Supplemental Notes and Figures for “Likely damaging de novo variants in congenital diaphragmatic hernia patients are associated with worse clinical outcomes”

Lu Qiao, PhD1,2,*, Julia Wynn, MS1,*, Lan Yu, PhD1, Rebecca Hernan, MS1, Xueya Zhou, PhD1,2, Vincent Duron, MD3, Gudrun Aspelund, MD MS3, Christiana Farkouh-Karoleski, MD1, Annette Zygumunt, PhD1, Usha S. Krishnan, MD1, Shannon Nees, MD1, Julie Khlevner, MD1, Foong Yen Lim, MD4, Timothy Crombleholme, MD4, Robert Cusick, MD5, Kenneth Azarow, MD6, Melissa Ellen Danko, MD7, Dai Chung, MD7, Brad W. Warner, MD8, George B. Mychaliska, MD MS9, Douglas Potoka, MD10, Amy J. Wagner, MD11, Samuel Soffer, MD12, David Schindel, MD13, David J. McCulley, MD14, Yufeng Shen, PhD2,15,16,#, Wendy K. Chung, MD PhD1,17,18,#
Supplementary Notes

Evaluation of pulmonary hypertension

Pulmonary hypertension (PH) severity was ascertained post-repair at one month and three months according to the methods detailed in a prior publication\(^1\). Briefly, echocardiograms were centrally read, by one of two cardiologists for the presence and degree of PH.

A right ventricular (RV) pressure estimate less than half of the systolic blood pressure at the time of echo, based on tricuspid regurgitation (TR) gradient was quantified as mild (or no) PH, one-half to two-thirds systemic blood pressure as moderate PH, and greater than two-thirds systemic blood pressure as severe PH. In the absence of TR, ventricular septal position and motion, RV function and RV size were used to estimate the severity of PH. The presence of bidirectional shunting or right to left shunting at the ductal or VSD level was also quantified as severe PH (systemic or suprasystemic pressures in the right heart).

For the analysis, none and mild PH and moderate and severe PH were collapsed to make a binary variable.

Identification of de novo variants

At the time the data was frozen for this analysis, 462 of all 647 neonatally enrolled probands met criteria and had trio sequencing data.

Exome sequencing was performed in 125 trios and genome sequencing in 394 trios. Of sequencing these, 91 trios had both exome and genome sequencing. The genome sequencing was performed in three batches: 157 trios at Baylor College of Medicine Human Genome Sequencing Center (named as, Baylor), 201 trios at Broad Institute Genomic Services (Broad1), and 39 trios at Broad Institute Genomic Services (Broad2). Three trios were replicated in Baylor and Broad1. The genomic libraries of Baylor, Broad1 and 38 Broad2 cases were prepared by the Illumina TruSeq DNA PCR-Free Library Prep Kit (Illumina), while other Broad2 (named as Broad3) by Illumina
TruSeq DNA PCR-Plus Library Prep Kit (Illumina), with average fragment length about 350 bp, and sequenced as paired-end of 150-bp on Illumina HiSeq X platform.

**De novo single nucleotide variants and small insertions/deletions.** The calling and annotation of de novo SNVs and indels were processed using a pipeline based on GATK Best Practice (version 3) followed with heuristic hard filters as described in our previous study (Qi, 2018). All coding variants were classified as synonymous, missense, inframe, or likely-gene-disrupting (LGD, which includes frameshift indels, canonical splice site, or nonsense variants). The missense variants were further predicted as deleterious missense variants (D-mis) by MPC\(^2\) score ≥2 and phred-scaled CADD\(^3\) v1.3 score ≥20, and plausible deleterious missense variants (PD-mis) by MPC score ≥1 and CADD score ≥20.

**De novo copy number variants.** Overview. We designed an inhouse pipeline of read depth-based algorithm for detecting and validating CNVs from genome sequencing in Supplementary Figure 1. Given that samples for this study were aggregated across sequencing sites or genome sequencing library preparation protocols, to assign filter thresholds separately for different sequencing batches in larger sample size, the workflow was performed in 737 trios, including 397 genome sequencing trios with clinical data in this study, plus an additional 340 genome sequencing trios without outcome data not included in this study, which totally spread among 195 trios in Baylor, 312 trios in Broad1, 213 trios in Broad2, 17 trios in Broad3.

**Sample preprocessing.** We assessed the genome sequencing data properties for all 737 trios prior to CNV discovery separately for each sequencing batch and measured 4 features per sample (Figure S2): (1) replication rate, i.e., PCR duplication rate, (2) pairwise alignment rate, i.e., properly paired alignment rate, (3) split reads rate, (4) insert size for pairs excluding reads marked as duplicates.

**CNV calling and genotyping.** We obtained variants predicted with read depth (RD) evidence by CNVnator\(^4\) v0.3.3 with the bin size set as 100bp. Variants detected were filtered against regions that had ≥80% overlap with assembly gaps, short tandem repeat (STR), centromere, or segmental duplication. We excluded CNV calls with ≥50% of zero mapping quality (q0) reads and normalized RD >4 because neutral RD was 1 in calling module of CNVnator. Then the genotyping
module was run jointly across the three members of each trio if CNV calls passed a range of quality score thresholds. For diploid chromosomes whose neutral normalized RD was 2 in genotyping module, we called de novo CNVs if (1) the normalized average RD signal (RD$_{norm}$) in the proband was <1.4 (for deletions) or >2.6 (for duplications), (2) RD$_{norm}$ in each parent was between 1.6 and 2.4. Putative heterozygous inherited CNVs for each trio were detected with the following criteria: (1) for deletions, RD$_{norm}$ in the proband and one parent was <1.4 and >0.6 while in another parent was between 1.6 and 2.4, (2) for duplications, RD$_{norm}$ in the proband and one parent was >2.6 and <3.4 while in another parent was between 1.6 and 2.4. Defining CNV transmission as heterozygous inherited CNV, we assessed whether the rate of transmission for CNVs was closer to the expected Mendelian rate of 50% and whether there were reasonable de novo rates and Mendelian error rate across a range of quality score thresholds. If a sample had larger number of de novo CNV segments, i.e., more than one SD away from sample mean, the CNVs were called by the additional pair-end/split-read (PE/SR) evidence Lumpy$^5$ v0.2.13 and genotyped by SVtyper$^6$ v0.1.4. Only the CNVs supported by both RD and PE/SR would be kept for downstream analysis.

**CNV refining and validating.** Using the recommended quality score cutoff that broadly conformed to expectations of random Mendelian segregation, we assigned copy number and re-genotyped CNVs trio-jointly. To increase sensitivity and specificity, de novo CNVs with a size >1kb and allele frequency <1% were included in this study. We clustered and merged de novo CNV segments manually to adjust the bounds for each sample. Any sample with larger number of de novo CNVs was called CNVs with additional PE/SR evidence and only CNV calls supported by RD and PE/SR were kept for further analysis. We generated a series of manual inspections and visualization information for each de novo CNV from orthogonal supporting evidences as follows: (1) normalized read depth (Figure S3), i.e., the RD$_{norm}$ was 0.5/1.5 in the offspring and about 1.0 in both parents for deletions/duplications, (2) discordant pairs (Figure S4), i.e., there were read depth peaks of discordant pairs at CNV boundaries, (3) split reads (Figure S5), i.e., there were read depth peaks of split reads at CNV boundaries, (4) the allelic copy ratio in terms of B allele frequency (BAF, Figure S6), i.e., allelic depth for alternate allele was divided by sum of allelic depths for the reference and alternate alleles. The heterozygous single nucleotide polymorphism (SNP) with BAF value approximated 0.5. A hemizygous deletion results in BAF values of 0 and 1, while a duplication results in approx. 0, 0.33, 0.66 and 1. De novo CNVs were called with at
least 2 confirmed evidences. CNV boundaries were refined based on the PE/SR evidence. We required the candidate *de novo* variants with population allele frequency in gnomAD-SV\(^7\) was less than 1%, as appropriate and 0% allele frequency in the healthy-parent cohort. All *de novo* CNVs were manually inspected in the Integrated Genomics Viewer ([http://software.broadinstitute.org/software/igv](http://software.broadinstitute.org/software/igv)). We performed experimental validation of 25 putative *de novo* genic CNVs including all 9 small CNVs (<5kb) using quantitative PCR (qPCR). All PCR primers were designed for the selected genes located within the *de novo* CNVs and synthesized by IdtDNA. All qPCR reactions were performed in a total of 10 μl volume, comprising 5 μl 2x SYBR Green I Master Mix (Promega), 1 μl 10nM of each primer and 2 μl of 1:20 diluted cDNA in 96-well plates using CFX Connect Real-Time PCR Detection System (Bio-Rad). All reactions were performed in triplicate, and the conditions were 5 minutes at 95°C, then 40 cycles of 95°C at 15 seconds and 60°C at 30 seconds. The relative copy numbers were calculated using the standard curve method relative to the β-actin housekeeping gene. Five-serial 4-fold dilutions of DNA samples were used to construct the standard curves for each primer. From these analyses, we found that 22/25 (88%) *de novo* CNV predictions were confirmed.

**CNV annotation.** We mapped *de novo* CNVs on GENCODE (v29) protein coding genes with at least 1bp in the shared interval. The GENCODE genes were annotated with mutation intolerance metric by gnomAD pLI\(^2\) and haploinsufficiency or triplosensitivity metrics from ClinGen genome dosage map.

**Statistical analysis of associations of genetic factors with clinical outcomes**

**Lasso feature selection.** We used Lasso to select a subset of clinical factors as covariates for regression analysis of outcomes. For each outcome \(y\) in \(n\) samples, we considered \(p\) clinical factors recorded before the time of the outcome \((x_1, x_2, ..., x_p)\) in a regression model:

\[
y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,
\]

where \(i = (1, ..., n)\), \((\beta_1, \beta_2, ..., \beta_p)\) were the regression coefficients for the \(p\) clinical factors, and \(\epsilon_i\) was an error term. We estimated the parameters by:
\[
\hat{\beta}_j = \arg\min \sum_{i=1}^{n} (y_i - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,
\]

where \(\lambda\) was the penalty term to control the strength of the regularization penalty.

We used cross-validation to choose the optimal \(\lambda\) that gave minimum mean cross-validated error. Specifically, we divided all 647 subjects into 3 folds; for every fold as test set, we built the trained regression model using the remaining 2 folds. After obtaining the prediction of outcomes using the trained model on test set, we computed the mean squared error (MSE). We calculated the average MSE across 3 folds with the optimal \(\lambda\) at which the minimal MSE was achieved. After repeating the cross-validation procedure 1000 times, we recorded the clinical factors with estimated non-zero coefficients using the optimal \(\lambda\) in each repetition and selected the clinical factors as confounding covariates if they were repeatedly selected in 700 or more repetitions. This feature selection was carried out for every outcome using glmnet package in R v3.4.1. The parameter ‘family’ in glmnet function was ‘binomial’ for binary outcomes, e.g. PH at 1 or 3 months or mortality prior to initial discharge. The ‘family’ of quantitative outcome e.g. development at 2 years of age was ‘gaussian’. The selected features from the entire CDH cohort were also taken as covariates of same outcome for analyses in isolated and complex CDH cohorts.

**Analysis in CDH cases excluded carriers with known syndromic abnormalities and/or large CNVs**

A total of 25 cases (10 isolated and 15 complex) have known syndromic abnormalities or large CNVs (Table S1). Excluding these cases, the mortality rate remained similar and complex cases had a higher mortality rate than isolated cases (30% vs. 13%, \(P=2\times10^{-5}\)). Compared to the non-LD group, mortality prior to initial discharge was higher in the LD group (\(P=1\times10^{-3}\), Figure 9A). The difference in mortality between LD and non-LD group was largely attributable to differences in the isolated cases. The isolated cases with LD variants had a mortality rate (35%) which was similar to complex cases and was much higher than isolated cases without LD (11%, \(P=3\times10^{-3}\)). Cases with a LD variant had marginally higher PH prevalence at 1 month (77% vs. 49%, \(P=1\times10^{-3}\)) compared to non-LD cases (Figures S10B and C).
After the exclusion of carriers with known syndromic abnormalities and/or large CNVs, the associations of poorer neurodevelopmental outcomes for LD cases compared to non-LD cases were consistent with largely similar effect size for some domains (Table S6 and Figure S11). In 437 CDH cases without known syndromic or large CNVs, compared to non-LD variants, LD variants had significantly lower BSID-III motor scores (β=9.2, 95% CI=3.1-15.3, \(P=4\times10^{-3}\)). Compared to all 462 CDH cases association analysis, the covariates were the same clinical factors for each domain and had the same effect size. The presence of oxygen at discharge was significantly associated with lower BSID-III motor scores (β=12.1, 95% CI=6.5-17.7, \(P=4\times10^{-5}\)) and lower VABS-II composite adaptive behavior scores (β=7.7, 95% CI=3.2-12.1, \(P=8\times10^{-4}\)). Even though there was still no significant association between LD variants and lower BSID-III language score, oxygen at discharge (β=10.3, 95% CI=4.2-16.3, \(P=1\times10^{-3}\)) and SES (β=+4.8, 95% CI=2.7-6.8, \(P=9\times10^{-6}\)) had the same power to BSID-III language domain scores as results in all 462 CDH cases association analysis.
Figure S1. Workflow of read depth-based algorithm for detecting and validating CNVs from genome sequencing. The blue shape represents the tools we used to call variants. Decision points are colored in yellow. Quality control and validations are colored in gray.
Figure S2. Sample preprocessing in CNV calling workflow. (A) replication rate, i.e., PCR duplication rate, (B) pairwise alignment rate, i.e., properly paired alignment rate, (C) split reads rate, (D) mean insert size for pairs.
Figure S3. Number and size of CNVs called by read depth-based workflow from each batch.
Figure S4. Distribution of BSID-III and VABS-II scores for each domain. Asterisk labels the mean +/- standard deviation of domain’s distribution.
Figure S5. Correlation of clinical factors and outcomes. Red means positive correlation, blue is the negative correlation.
Figure S6. Effective number of association tests. Variance is explained by every eigenvalue. The vertical dash line shows the number of independent tests based on the cumulative proportion that explain 99% of variation.
Figure S7. Association analyses for genetic groups with growth outcomes at 2 years of age. (A) Height, (B) head circumference. The black line in the middle of the box is the mean value for each group. The vertical size of the box is the interquartile range (IQR). The whisker is 1.5xIQR. P value is given by Student’s t-test with bold as significant ($P<5 \times 10^{-3}$) after correcting multiple tests.
Figure S8. Effect size for likely damaging variants on growth outcomes at 2 years of age comparing to non-LD variants. (A) Height, (B) head circumference. Effect size is the coefficient/beta, which aligns with outcome on the y-axis, with the vertical dashed line (v=0) separating positive and negative direction of association. Significant $P$ value after correcting multiple tests ($P<5\times10^{-3}$) is noted. LD: likely damaging variants.
Figure S9. Effect size for likely damaging variants on growth outcomes at 2 years of age comparing to non-LD variants in CDH cases without known syndromic abnormalities or CNVs. (A) Height, (B) head circumference. Effect size is the coefficient/beta, which aligns with outcome on the y-axis, with the vertical dashed line (v=0) separating positive and negative direction of association. Significant P value after correcting multiple tests (P<5×10^{-3}) is noted. LD: likely damaging variants; SES: socioeconomic status; PH: pulmonary hypertension; ECMO: extracorporeal membrane oxygenation.
Figure S10. Association analyses for genetic groups with mortality and pulmonary hypertension in CDH cases without known syndromic abnormalities or large CNVs. (A) Mortality prior to initial discharge. (B) Pulmonary hypertension at 1 month. (C) Pulmonary hypertension at 3 months. *P value is given by Fisher’s exact test after correcting multiple tests (P<5×10⁻³). **: *P<0.001; *: *P<0.005”.
Figure S11. Effect size for likely damaging variants on neurodevelopmental outcomes at 2yr comparing to non-LD variants in CDH cases without known syndromic abnormalities or CNVs. (A) BSID-III language, (B) BSID-III cognition, (C) BSID-III motor, (D) VABS-II adaptive behavior. Effect size is the coefficient/beta, which aligns with outcome on the y-axis, with the vertical dashed line (v=0) separating positive and negative direction of association. Significant P value after correcting multiple tests (P<5×10^{-3}) is noted. LD: likely damaging variants; SES: socioeconomic status; PH: pulmonary hypertension; ECMO: extracorporeal membrane oxygenation.
Figure S12. Association of clinical factors with the presence of LD variants. (A) Odds ratios for binary factors. Positive odds ratios are associated with increased risk of LD when the factor is present. *P* values are given (Fisher’s exact test). (B) Beta coefficients from logistic regression of quantitative factors versus presence of LD variant. SES: socioeconomic status; ECMO: extracorporeal membrane oxygenation.
Supplementary references

1. Wynn J, Krishnan U, Aspelund G, et al. Outcomes of congenital diaphragmatic hernia in the modern era of management. *J Pediatr.* 2013;163(1):114-119 e111.

2. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet.* 2014;46(9):944-950.

3. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886-D894.

4. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 2011;21(6):974-984.

5. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 2014;15(6):R84.

6. Chiang C, Layer RM, Faust GG, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods.* 2015;12(10):966-968.

7. Collins RL, Brand H, Karczewski KJ, et al. An open resource of structural variation for medical and population genetics. *bioRxiv.* 2019.