than whole exome sequencing with detection of a broader genetic variation types, including not only single-nucleotide variant (snv), small insertion and deletions (indels), but also structural variants such as copy number variants (CNV) as well as large translocations[183]. With a large cohorts and robust statistical methods, we can identify disease-associated regulatory elements and reveal the contributions of mutations in noncoding regions to developmental disorders.

References

1. Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S. & Robinson, G.E. Big data: astronomical or genomical? *PLoS biology* **13**, e1002195 (2015).
2. Lander, E.S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187 (2011).
3. Veltman, J.A. & Brunner, H.G. De novo mutations in human genetic disease. *Nature Reviews Genetics* **13**, 565 (2012).
4. Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. A vision for the future of genomics research. *Nature* **422**, 835 (2003).
5. Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J.D., Romano-Adesman, A., Bjornson, R.D., Breitbart, R.E. & Brown, K.K. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223 (2013).
6. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M. & Walker, S. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209 (2014).
7. Yu, L., Sawle, A.D., Wynn, J., Aspelund, G., Stolar, C.J., Arkovitz, M.S., Potoka, D., Azarow, K.S., Mychaliska, G.B. & Shen, Y. Increased burden of de novo predicted deleterious variants in complex congenital diaphragmatic hernia. *Human molecular genetics* **24**, 4764-4773 (2015).
8. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics* **33**, 228 (2003).
9. MacArthur, D., Manolio, T., Dimmock, D., Rehm, H., Shendure, J., Abecasis, G., Adams, D., Altman, R., Antonarakis, S. & Ashley, E. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469 (2014).
10. Consortium, I.S. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748 (2009).
11. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R. & Chakravarti, A. Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009).
12. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., Van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M. & Bayzetinova, T. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet* **385**, 1305-1314 (2015).
13. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N. & Alvi, M. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433 (2017).
14. Van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in genetics* **30**, 418-426 (2014).
15. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics* **38**, 95-109 (2011).

16. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. Comparison of next-generation sequencing systems. *BioMed Research International* **2012**(2012).
17. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S. & Sanjad, S. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 19096-19101 (2009).
18. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N. & Stein, J.L. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237 (2012).
19. Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J. & Snyder, M. Performance comparison of exome DNA sequencing technologies. *Nature biotechnology* **29**, 908 (2011).
20. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C. & Owen, M.J. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics* **91**, 597-607 (2012).
21. Cook Jr, E.H. & Scherer, S.W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919 (2008).
22. Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R. & Hannig, V. A copy number variation morbidity map of developmental delay. *Nature genetics* **43**, 838 (2011).
23. Micale, L., Augello, B., Maffeo, C., Selicorni, A., Zucchetti, F., Fusco, C., De Nittis, P., Pellico, M.T., Mandriani, B. & Fischetto, R. Molecular analysis, pathogenic mechanisms, and readthrough therapy on a large cohort of Kabuki syndrome patients. *Human mutation* **35**, 841-850 (2014).
24. Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B. & Han, Y. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217 (2013).
25. Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., Lyon, E. & Ward, B.E. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in Medicine* **10**, 294 (2008).
26. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E. & Spector, E. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine* **17**, 405 (2015).
27. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. & Lander, E.S. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455-E464 (2014).
28. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315 (2014).

29. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC genomics* **14**, S3 (2013).
30. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. & Liu, X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125-2137 (2014).
31. Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper, D.N., Bernstein, J.A. & Bejerano, G. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics* **48**, 1581-1586 (2016).
32. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E. & Karyadi, D. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
33. Chandrasekharan, P.K., Rawat, M., Madappa, R., Rothstein, D.H. & Lakshminrusimha, S. Congenital diaphragmatic hernia–A review. *Maternal health, neonatology and perinatology* **3**, 6 (2017).
34. Adam, M., Ardinger, H., Pagon, R., Wallace, S., Bean, L., Mefford, H., Stephens, K., Amemiya, A. & Ledbetter, N. Congenital Diaphragmatic Hernia Overview--GeneReviews®.
35. Pober, B.R. Overview of epidemiology, genetics, birth defects, and chromosome abnormalities associated with CDH. in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* Vol. 145 158-171 (Wiley Online Library, 2007).
36. Danzer, E., Gerdes, M., Bernbaum, J., D'Agostino, J., Bebbington, M.W., Siegle, J., Hoffman, C., Rintoul, N.E., Flake, A.W. & Adzick, N.S. Neurodevelopmental outcome of infants with congenital diaphragmatic hernia prospectively enrolled in an interdisciplinary follow-up program. *Journal of pediatric surgery* **45**, 1759-1766 (2010).
37. Holder, A., Klaassens, M., Tibboel, D., de Klein, A., Lee, B. & Scott, D. Genetic factors in congenital diaphragmatic hernia. *The American Journal of Human Genetics* **80**, 825-845 (2007).
38. Wynn, J., Yu, L. & Chung, W.K. Genetic causes of congenital diaphragmatic hernia. in *Seminars in Fetal and Neonatal Medicine* Vol. 19 324-330 (Elsevier, 2014).
39. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L. & Patterson, K.E. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
40. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S. & Kirby, A. A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944 (2014).
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
42. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. & Daly, M. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
43. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164 (2010).

44. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E. & Cummings, B.B. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research* **45**, D840-D845 (2016).
45. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. & Mesirov, J.P. Integrative genomics viewer. *Nature biotechnology* **29**, 24 (2011).
46. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences* **74**, 5463-5467 (1977).
47. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H. & Gorham, J. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015).
48. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A. & Jonasdottir, A. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471 (2012).
49. Neale, B.M., Kou, Y., Liu, L., Ma'Ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S. & Makarov, V. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242 (2012).
50. Graziano, J.N. Cardiac anomalies in patients with congenital diaphragmatic hernia and their prognosis: a report from the Congenital Diaphragmatic Hernia Study Group. *Journal of pediatric surgery* **40**, 1045-1050 (2005).
51. Skari, H., Bjornland, K., Haugen, G., Egeland, T. & Emblem, R. Congenital diaphragmatic hernia: a meta-analysis of mortality factors. *Journal of pediatric surgery* **35**, 1187-1197 (2000).
52. Zalla, J.M., Stoddard, G.J. & Yoder, B.A. Improved mortality rate for congenital diaphragmatic hernia in the modern era of management: 15 year experience in a single institution. *Journal of pediatric surgery* **50**, 524-527 (2015).
53. Leeuwen, L., Mous, D.S., van Rosmalen, J., Olieman, J.F., Andriessen, L., Gischler, S.J., Joosten, K.F., Wijnen, R.M., Tibboel, D. & IJsselstijn, H. Congenital Diaphragmatic Hernia and Growth to 12 Years. *Pediatrics*, e20163659 (2017).
54. Ackerman, K.G. & Greer, J.J. Development of the diaphragm and genetic mouse models of diaphragmatic defects. in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* Vol. 145 109-116 (Wiley Online Library, 2007).
55. Merrell, A.J., Ellis, B.J., Fox, Z.D., Lawson, J.A., Weiss, J.A. & Kardon, G. Muscle connective tissue controls development of the diaphragm and is a source of congenital diaphragmatic hernias. *Nature genetics* **47**, 496 (2015).
56. Jay, P.Y., Bielinska, M., Erlich, J.M., Mannisto, S., Pu, W.T., Heikinheimo, M. & Wilson, D.B. Impaired mesenchymal cell function in Gata4 mutant mice leads to diaphragmatic hernias and primary lung defects. *Developmental biology* **301**, 602-614 (2007).
57. Castiglia, L., Fichera, M., Romano, C., Galesi, O., Grillo, L., Sturnio, M. & Failla, P. Narrowing the candidate region for congenital diaphragmatic hernia in chromosome 15q26: contradictory results. *American journal of human genetics* **77**, 892 (2005).
58. Klaassens, M., van Dooren, M., Eussen, H., Douben, H., Den Dekker, A., Lee, C., Donahoe, P., Galjaard, R.-J., Goemaere, N. & De Krijger, R. Congenital diaphragmatic hernia and chromosome 15q26: determination of a candidate region by use of fluorescent

in situ hybridization and array-based comparative genomic hybridization. *The American Journal of Human Genetics* **76**, 877-882 (2005).
59. Shimokawa, O., Miyake, N., Yoshimura, T., Sosonkina, N., Harada, N., Mizuguchi, T., Kondoh, S., Kishino, T., Ohta, T. & Remco, V. Molecular characterization of del (8)(p23. 1p23. 1) in a case of congenital diaphragmatic hernia. *American Journal of Medical Genetics Part A* **136**, 49-51 (2005).
60. Wat, M.J., Shchelochkov, O.A., Holder, A.M., Breman, A.M., Dagli, A., Bacino, C., Scaglia, F., Zori, R.T., Cheung, S.W. & Scott, D.A. Chromosome 8p23. 1 deletions as a cause of complex congenital heart defects and diaphragmatic hernia. *American Journal of Medical Genetics Part A* **149**, 1661-1677 (2009).
61. You, L.-R., Takamoto, N., Yu, C.-T., Tanaka, T., Kodama, T., DeMayo, F.J., Tsai, S.Y. & Tsai, M.-J. Mouse lacking COUP-TFII as an animal model of Bochdalek-type congenital diaphragmatic hernia. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16351-16356 (2005).
62. Longoni, M., High, F.A., Qi, H., Joy, M.P., Hila, R., Coletti, C.M., Wynn, J., Loscertales, M., Shan, L. & Bult, C.J. Genome-wide enrichment of damaging de novo variants in patients with isolated and complex congenital diaphragmatic hernia. *Human genetics* **136**, 679-691 (2017).
63. Yu, L., Wynn, J., Ma, L., Guha, S., Mychaliska, G.B., Crombleholme, T.M., Azarow, K.S., Lim, F.Y., Chung, D.H. & Potoka, D. De novo copy number variants are associated with congenital diaphragmatic hernia. *Journal of medical genetics* **49**, 650-659 (2012).
64. Wynn, J., Aspelund, G., Zygmunt, A., Stolar, C.J., Mychaliska, G., Butcher, J., Lim, F.-Y., Gratton, T., Potoka, D. & Brennan, K. Developmental outcomes of children with congenital diaphragmatic hernia: a multicenter prospective study. *Journal of pediatric surgery* **48**, 1995-2004 (2013).
65. Hinton, C.F., Siffel, C., Correa, A. & Shapira, S.K. Survival Disparities Associated with Congenital Diaphragmatic Hernia. *Birth defects research* **109**, 816-823 (2017).
66. Ware, J.S., Samocha, K.E., Homsy, J. & Daly, M.J. Interpreting de novo variation in human disease using denovolyzeR. *Current Protocols in Human Genetics*, 7.25. 1-7.25. 15 (2015).
67. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J. & Cummings, B.B. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
68. Clugston, R.D., Zhang, W. & Greer, J.J. Gene expression in the developing diaphragm: significance for congenital diaphragmatic hernia. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **294**, L665-L675 (2008).
69. Russell, M.K., Longoni, M., Wells, J., Maalouf, F.I., Tracy, A.A., Loscertales, M., Ackerman, K.G., Pober, B.R., Lage, K. & Bult, C.J. Congenital diaphragmatic hernia candidate genes derived from embryonic transcriptomes. *Proceedings of the National Academy of Sciences* **109**, 2978-2983 (2012).
70. Carmona, R., Cañete, A., Cano, E., Ariza, L., Rojas, A. & Muñoz-Chápuli, R. Conditional deletion of WT1 in the septum transversum mesenchyme causes congenital diaphragmatic hernia in mice. *Elife* **5**(2016).
71. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W. & Sierant, M.C. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature genetics* **49**, ng. 3970 (2017).

72. Azarow, K.S., Cusick, R., Wynn, J., Chung, W., Mychaliska, G.B., Crombleholme, T.M., Chung, D.H., Lim, F.Y., Potoka, D. & Warner, B.W. The association between congenital diaphragmatic hernia and undescended testes. *Journal of pediatric surgery* **50**, 744-745 (2015).
73. Bujalka, H., Koenning, M., Jackson, S., Perreau, V.M., Pope, B., Hay, C.M., Mitew, S., Hill, A.F., Lu, Q.R. & Wegner, M. MYRF is a membrane-associated transcription factor that autoproteolytically cleaves to directly activate myelin genes. *PLoS biology* **11**, e1001625 (2013).
74. Kim, D., Choi, J.-o., Fan, C., Shearer, R.S., Sharif, M., Busch, P. & Park, Y. Homo-trimerization is essential for the transcription factor function of Myrf for oligodendrocyte differentiation. *Nucleic acids research* **45**, 5112-5125 (2017).
75. Hornig, J., Fröb, F., Vogl, M.R., Hermans-Borgmeyer, I., Tamm, E.R. & Wegner, M. The transcription factors Sox10 and Myrf define an essential regulatory network module in differentiating oligodendrocytes. *PLoS genetics* **9**, e1003907 (2013).
76. McKenzie, I.A., Ohayon, D., Li, H., De Faria, J.P., Emery, B., Tohyama, K. & Richardson, W.D. Motor skill learning requires active central myelination. *Science* **346**, 318-322 (2014).
77. Xiao, L., Ohayon, D., McKenzie, I.A., Sinclair-Wilson, A., Wright, J.L., Fudge, A.D., Emery, B., Li, H. & Richardson, W.D. Rapid production of new oligodendrocytes is required in the earliest stages of motor-skill learning. *Nature neuroscience* **19**, 1210 (2016).
78. Robinson, E.B., Lichtenstein, P., Anckarsäter, H., Happé, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proceedings of the National Academy of Sciences* **110**, 5258-5262 (2013).
79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
80. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904 (2006).
81. Consortium, I.H. The international HapMap project. *Nature* **426**, 789 (2003).
82. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I. & Daly, M.J. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).
83. Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B. & Olshen, A.B. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology* (2015).
84. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E. & Ruef, B.J. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome research* **19**, 1316-1323 (2009).
85. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A. & Schenck, A. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344 (2014).

86. Fitzgerald, T., Gerety, S., Jones, W., Van Kogelenberg, M., King, D., McRae, J., Morley, K., Parthiban, V., Al-Turki, S. & Ambridge, K. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223 (2015).
87. Waite, K.A. & Eng, C. From developmental disorder to heritable cancer: it's all in the BMP/TGF-β family. *Nature Reviews Genetics* **4**, 763-773 (2003).
88. Schubbert, S., Shannon, K. & Bollag, G. Hyperactive Ras in developmental disorders and cancer. *Nature Reviews Cancer* **7**, 295-308 (2007).
89. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S. & Geschwind, D.H. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013).
90. Yuen, R.K., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., Chrysler, C., Nalpathamkalam, T., Pellecchia, G. & Liu, Y. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature medicine* **21**, 185-191 (2015).
91. O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C. & Ankenman, K. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012).
92. Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Miliaresis, C., Rodgers, L. & McCombie, R. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *science* **275**, 1943-1947 (1997).
93. Tartaglia, M., Niemeyer, C.M., Fragale, A., Song, X., Buechner, J., Jung, A., Hählen, K., Hasle, H., Licht, J.D. & Gelb, B.D. Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nature genetics* **34**, 148 (2003).
94. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R. & Larsson, E. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1 (2013).
95. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. & Network, C.G.A.R. The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113 (2013).
96. De Ligt, J., Willemsen, M.H., Van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., De Vries, P. & Gilissen, C. Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine* **367**, 1921-1929 (2012).
97. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J. & Di Donato, N. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).
98. Hamdan, F.F., Srour, M., Capo-Chichi, J.-M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D. & Diallo, O. De novo mutations in moderate or severe intellectual disability. *PLoS genetics* **10**, e1004772 (2014).
99. Darbro, B.W., Singh, R., Zimmerman, M.B., Mahajan, V.B. & Bassuk, A.G. Autism linked to increased oncogene mutations but decreased cancer rate. *PloS one* **11**, e0149041 (2016).
100. Ronan, J.L., Wu, W. & Crabtree, G.R. From neural development to cognition: unexpected roles for chromatin. *Nature Reviews Genetics* **14**, 347 (2013).

101. Roelfsema, J.H., White, S.J., Ariyürek, Y., Bartholdi, D., Niedrist, D., Papadia, F., Bacino, C.A., den Dunnen, J.T., van Ommen, G.-J.B. & Breuning, M.H. Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease. *The American Journal of Human Genetics* **76**, 572-580 (2005).
102. Robinson, E.B., Samocha, K.E., Kosmicki, J.A., McGrath, L., Neale, B.M., Perlis, R.H. & Daly, M.J. Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proceedings of the National Academy of Sciences* **111**, 15161-15165 (2014).
103. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L. & Larsson, E. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. (AACR, 2012).
104. Abbott, K.L., Nyre, E.T., Abrahante, J., Ho, Y.-Y., Isaksson Vogel, R. & Starr, T.K. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic acids research* **43**, D844-D848 (2014).
105. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C. & Ward, S. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805-D811 (2014).
106. Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J. & Monaghan, K.G. Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine* **18**, 696 (2016).
107. Cheng, F., Zhao, J. & Zhao, Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in bioinformatics* **17**, 642-656 (2015).
108. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H. & Roberts, S.A. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214 (2013).
109. He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J. & Buxbaum, J.D. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics* **9**, e1003671 (2013).
110. Lobry, C., Oh, P. & Aifantis, I. Oncogenic and tumor suppressor functions of Notch in cancer: it's NOTCH what you think. *Journal of Experimental Medicine* **208**, 1931-1935 (2011).
111. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., van der Burgt, I., Crosby, A.H., Ion, A. & Jeffery, S. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nature genetics* **29**, 465 (2001).
112. Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J. & Elledge, S.J. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962 (2013).
113. Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B. & Olshen, A.B. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology* **34**, 155 (2016).

114. Yang, F., Petsalaki, E., Rolland, T., Hill, D.E., Vidal, M. & Roth, F.P. Protein domain-level landscape of cancer-type-specific somatic mutations. *PLOS Comput Biol* **11**, e1004147 (2015).
115. Rodriguez-Viciana, P., Tetsu, O., Tidyman, W.E., Estep, A.L., Conger, B.A., Santa Cruz, M., McCormick, F. & Rauen, K.A. Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. *Science* **311**, 1287-1290 (2006).
116. Jones, S., Wang, T.-L., Shih, I.-M., Mao, T.-L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L.A. & Vogelstein, B. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231 (2010).
117. De Keersmaecker, K., Atak, Z.K., Li, N., Vicente, C., Patchett, S., Girardi, T., Gianfelici, V., Geerdens, E., Clappier, E. & Porcu, M. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nature genetics* **45**, 186 (2013).
118. Morin, P.J. β-catenin signaling and cancer. *Bioessays* **21**, 1021-1030 (1999).
119. Tucci, V., Kleefstra, T., Hardy, A., Heise, I., Maggi, S., Willemsen, M.H., Hilton, H., Esapa, C., Simon, M. & Buenavista, M.-T. Dominant β-catenin mutations cause intellectual disability with recognizable syndromic features. *The Journal of clinical investigation* **124**, 1468-1482 (2014).
120. Medina, P.P., Romero, O.A., Kohno, T., Montuenga, L.M., Pio, R., Yokota, J. & Sanchez-Cespedes, M. Frequent BRG1/SMARCA4–inactivating mutations in human lung cancer cell lines. *Human mutation* **29**, 617-622 (2008).
121. Vandeweyer, G., Helsmoortel, C., Van Dijck, A., Vulto-van Silfhout, A.T., Coe, B.P., Bernier, R., Gerdts, J., Rooms, L., Van Den Ende, J. & Bakshi, M. The transcriptional regulator ADNP links the BAF (SWI/SNF) complexes with autism. in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* Vol. 166 315-326 (Wiley Online Library, 2014).
122. Tsurusaki, Y., Okamoto, N., Ohashi, H., Kosho, T., Imai, Y., Hibi-Ko, Y., Kaname, T., Naritomi, K., Kawame, H. & Wakui, K. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nature genetics* **44**, 376 (2012).
123. Shang, L., Cho, M.T., Retterer, K., Folk, L., Humberson, J., Rohena, L., Sidhu, A., Saliganan, S., Iglesias, A. & Vitazka, P. Mutations in ARID2 are associated with intellectual disabilities. *Neurogenetics* **16**, 307-314 (2015).
124. Albert, T.K., Lemaire, M., van Berkum, N.L., Gentz, R., Collart, M.A. & Timmers, H.T.M. Isolation and characterization of human orthologs of yeast CCR4–NOT complex subunits. *Nucleic acids research* **28**, 809-817 (2000).
125. Chen, J., Chiang, Y.C. & Denis, C.L. CCR4, a 3′–5′ poly (A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. *The EMBO journal* **21**, 1414-1426 (2002).
126. Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S. & Getz, G. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences* **112**, E5486-E5495 (2015).
127. Miller, M.S., Rialdi, A., Ho, J.S.Y., Tilove, M., Martinez-Gil, L., Moshkina, N.P., Peralta, Z., Noel, J., Melegari, C. & Maestre, A.M. Senataxin suppresses the antiviral

transcriptional response and controls viral biogenesis. *Nature immunology* **16**, 485-494 (2015).
128. Hanahan, D. & Weinberg, R. Hallmarks of cancer: the next generation. cell 144 (5): 646–674. *CAS PubMed Article* (2011).
129. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C.C. p53 mutations in human cancers. *Science* **253**, 49-53 (1991).
130. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A. & Børresen-Dale, A.-L. Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013).
131. Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L. & Saksena, G. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature genetics* **45**, 970 (2013).
132. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*, (MIT press Cambridge, 2016).
133. Dorschner, M.O., Amendola, L.M., Turner, E.H., Robertson, P.D., Shirts, B.H., Gallego, C.J., Bennett, R.L., Jones, K.L., Tokita, M.J. & Bennett, J.T. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *The American Journal of Human Genetics* **93**, 631-640 (2013).
134. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* **6**, e1001154 (2010).
135. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M. & Daly, M.J. Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353 (2017).
136. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770-778 (2016).
137. Nair, P.S. & Vihinen, M. VariBench: a benchmark database for variations. *Human mutation* **34**, 42-49 (2013).
138. Chang, M.T., Bhattarai, T.S., Schram, A.M., Bielski, C.M., Donoghue, M.T., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C. & Penson, A. Accelerating discovery of functional mutant alleles in cancer. *Cancer discovery* (2017).
139. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C.V., Staples, J. & Gonzaga-Jauregui, C. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
140. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D. & Hoover, J. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862-D868 (2015).
141. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D. & Cooper, D.N. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, 1-13 (2017).
142. Consortium, U. Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* **39**, D214-D219 (2011).
143. Kosmicki, J.A., Samocha, K.E., Howrigan, D.P., Sanders, S.J., Slowikowski, K., Lek, M., Karczewski, K.J., Cutler, D.J., Devlin, B. & Roeder, K. Refining the role of de novo

protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nature genetics* **49**, 504 (2017).
144. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915-10919 (1992).
145. Dayhoff, M.O. A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**, 89-99 (1972).
146. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20**, 110-121 (2010).
147. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. & Sidow, A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**, 901-913 (2005).
148. Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N. & Xie, X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-i62 (2009).
149. Hubisz, M.J., Pollard, K.S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in bioinformatics* **12**, 41-51 (2010).
150. McGuffin, L.J., Bryson, K. & Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405 (2000).
151. Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P. & Parzen, H. Architecture of the human interactome defines protein communities and disease networks. *Nature* (2017).
152. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. & Mewes, H.-W. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids research* **38**, D497-D501 (2009).
153. Zhang, Q.C., Petrey, D., Garzon, J.I., Deng, L. & Honig, B. PrePPI: a structure-informed database of protein–protein interactions. *Nucleic acids research* **41**, D828-D833 (2012).
154. Lee, T.-Y., Huang, H.-D., Hung, J.-H., Huang, H.-Y., Yang, Y.-S. & Wang, T.-H. dbPTM: an information repository of protein post-translational modification. *Nucleic acids research* **34**, D622-D627 (2006).
155. Zhao, Q., Xie, Y., Zheng, Y., Jiang, S., Liu, W., Mu, W., Liu, Z., Zhao, Y., Xue, Y. & Ren, J. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic acids research* **42**, W325-W330 (2014).
156. Xue, Y., Liu, Z., Cao, J., Ma, Q., Gao, X., Wang, Q., Jin, C., Zhou, Y., Wen, L. & Ren, J. GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Engineering, Design & Selection* **24**, 255-260 (2010).
157. Chen, X., Qiu, J.-D., Shi, S.-P., Suo, S.-B., Huang, S.-Y. & Liang, R.-P. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics* **29**, 1614-1622 (2013).
158. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709 (2013).
159. Quinodoz, M., Royer-Bertrand, B., Cisarova, K., Di Gioia, S.A., Superti-Furga, A. & Rivolta, C. DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders. *The American Journal of Human Genetics* **101**, 623-629 (2017).
160. Cassa, C.A., Weghorn, D., Balick, D.J., Jordan, D.M., Nusinow, D., Samocha, K.E., O'Donnell-Luria, A., MacArthur, D.G., Daly, M.J. & Beier, D.R. Estimating the selective

effects of heterozygous protein-truncating variants from human exome data. *Nature genetics* **49**, 806 (2017).
161. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3. 0: A One‐Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice‐Site SNVs. *Human mutation* **37**, 235-241 (2016).
162. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics* **48**, 214-220 (2016).
163. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature methods* **11**, 361-362 (2014).
164. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen‐2. *Current protocols in human genetics*, 7.20. 1-7.20. 41 (2013).
165. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073 (2009).
166. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PloS one* **7**, e46688 (2012).
167. Shihab, H.A., Gough, J., Mort, M., Cooper, D.N., Day, I.N. & Gaunt, T.R. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human genomics* **8**, 11 (2014).
168. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118-e118 (2011).
169. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome research* **19**, 1553-1561 (2009).
170. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 315-323 (2011).
171. Han, J. & Moraga, C. The influence of the sigmoid function parameters on the speed of backpropagation learning. in *International Workshop on Artificial Neural Networks* 195-201 (Springer, 1995).
172. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
173. Bousquet, O. & Bottou, L. The tradeoffs of large scale learning. in *Advances in neural information processing systems* 161-168 (2008).
174. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763 (2014).
175. Paige, S.L., Saha, P. & Priest, J.R. Beyond Gene Panels: Whole Exome Sequencing for Diagnosis of Congenital Heart Disease. (Am Heart Assoc, 2018).
176. Sifrim, A., Hitz, M.-P., Wilsdon, A., Breckpot, J., Al Turki, S.H., Thienpont, B., McRae, J., Fitzgerald, T.W., Singh, T. & Swaminathan, G.J. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature genetics* **48**, 1060 (2016).
177. Cordell, H.J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., Cosgrove, C., Blue, G., Granados-Riveron, J. & Setchfield, K. Genome-wide association study of

multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nature genetics* **45**, 822 (2013).
178. Gelb, B.D. & Chung, W.K. Complex genetics and the etiology of human congenital heart disease. *Cold Spring Harbor perspectives in medicine* **4**, a013953 (2014).
179. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
180. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M. & Lee, B.T. ENCODE data at the ENCODE portal. *Nucleic acids research* **44**, D726-D732 (2015).
181. Spitz, F. & Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics* **13**, 613 (2012).
182. Ward, L.D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology* **30**, 1095 (2012).
183. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R. & Barrett, J.C. De novo mutations in regulatory elements cause neurodevelopmental disorders. *bioRxiv*, 112896 (2017).