Meta-analysis reveals transcription factors and DNA binding domain variants associated with congenital heart defect and orofacial cleft

Raehoon Jeong[1,2] and Martha L. Bulyk[1,2,3†]

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

[2] Bioinformatics and Integrative Genomics Graduate Program, Harvard University, Cambridge, MA 02138, USA.

[3] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

[†] Correspondence: mlbulyk@genetics.med.harvard.edu

Abstract

Many structural birth defect patients lack genetic diagnoses because there are many disease genes as yet to be discovered. We applied a gene burden test incorporating *de novo* predicted-loss-of-function (pLoF) and likely damaging missense variants together with inherited pLoF variants to a collection of congenital heart defect (CHD) and orofacial cleft (OC) parent-offspring trio cohorts (n = 3,835 and 1,844, respectively). We identified 17 novel candidate CHD genes and 10 novel candidate OC genes, of which many were known developmental disorder genes. Shorter genes were more powered in a "*de novo* only" analysis as compared to analysis including inherited pLoF variants. TFs were enriched among the significant genes; 14 and 8 transcription factor (TF) genes showed significant variant burden for CHD and OC, respectively. In total, 30 affected children had a *de novo* missense variant in a DNA binding domain of a known CHD, OC, and other developmental disorder TF genes. Our results suggest candidate pathogenic variants in CHD and OC and their potentially pleiotropic effects in other developmental disorders.

Introduction

Various structural birth defects, ranging from congenital heart defect (CHD) to orofacial cleft (OC), affect approximately 3% of births each year in the United States (1) and account for about 20% of infant mortality (2). CHD patients have abnormalities in the structure of the heart at birth (3), while OC patients have an opening in their lips or palates (4). Improved understanding of their genetic etiology will improve the accuracy of genetic diagnoses and guide potential disease-specific treatment strategies.

Transcription factors (TFs) play key roles in orchestrating differentiation and establishing cell identity during development (5,6). Genetic variants that damage TF function can cause various developmental disorders (7). Sequence-specific TFs control gene expression programs by binding to recognition sites in the genome and regulating the expression of their target genes. Missense variants in the DNA binding domains (DBDs) of TFs can alter DNA binding activity and cause a wide range of diseases, including Mendelian diseases (8). For example, many of the pathogenic variants in *NKX2-5* and *TBX5* for CHD, and *IRF6* for OC, are found in their DNA binding domains (9,10). We thus hypothesized that DBD variants in other TF genes might also cause CHD or OC. Furthermore, we hypothesized that DBD variants not yet found to be pathogenic but that occur in TFs with DBD variants previously found to cause CHD or OC, might also cause CHD or OC.

Searching for genetic causes underlying structural birth defects requires genetic data from patients. In recent years, the Gabriella Miller Kids First pediatric research program ("Kids First" from here on) funded efforts to sequence the genomes of patients as well as the family trios. Such family trio studies have been a primary strategy to discover disease genes for structural birth defects (11–13). The trio design is crucial in detecting *de novo* variants in probands and ascertaining rare pathogenic variants, as demonstrated by the Deciphering Developmental Disorders (DDD) study (14). Most probands for CHD and OC are sporadic cases with unaffected parents (100% for CHD cohorts and 95.3% for OC cohorts in this study). Therefore, in this study, we searched for *de novo* variants and rare inherited variants in the probands.

The aim of our study was two-fold. First, we sought to discover novel disease genes in CHD and OC since more causal genes likely remain to be found (8,12,13,15). We boosted power to discover novel disease genes by combining data from multiple cohorts across the spectrum of syndromic and non-syndromic cases for CHD and OC, respectively (12,13,16–18). We utilized the PrimateAI variant effect prediction tool (19) to identify missense variants likely to be

pathogenic more precisely than earlier studies (12,13). Furthermore, we applied the Transmission And *De novo* Association (TADA) (20) test to identify genes that show enrichment of putative damaging *de novo* inherited variants across different types of variant classes, such as missense and predicted loss-of-function (pLoF) variants (*i.e.*, nonsense, canonical splicing, and frameshift variants). This method has been successfully applied to discover potential autism genes (21).

Second, focusing on TFs because of their key roles in development and Mendelian diseases, we surveyed TFs and TF DBD variants for their potential association with CHD and OC. The resulting list of TFs and DBD variants are provided as a resource for future studies to evaluate whether they alter DNA binding activity (8,15).

Results

Genetic variants identified from multiple family trio cohorts of CHD and OC

To maximize power to discover novel disease genes, we combined genetic data from multiple CHD and, separately, OC cohorts. For CHD, we collected a non-redundant list of *de novo* variants and heterozygous predicted loss-of-function (pLoF) variants (*i.e.*, nonsense, canonical splicing, and frameshift variants) in probands from three prior studies (12,16,17), one of which is part of the Kids First program (17). In total, our list included variants from 3,835 family trios with a proband with CHD (**Supplementary Table 1**). For OC, we assembled genetic data from four Kids First cohorts (13,22) and the Deciphering Developmental Disorders (DDD) study (18), totaling 1,844 family trios (**Supplementary Table 1**). We combined those data with a list of *de novo* variants found in 757 family trios from Bishop *et al*. (11) and 603 family trios from Wilson *et al*. (18). For the Kids First cohort samples not analyzed in these two studies, we identified *de novo* variants from the whole-genome sequencing data using the slivar tool (23) (**Methods**).

Missense variant effect prediction methods prioritized putatively damaging variants

Missense variant effect prediction methods aim to score missense variants according to their likelihood of being benign or pathogenic (24–32). Disease genes are expected to be enriched for damaging, and not neutral, variants. Therefore, we compared ten variant effect prediction tools in order to select one that best differentiates potentially damaging variants from neutral ones in the context of structural birth defects. For this, we scored *de novo* variants in known CHD genes

(**Supplementary Table 2**) from CHD patients (12) (3,835 families with 113 variants) and unaffected siblings from an autism study (33) (2,179 families with 26 variants). We included unaffected siblings from an autism study because CHD cohorts did not have any genetic data from unaffected siblings and we can expect that unaffected siblings from an autism study likely did not have CHD diagnoses. Although these variants' pathogenicity has not all been resolved, we nonetheless expect many of the *de novo* variants from CHD patients to be pathogenic and most of those from the unaffected children in the autism study to be benign for CHD.

We compared the performance of the ten tools in discriminating the two sets of variants at various score thresholds (**Figure 1A**). We aimed to select a method that highly enriches potentially pathogenic variants at the top quantile. Overall, PrimateAI (19) showed the highest area under the curve metric for both receiver operator characteristic (ROC) and precision-recall (**Supplementary Figure 1**). Although Missense Variant Pathogenicity (MVP) (25) performed similarly well, the number of variants from unaffected children that were falsely classified as pathogenic was higher than that using PrimateAI. For instance, there were 13 and 4 predicted pathogenic variants out of 26 *de novo* variants from unaffected children over the score percentile threshold of 0.75, using MVP and PrimateAI, respectively. Moreover, since PrimateAI does not use any disease association information in model training, we anticipate it is less likely to show overfitting. Therefore, we used PrimateAI to infer the likelihood of missense variant pathogenicity in all subsequent analyses in this study.

Next, we determined score thresholds to classify all *de novo* missense variants. Based on the total missense mutation rate (~0.68 per generation), we inferred the expected number of *de novo* missense mutations in each 5% PrimateAI score bin. Then, we derived the enrichment of *de novo* missense variants in CHD versus control samples for each score bin **(Figure 1B**). The enrichment was more pronounced at the higher score bins. Therefore, we set two score thresholds: a stringent threshold of 0.9, and a more permissive, albeit still highly enriching, threshold of 0.75, to derive two groups of putatively damaging missense variants (PrimateAI $\geq$ 0.9 as MissenseA (MisA) and 0.75 $\leq$ PrimateAI $<$ 0.9 as MissenseB (MisB)). These two subsets were enriched among CHD samples but depleted among control samples (**Supplementary Figure 2**). Variants with lower PrimateAI scores showed neither enrichment nor depletion in these samples. This is consistent with enrichment of *de novo* missense variants predicted to be damaging in patients of CHD and autism (11,33). From here on, we considered *de novo* and inherited pLoF, *de novo* MisA,

and *de novo* MisB variants as putatively damaging. We used the same score thresholds for the analysis of the OC patient cohorts.

Detection of genes with enrichment of putatively damaging *de novo* and rare variants

Next, to identify candidate CHD and OC genes, we analyzed the *de novo* pLoF, MisA, and MisB variants and rare inherited pLoF variants using the transmission and *de novo* association (TADA) model (20). This model integrates enrichment of *de novo* variants based on a mutational model (34) and the enrichment of variants from cases compared to those from controls. The test calculates a Bayes factor that captures the enrichment of putatively damaging variants of different types. We considered 3,578 unaffected parents in an autism cohort as controls (12,35).

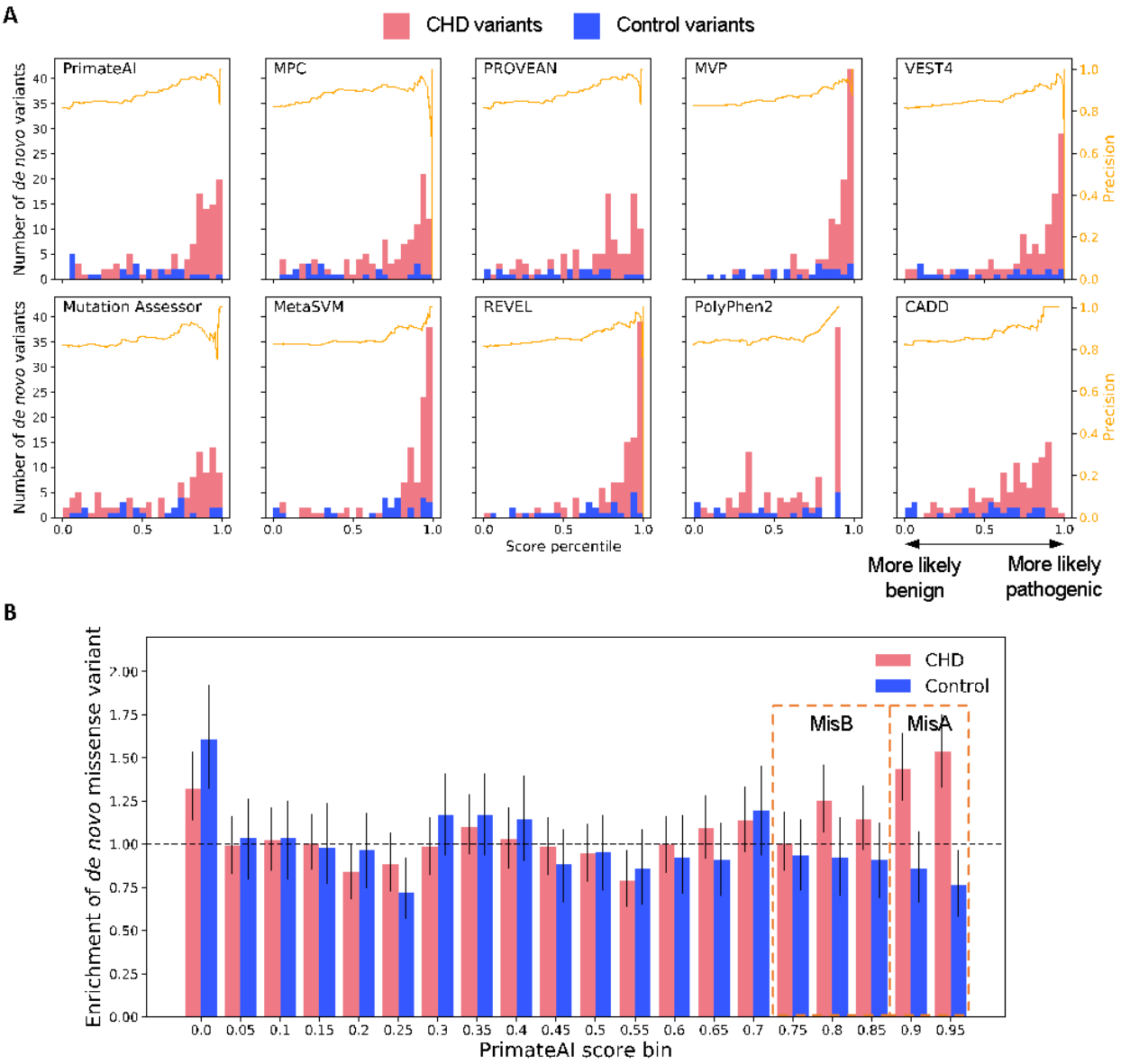
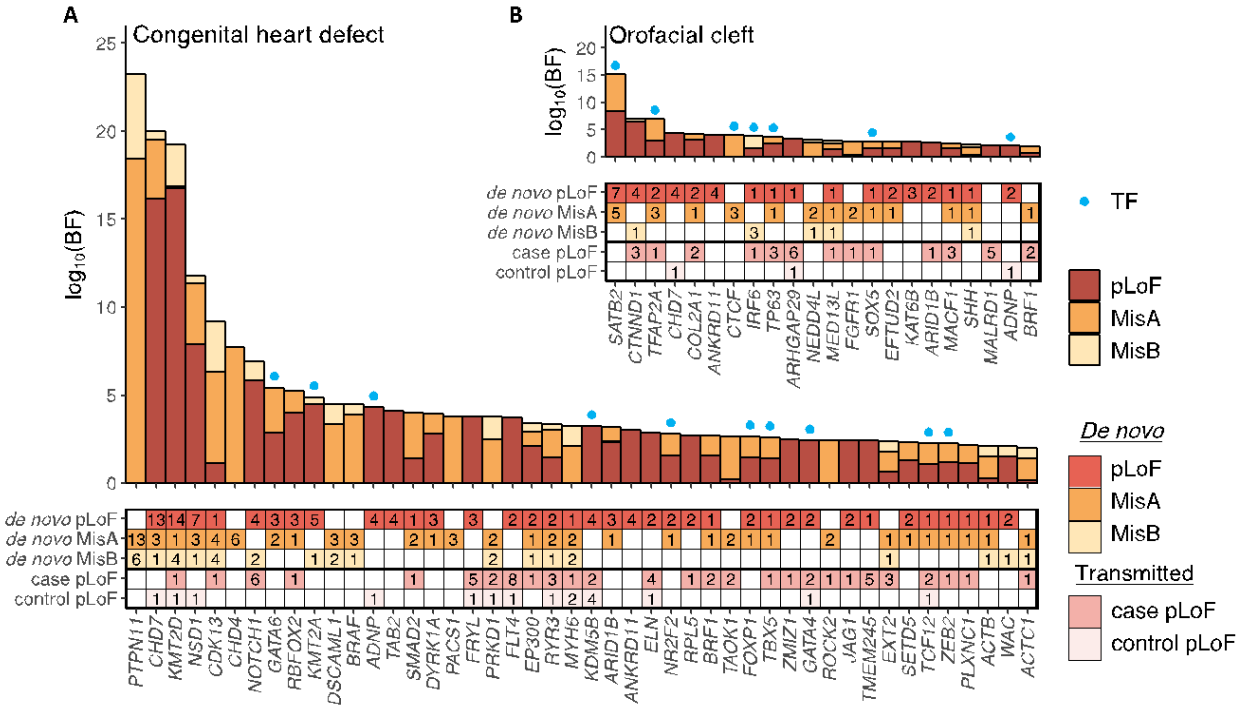


Figure 1. Comparison of missense variant prediction methods. (**A**) Number of variants in each score percentile bin, which corresponds to 5% increments, for ten missense variant effect predictions. Only *de novo* variants in 225 human CHD genes, which are listed in (**Supplementary Table 2**), are considered. The orange line depicts the precision at each percentile threshold. (**B**) Enrichment of missense variants in 5% PrimateAI score bins for all *de novo* variants in CHD patients and unaffected children. The error bars are 95% bootstrap confidence intervals. MisA, missense class A (PrimateAI ≥ 0.9); MisB, missense class B (0.75 < PrimateAI ≤ 0.9).

We detected 46 and 22 significant genes (q value < 0.1) for CHD and OC, respectively, of which some are known CHD or OC genes (**Supplementary Tables 2-5**). Since genes with no depletion of pLoF variants in a healthy population are not likely to be structural birth defect genes, we

137 excluded genes with a gnomAD (36) loss-of-function observed/expected upper bound fraction
138 (LOEUF) > 1. Most candidate genes had both pLoF and missense variants contributing to the
139 enrichment (**Figure 2**). Thus, integrating the variant types was useful in detecting candidate
140 disease genes.



141
142 **Figure 2. Bayes factor for each variant type's enrichment in candidate disease genes.** (**A** and
143 **B**) (Top) Bayes factor contribution by MisA, MisB, and pLoF variants in TADA for **(A)** CHD and
144 **(B)** OC in the "*de novo* + case/control" setting. Only positive Bayes factor contributions in
145 candidate genes (q value < 0.1) with LOEUF < 1 are displayed (CHD: 46 genes, OC: 22 genes).
146 (Bottom) Number of variants in each category. BF, Bayes factor; TF, transcription factor.
147

148 17 of the 46 genes identified in the CHD analysis cohorts were not known CHD genes (*i.e.*, not
149 significant in studies of individual cohorts and not annotated as CHD genes). 8 of the 22 genes
150 identified in the OC analysis cohorts were not known OC genes; known OC genes were taken from
151 the Genomics England PanelApp (37) 'Clefting' version 4.0 list. CHD and OC patients are at
152 higher risk for other congenital anomalies (38,39). Indeed, several of these genes are
153 developmental disorder genes, such as *TAOK1*, *WAC*, *PACS1*, *FOXP1*, *BRAF*, *SETD5*, and *ZMIZ1*
154 (phenotype MIM numbers: 619575, 616708, 615009, 613670, 613706, 615761, and 618659,
155 respectively). In a recent study on CHD (40), a *de novo* variant in *SETD5* was considered to be a
156 positive diagnosis. However, that study did not perform an enrichment analysis to identify novel

disease genes. Similarly, 7 of the 8 novel candidate OC genes – *MED13L*, *SOX5*, *KAT6B*, *ARID1B*, *MACF1*, *ADNP*, and *BRF1* – are linked to various developmental disorders (phenotype MIM numbers: 6616789, 616803, 616170, 135900, 618325, 615873, and 616202, respectively). These results are consistent with the known associations of CHD and OC with neurodevelopmental disorders (41,42).

More than half of the significant genes in CHD and OC showed probands with an inherited pLoF variant in the candidate disease gene (27 out of 46 for CHD and 13 out of 22 for OC). Two of the OC family trios (one with a *CTNND1* pLoF variant and another with an *AFHGAP29* pLoF variant) had an affected parent who passed on the pLoF variant. However, most inherited pLoF variants in candidate and known disease genes were inherited from unaffected parents, suggesting the possibility of incomplete penetrance.

***De novo* missense variants in CHD and OC genes**

Predicting the pathogenic effects of missense variants is challenging, and many are classified as variants of uncertain significance (VUSs) in ClinVar (43). Although we selected PrimateAI for this study, predictions by other methods can also be informative. As a resource for clinical researchers, we provide a table of predictions for the *de novo* missense variants identified in CHD and OC genes (**Supplementary Tables 6** and **7**). These tables include *de novo* missense variants in known CHD or OC genes (**Supplementary Table 2** and **3**) and candidate CHD or OC genes in the respective cohorts. In addition to scores from the tools we compared in **Figure 1**, we also include scores from the more recent AlphaMissense tool (44).

Coding sequence length affects which TADA model detects enrichment in a gene

To evaluate the utility of incorporating inherited pLoF variants in the case/control setting (*i.e.,* "*de novo* & case/control"), we compared against the enrichment obtained using just *de novo* variants with TADA (*i.e.*, "*de novo* only"). Surprisingly, using just the *de novo* variants yielded more candidate CHD genes (**Supplementary Table 4**) than using the "*de novo* & case/control" setting; 24 and 10 genes were exclusively significant in "*de novo* only" and "*de novo* & case/control" settings, respectively. The 24 genes that were significant (*i.e.*, TADA q value < 0.1 and LOEUF < 1) only in the "*de novo* only" setting had no rare inherited pLoF variants in the cohorts, which lowered the Bayes factor estimates when case/control data were incorporated. Since approximately

90% of these genes are highly constrained with LOEUF < 0.3 (*i.e.*, in approximately the top 10% of all protein-coding genes), pLoF variants in these genes are expected to be extremely rare in unaffected individuals. Since longer genes are expected to have more pLoF variants on average, we compared the lengths of genes unique to each setting. The coding sequence lengths of the 10 genes that were uniquely significant in the "*de novo* & case/control" model were significantly longer than those of the 24 genes uniquely significant in the "*de novo* only" model ($p = 0.019$, one-sided Wilcoxon rank-sum test; **Figure 3**). The LOEUF estimates of genes in the two sets were not significantly different ($P > 0.05$, Wilcoxon rank-sum test). We observed similar trends for candidate OC genes (**Supplementary Table 5** and **Supplementary Figure 3**). Altogether, these results demonstrate that the coding sequence length of genes affects their identification as significant disease genes by the "*de novo* only" versus the "*de novo* & case/control" TADA model. This effect is likely because longer genes have a greater chance that pLoF variants are present in a population and inherited, thereby contributing to increased enrichment in the "*de novo* & case/control" setting; in contrast, shorter genes have a lower expected mutation rate for pLoF variants, thus each *de novo* variant contributes to a greater amount of enrichment.

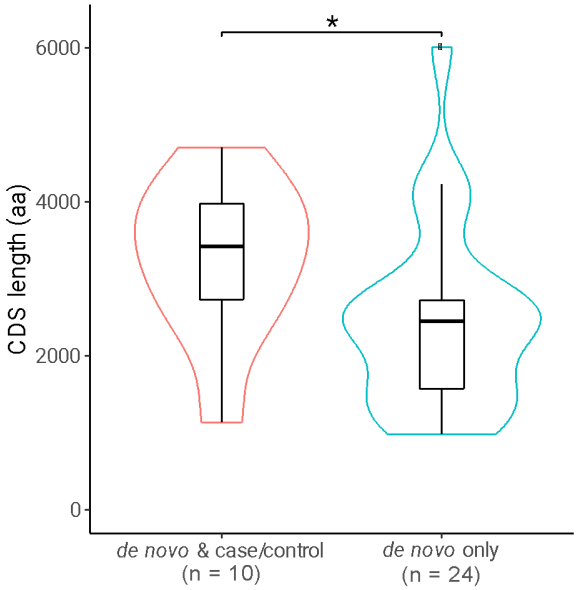


Figure 3. Coding sequence length of significant CHD genes by discovery model. Distribution of coding sequence length for the significant genes unique to the "*de novo* & case/control" model and "*de novo* only" model. The number of genes is labeled below each category. CDS, coding sequence; aa, amino acid. * $P < 0.05$, one-sided Wilcoxon rank-sum test.

Table 1. Transcription factors significantly enriched for predicted deleterious *de novo* variants.

LOEUF, loss-of-function observed/expected upper bound fraction (36); pLoF, predicted loss-of-function; MisA, PrimateAI > 0.9; MisB, PrimateAI 0.75-0.9. [a] novel candidate CHD genes. [b] novel candidate OC genes.

Gene	LOEUF	*de novo variants*			Inherited variants		*de novo* & case/control q value	*de novo* only q value
		pLoF	MisA	MisB	Case pLoF	Control pLoF		
Congenital heart defect								
GATA6	0.174	3	2	0	0	0	2.5×10^{-5}	3.5×10^{-6}
KMT2A	0.065	5	0	1	0	0	3.2×10^{-4}	5.0×10^{-5}
ADNP	0.123	4	0	0	0	1	5.8×10^{-4}	3.1×10^{-4}
KDM5B[a]	0.572	4	0	0	2	4	0.012159	5.2×10^{-4}
NR2F2	0.217	2	1	0	0	0	0.014122	0.002103
FOXP1[a]	0.175	2	1	0	0	0	0.029404	0.003100
TBX5	0.135	1	1	0	1	0	0.031389	0.053298
GATA4	0.527	2	0	0	2	1	0.040077	0.050796
TCF12	0.372	1	1	0	2	1	0.051542	0.068473
ZEB2	0.107	1	1	0	1	0	0.058741	0.131609
KLF2[a]	0.710	1	1	0	0	0	0.204573	0.032261
SMAD4	0.222	0	2	0	0	0	0.209108	0.035057
MEIS2[a]	0.184	2	0	0	0	0	0.271015	0.055872
CTCF[a]	0.148	0	2	0	0	0	0.278293	0.058374
Orofacial cleft								
SATB2	0.091	7	5	0	0	0	3.86×10^{-14}	5.77×10^{-15}
TFAP2A	0.261	2	3	0	1	0	2.84×10^{-6}	7.70×10^{-6}
CTCF[b]	0.148	0	3	0	0	0	0.011737	0.001484
IRF6	0.132	1	0	3	1	0	0.002951	0.007637
TP63	0.267	1	1	0	3	0	0.003631	0.072430
SOX5[b]	0.188	1	1	0	1	0	0.018728	0.058691
ADNP[b]	0.123	2	0	0	0	1	0.092444	0.088307
GRHL2[b]	0.270	2	0	0	0	0	0.328840	0.076571

TF DBD variants identified in candidate CHD and OC disease genes

Because of the known role of TFs in CHD (45) and OC (46), we examined how many significant genes from our analysis were TFs (47). For CHD, there were 14 TFs that showed significant enrichment in either "*de novo* & case/control" or "*de novo* only" analysis (**Table 1**, **Figure 2**). For OC, 7 TFs showed significant enrichment (**Table 1**, **Figure 2**). For both CHD and OC, TFs were

219 significantly enriched among the significant genes (p = 0.006 and p = 0.016, respectively, one-
220 sided Fisher's exact test).

221 There were 5 and 3 candidate CHD and OC TF genes, respectively, that are not yet
222 established CHD or OC disease genes. For CHD, we identified *KDM5B*, *FOXP1*, *KLF2*, *MEIS2*,
223 and *CTCF*. For OC, we identified *SOX5, ADNP,* and *GRHL2*. Two candidate CHD TF genes –
224 *KDM5B* and *FOXP1* – were also statistically implicated in a similar CHD study (48) that
225 aggregated *de novo* variants from two (12,16) of the 3 studies that we analyzed. Nevertheless,
226 *KDM5B*, *FOXP1*, *MEIS2*, and *CTCF* are known developmental disorder genes (phenotype MIM
227 numbers: 618109, 613670, 600987, and 615502, respectively). Some children with mutations in
228 these genes have been reported to show heart defects (49–52). *KLF2* has not been directly
229 associated with CHD, but its zebrafish homologue *klf2* is required for heart valve formation (53).
230 A non-coding variant that causes over-expression of *Grhl2* in mice led to orofacial cleft
231 phenotypes (54).

232 Since DNA binding activity plays a crucial role in TF function, we searched for TF DBD
233 missense variants in known developmental disorder genes. We developed a pipeline to filter for
234 missense variants in the TF DBDs based on a set of 62 DBD classes in the Pfam database (55)
235 (**Supplementary Table 8**) and the protein domain prediction model HMMer (56). Without
236 filtering for disease genes, there were 46 and 11 *de novo* TF DBD missense variants in the CHD
237 and OC cohorts, respectively (**Supplementary Table 9**); with filtering, there were 17 and 13 DBD
238 missense variants, respectively (**Table 2**). Some of these variants are in CHD, OC, and other
239 developmental disorder genes that are mostly haploinsufficient, characterized by low LOEUF
240 estimates (**Table 2**). Based on PrimateAI, they were all predicted to be pathogenic (PrimateAI
241 rank score > 0.8). We hypothesize that these variants damage the TFs' DNA binding activity.
242

Table 2. *De novo* TF DBD missense variants in genes associated with CHD, OC, or developmental disorder genes.

The table lists *de novo* TF DBD variants from our analysis in genes that are either significantly enriched in our study (marked with an asterisk [*]) or are reported as CHD, OC, or developmental disorder genes. For developmental disorders, the specific syndrome is written in parentheses. PrimateAI rank score is a percentile score (range 0-1) based on the raw PrimateAI score. CAKUT, congenital anomalies of kidney and urinary tract; CDH, congenital diaphragmatic hernia; ETS, erythroblast transformation specific; IRF, interferon regulatory factor; AP-2, activator protein 2; EEC, Ectrodactyly, ectodermal dysplasia, and cleft lip/palate. [a] Candidate CHD gene based on damaging variant enrichment. * Significant enrichment of damaging variants in this study.

Developmental disorder	Gene	LOEUF	Amino acid change	PrimateAI rank score	Variant	DBD (Pfam ID)
Congenital heart defect						
CHD	*FOXP1**	0.175	F499L	0.99469	3:70976974:A:T	Forkhead (PF00250)
CHD (Axenfeld-Rieger syndrome)	*FOXC1*	0.311	T88I	0.94564	6:1610708:C:T	Forkhead (PF00250)
CHD (Wiedemann-Steiner syndrome)	*KMT2A*	0.065	K1186E	0.87072	11:118478188:A:G	CXXC zinc finger (PF02008)
CHD (Holt-Oram syndrome)	*TBX5**	0.135	I227T	0.98142	12:114385551:A:G	T-box (PF00907)
CHD	*TCF12**	0.372	H631Q	0.98114	15:57273177:C:G	Helix-loop-helix (PF00010)
CHD	*NR2F2**	0.217	C96F	0.98292	15:96332392:G:T	C4 zinc finger (PF00105)
CHD	*GATA6**	0.174	R456G	0.90881	18:22181516:C:G	GATA zinc finger (PF00320)
CHD	*GATA6**	0.174	R456H	0.92717	18:22181517:G:A	GATA zinc finger (PF00320)
CHD[a]	*KLF2**	0.71	C334Y	0.99874	19:16326964:G:A	C2H2 zinc finger (PF00096)
CHD (DiGeorge syndrome)	*TBX1*	0.427	L293F	0.98054	22:19765767:C:T	T-box (PF00907)
CAKUT	*PBX1*	0.255	R235Q	0.95192	1:164807544:G:A	Homeodomain (PF00046)
CAKUT	*TBX18*	0.193	T305A	0.81286	6:84747946:T:C	T-box (PF00907)
CDH (Cardiac-urogenital syndrome)	*MYRF*	0.117	Q403H	0.8663	11:61774060:G:C	NDT80 / PhoG (PF05224)
CDH (Cardiac-urogenital syndrome)	*MYRF*	0.117	L479V	0.86641	11:61776368:C:G	NDT80 / PhoG (PF05224)
Den Hoed-de Boer-Voisin syndrome	*SATB1*	0.293	E547K	0.96969	3:18352132:C:T	CUT (PF02376)
Speech language disorder	*FOXP2*	0.219	R553H	0.9789	7:114662075:G:A	Forkhead (PF00250)
Craniosynostosis	*ERF**	0.261	K96N	0.96845	19:42249912:C:A	ETS (PF00178)
Orofacial cleft						

OC (van der Woude syndrome)	*IRF6**	0.132	N88D	0.86413	1:209796465:T:C	IRF (PF00605)
OC (van der Woude syndrome)	*IRF6**	0.132	R84H	0.84067	1:209796476:C:T	IRF (PF00605)
OC (Glass syndrome)	*SATB2**	0.091	R667G	0.94148	2:199272414:G:C	Homeodomain (PF00046)
OC (Glass syndrome)	*SATB2**	0.091	R399H	0.90829	2:199328888:C:T	CUT (PF02376)
OC (Glass syndrome)	*SATB2**	0.091	L394S	0.95427	2:199328903:A:G	CUT (PF02376)
OC (Glass syndrome)	*SATB2**	0.091	R389L	0.9951	2:199348708:C:A	CUT (PF02376)
OC (Glass syndrome)	*SATB2**	0.091	R389C	0.99811	2:199348709:G:A	CUT (PF02376)
Lamb-Shaffer syndrome	*SOX5*	0.188	H582Y	0.9764	12:23543238:G:A	HMG_box (PF00505)
OC	*TFAP2A**	0.261	R256Q	0.921	6:10404511:C:T	AP-2 (PF03299)
OC	*TFAP2A**	0.261	S249L	0.98055	6:10404532:G:A	AP-2 (PF03299)
OC (EEC syndrome)	*TP63*	0.267	C347F	0.94957	3:189868627:G:T	P53 (PF00870)
Holoprosencephaly	*SIX3*	0.323	W253R	0.99697	2:44942861:T:A	Homeodomain (PF00046)
Ayme-Gripp syndrome	*MAF*	0.537	R294W	0.99834	16:79599023:G:A	bZIP_MAF (PF03131)

Discussion

We aggregated multiple parent-offspring trio cohorts of CHD and OC to detect 46 and 22 genes, respectively, with enrichment of damaging *de novo* variants and inherited pLoF variants. Of those, 17 were novel candidate CHD genes and 10 were novel candidate OC genes (**Supplementary Tables 3** and **4**). Further studies are needed to validate which of these are true disease genes for CHD and OC. Increasing the sample sizes of family trio cohorts will be key to discovering more candidate disease genes; however, thousands of family trios are still insufficient to discover most of the disease genes. As there are likely hundreds of genes causing these structural birth defects, the likelihood of observing multiple cases with damaging *de novo* variants in the same gene is still low. Kaplanis and colleagues estimated that sequencing hundreds of thousands of parent-offspring trios will be necessary to reach sufficient power to detect about 80% of developmental disorder genes based on analysis of *de novo* variants (14).

We evaluated the performance of multiple missense variant effect prediction methods to prioritize candidate pathogenic variants. While most methods were able to discriminate *de novo* missense variants in CHD genes found in CHD patients from those found in unaffected children,

PrimateAI was the most effective and led to the identification of more *de novo* missense variants. We also provide a list of *de novo* missense variants in known and candidate CHD and OC genes as a resource (**Supplementary Tables 6** and **7**).

Incorporating the number of inherited pLoF variants in cases and controls into enrichment analyses led to some significant genes not reaching significance with *de novo* variants alone. However, in the current sample size, there were many genes with no inherited pLoF variants, and many of them were only significant in the "*de novo* only" analysis. These genes were generally shorter than the genes identified uniquely by the "*de novo* & case/control" analysis, suggesting that gene length affects which model may be better powered. Moreover, applying both the "*de novo* only" and the "*de novo* & case/control" model is useful for detecting as many candidate disease genes as possible.

In this study, we analyzed only pLoF and missense variants. Copy number variations (CNVs) that increase or decrease gene dosage also play a role in structural birth defects (57). Therefore, calling *de novo* and inherited CNVs in the affected children and testing their enrichment in individual genes will increase the chance of disease gene discovery in future studies (21). In terms of inherited variants, we considered only pLoF variants because the effects of missense variants are more difficult to predict. Including inherited missense variants in the model may potentially increase power, but ensuring high precision in pathogenicity prediction will be essential.

TFs were enriched among the identified genes. We identified many *de novo* TF DBD missense variants in genes that were significantly enriched in CHD or OC or that are known CHD, OC, or developmental disorder genes. The identified variants were predicted to be pathogenic by PrimateAI. Some of the TFs with TF DBD variants in the CHD cohort are known to cause other developmental disorders, such as congenital diaphragmatic hernia and congenital anomalies of kidneys and urinary tract (58,59). These results suggest that these TFs are pleiotropic and that other mutations in them may cause heart defects in some patients.

Variant effect prediction tools are only moderately accurate, at best, in distinguishing TF DBD missense variants with altered DNA binding activity (15). Future studies using DNA binding assays, such as protein binding microarrays (PBMs) (8,60), will be needed to determine which of the identified CHD and OC variants alter DNA binding activity and in what manner they do so.

Methods

Genetic data from family trio cohorts of CHD and OC

We aggregated multiple datasets to maximize statistical power to detect disease genes. For CHD, we downloaded *de novo* variant data from two exome-sequencing studies (12,16) and one genome-sequencing study (17). We also downloaded the list of rare inherited pLoF variants from Jin *et al*. (12). We identified overlapping samples by comparing the set of *de novo* variants from each proband. After removing duplicate samples, there were a total of 3,835 unique family trios.

For OC, we downloaded genotype data from 4 cohorts from the Gabriella Miller Kids First data portal (61). Their database of Genotypes and Phenotypes (dbGaP) IDs were phs001168 (n = 376 trios), phs001997 (n = 404 trios), phs001420 (n = 262 trios), and phs002595 (n = 351 trios). In addition, we downloaded a list of *de novo* variants from 374 European (phs001168), 267 Colombian (phs001420), and 116 Taiwanese (phs001997) family trios from Table S3 of Bishop *et al*. (13). We also downloaded a list of *de novo* variants from 603 family trios from Table S4 of Wilson *et al*. (18). We downloaded *de novo* variant data from unaffected siblings in families in an autism cohort (33) to compare variant enrichment statistics. Lastly, we downloaded heterozygous pLoF variants from 3,578 unaffected parents in an autism cohort as controls (12,35). We analyzed all genetic variants based on the GRCh38 human reference genome. The downloaded variants in hg19 were lifted over to the GRCh38 human reference. We performed variant calling and curation just for the 484 OC samples not included in Bishop *et al*. (13).

Identifying *de novo* variants and rare inherited variants in the OC cohorts

For the samples not included in Bishop *et al*. (13) (n = 484), we applied different strategies for identifying *de novo* predicted-loss-of-function (pLoF) and missense variants. pLoF variants consist of nonsense, splice site, and frameshift variants. Since trio-based variant calls (*i.e.*, vcf files) provided in the Gabriella Miller Kids First data portal (61) showed false negatives in *de novo* single nucleotide variants (SNVs), we derived *de novo* SNVs based on the gvcf files of the three family members in each trio.

For SNVs, which span pLoF and missense variants, we identified *de novo* variants by 1) merging gvcf files of the three family members in each trio using GLNexus (62) with the ‘gatk’ setting and 2) using slivar (23) to filter for variants that are heterozygous in the proband but

homozygous reference in the two parents. We further filtered for those with the maximum population allele frequency in gnomAD (36) of less than 5×10^{-5}, no homozygous individuals in gnomAD, and TOPMed (63) allele frequency of less than 5×10^{-5}.

In contrast, we used *de novo* insertions and deletions (indels) identified in the trio-based variant calls. For indel pLoF variants, we 1) downloaded the family-based vcf files from the Gabriella Miller Kids First data portal and 2) filtered for variants that are heterozygous in the proband but homozygous reference in the two parents using slivar (23). The variants were filtered for having genotype quality (GQ) greater than 20 and read depth (DP) greater than 6. We also filtered for those with a maximum population allele frequency in gnomAD (36) of less than 5×10^{-5}, no homozygous individuals in gnomAD, and TOPMed (63) allele frequency of less than 5×10^{-5}.

For all OC samples, we identified rare inherited pLoF variants by filtering for variants with a heterozygous genotype in the proband and only one parent with a heterozygous genotype using the family-based vcf files from the Gabriella Miller Kids First data portal. We also filtered for those with the maximum population allele frequency in gnomAD (36) of less than 5×10^{-5}, no homozygous individuals in gnomAD, and TOPMed (63) allele frequency of less than 5×10^{-5}.

Comparison of missense variant effect prediction methods

We compared the performance of ten missense variant effect prediction methods: PrimateAI (19), MPC (31), PROVEAN (26), MVP (25), VEST4 (30), MutationAssessor (32), MetaSVM (28), REVEL (29), PolyPhen2 (24), and CADD (27). These tools' scores for missense variants were accessed from the database for nonsynonymous SNPs' functional predictions (dbNSFP) version 4.5 (64). To compare between scores easily, we utilized the rank scores, which range from 0 to 1 and correspond to the percentile among missense variants. We compared their performance in discriminating *de novo* missense variants in CHD genes (**Supplementary Table 2**) from CHD patients from those from unaffected children. There were a total of 3,836 CHD family trios (12,16,17) and 2,179 control family trios (33) that carried 113 and 26 *de novo* variants in CHD genes, respectively. We computed their area under the curve for receiver operator characteristic (ROC) and precision-recall to compare their performance.

Next, we determined the appropriate PrimateAI score thresholds for potentially damaging variants. Across all genes, we estimated the enrichment of *de novo* missense variants for CHD

356 families and control families in each of the 5% score bins. The expected number of *de novo*
357 missense variants per family was the sum of all missense mutation rates (~ 0.68 per generation).
358 Then, we bootstrapped sampled CHD and control families to establish the respective 95%
359 confidence intervals of the enrichment estimates. Ultimately, based on **Figure 1B**, we selected
360 PrimateAI ≥ 0.9 and $0.75 \leq$ PrimateAI < 0.9 as the two missense variant groups – MisA and MisB.

361 **Testing enrichment of damaging *de novo* and rare inherited variants**

362 We used the TADA model (20) to detect genes with an enrichment of potentially damaging
363 variants (*i.e.* predicted-loss-of-function (pLoF), missense with PrimateAI (19) rank score ≥ 0.9
364 (MisA), or missense with PrimateAI rank score 0.75~0.9 (MisB)) from the number of *de novo*
365 variants and mutation rate estimates. We derived the per-gene mutation rates for MisA, MisB, and
366 pLoF based on estimates in Samocha *et al.* (34) and gnomAD (36). We multiplied the per-gene
367 missense mutation rate $\mu_{\text{Mis, gene}}$ by 0.1 and 0.15, to derive $\mu_{\text{MisA, gene}}$ and $\mu_{\text{MisB, gene}}$, respectively, as
368 all possible MisA and MisB variants are expected be 0.1 and 0.15 of all missense variants. We
369 added the per-gene nonsense, splice site and frameshift mutation rates to derive the per-gene pLoF
370 mutation rates.

371 We applied TADA to 17,488 autosomal genes with LOEUF estimates in gnomAD (36).
372 We performed the test once including inherited pLoF variants and once without to compare the
373 effect of inherited variants. Multiple hypothesis correction across all genes was applied using the
374 q value estimates. We considered genes with q value < 0.1 and gnomAD's LOEUF < 1 to be
375 significant. We excluded genes with LOEUF ≥ 1 because it suggests that there is negligible
376 selective constraint against predicted-loss-of-function variants in those genes.

377 **Identifying TF DBD variants in candidate disease genes**

378 We identified disease-associated TF genes based on a list of 1,639 TFs (47). Then, we
379 determined the location of the DBDs using a set of 62 DBD classes in the Pfam database version
380 35.0 (55) (**Supplementary Table 5**) and the protein domain prediction model HMMer (56). We
381 considered only canonical transcripts and amino acid sequences based on GENCODE (65) in
382 annotating whether the missense variants fall within a DBD.

383 **Data availability**

For CHD, we downloaded *de novo* variant data from two exome-sequencing studies (12,16) and one genome-sequencing study (17). We also downloaded the list of rare inherited pLoF variants from Jin *et al*. (12). For OC, we downloaded genotype data from 4 cohorts from the Gabriella Miller Kids First data portal (61). Their database of Genotypes and Phenotypes (dbGaP) IDs were phs001168, phs001997, phs001420, and phs002595.

Code availability

Code and data for generating the figures is available at https://github.com/BulykLab/CHD-OC-manuscript-figures.

Acknowledgments

We thank members of the Bulyk lab for helpful discussion. This work was funded by NIH grants R03 HD099358 and R01 HG010501 to M.L.B.

Author Contributions

R.J. and M.L.B. conceived and designed the research project. R.J. performed all analyses and prepared the figures. M.L.B. supervised the research. R.J. and M.L.B. wrote the manuscript. Both authors reviewed the manuscript.

Ethics Declarations

The authors declare no competing interests.

References

1. Centers for Disease Control and Prevention (CDC) (2008) Update on overall prevalence of major birth defects--Atlanta, Georgia, 1978-2005. *MMWR. Morb. Mortal. Wkly. Rep.*, **57**, 1–5.

2. Ely, D.M. and Driscoll, A.K. (2023) Infant Mortality in the United States, 2021: Data From the Period Linked Birth/Infant Death File. *Natl. Vital Stat. Rep.*, **72**, 1–19.

3. Mitchell, S.C., Korones, S.B. and Berendes, H.W. (1971) Congenital heart disease in 56,109 births. Incidence and natural history. *Circulation*, **43**, 323–32.

4. Watkins, S.E., Meyer, R.E., Strauss, R.P. and Aylsworth, A.S. (2014) Classification, epidemiology, and genetics of orofacial clefts. *Clin. Plast. Surg.*, **41**, 149–63.

5. Spitz, F. and Furlong, E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–26.

6. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.

7. Seidman, J.G. and Seidman, C. (2002) Transcription factor haploinsufficiency: when half a loaf is not enough. *J. Clin. Invest.*, **109**, 451–5.

8. Barrera, L.A., Vedenko, A., Kurland, J. V., Rogers, J.M., Gisselbrecht, S.S., Rossin, E.J., Woodard, J., Mariani, L., Kock, K.H., Inukai, S., *et al.* (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, **351**, 1450–1454.

9. Su, W., Zhu, P., Wang, R., Wu, Q., Wang, M., Zhang, X., Mei, L., Tang, J., Kumar, M., Wang, X., *et al.* (2017) Congenital heart diseases and their association with the variant distribution features on susceptibility genes. *Clin. Genet.*, **91**, 349–354.

10. Kondo, S., Schutte, B.C., Richardson, R.J., Bjork, B.C., Knight, A.S., Watanabe, Y., Howard, E., de Lima, R.L.L.F., Daack-Hirsch, S., Sander, A., *et al.* (2002) Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat. Genet.*, **32**, 285–9.

11. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R.,

McKean, D., Wakimoto, H., Gorham, J., *et al.* (2015) De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science*, **350**, 1262–6.

12. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., *et al.* (2017) Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.*, **49**, 1593–1601.

13. Bishop, M.R., Diaz Perez, K.K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., Hetmanski, J.B., Taub, M.A., Moreno-Uribe, L.M., Valencia-Ramirez, L.C., *et al.* (2020) Genome-wide Enrichment of De Novo Coding Mutations in Orofacial Cleft Trios. *Am. J. Hum. Genet.*, **107**, 124–136.

14. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., *et al.* (2020) Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, **586**, 757–762.

15. Kock, K.H., Kimes, P.K., Gisselbrecht, S.S., Inukai, S., Phanor, S.K., Anderson, J.T., Ramakrishnan, G., Lipper, C.H., Song, D., Kurland, J. V, *et al.* (2024) DNA binding analysis of rare variants in homeodomains reveals homeodomain specificity-determining residues. *Nat. Commun.*, **15**, 3110.

16. Sifrim, A., Hitz, M.-P., Wilsdon, A., Breckpot, J., Turki, S.H. Al, Thienpont, B., McRae, J., Fitzgerald, T.W., Singh, T., Swaminathan, G.J., *et al.* (2016) Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.*, **48**, 1060–5.

17. Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., *et al.* (2020) Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.*, **52**, 769–777.

18. Wilson, K., Newbury, D.F. and Kini, U. (2023) Analysis of exome data in a UK cohort of 603 patients with syndromic orofacial clefting identifies causal molecular pathways. *Hum. Mol. Genet.*, **32**, 1932–1942.

19. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., *et al.* (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.*, **50**, 1161–1170.

20. 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., *et al.* (2013) Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *PLoS Genet.*, **9**.

21. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., *et al.* (2022) Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.*, **54**, 1320–1331.

22. Awotoye, W., Mossey, P.A., Hetmanski, J.B., Gowans, L.J.J., Eshete, M.A., Adeyemo, W.L., Alade, A., Zeng, E., Adamson, O., Naicker, T., *et al.* (2022) Whole-genome sequencing reveals de-novo mutations associated with nonsyndromic cleft lip/palate. *Sci. Rep.*, **12**, 11743.

23. Pedersen, B.S., Brown, J.M., Dashnow, H., Wallace, A.D., Velinder, M., Tristani-Firouzi, M., Schiffman, J.D., Tvrdik, T., Mao, R., Best, D.H., *et al.* (2021) Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ genomic Med.*, **6**, 60.

24. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–9.

25. Qi, H., Zhang, H., Zhao, Y., Chen, C., Long, J.J., Chung, W.K., Guan, Y. and Shen, Y. (2021) MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.*, **12**.

26. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, **7**, e46688.

27. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

28. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. and Liu, X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–37.

29. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., *et al.* (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.*, **99**, 877–885.

30. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. and Karchin, R. (2013) Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, **14**.

31. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M. and Daly, M.J. (2017) Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353.

32. Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

33. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., *et al.* (2020) Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, **180**, 568-584.e23.

34. 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., *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.

35. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.-X., *et al.* (2015) Excess of rare, inherited truncating mutations in autism. *Nat. Genet.*, **47**, 582–8.

36. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

37. Martin, A.R., Williams, E., Foulger, R.E., Leigh, S., Daugherty, L.C., Niblock, O., Leong, I.U.S., Smith, K.R., Gerasimenko, O., Haraldsdottir, E., *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.

38. Egbe, A., Lee, S., Ho, D., Uppu, S. and Srivastava, S. (2014) Prevalence of congenital anomalies in newborns with congenital heart disease diagnosis. *Ann. Pediatr. Cardiol.*, **7**, 86–91.

39. Stoll, C., Alembik, Y. and Roth, M.-P. (2022) Co-occurring anomalies in congenital oral clefts. *Am. J. Med. Genet. A*, **188**, 1700–1715.

40. Hartill, V., Kabir, M., Best, S., Shaikh Qureshi, W.M., Baross, S.L., Lord, J., Yu, J., Sasaki, E.,

Needham, H., Shears, D., *et al.* (2024) Molecular diagnoses and candidate gene identification in the congenital heart disease cohorts of the 100,000 genomes project. *Eur. J. Hum. Genet.*

41. Marino, B.S., Lipkin, P.H., Newburger, J.W., Peacock, G., Gerdes, M., Gaynor, J.W., Mussatto, K.A., Uzark, K., Goldberg, C.S., Johnson, W.H., *et al.* (2012) Neurodevelopmental outcomes in children with congenital heart disease: evaluation and management: a scientific statement from the American Heart Association. *Circulation*, **126**, 1143–72.

42. Tillman, K.K., Hakelius, M., Höijer, J., Ramklint, M., Ekselius, L., Nowinski, D. and Papadopoulos, F.C. (2018) Increased Risk for Neurodevelopmental Disorders in Children With Orofacial Clefts. *J. Am. Acad. Child Adolesc. Psychiatry*, **57**, 876–883.

43. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. and Maglott, D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980-5.

44. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., *et al.* (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, **381**, eadg7492.

45. Clark, K.L., Yutzey, K.E. and Benson, D.W. (2006) Transcription factors and congenital heart defects. *Annu. Rev. Physiol.*, **68**, 97–121.

46. Moretti, F., Marinari, B., Lo Iacono, N., Botti, E., Giunta, A., Spallone, G., Garaffo, G., Vernersson-Lindahl, E., Merlo, G., Mills, A.A., *et al.* (2010) A regulatory feedback loop involving p63 and IRF6 links the pathogenesis of 2 genetically different human ectodermal dysplasias. *J. Clin. Invest.*, **120**, 1570–7.

47. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R. and Weirauch, M.T. (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

48. Ji, W., Ferdman, D., Copel, J., Scheinost, D., Shabanova, V., Brueckner, M., Khokha, M.K. and Ment, L.R. (2020) De novo damaging variants associated with congenital heart diseases contribute to the connectome. *Sci. Rep.*, **10**, 7046.

49. Douglas, G., Cho, M.T., Telegrafi, A., Winter, S., Carmichael, J., Zackai, E.H., Deardorff, M.A., Harr, M., Williams, L., Psychogios, A., *et al.* (2018) De novo missense variants in MEIS2

recapitulate the microdeletion phenotype of cardiac and palate abnormalities, developmental delay, intellectual disability and dysmorphic features. *Am. J. Med. Genet. A*, **176**, 1845–1851.

50. Konrad, E.D.H., Nardini, N., Caliebe, A., Nagel, I., Young, D., Horvath, G., Santoro, S.L., Shuss, C., Ziegler, A., Bonneau, D., *et al.* (2019) CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum. *Genet. Med.*, **21**, 2723–2733.

51. Faundes, V., Newman, W.G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., Davies, S.J., Demos, M.K., Goldman, A., Gill, H., *et al.* (2018) Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. *Am. J. Hum. Genet.*, **102**, 175–187.

52. Chang, S.-W., Mislankar, M., Misra, C., Huang, N., Dajusta, D.G., Harrison, S.M., McBride, K.L., Baker, L.A. and Garg, V. (2013) Genetic abnormalities in FOXP1 are associated with congenital heart defects. *Hum. Mutat.*, **34**, 1226–30.

53. Goddard, L.M., Duchemin, A.-L., Ramalingan, H., Wu, B., Chen, M., Bamezai, S., Yang, J., Li, L., Morley, M.P., Wang, T., *et al.* (2017) Hemodynamic Forces Sculpt Developing Heart Valves through a KLF2-WNT9B Paracrine Signaling Axis. *Dev. Cell*, **43**, 274-289.e5.

54. Crane-Smith, Z., De Castro, S.C.P., Nikolopoulou, E., Wolujewicz, P., Smedley, D., Lei, Y., Mather, E., Santos, C., Hopkinson, M., Pitsillides, A.A., *et al.* (2023) A non-coding insertional mutation of Grhl2 causes gene over-expression and multiple structural anomalies including cleft palate, spina bifida and encephalocele. *Hum. Mol. Genet.*, **32**, 2681–2692.

55. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.

56. Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*

57. Southard, A.E., Edelmann, L.J. and Gelb, B.D. (2012) Role of copy number variants in structural birth defects. *Pediatrics*, **129**, 755–63.

58. Pinz, H., Pyle, L.C., Li, D., Izumi, K., Skraban, C., Tarpinian, J., Braddock, S.R., Telegrafi, A., Monaghan, K.G., Zackai, E., *et al.* (2018) De novo variants in Myelin regulatory factor (MYRF)

as candidates of a new syndrome of cardiac and urogenital anomalies. *Am. J. Med. Genet. A*, **176**, 969–972.

59. Vivante, A., Kleppa, M.-J., Schulz, J., Kohl, S., Sharma, A., Chen, J., Shril, S., Hwang, D.-Y., Weiss, A.-C., Kaminski, M.M., *et al.* (2015) Mutations in TBX18 Cause Dominant Urinary Tract Malformations via Transcriptional Dysregulation of Ureter Development. *Am. J. Hum. Genet.*, **97**, 291–301.

60. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W. and Bulyk, M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

61. Gabriella Miller Kids First Data Resource Center. https://kidsfirstdrc.org/.

62. Yun, T., Li, H., Chang, P.C., Lin, M.F., Carroll, A. and McLean, C.Y. (2020) Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, **36**, 5582–5589.

63. TOPMed Consortium, T., Taliun, D., Zachary, A., Torres, R., Taliun, S.A.G., Corvelo, A., Stephanie, M., Albert, C., Alonso, A., Ardlie, K.G., *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*.

64. Liu, X., Li, C., Mou, C., Dong, Y. and Tu, Y. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.*, **12**, 103.

65. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., *et al.* (2021) GENCODE 2021. *Nucleic Acids Res.*, **49**, D916–D923.