Molecular pathogenesis of congenital diaphragmatic hernia revealed by exome sequencing, developmental data, and bioinformatics

Mauro Longoni, Frances A. High, Meaghan K. Russell, Alireza Kashani, Adam A. Tracy, Caroline M. Coletti, Regis Hila, Ahmed Shamia, Julie Wells, Kate G. Ackerman, Jay M. Wilson, Carol J. Bult, Charles Lee, Kasper Lage, Barbara R. Pober, and Patricia K. Donahoe

Contributed by Patricia K. Donahoe, July 7, 2014 (sent for review March 6, 2014; reviewed by Wendy K. Chung and Ramneek Gupta)

Congenital diaphragmatic hernia (CDH) is a common and severe birth defect. Despite its clinical significance, the genetic and developmental pathways underlying this disorder are incompletely understood. In this study, we report a catalog of variants detected by a whole exome sequencing study on 275 individuals with CDH. Predicted pathogenic variants in genes previously identified in either humans or mice with diaphragm defects are enriched in our CDH cohort compared with 120 size-matched random gene sets. This enrichment was absent in control populations. Variants in these critical genes can be found in up to 30.9% of individuals with CDH. In addition, we filtered variants by using genes derived from regions of recurrent copy number variations in CDH, expression profiles of the developing diaphragm, protein interaction networks expanded from the known CDH-causing genes, and prioritized genes that likely contribute to the CDH phenotype. These data are valuable for comparison of candidate genes generated from whole exome sequencing of other CDH cohorts or multiplex kindreds and provide ideal candidates for further functional studies. Furthermore, we propose that these genes and pathways will enhance our understanding of the heterogeneous molecular etiology of CDH.

Significance

Congenital diaphragmatic hernia (CDH) is a common birth defect associated with high morbidity and mortality. Focusing on the coding sequence of 51 genes, discovered in human studies and in mouse models, we studied 275 CDH patients and identified multiple variants in CDH-causing genes. Information on gene expression in embryonic mouse diaphragms and protein interactions allowed us to prioritize additional compelling CDH-associated genes. We believe that an improved understanding of the genetics of CDH will be important to design new therapeutic strategies for patients with diaphragmatic defects.


Reviewers included: W.K.C., New York-Presbyterian/Columbia; and R.G., Technical University of Denmark.

The authors declare no conflict of interest.

Data deposition: Sequences have been deposited in the database of Genotypes and Phenotypes (dbGaP) (accession no. phs000783.v1.p1).

1M.L., F.A.H., M.K.R., and A.K. contributed equally to this work.

2To whom correspondence should be addressed. Email: pdonahoe@partners.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412509111/-/DCSupplemental.
additional developmental pathways must also be important. A comprehensive transcriptome analysis of the developing diaphragm, which includes genes in the RA pathway, was made available recently by our laboratory (10) and can serve as a useful tool for evaluating CDH candidate genes.

Improved understanding of the genetic pathways that contribute to CDH is critical to improving survival and reducing complications for infants born with CDH. To this end, we report the results of whole-exome sequencing on 275 individuals with CDH, focusing on (i) rare variants in CDH-causing genes, identified in mice and/or humans, (ii) chromosomal hotspots of recurring deletion or duplications, (iii) candidates prioritized by embryonic diaphragm transcriptome and protein-interaction networks, and (iv) ultrarare and highly disruptive variants in these prioritized categories. Candidate gene analysis and the other strategies used point toward molecular pathways and a set of high priority targets for further studies with the ultimate aim of devising novel treatment paradigms.

Results

Variant Filtering. Whole exome sequencing was performed on a cohort of 275 CDH patients, including both isolated and complex cases. Ninety-four percent of the patients had no family history of CDH, and all analyzed exomes were from unrelated individuals. Demographic and phenotypic details of this population can be found in Dataset S1. The average read depth for the targeted exome was 59x, with 81% covered at greater than 30x, and 96% covered at greater than 10x. Sequence alignment and variant calling revealed a total of 551,781 variants in 18,992 genes. Common variants, determined by a minor allele frequency (MAF) > 1% in any of the 1000 Genomes, National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) exomes, or Complete Genomics databases were excluded. This relatively generous cutoff of 1% was chosen to allow for the likelihood of low penetrance genes. After filtering out low quality and common variants, there were 219,889 in 18,458 genes remaining. Variants were then filtered based on predicted pathogenicity, including only those variants introducing stop codons, frameshifts, in-frame insertions and deletions, start site losses, splice site variants within two bases of an intron-exon boundary, and missense variants that were nonsynonymous and predicted to be pathogenic based on previous reports in the literature or by either the sorting intolerant from tolerant (SIFT) or PolyPhen-2 algorithms. This filtering resulted in a total of 34,973 variants in 12,551 genes, an average of 127 rare, predicted
pathogenic variants per individual, which is consistent with the number reported for the 1000 Genomes Project (11).

**Rare Variants in CDH-Causing Genes.** Given the heterogeneous nature of CDH genetics, we implemented a top-down strategy based on existing genetic evidence for CDH. Specifically, we began by searching for variants in 51 CDH-causing genes selected by stringent criteria, indicating involvement in CDH in humans \( (n = 11) \), and/or by the presence of a significant diaphragmatic phenotype (either hernia or thinning of the diaphragm muscle) in mouse models \( (n = 43) \) (**Dataset S2**). The targeted exome sequencing over these selected genes was high quality, with an average read depth of 66x. This analysis revealed a total of 154 rare, heterozygous, predicted pathogenic variants in 39 CDH-causing genes that were present in at least one of the 275 probands. Detailed information on each validated variant and proband phenotype is available in **Dataset S3**. The validation rate of selected variants was 84% by Sanger sequencing; false positives were primarily insertion/deletion calls (**Dataset S4**). The validated variants mapped to 34 genes, including six previously reported human CDH-causing genes and 28 CDH genes identified in mice. The number of variants detected in each of these genes and the evidence for their pathogenicity is depicted in **Fig. 1D**. Many were novel or ultrarare (MAF < 0.1%) (**Fig. 1F**). Additional algorithms supported the prediction of pathogenicity (**Dataset S4**). Functional domain information for protein variants is provided in **Dataset S3**.

We concluded that 85 of 275 CDH probands (30.8%) had at least one predicted pathogenic variant in one of the 51 CDH-causing genes. Eighteen of 275 (6.5%) had two, 6 (2%) had three, and 1 (0.4%) had four variants in CDH-causing genes (**Fig. 1C**).

**Variant Enrichment Analysis.** We determined whether the set of known CDH-causing genes in our cohort was enriched for predicted pathogenic variants compared with size-matched random gene sets. To control for different ancestries in our CDH cohort and control populations, we performed principal component analysis (PCA) on CDH cases and 1000 Genomes controls. Based on European ancestry assigned by common clustering, 195 CDH cases and 286 controls were selected (**Fig. 1F**).

We assessed the European subset of our CDH cohort for nonsense and predicted pathogenic missense variants (defined by a positive score by either SIFT or PolyPhen-2) in the 51 CDH-causing genes and compared them with 120 randomly generated gene sets that were size-matched based on the coding region of the longest predicted isoform. The burden of pathogenic variants in the CDH-causing genes was 3.5 SDs greater than the mean number of pathogenic variants found in the random gene sets (**Fig. 1E** \( P < 0.01 \), based on empirical distribution). The receptor LRP2 is encoded by a large and variable gene in normal populations. Therefore, we repeated the analysis excluding LRP2 and confirmed enrichment in the remaining genes (2.4 SD, \( P < 0.05 \)).

As a control, we performed the same analysis in individuals from the 1000 Genomes Project matched by ancestry and showed that the number of predicted pathogenic variants in CDH-causing genes was not significantly different from the random gene sets (**Fig. 1F**). Similarly, no enrichment was detected in individuals with self-reported European ancestry in the NHLBI ESP group \( (n = 4,300) \) (**Fig. 1G**).

**Inheritance Pattern.** Exome data were not available for unaffected family members. For this reason, the inheritance pattern of selected variants was determined by Sanger sequencing of parental DNA, available for ~55% of the probands. For variants in the 51 CDH-causing genes, 19 de novo variants were detected, and the remaining 63 variants were inherited from an unaffected parent (**Dataset S3**). The first de novo variant was in the zinc finger protein ZFP522 (FOG2) (c.89A > G; p.E30G), also described in a focused report (**Fig. 12**). The second was detected in the transcriptional coactivator EYA1 (c.164C > T; p.T55M), a previously reported pathogenic mutation causing Branchio-oto-renal syndrome 1 [BOR1; Online Mendelian Inheritance in Man (OMIM) no. 113650] (**Dataset S3**).

**Chromosomal Hotspots.** We filtered our exome results for variants in genes mapped within “hotspots” for chromosomal deletions in patients with CDH, i.e., regions reported in the literature to be deleted in two or more individuals with CDH. The following seven chromosomal regions were given priority because they contained breakpoints precisely defined by molecular cytogenetics: 1q41-q42.12, 4p16.3, 6p25.2-p25.3, 8p21.1, 8q23.3-q23.1, 15q26.1-q26.3, and 16p11.2 (13–17). Fifteen genes in these critical regions have been proposed to play a role in CDH, seven of which were also included in the analysis above as CDH-causing genes. We found rare and predicted pathogenic variants (by SIFT or PolyPhen-2) in all eight of the additional candidate genes from these regions (DISP1, FOX2, FOXC1, NEIL2, MEF2A, TBX6, ARRD4, and IGF1R), further substantiating their potential role in pathogenesis of CDH (**Table 1**).

<table>
<thead>
<tr>
<th>Pt</th>
<th>Gene</th>
<th>Variant</th>
<th>Co-occurring</th>
</tr>
</thead>
<tbody>
<tr>
<td>12†</td>
<td>DISP1</td>
<td>p.M1096T</td>
<td>—</td>
</tr>
<tr>
<td>148</td>
<td>DISP1</td>
<td>p.M1096T</td>
<td>CTBP2</td>
</tr>
<tr>
<td>215†</td>
<td>DISP1</td>
<td>p.M1096T</td>
<td>—</td>
</tr>
<tr>
<td>246</td>
<td>DISP1</td>
<td>p.M1096T</td>
<td>—</td>
</tr>
<tr>
<td>267†</td>
<td>DISP1</td>
<td>p.M1096T</td>
<td>LRP2</td>
</tr>
<tr>
<td>108†</td>
<td>DISP1</td>
<td>p.R1132Q</td>
<td>—</td>
</tr>
<tr>
<td>124†</td>
<td>DISP1</td>
<td>p.R1132Q</td>
<td>—</td>
</tr>
<tr>
<td>62</td>
<td>FOXF2</td>
<td>p.V210M</td>
<td>—</td>
</tr>
<tr>
<td>121</td>
<td>FOXF2</td>
<td>p.G425E</td>
<td>—</td>
</tr>
<tr>
<td>226†</td>
<td>FOXC1</td>
<td>p.A488_A489del</td>
<td>CHD7, NEDD4</td>
</tr>
<tr>
<td>105</td>
<td>NEIL2</td>
<td>p.R88fs</td>
<td>—</td>
</tr>
<tr>
<td>57†</td>
<td>NEIL2</td>
<td>p.S115C</td>
<td>—</td>
</tr>
<tr>
<td>128†</td>
<td>NEIL2</td>
<td>p.W142R</td>
<td>—</td>
</tr>
<tr>
<td>77</td>
<td>MEF2A</td>
<td>p.P279L</td>
<td>TBX6</td>
</tr>
<tr>
<td>77</td>
<td>TBX6</td>
<td>p.G162S</td>
<td>MEF2A</td>
</tr>
<tr>
<td>260</td>
<td>TBX6</td>
<td>p.R272Q</td>
<td>MYOD1, TBX6</td>
</tr>
<tr>
<td>260</td>
<td>TBX6</td>
<td>p.G162S</td>
<td>MYOD1, TBX6</td>
</tr>
<tr>
<td>90†</td>
<td>ARRDC4</td>
<td>p.Q139H</td>
<td>PBX3</td>
</tr>
<tr>
<td>186†</td>
<td>ARRDC4</td>
<td>p.E217K</td>
<td>—</td>
</tr>
<tr>
<td>127</td>
<td>IGF1R</td>
<td>p.D433Y</td>
<td>—</td>
</tr>
<tr>
<td>75†</td>
<td>IGF1R</td>
<td>p.V1021M</td>
<td>—</td>
</tr>
<tr>
<td>103</td>
<td>IGF1R</td>
<td>p.V1201I</td>
<td>—</td>
</tr>
</tbody>
</table>

Pt, patient. Additional information available as **Dataset S3**.

**Ultrarare and Highly Disruptive Variants.** Because mutations causing birth defects are likely to affect reproductive fitness, we analyzed...
our exome data for ultrarare (MAF < 0.1%) and highly disruptive variants and identified ultrarare nonsense (n = 876), splice site (n = 553), and frameshift (n = 1,226) SNVs in 2,815 genes. Of these genes, 30% also had missense variants that were predicted pathogenic by SIFT or PolyPhen-2 in two or more additional patients (representing at least 1% of the entire CDH cohort), increasing the likelihood for their involvement in CDH. Fourteen also overlapped by either the IBAS network or the PPF expression genes. Three were known to be associated with CDH (PDGFRα, ZFPM2, ILF3) and one is mapped to a CNV hotspot (NEIL2), whereas the remaining 10 represent previously unreported, or novel, CDH candidates (Fig. 2A and B). These ultrarare and highly disruptive variants were confirmed with Sanger sequencing.

Using the approaches described in this manuscript, we were able to identify a potentially relevant variant in ~42.2% of our patient cohort. These variants map to known CDH genes (30.9%), CNV candidate genes (5.8%), and candidates identified through the intersection of protein interaction networks and diaphragm expression profiles (3.3%), and genes with highly disruptive SNVs (2.2%) (Fig. 2C). Further, IBAS identified several coherent subnetworks of direct interactors predicted to be important in CDH, including muscle development, cell migration and signaling, and transcription factors with a specific role in cardiac development and blood vessel formation (Fig. 3).

**CDH-Causing Genes Are Enriched for Damaging Variants.** We report a catalog of rare and predicted pathogenic variants within CDH-causing genes, which are present in 30.9% of our study cohort, further substantiating their role in diaphragm development. Patients showed a significant enrichment for predicted pathogenic variation within these CDH-causing genes compared with random gene sets, unlike a control cohort matched by ancestry. This enrichment analysis was designed to minimize the impact of population stratification, because the CDH-causing gene set is compared against 120 size-matched random gene sets in the same cohort, acting as internal controls.

**Insights in the Genetics of CDH.** Our data are consistent with several proposed hypotheses about the genetic origins of CDH. First, we found multiple different genes affected across the CDH population, supporting the heterogeneous nature of the disease. Second, we identified rare and predicted pathogenic variants in two or more genes in multiple patients, raising the possibility of polygenic inheritance patterns in at least a subset of patients. However, this study does not allow us to assess the relative contribution of each variant within a given individual. Third, the majority of variants identified were inherited from an apparently unaffected parent, consistent with decreased penetrance for CDH. Additionally, we describe variants in eight CDH candidate genes mapped to CDH-associated chromosomal hotspots in patients. Finally, we used a bottom-up approach, including transcriptome and protein interaction data, and identified rare and predicted pathogenic variants in 14 novel CDH candidate genes.

**SNVs in CDH-Causing Genes Associated with Noncanonical Phenotypes.** Rare and predicted pathogenic variants were identified in the DNA-binding protein CHD7, which is responsible for the CHARGE syndrome (OMIM no. 214800), in seven patients who did not match the classical clinical phenotype. CHD7 variants in our cohort differ from those reported in known CHARGE syndrome cases, suggesting that different variants in CHD7 may result in either the complex syndrome or an increased likelihood of developing CDH.

We also report multiple heterozygous rare and predicted pathogenic variants in LRP2, a gene that causes the autosomal recessive Donnai-Barrow syndrome (DBS; OMIM no. 222448), characterized by CDH, corpus callosum agenesis, and eye and kidney phenotypes (19). None of the 14 individuals with heterozygous LRP2 variants had other findings consistent with a diagnosis of DBS. Therefore, we suggest that variable levels of LRP2 activity may result in a spectrum of disorders, with complete loss of both alleles causing DBS, whereas other heterozygous variants in LRP2 may result in a milder phenotype manifested as an increased susceptibility to CDH.
microtia, a phenotype consistent with BOR1 syndrome. We also reported in CDH patients (22). In particular, this proband had mutations in the cardiac, skeletal, and central nervous systems (24).

Kif7, a cilia-associated protein that modulates GLI transcription factor activity (23). Mice harboring homozygous mutations in the gene, affecting the GLI transcription factor activity (34, 35). Other affected pathways are muscle differentiation and GLI transcription activity (34, 35).

Diaphragm Developmental Expression Profiles and Protein Interaction Networks Identify CDH Candidate Genes. In the gene discovery phase of this study, we identified additional candidate genes by integrating exome sequencing data with (i) PFgene expression profiles and (ii) IBAS protein interaction networks based on known CDH genes. We further prioritized these genes by focusing on ultrarare (MAF < 0.1%) variants present in multiple individuals, one of which had to be a highly disruptive mutation. This approach, designed to enrich for genes more likely to be functionally relevant, uncovered candidate genes, which should be given high priority in functional studies to assess their role in diaphragm development (Figs. 2A and 3).

Network Analyses. Our network analyses also reveal several functional nodes that are likely to play key roles in the pathogenesis of CDH. RA signaling is a pathway known to be central to the pathogenesis of diaphragmatic defects (9), and many of the genes in which we describe rare and predicted pathogenic variants have been shown to interact directly or indirectly with the RA pathway. These candidates include Tgif1, PbX3, Runx1, and Zfhx4 (31–34).

The genes delineated in this study harbor rare variants in 42.2% in CDH patients, thus establishing an important foundation on which future human and animal model CDH research can be built. Integrated sequencing, developmental, and bioinformatics strategies could direct future functional studies on CDH, could be applied to cohorts and consortia for CDH and other birth defects, and could pave the way for potential therapies by providing molecular targets for drug discovery.

Methods

Patient Recruitment. Probands and family members were enrolled in the “Gene Mutation and Rescue in Human Diaphragmatic Hernia” study. Informed consent, blood, and tissue samples were obtained according to Partners Human Research Committee and Boston Children’s Hospital clinical investigation standards (Protocol 2000P000372 and 05-07-105R, respectively). Whenever possible, consented individuals underwent a physical examination by a geneticist and review of medical records.

Whole-Exome Sequencing. Samples were sequenced at the University of Washington, Seattle, by the NHLBI Resequencing and Genotyping Service (n = 92), at the Yale Center for Genome Analysis (n = 169), and at the Broad Institute (n = 14), on Illumina HiSeq. 2000 platforms. Sample libraries were captured on the SeqCap EZ Human Exome Library v2.0 (Roche NimbleGen).

Sequencing Data Analysis. Raw sequencing data for each individual were aligned to the human reference genome (build hg19) by using Burrows-Wheeler Aligner (BWA 0.7.5a). The alignment files were converted from a sequence alignment map (SAM) format to a sorted, indexed, binary
alignement map (BAM) file (SAMtools version 0.1.19; samtools.sourceforge.net) (36). To improve alignments and genotype calling, BAM files were realigned with The Genome Analysis Toolkit or GATK IndelRealigner (Broad Institute) (37, 38). Base quality scores were recalibrated, and duplicate reads were removed by the GATK base quality recalibration tool. SNP and indel detection were performed across all 275 samples using GATK for those variants that have minimum probability of incorrect base call equal to 1 in 100,000 (Phred Quality Score equal or greater than 50). The resulting Variant Call Format (VCF; version 4.0) files were imported to the Ingenuity Variant Analysis platform for further filtering.

**Principal Component Component Analysis**. Multidimensional scaling function, implemented in PLINK, was used to select ethnically matched individuals in the CDH and 1000 Genomes Project cohorts, using SNPs present in both groups, for enrichment analyses (39).

**Enrichment Analysis**. The enrichment of sequence variants in CDH-causing genes was determined against 120 random in silico-size-matched control gene sets (coding sequence ±10% as reported by ENSEMBL (www.ensembl.org)), using an approach that we call X RANGE (eXome RANDom Genesets Enrichment). Pathogenicity of missense variants was determined by using the Variant Effect Predictor version 73 [vepred] to run SIFT and Polyphen-2 (40, 41). The number of predicted pathogenic variants (either of these of these algorithms) was counted for the CDH-causing genes and each random control gene set. All variants, regardless of population frequency, were included in this analysis to eliminate bias due to the fact that population frequencies are based on the 1000 Genomes and NHLBI ESP datasets that were used as controls for the analysis.


**Network Analysis**. To uncover the proteins that interact with a list of known CDH-causing genes, we used iBAS (18), which is based on the updated human protein interaction network (inWeb) of ~430,000 interactions among 23,000 human proteins (42). First, iBAS was trained with 51 known CDH-causing genes curated from the literature. Parameters were optimized to predict interacting phenotype-causing proteins by testing each one against 10,000 random proteins. Furthermore, using these parameters, all candidates (i.e., 23,000 proteins) covered by interaction data were scored and ranked, and permutation tests were used for determining the significance of the observed scores.

DNA sample preparation and variant confirmation were performed as described in SI Methods.

**ACKNOWLEDGMENTS**. We thank The Association of Congenital Diaphragmatic Hernia Research, Awareness and Support (CHERUBS) for additional support; L. Keilmendorf, J. Kim, L. Luque-Bustamante, H. Al-Turkmani for technical assistance; and, for careful recruitment of patients, the surgeons at MassGeneral Hospital for Children and Boston Children’s Hospital: T. Buchwiller, C. C. Chen, D. Doody, S. J. Fishman, A. Goldstein, L. Holmes, T. Jakasic, R. Jennings, C. Kelleher, D. Lawlor, C.W. Lillehei, P. Masiakos, D. P. Mooney, K. Papadakis, R. Pieretti, M. Puder, D. P. Ryan, R. C. Shamberger, C. Smithers, J. Vacanti, and C. Weldon. We are grateful to the CDH support groups, CHERUBS and Breath of Hope, for providing families and caregivers information about our research study. Funding was provided by National Institute of Child Health and Human Development P01 HD08250-03 (to P.K.A.) and National Research Service Award 2T32GM007748-35 (to F.A.H.). Sequencing services were provided through the Resequencing and Genotyping Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under US Federal Government contract no. HHSN28210100037C from the NHLBI.
Supporting Information

Longoni et al. 10.1073/pnas.1412509111

SI Methods

DNA Sample Preparation. DNA samples were obtained by direct extraction from K3 EDTA preserved whole blood (QIAamp DNA Blood Maxi Kit; Qiagen), skin fibroblast primary cultures (TRIzol Reagent; Life Technologies), or EBV-transformed lymphoblastoid lines (QIAamp DNA Mini Kit; Qiagen), generated as described (1). DNA was concentrated and measured by picogreen (Quant-iT dsDNA Assay Kit, broad range and Qubit Fluorometric Quantitation; Life Technologies). Samples showing fragmented genomic DNA by electrophoresis in 1% agarose gels were purified and concentrated (DNA Clean & Concentrator-5; Zymo Research).

Variant Validation. Selected sequence variants were validated by using Sanger sequencing on proband and parental DNA samples, where available. Variant-specific primers were designed by using the National Center for Biotechnology Information Primer-BLAST (ncbi.nlm.nih.gov/tools/primer-blast). Sequencing products were resolved on the Applied Biosystems 3730xl DNA Analyzer.


Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)