Genetic analysis of *de novo* variants reveals sex differences in complex and isolated congenital diaphragmatic hernia and indicates *MYRF* as a candidate gene

Hongjian Qi [1,2,*], Lan Yu [3,*], Xueya Zhou[1,3*], Alexander Kitaygorodsky[1,4], Julia Wynn[3], Na Zhu[1,3], Gudrun Aspelund[5], Foong Yen Lim[6], Timothy Crombleholme[6], Robert Cusick[7], Kenneth Azarow[8], Melissa Ellen Danko[9], Dai Chung[9], Brad W. Warner[10], George B. Mychaliska[11], Douglas Potoka[12], Amy J. Wagner[13], Mahmoud ElFiky[14], Deborah A. Nickerson[15], Michael J. Bamshad[15], Jay M. Wilson[16,17], Frances A. High[16,17,18], Mauro Longoni[17,18], Patricia Donahoe[17,18], Wendy K. Chung [3,19,20,#], Yufeng Shen [1,4,21,#]

1. Department of Systems Biology, Columbia University, New York, NY, USA
2. Department of Applied Mathematics and Applied Physics, Columbia University, New York, NY, USA
3. Department of Pediatrics, Columbia University, New York, NY, USA
4. Department of Biomedical Informatics, Columbia University, New York, NY 10032
5. Department of Surgery, Columbia University Medical Center, New York, NY, USA
6. Cincinnati Children's Hospital, Cincinnati, OH, USA
7. Children's Hospital & Medical Center of Omaha, University of Nebraska College of Medicine, Omaha, NE, USA
8. Department of Surgery, Oregon Health&Science University, Portland, OR, USA
9. Monroe Carell Jr. Children's Hospital, Vanderbilt University Medical Center, Nashville, TN, USA
10. Washington University, St. Louis Children's Hospital, St. Louis, MO, USA
11. University of Michigan, CS Mott Children's Hospital, Ann Arbor, MI, USA
12. Children's Hospital of Pittsburgh, Pittsburgh, PA, USA
13. Medical College of Wisconsin, Milwaukee, WI, USA
14. Department of Pediatric Surgery, Faculty of Medicine, Cairo University, Cairo, Egypt
15. Center for Mendelian Genomics, University of Washington, Seattle, Washington, USA
16. Department of Surgery, Boston Children’s Hospital, Boston, MA, USA.
17. Department of Surgery, Harvard Medical School, MA, USA
18. Pediatric Surgical Research Laboratories, Department of Surgery, Massachusetts General Hospital, Boston, MA, USA
19. Department of Medicine, Columbia University, New York, NY, USA
20. Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA
21. JP Sulzberger Columbia Genome Center, Columbia University, New York, NY, USA

* These authors contributed to the work equally

# Corresponding authors:

W.K.C. (wkc15@columbia.edu)

Y.S. (ys2411@cumc.columbia.edu)

Abstract

Congenital diaphragmatic hernia (CDH) is one of the most common and lethal birth defects. Previous studies using exome sequencing support a significant contribution of coding *de novo* variants in complex CDH cases with additional anomalies and likely gene-disrupting (LGD) variants in isolated CDH cases. To further investigate the genetic architecture of CDH, we performed exome or genome sequencing in 283 proband-parent trios. Combined with data from previous studies, we analyzed a total of 357 trios, including 148 complex and 209 isolated cases. Complex and isolated cases both have a significant burden of deleterious *de novo* coding variants (1.7~fold, p= 1.2x10^{-5} for complex, 1.5~fold, p= 9.0x10^{-5} for isolated). Strikingly, in isolated CDH, almost all of the burden is carried by female cases (2.1~fold, p=0.004 for likely gene disrupting and 1.8~fold, p= 0.0008 for damaging missense variants); whereas in complex CDH, the burden is similar in females and males. Additionally, *de novo* LGD variants in complex cases are mostly enriched in genes highly expressed in developing diaphragm, but distributed in genes with a broad range of expression levels in isolated cases. Finally, we identified a new candidate risk gene *MYRF* (4 *de novo* variants, p-value=2x10^{-10}), a transcription factor with intolerant of mutations.

Main text

Congenital diaphragmatic hernia (CDH) affects approximately 1 in 3000 live births and is often lethal[1,2]. It can be isolated (50-60%) or associated with other anomalies including cardiac, brain, skeletal, gastrointestinal and genitourinary malformations [3]. Most genes implicated in CDH have been identified through recurrent chromosomal anomalies and mutant mice[4-10]. 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. We and others have reported an enrichment of *de novo* genetic events in sporadic CDH patients[11-13], especially 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[11,12], we analyzed a total of 357 trios (Supplementary Table 1), including 148 complex and 209 isolated cases.

Patients were recruited from the multicenter, longitudinal DHREAMS study [14] 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 [15,16]. Among the 148 complex cases, the most frequent anomalies were congenital heart disease (41%), but neurodevelopmental delay, gastrointestinal, and other malformations were common (Table 1 and Supplementary Table 2). A total of 209 (59%) patients had isolated CDH without additional anomalies at last contact[13]. In the DHREAMS cohort (Online Methods) 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 (Supplementary Table 2).

Characteristics	Number	percentage(%)
Male/Female	210/147	59/42
Left/Right/Other CDH location	269/56/32	75/16/9
White/Asian/Black/Other or unknown	240/13/10/94	67/4/3/27
Isolated cases	209	59
Complex cases	148	42
congenital heart disease	60	41
gastrointestinal anomaly	14	10
structural brain anomaly	15	10
other congenital malformations	67	45
neurodevelopmental delay	14	10

Table 1. Clinical and phenotypic summary of CDH patients (n=357)

We identified 461 protein-coding *de novo* variants (Supplementary Table 3) (~1.29 per patient), including 190 damaging *de novo* variants in LGD and predicted deleterious missense variants ("D-mis" defined as CADD score ≥ 25, Supplementary Table 4). 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=4.2x10^{-4} for LGD, and FE=1.5, P=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.[17,18](Table 2, Online Methods). 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 2). 22% of complex and 16% of isolated cases are explained by damaging *de novo* variants.

We then tested whether the burden of damaging *de novo* variants were concentrated in constrained genes (defined as ExAC [19] pLI≥0.5)[19] 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) (Supplementary Table 5 and 6). 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.

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**

Table 2. Enrichment of *de novo* variants in 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[17,18].

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 2). We estimated that 27% of females can 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 (Supplementary Table 7). 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 (Fig. 1a and Supplementary Table 7). Furthermore, in isolated female cases, LGD variants are mainly enriched in constrained genes (FE=3.3, P=0.001, Fig.1a), and D-mis variants were mainly in other genes (FE=2.2, P=0.0002) (Supplementary Table 8, Fig.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 (Supplementary Table 8, Fig.1b).

Genes associated with CDH are often expressed in pleuroperitoneal folds (PPF), an early structure critical in the developing diaphragm[20,21]. We analyzed the expression patterns of genes with LGD and D-mis variants using a mouse E11.5 PPF data set[22]. 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=$7x10^{-7}$)

(Supplementary Table 9, Fig. 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 (Supplementary Table 9 and 10, Fig. 2).

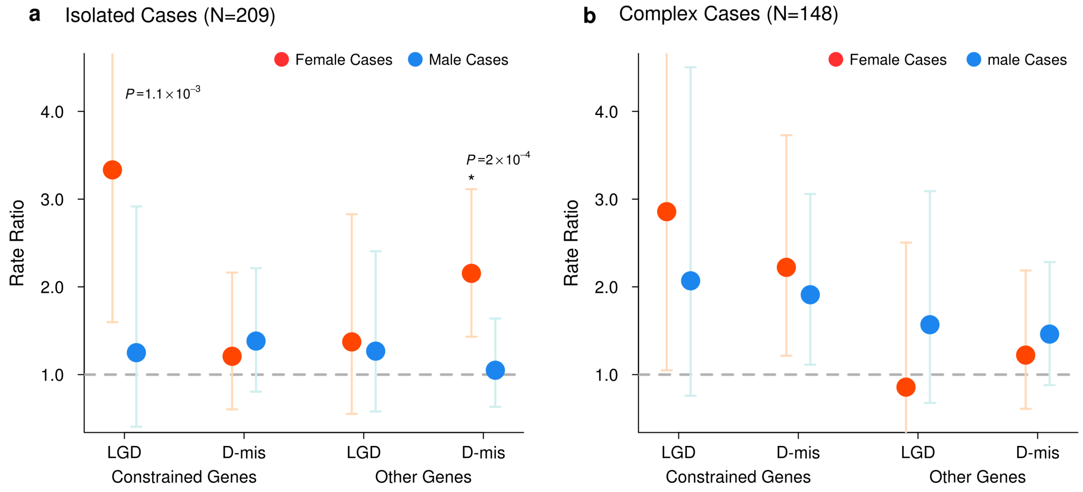

Figure 1. Female and male CDH cases have different enrichment rate of damaging de novo variants. (a) Enrichment of LGD variants and D-mis in constrained or other genes in isolated female and male cases. Constrained genes with LGD variants and other genes with D-mis variants are mainly enriched in female 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.

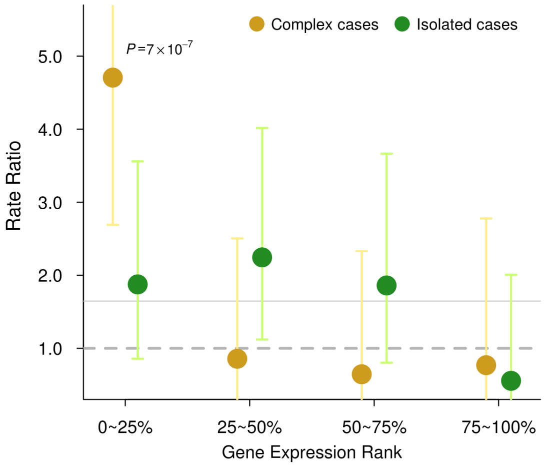

Figure 2. Isolated and complex cases have different enrichment patterns of LGD *de novo* variants. Enrichment rate of LGD *de novo* variants are shown in gene sets grouped by expression rank in E11.5 pleuroperitoneal folds (PPFs). In complex CDH cases, LGD *de novo* variants are dramatically enriched 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.

Two genes are observed with multiple damaging *de novo* variants. Wilms tumor 1 (*WT1)* has been previously implicated in CDH[23] and has two D-mis variants. Myelin Regulatory Factor (*MYRF)*, a transcription factor, has one *de novo* LGD and three D-mis variants (Fig. 3a) in four complex CDH patients (p=2x10^{-10}, based on comparison to expectation from background mutations [17,18]) (Table 3). A recent study of congenital heart disease (CHD) [24,25] reported three additional damaging *de novo* missense variants (p.F387S, p.Q403H and p.L479V) in *MYRF* (Table 3, Fig 3a). All four CDH patients had CHD (Table 3). 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[19] 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 [22] and E14.5 heart [26], respectively). Genital malformation may share developmental processes[27] because PPF is physically connected dorsally to urogenital ridge.

Sample ID	Gender	Diaphragm defect	Heart defect	Genital defect	Other malformations	Protein	CADD
01-1008	Male	CDH	ASD,VSD,TOF	bilateral undescended testes	NA	p.G81Wfs*45	27.3
01-0429	Female	CDH	VSD	no internal genital organs, external blind-ending vagina	accessory spleen	p.G435R	32
04-0042	Male	CDH	ASD,VSD	NA	NA	p.V679A	25.9
05-0050	Male	CDH	hypoplastic left heart syndrome	ambiguous genitalia	intellectual disability	p.R695H	34
1-02264	Male	NA	abnormal aorta	ambiguous genitalia, hypospadias, undescended testis	NA	p.F387S	27.9
1-03160	Male	right hemidiaphragm eventration	abnormal atrial septum, pulmonary vein, systemic vein, aorta, aortic valve, mitral valve, pulmonary arteries, ventricular septum	undescended testis	lung hypoplasia, abdominal abnormalities	p.Q403H	27.6
1-07403	Female	NA	abnormal aorta and aortic valve	NA	skeletal abnormalities, short stature	p.L479V	23.9

Table 3. *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 three variants identified in CHD patients and p.G435R are located in the conserved DNA binding domain (DBD) of *MYRF* (Fig. 3), and could alter DNA binding[28]. 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[29]. 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[29] which is important for the transcription factor function[30]. MYRF is thought to be an essential transcription factor for oligodendrocyte differentiation and myelination[31]. Conditional deletion of *Myrf* impaired motor learning[32,33] and the individual with the p.V679A variant we assessed at two years old had intellectual disability.

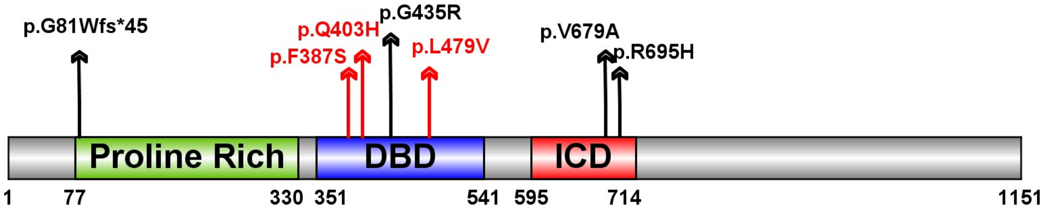

Figure 3. *De novo* variants identified in *MYRF*. Schematic of the MYRF protein with predicted sequence features, including N-terminal Proline Rich region, DNA-binding domain (DBD) and intramolecular chaperone domain (ICD). Variants identified in CDH indicated as black arrow, variants identified in congenital heart disease cases indicated with red arrows.

Our study suggests for the first time that isolated male and female CDH may have a different genetic architecture. Damaging *de novo* variants with large effect have a substantial contribution to isolated female cases but little to isolated male cases. Given the male bias in 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 cases[34], additional highly penetrant *de novo* variants are required to cause CDH in females. This is similar to autism [35]. Since there is a similar male/female ratio in overall cohort and neonatal cohort (1.4:1), this difference is unlikely due to ascertainment 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 enrichment of LGD and D-mis variants in genes highly expressed in diaphragm structure (PPF) in early embryonic development is consistent with the pleiotropic effects 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 expression data from PPFs is the sum of different cell types[36], the lack of correlation of LGD enrichment and expression level in PPF suggests a substantial portion of the implicated genes in isolated cases could be expressed only in sub-populations of cells in PPF. Single-cell mRNA-sequencing will be necessary to analyze gene expression pattern in specific cell types and further assess the etiologies of isolated CDH. Finally, the four damaging *de novo* variants in *MYRF* were identified in complex CDH patients with congenital heart disease and genitourinary anomalies and likely represent a novel syndrome.

METHODS

Patients

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)[11] (Boston Cohort) and 39 trios from a previous study [12] (Supplementary Table 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/)[13]. 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[12,13]). 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[11]. 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.

Whole Exome/Genome Sequencing

We included previously two sets of WES data for analysis[11,12]. We performed at the University of Washington whole exome sequencing (WES) 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[12]. 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.

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)[37] 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 [38], mark PCR duplicates using Picard (v1.67), performed local realignment and quality recalibration using GATK [39]. 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 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 (Supplementary Fig. 2).

We performed principal component analysis of common variants (allele frequency >5%) using Eigenstrat [40] to determine the population structure and ancestry of both cases and controls, with HapMap 3 sample collection data [41] as a reference.

Detection of *de novo* SNVs and indels.

We used Plink[42] (http://pngu.mgh.harvard.edu/purcell/plink/) to estimate Identity by Descent (IBD)[43] 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 [12,17,44]. 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 [12].

Annotation of variants.

We used ANNOVAR[45] 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[46] phred-scale score ≥25.

Statistical analysis

We performed statistical analyses using R package from the Comprehensive R Archive Network, and the denovolyzerR [18] package.

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

We calibrated the expected number of *de novo* variants in patients in each variant class in each gene based on the 3-nucleotide context-specific mutation rate estimated by Samocha et al.[17,18].

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.

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.

Expression profile during diaphragm development

Mouse developing diaphragm (MDD) gene expression datasets from the pleuroperitoneal folds (PPFs)[22] 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[12].

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[47].

ACKNOWLEDGEMENTS

We would like to thank the patients and their families for their generous contribution. We are grateful for the technical assistance provided by Patricia Lanzano, Jiancheng Guo, and Liyong Deng from Columbia University, Jessica Kim at Boston Children's Hospital, and Caroline Coletti and Pooja Bhayani at Massachusetts General Hospital. We thank our clinical coordinators across the DHREAMS centers: Trish Burns at Cincinnati Children's Hospital, Sheila Horak at Children's Hospital & Medical Center of Omaha, Brandy Gonzales at Oregon Health and Science University, Karen Lukas at St. Louis Children's Hospital, Jeannie Kreutzman at CS Mott Children's Hospital, Min Shi at Children's Hospital of Pittsburgh, Michelle Knezevich and Cheryl Kornberg at Medical College of Wisconsin. We thank University of Washington Center for Mendelian Genomics (UW-CMG) team for generating part of the WES data. The WES data generation at UW-CMG was funded by the National Human Genome Research Institute and the National Heart, Lung and Blood Institute grant HG006493 to Drs. Debbie Nickerson, Michael Bamshad, and Suzanne Leal. The whole genome sequencing data were generated by NIH Gabriella Miller Kids First Pediatric Research Program (X01HL132366). This work was supported by NIH grants R01HD057036 (L.Y., J.W., W.K.C.), R03HL138352 (A.K., W.K.C., Y.S.), R01GM120609 (H.Q., Y.S.), UL1 RR024156 (W.K.C.), and 1P01HD068250 (P.K.D, M.L., F.A.H., J.M.W., W.K.C.) Additional funding support was provided by grant from CHERUBS, a grant from the National Greek Orthodox Ladies Philoptochos Society, Inc. and generous donations from The Wheeler foundation, Vanech Family Foundation, Larsen Family, Wilke Family and many other families.

Author Contributions

W.K.C and Y.S. conceived the study. F.A.H., J.M.W., and P.K.D. provided genomic data. H.Q., X.Z., Y.L., A.K., W.K.C., and Y.S. analyzed and interpreted the data. Y.L., X.Z., H.Q., W.K.C., and Y.S. wrote the manuscript. J.W., G.A., F.L., T.C., R.C., K.A., M.E.D., D.C., B.W.W., G.B.M., D.P., A.J.W., M.E., F.A.H., M.L., J.M.W., P.K.D. collected samples and clinical information. All authors contributed and discussed the results and critically reviewed the manuscript.

Competing Financial Interests statement. None declared

REFERENCES

1. Chandrasekharan, P.K., Rawat, M., Madappa, R., Rothstein, D.H. & Lakshminrusimha, S. Congenital Diaphragmatic hernia - a review. *Matern Health Neonatol Perinatol* **3**, 6 (2017).
2. Pober, B.R., Russell, M.K. & Ackerman, K.G. Congenital Diaphragmatic Hernia Overview. in *GeneReviews(R)* (eds. Pagon, R.A. *et al.*) (Seattle (WA), 1993).
3. Pober, B.R. Overview of epidemiology, genetics, birth defects, and chromosome abnormalities associated with CDH. *Am J Med Genet C Semin Med Genet* **145C**, 158-71 (2007).
4. Jay, P.Y. *et al.* Impaired mesenchymal cell function in Gata4 mutant mice leads to diaphragmatic hernias and primary lung defects. *Dev Biol* **301**, 602-14 (2007).
5. Ackerman, K.G. *et al.* Fog2 is required for normal diaphragm and lung development in mice and humans. *Plos Genetics* **1**, 58-65 (2005).
6. Castiglia, L. *et al.* Narrowing the candidate region for congenital diaphragmatic hernia in chromosome 15q26: Contradictory results. *American Journal of Human Genetics* **77**, 892-894 (2005).
7. Klaassens, M. *et al.* Congenital Diaphragmatic hernia and chromosome 15q26: Determination of a candidate region by use of fluorescent in situ hybridization and array-based comparative genomic hybridization. *American Journal of Human Genetics* **76**, 877-882 (2005).
8. Shimokawa, O. *et al.* Molecular characterization of del(8)(p23.1p23.1) in a case of congenital diaphragmatic hernia. *American Journal of Medical Genetics Part A* **136A**, 49-51 (2005).
9. Wat, M.J. *et al.* Chromosome 8p23.1 Deletions as a Cause of Complex Congenital Heart Defects and Diaphragmatic Hernia. *American Journal of Medical Genetics Part A* **149A**, 1661-1677 (2009).
10. You, L.R. *et al.* 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).
11. Longoni, M. *et al.* Genome-wide enrichment of damaging de novo variants in patients with isolated and complex congenital diaphragmatic hernia. *Hum Genet* **136**, 679-691 (2017).
12. Yu, L. *et al.* Increased burden of de novo predicted deleterious variants in complex congenital diaphragmatic hernia. *Hum Mol Genet* **24**, 4764-73 (2015).
13. Yu, L. *et al.* De novo copy number variants are associated with congenital diaphragmatic hernia. *J Med Genet* **49**, 650-9 (2012).
14. DHREAMS: Diaphragmatic Hernia Research & Exploration; Advancing Molecular Science, http://www.cdhgenetics.com. (2009).
15. Hinton, C.F., Siffel, C., Correa, A. & Shapira, S.K. Survival Disparities Associated with Congenital Diaphragmatic Hernia. *Birth Defects Res* **109**, 816-823 (2017).
16. Leeuwen, L. *et al.* Congenital Diaphragmatic Hernia and Growth to 12 Years. *Pediatrics* **140**(2017).

17. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
18. Ware, J.S., Samocha, K.E., Homsy, J. & Daly, M.J. Interpreting de novo Variation in Human Disease Using denovolyzeR. *Curr Protoc Hum Genet* **87**, 7 25 1-7 25 15 (2015).
19. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
20. Clugston, R.D., Zhang, W. & Greer, J.J. Gene expression in the developing diaphragm: significance for congenital diaphragmatic hernia. *Am J Physiol Lung Cell Mol Physiol* **294**, L665-75 (2008).
21. Merrell, A.J. *et al.* Muscle connective tissue controls development of the diaphragm and is a source of congenital diaphragmatic hernias. *Nat Genet* **47**, 496-504 (2015).
22. Russell, M.K. *et al.* Congenital diaphragmatic hernia candidate genes derived from embryonic transcriptomes. *Proc Natl Acad Sci U S A* **109**, 2978-83 (2012).
23. Carmona, R. *et al.* Conditional deletion of WT1 in the septum transversum mesenchyme causes congenital diaphragmatic hernia in mice. *Elife* **5**(2016).
24. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (New York, N.Y.)* **350**, 1262-6 (2015).
25. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature Genetics* **In press**(2017).
26. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).
27. Azarow, K.S. *et al.* The association between congenital diaphragmatic hernia and undescended testes. *J Pediatr Surg* **50**, 744-5 (2015).
28. Bujalka, H. *et al.* MYRF is a membrane-associated transcription factor that autoproteolytically cleaves to directly activate myelin genes. *PLoS Biol* **11**, e1001625 (2013).
29. Li, Z., Park, Y. & Marcotte, E.M. A Bacteriophage tailspike domain promotes self-cleavage of a human membrane-bound transcription factor, the myelin regulatory factor MYRF. *PLoS Biol* **11**, e1001624 (2013).
30. Kim, D. *et al.* Homo-trimerization is essential for the transcription factor function of Myrf for oligodendrocyte differentiation. *Nucleic Acids Res* **45**, 5112-5125 (2017).
31. Hornig, J. *et al.* The transcription factors Sox10 and Myrf define an essential regulatory network module in differentiating oligodendrocytes. *PLoS Genet* **9**, e1003907 (2013).
32. McKenzie, I.A. *et al.* Motor skill learning requires active central myelination. *Science* **346**, 318-22 (2014).
33. Xiao, L. *et al.* Rapid production of new oligodendrocytes is required in the earliest stages of motor-skill learning. *Nat Neurosci* **19**, 1210-1217 (2016).
34. Robinson, E.B., Lichtenstein, P., Anckarsater, H., Happe, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc Natl Acad Sci U S A* **110**, 5258-62 (2013).
35. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).
36. Clugston, R.D. & Greer, J.J. Diaphragm development and congenital diaphragmatic

hernia. *Semin Pediatr Surg* **16**, 94-100 (2007).

37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

38. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-8 (2008).

39. GATK-Team. GATK best practice. (2016).

40. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).

41. International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).

42. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

43. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

44. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-6 (2015).

45. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

46. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

47. Pruitt, K.D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316-23 (2009).