npj | genomic medicine

Article

Published in partnership with CEGMR, King Abdulaziz University

https://doi.org/10.1038/s41525-025-00525-0

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

Raehoon Jeong[1,2] & Martha L. Bulyk[1,2,3] ✉

Many congenital anomaly 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 (OFC) parent-offspring trio cohorts (n = 3835 and 1844, respectively). We identified 17 novel candidate CHD genes and 8 novel candidate OFC genes, of which many were known developmental disorder genes. TFs were enriched among the significant genes; 14 and 8 transcription factor (TF) genes showed significant variant burden for CHD and OFC, respectively. In total, 30 affected children had a de novo missense variant in a DNA binding domain of a known CHD, OFC, and other developmental disorder TF genes. Our results suggest candidate pathogenic variants in CHD and OFC and their potentially pleiotropic effects in other developmental disorders.

Various congenital anomalies, ranging from congenital heart defect (CHD) to orofacial cleft (OFC), 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 OFC patients have incomplete fusions of embryonic tissues 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 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 OFC, are found in their DNA binding domains[9,10]. We thus hypothesized that DNA binding domain variants in other TF genes might also cause these congenital anomalies. Furthermore, we hypothesized that DNA binding domain variants not yet found to be pathogenic but that occur in TFs with DNA binding domain variants previously found to cause CHD or OFC might also cause these conditions.

Searching for genetic causes underlying congenital anomalies 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 congenital anomalies[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 OFC are sporadic cases with unaffected parents (100% for CHD cohorts and 95.3% for OFC cohorts in this study). Therefore, in this study, we searched for de novo variants and rare inherited variants in the probands.

As many TFs are essential, and their haploinsufficiency cause Mendelian diseases[7], there is selective pressure acting against damaging variants in essential TF genes. Therefore, damaging variants in TFs in humans are expected to be present as de novo variants, which have yet to undergo negative selection. These individuals can carry genetic conditions, like CHD and OFC, which are often caused by de novo variants. Recently, DNA binding domain variants in three distinct TFs found in ocular congenital

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Bioinformatics and Integrative Genomics Graduate Program, Harvard University, Cambridge, MA, USA. [3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. ✉e-mail: mlbulyk@genetics.med.harvard.edu

cranial dysinnervation disorders were shown to affect DNA binding affinity[15]. Such findings support the likelihood that analyzing data from cohorts of congenital anomalies, like CHD and OFC, can uncover damaging variants in TFs that are causative.

The aim of our study was two-fold. First, we sought to discover novel disease genes in CHD and OFC because more causal genes likely remain to be found[8,12,13,16]. While CHD and OFC are distinct congenital anomalies, here we analyzed data for these two congenital anomalies because: (1) they are largely genetic conditions, (2) de novo variants explain a significant proportion of the patients' molecular cause, (3) TF genes have been implicated as disease genes, and (4) there were large cohort data available from multiple studies to increase power of disease gene discovery. 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 OFC, respectively[12,13,17–19]. We utilized the PrimateAI variant effect prediction tool[20] 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)[21] 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[22].

Second, focusing on TFs because of their key roles in development and Mendelian diseases, we surveyed TFs and TF DNA binding domain variants for their potential association with CHD and OFC. The resulting list of TFs and DNA binding domain variants is provided as a resource for future studies to evaluate whether they alter DNA binding activity[8,16].

Results

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

To maximize power to discover novel disease genes, we combined genetic data from multiple CHD and, separately, OFC 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[12,17,18], one of which is part of the Kids First program[18]. In total, our list included variants from 3835 family trios with a proband with CHD (Supplementary Data 1). For OFC, we assembled genetic data from four Kids First cohorts[13,23] and the Deciphering Developmental Disorders (DDD) study[19], totaling 1844 family trios (Supplementary Data 1). We combined those data with a list of de novo variants found in 757 family trios from Bishop et al.[13], and 603 family trios from Wilson et al.[19]. For the Kids First cohort samples not analyzed by these two studies, we identified de novo variants from the whole-genome sequencing data using the slivar tool[24] (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[25–33]. 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 congenital anomalies. For this, we scored de novo variants in known CHD genes (Supplementary Data 2) from CHD patients[12] (3835 families with 113 variants) and unaffected siblings from an autism study[34] (2179 families with 26 variants). The autism study was unique in that four members of an autism proband family were sequenced: 2 unaffected parents, 1 unaffected sibling, and 1 proband. This enabled deriving a set of de novo variants that are likely benign in the unaffected siblings. In contrast, the 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 siblings 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 (Fig. 1A). We aimed to select a method that highly enriches potentially pathogenic variants at the top quantile. Overall, PrimateAI[20] showed the highest area under the curve metric for both receiver operator characteristic (ROC) and precision-recall (Supplementary Fig. 1). Although Missense Variant Pathogenicity (MVP)[26] 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. If we add up the mutation rate per generation for all possible missense variants in the human genome, the total missense mutation rate is approximately 0.68 per generation[35,36]. Then, we inferred the expected number of de novo missense mutations in each 5% PrimateAI score bin (i.e., 0.68×0.05). Based on this expected rate, we derived the enrichment of de novo missense variants in CHD versus control samples for each score bin (Fig. 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 ≥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 Fig. 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[11,34]. 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 OFC patient cohorts.

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

Next, to identify candidate CHD and OFC 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[21]. This model integrates enrichment of de novo variants based on a mutational model[35] and the enrichment of inherited 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 (i.e., higher Bayes factor indicates more statistically significant enrichment). We considered 3578 unaffected parents in an autism cohort as controls because we can expect that they likely do not have CHD or OFC[12,37]. This approach was used in an earlier study for CHD that aimed to discover genes with enrichment of putatively damaging variants[12].

We detected 46 and 22 significant genes for CHD and OFC, respectively (q value < 0.1, Supplementary Data 3 and 4). Since genes with no depletion of pLoF variants in a healthy population are not likely to be congenital anomaly genes, we excluded genes with gnomAD's loss-of-function observed/expected upper bound fraction (LOEUF) > 1[36]. Most candidate genes had both pLoF and missense variants contributing to the enrichment (Fig. 2). Thus, integrating the variant types was useful in detecting candidate disease genes.

17 of the 46 genes identified in the CHD analysis cohorts were not known CHD genes (i.e., not significant in studies of individual cohorts and not annotated as CHD genes; Table 1). 8 of the 22 genes identified in the OFC analysis cohorts were not known OFC genes; known OFC genes were taken from Genomics England PanelApp[38] "Clefting" version 4.0 list

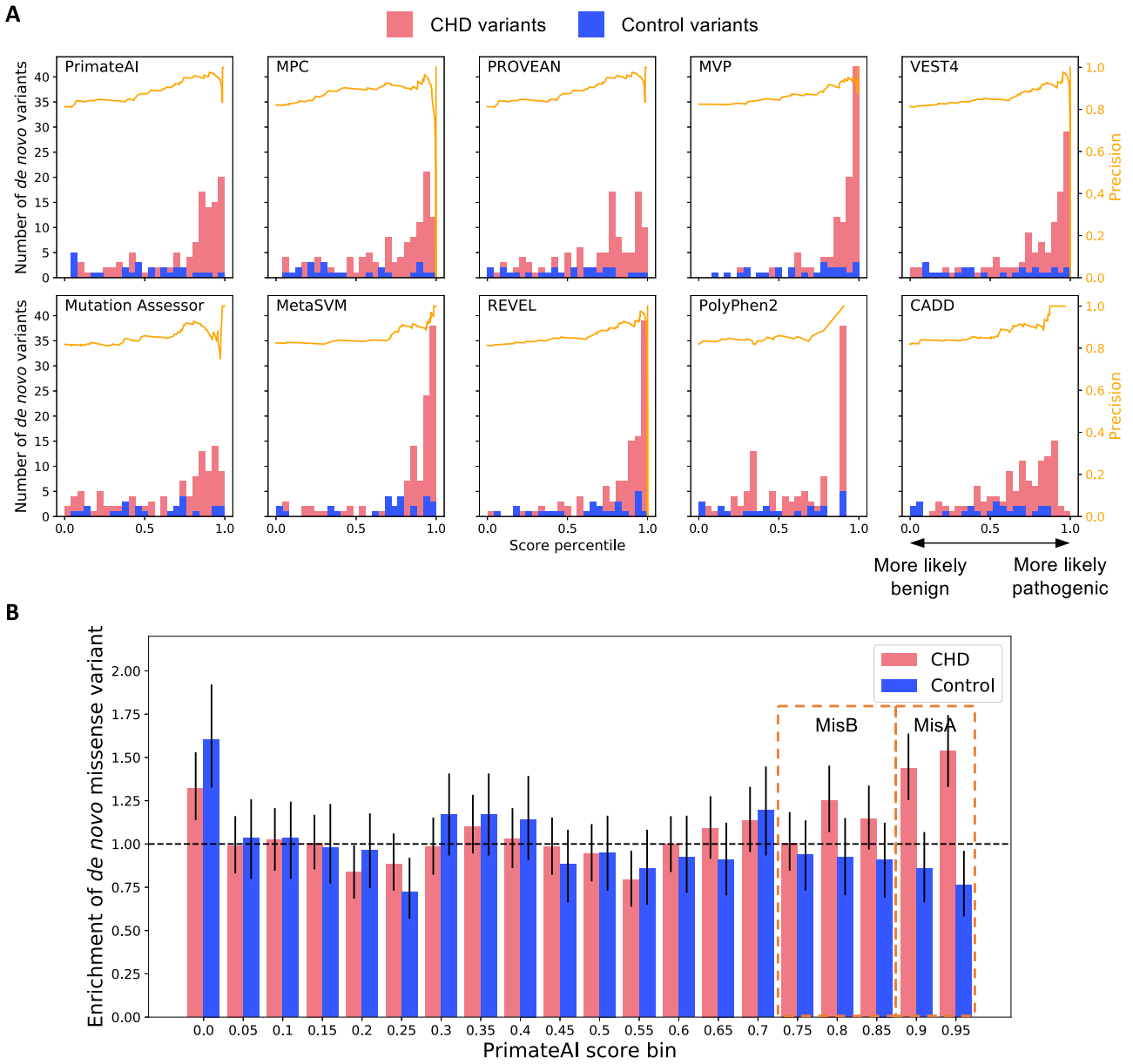

Fig. 1 | Comparison of missense variant prediction methods. A Number of variants in each score percentile bin, which corresponds to 5% increments, for ten missense variant effect predictions. Only de novo variants in 225 human CHD genes, which are listed in (Supplementary Data 2), are considered. The 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).

(Supplementary Data 5). CHD and OFC patients are at higher risk for other congenital anomalies[39,40]. Indeed, several of these genes are developmental disorder genes, such as *TAOK1*, *WAC*, *PACS1*, *FOXP1*, *BRAF*, *SETD5*, and *ZMIZ1* (phenotype MIM numbers: 619575, 616708, 615009, 613670, 613706, 615761, and 618659, respectively). In a recent study on CHD[41], a de novo variant in *SETD5* was considered as a positive diagnosis. Similarly, 7 of the 8 novel candidate OFC 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 OFC with neurodevelopmental disorders[42,43].

More than half of the significant genes in CHD and OFC showed probands with an inherited pLoF variant in the candidate disease gene (27 out of 46 and 13 out of 22 for CHD and OFC, respectively). Two of the OFC family trios (one with a *CTNND1* pLoF variant and another with an *ARHGAP29* 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. For both CHD and OFC, the contribution of inherited variants from unaffected parents has been documented, consistent with our observation[17,44]

De novo missense variants in CHD and OFC genes

Predicting the pathogenic effects of missense variants is challenging, and many are classified as variants of uncertain significance (VUSs) in ClinVar[45]. 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 OFC genes (Supplementary Data 6 and 7). These tables include de novo missense variants in known CHD or OFC genes (Supplementary Data 2 and 3) and candidate CHD or OFC genes in the respective cohorts. In

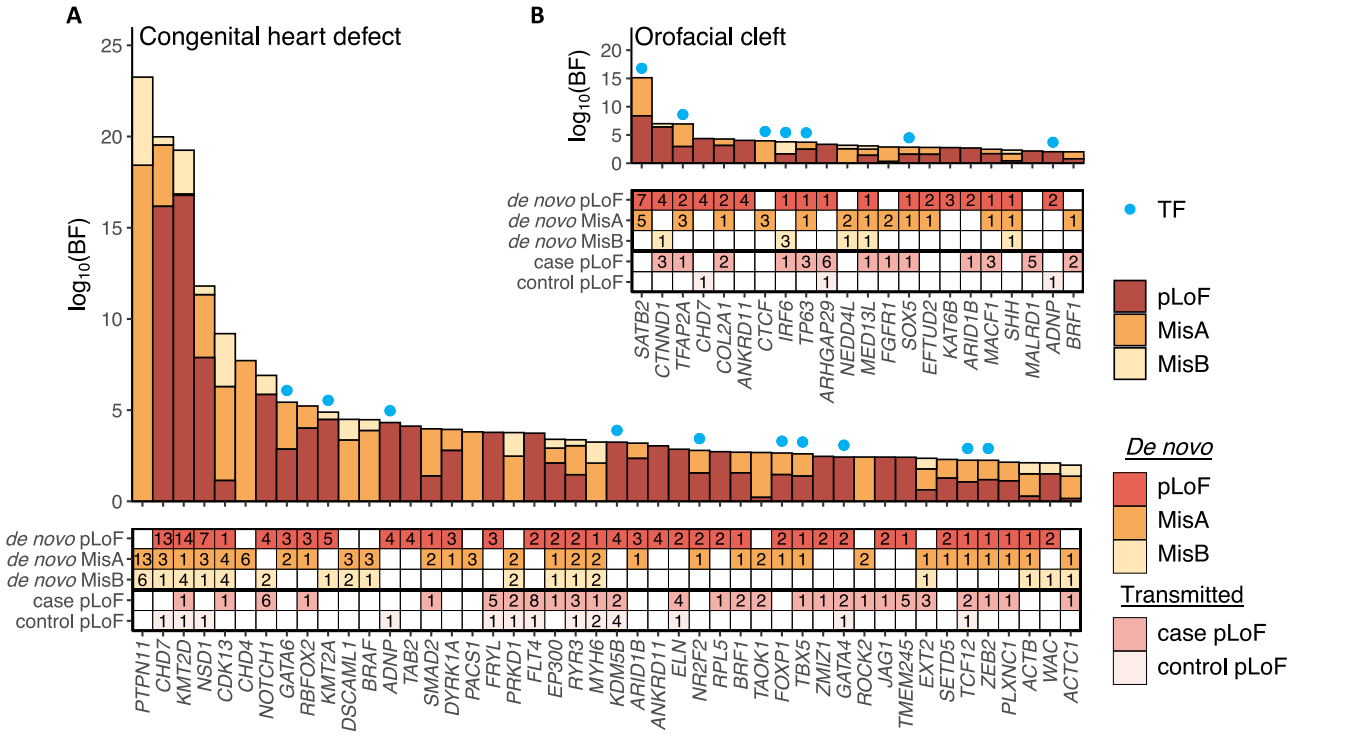

Fig. 2 | Bayes factor for each variant type's enrichment in candidate disease genes. (Top) Bayes factor contribution by MisA, MisB, and pLoF variants in TADA for **A** CHD and **B** OFC in the "de novo + case/control" setting. Only positive Bayes factor contributions in candidate genes (q value < 0.1)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

Table 1 | List of novel candidate disease genes for CHD and OFC

Condition	Novel genes (supporting reference, if available)
Congenital heart defect	*RYR3*[69], *TAOK1, EXT2, TMEM245, ROCK2, ZMIZ1, PLXNC1, ACTC1*[70], *FRYL*[71], *KDM5B*[72], *BRF1, PACS1, MSLNL, SETD5, TCF12, WAC, ZDHHC18*
Orofacial cleft	MED13L, SOX5, KAT6B, ADNP, ARID1B, MACF1, MALRD1, BRF1

We considered a gene novel if they were not listed in Supplementary Data 3 and 4 for CHD and OFC, respectively.

addition to scores from the tools we compared in Fig. 1, we also include scores from the more recent AlphaMissense tool[46].

The coding sequence length affects which TADA model detects enrichment in the gene

To evaluate the utility of incorporating inherited pLoF variants in the case/control setting (i.e., "de novo and 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 Data 3) than using the "de novo and case/control" setting; 24 and 10 genes were exclusively significant in "de novo only" and "de novo and 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 and 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; Fig. 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 OFC genes (Supplementary Data 4 and Supplementary Fig. 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 and case/control" TADA model. It 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 and case/control" setting. On the other hand, shorter genes have lower expected mutation rate for pLoF variants, so each de novo variant contributes to greater amount of enrichment.

TF DNA binding domain variants identified in candidate CHD and OFC disease genes

Because of the known role of TFs in CHD[47] and OFC[48], we examined how many significant genes from our analysis were TFs[49]. For CHD, there were 14 TFs that showed significant enrichment in either "de novo and case/control" or "de novo only" analysis (Table 2 and Fig. 2). For OFC, 7 TFs showed significant enrichment (Table 2 and Fig. 2). For both CHD and OFC, TFs were significantly enriched among the significant genes ($p = 0.006$ and $p = 0.016$, respectively, one-sided Fisher's exact test).

There were 5 and 3 candidate CHD and OFC TF genes, respectively, that are not yet established CHD or OFC disease genes. For CHD, we identified *KDM5B*, *FOXP1*, *KLF2*, *MEIS2*, and *CTCF*. For OFC, we

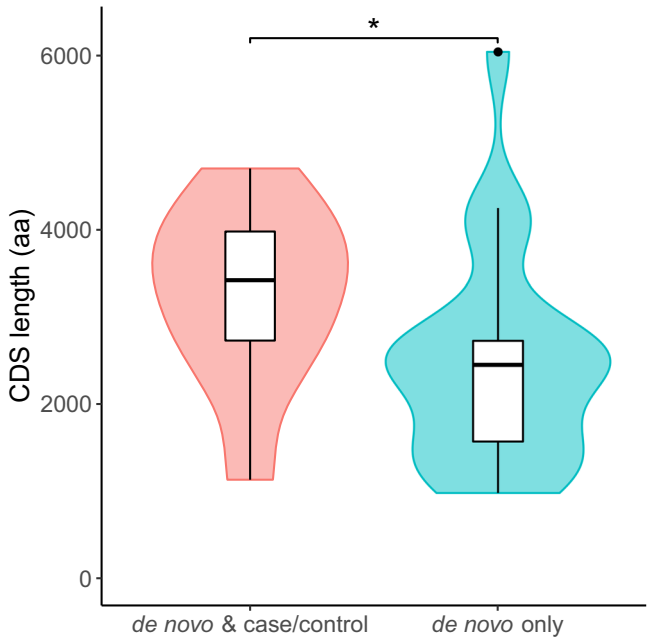

Fig. 3 | Coding sequence length of significant CHD genes by discovery model. Distribution of coding sequence length for the significant genes unique to the "de novo and case/control" model and "de novo only" model. The number of genes is labeled below each category. CDS, coding sequence; aa, amino acid.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

identified *SOX5, ADNP*, and *GRHL2*. Two candidate CHD TF genes—*KDM5B* and *FOXP1*—were also statistically implicated in a similar CHD study[50] that aggregated de novo variants from two[12,17] of the 3 studies that we analyzed. Nevertheless, *KDM5B*, *FOXP1*, *MEIS2*, and *CTCF* are known developmental disorder genes (phenotype MIM numbers: 618109, 613670, 600987, and 615502, respectively). Some children with mutations in these genes have been reported to show heart defects[51–54]. *KLF2* has not been directly associated with CHD, but its zebrafish homologue *klf2* is required for heart valve formation[55]. A non-coding variant that causes over-expression of *Grhl2* in mice led to orofacial cleft phenotypes[56].

Since DNA binding activity plays a crucial role in TF function, we searched for TF DNA binding domain missense variants in known developmental disorder genes. We developed a pipeline to filter for missense variants in the TF DNA binding domains based on a set of 62 DNA binding domain classes in the Pfam database[57] (Supplementary Data 8) and the protein domain prediction model HMMer[58]. Without filtering for disease genes, there were 46 and 11 de novo TF DNA binding domain missense variants in the CHD and OFC cohorts, respectively (Supplementary Data 9); with filtering, there were 17 and 13 DNA binding domain missense variants, respectively (Table 3). Some of these variants are in CHD, OFC, and other developmental disorder genes that are mostly haploinsufficient, characterized by low LOEUF estimates (Table 3). Based on PrimateAI, they were all predicted to be pathogenic (PrimateAI rank score > 0.8). We hypothesize that these variants damage the TFs' DNA binding activity.

Discussion

We aggregated multiple parent-offspring trio cohorts of CHD and OFC to detect 46 and 22 genes, respectively, with enrichment of damaging de novo variants and inherited pLoF variants. 17 were novel candidate CHD genes and 8 were novel candidate OFC genes (Supplementary Data 3 and 4). It is challenging to unambiguously define a list of known CHD and OFC genes. We defined them based on the list from the Seidman lab and Genomics England PanelApp, but they may still miss some genes with supporting evidence in the literature. In fact, some 'novel' genes have support from existing literature, while others do not (Table 1). This means that further studies are needed to validate which of these are true disease genes for CHD and OFC. Moreover, 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 congenital anomalies, 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. De novo variant data from unaffected siblings in autism studies was critical for this analysis, as these siblings are most likely not CHD patients (Fig. 1). We also provide a list of de novo missense variants in known and candidate CHD and OFC genes as a resource (Supplementary Data 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. We point out that the control samples for the variant enrichment analysis were unaffected parents from the autism cohort. They may carry autism-related variants, but they are not likely to carry pathogenic variants in CHD or OFC genes. Despite aggregating data from multiple studies, 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 and 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 and 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 congenital anomalies[59]. 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[22]. 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.

The contribution of inherited variants to risk of CHD and OFC is consistent with earlier reports. For instance, Sifrim et al. described that non-syndromic CHD cases had contributions of inherited damaging variants from unaffected parents, suggesting incomplete penetrance[17]. Our work does not directly address the reasons for incomplete penetrance, but understanding any genetic or environmental factors affecting the penetrance would be important for patient diagnosis and prognosis. One possible explanation is mosaicism, as a multiplex family study on OFC hypothesized[60].

In this study, TFs were enriched among the identified genes. We identified many de novo TF DNA binding domain missense variants in genes that were significantly enriched in CHD or OFC or that are known CHD, OFC, or developmental disorder genes. The identified variants were predicted to be pathogenic by PrimateAI. Some of the TFs with TF DNA binding domain 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[61,62]. 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 DNA binding domain missense variants with altered DNA binding activity[16]. Future studies using DNA binding assays, such as protein binding microarrays (PBMs)[8,63], will be needed to determine which

Table 2 | Transcription factors significantly enriched for predicted deleterious de novo variants

Gene	LOEUF	de novo *variants*			Inherited variants		de novo and 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	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**

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.
Q values less than 0.1 are bolded.
[a]Novel candidate CHD genes.
[b]Novel candidate OFC genes.

of the identified CHD and OFC variants alter DNA binding activity and in what manner they do so.

There are several limitations to this study. First, because we directly aggregated de novo variant data from multiple studies, multiple pipelines were used to call these variants. Calling de novo variants altogether would be a more involved but more consistent approach to identify de novo variants for meta-analysis. Second, we relied on computational predictions to stratify missense variants by their importance. Even though we selected PrimateAI as a tool that best discriminated de novo variants found in CHD patients from those found in unaffected siblings in an autism cohort, it is by no means a perfect tool. Moreover, not all de novo coding variants found in CHD patients are pathogenic, nor are all of those found in unaffected children benign. However, this kind of comparison is frequently made to evaluate variant effect prediction tools[20,46]. Lastly, as noted above, definitive lists of 'known' CHD and OFC genes are elusive. However, we followed the classification of CHD genes from a lab leading the efforts to find CHD genes (i.e., Seidman lab) and that of OFC genes from a panel curated by Genomics England that is running a large-scale genomic study on rare disease (i.e., 100,000 Genomes Project). All in all, future studies can benefit from a harmonized de novo variant database, more accurate variant effect prediction tools, and well-curated disease gene lists.

Methods

Genetic data from family trio cohorts of CHD and OFC

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[12,17] and one genome-sequencing study[18]. 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 3835 unique family trios.

For OFC, we analyzed data from 4 cohorts from the Gabriella Miller Kids First program[64] and an additional cohort from the United Kingdom[19] (Supplementary Data 1). For the 4 cohorts from Kids First, 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). Of these, data from 374 European (phs001168), 267 Colombian (phs001420), and 116 Taiwanese (phs001997) family trios were analyzed in Bishop et al.[13]. For these 757 family trios, we downloaded a list of de novo variants in probands from Table S3 of Bishop et al.[13]. The 113 of the trios in phs001997 data that were not analyzed in Bishop et al. were from the African Craniofacial Anomalies Network, and 351 trios in phs002595 were from a cohort in the Philippines. We analyzed data from these 484 trios using the genotype calls provided by the Kids First data portal. Lastly, we downloaded a list of de novo variants in probands of 603 family trios in the United Kingdom from Table S4 of Wilson et al.[19].

We considered unaffected siblings or parents of probands in an autism cohort as controls without CHD or OFC. We downloaded de novo variant data from unaffected siblings of probands in an autism cohort[34] to compare variant effect predictions (Fig. 1). Lastly, we downloaded heterozygous pLoF variants from 3578 unaffected parents in an autism cohort as controls[12,37], which we used to test enrichment of putatively damaging variants (Fig. 2).

Table 3 | De novo TF DNA binding domain missense variants in genes associated with CHD, OFC, or developmental disorder genes

Developmental disorder	Gene	LOEUF	Amino acid change	PrimateAI rank score	Variant	DBD (Pfam ID)
Congenital heart defect						
CHD	*FOXP1*[a]	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*[a]	0.135	I227T	0.98142	12:114385551:A:G	T-box (PF00907)
CHD	*TCF12*[a]	0.372	H631Q	0.98114	15:57273177:C:G	Helix-loop-helix (PF00010)
CHD	*NR2F2*[a]	0.217	C96F	0.98292	15:96332392:G:T	C4 zinc finger (PF00105)
CHD	*GATA6*[a]	0.174	R456G	0.90881	18:22181516:C:G	GATA zinc finger (PF00320)
CHD	*GATA6*[a]	0.174	R456H	0.92717	18:22181517:G:A	GATA zinc finger (PF00320)
CHD[a]	*KLF2*[a]	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*[a]	0.261	K96N	0.96845	19:42249912:C:A	ETS (PF00178)
Orofacial cleft						
OFC (van der Woude syndrome)	*IRF6*[a]	0.132	N88D	0.86413	1:209796465:T:C	IRF (PF00605)
OFC (van der Woude syndrome)	*IRF6*[a]	0.132	R84H	0.84067	1:209796476:C:T	IRF (PF00605)
OFC (Glass syndrome)	*SATB2*[a]	0.091	R667G	0.94148	2:199272414:G:C	Homeodomain (PF00046)
OFC (Glass syndrome)	*SATB2*[a]	0.091	R399H	0.90829	2:199328888:C:T	CUT (PF02376)
OFC (Glass syndrome)	*SATB2*[a]	0.091	L394S	0.95427	2:199328903:A:G	CUT (PF02376)
OFC (Glass syndrome)	*SATB2*[a]	0.091	R389L	0.9951	2:199348708:C:A	CUT (PF02376)
OFC (Glass syndrome)	*SATB2*[a]	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)
OFC	*TFAP2A*[a]	0.261	R256Q	0.921	6:10404511:C:T	AP-2 (PF03299)
OFC	*TFAP2A*[a]	0.261	S249L	0.98055	6:10404532:G:A	AP-2 (PF03299)
OFC (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)

The table lists de novo TF DNA binding domain variants from our analysis in genes that are either significantly enriched in our study (marked with an asterisk [*]) or are reported as CHD, OFC, 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. *DBD* DNA binding domain, *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.
[a]Significant enrichment of damaging variants in this study.

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 OFC samples not included in Bishop et al.[13].

Identifying de novo variants and rare inherited variants in the OFC 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[64] 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[65] with the 'gatk' setting and (2) using slivar[24] 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[66] 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[24]. 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[66] allele frequency of less than 5×10^{-5}.

For all OFC 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[66] 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[20], PolyPhen2[25], MVP[26], PROVEAN[27], CADD[28], MetaSVM[29], REVEL[30], VEST4[31], MPC[32], and MutationAssessor[33]. These tools' scores for missense variants were accessed from the database for nonsynonymous SNPs' functional predictions (dbNSFP) version 4.5[67]. 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 Data 2) from CHD patients from those from unaffected children. There were a total of 3836 CHD family trios[12,17,18] and 2179 control family trios[34] 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 families and control families in each of the 5% score bins. The expected number of de novo missense variants per family was the sum of all missense mutation rates (~0.68 per generation). Then, we bootstrapped sampled CHD and control families to establish the respective 95% confidence intervals of the enrichment estimates. Ultimately, based on Fig. 1B, we selected PrimateAI ≥ 0.9 and $0.75 \leq$ PrimateAI < 0.9 as the two missense variant groups— MisA and MisB.

Testing enrichment of damaging de novo and rare inherited variants

We used the TADA model[21] to detect genes with an enrichment of potentially damaging variants (i.e. predicted-loss-of-function (pLoF), missense with PrimateAI[20] rank score ≥ 0.9 (MisA), or missense with PrimateAI rank score 0.75–0.9 (MisB)) from the number of de novo variants and mutation rate estimates. We derived the per-gene mutation rates for MisA, MisB, and pLoF based on estimates in Samocha et al.[35] and gnomAD[36]. We multiplied the per-gene 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 all possible MisA and MisB variants are expected be 0.1 and 0.15 of all missense variants. We added the per-gene nonsense, splice site, and frameshift mutation rates to derive the per-gene pLoF mutation rates.

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

Identifying TF DNA binding domain variants in candidate disease genes

We identified disease-associated TF genes based on a list of 1639 TFs[49]. Then, we determined the location of the DNA binding domains using a set of 62 DNA binding domain classes in the Pfam database version 35.0[57] (Supplementary Data 5) and the protein domain prediction model HMMer[58]. We considered only canonical transcripts and amino acid sequences based on GENCODE[68] in annotating whether the missense variants fall within a DNA binding domain.

Compliance with ethical regulations

This study complied with all relevant ethical regulations including the Declaration of Helsinki. The research described in this study did not require review by an institutional review board (IRB). We did not directly interact with patients, nor did we collect patient data for this study. All the dbGaP studies, from which we obtained data that we analyzed, state that IRB approval is not required. The data from Jin et al., Richter et al., Sifrim et al., Wilson et al., Bishop et al., etc., are all from the supplementary tables published and made freely, publicly available as part of those papers.

Data availability

For CHD, we downloaded de novo variant data from two exome-sequencing studies[12,17] and one genome-sequencing study[18]. We also downloaded the list of rare inherited pLoF variants from Jin et al.[12]. For OFC, we downloaded genotype data from 4 cohorts from the Gabriella Miller Kids First data portal[64]. Their database of Genotypes and Phenotypes (dbGaP) IDs were phs001168, phs001997, phs001420, and phs002595. No new data were generated as part of this study.

Code availability

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

Received: 30 January 2025; Accepted: 9 September 2025;
Published online: 29 September 2025

References

1. Centers for Disease Control and Prevention (CDC) Update on overall prevalence of major birth defects—Atlanta, Georgia, 1978-2005. *Mmwr. Morb. Mortal. Wkly. Rep.* **57**, 1–5 (2008).
2. Ely, D. M. & Driscoll, A. K. Infant mortality in the United States, 2021: data from the period linked birth/infant death file. *Natl. Vital. Stat. Rep.* **72**, 1–19 (2023).
3. Mitchell, S. C., Korones, S. B. & Berendes, H. W. Congenital heart disease in 56,109 births. Incidence and natural history. *Circulation* **43**, 323–332 (1971).
4. Watkins, S. E., Meyer, R. E., Strauss, R. P. & Aylsworth, A. S. Classification, epidemiology, and genetics of orofacial clefts. *Clin. Plast. Surg.* **41**, 149–163 (2014).
5. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
6. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell* **38**, 576–589 (2010).
7. Seidman, J. G. & Seidman, C. Transcription factor haploinsufficiency: when half a loaf is not enough. *J. Clin. Investig.* **109**, 451–455 (2002).
8. Barrera, L. A. et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450–1454 (2016).
9. Su, W. et al. Congenital heart diseases and their association with the variant distribution features on susceptibility genes. *Clin. Genet.* **91**, 349–354 (2017).
10. Kondo, S. et al. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat. Genet.* **32**, 285–289 (2002).
11. Homsy, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–1266 (2015).

12. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
13. Bishop, M. R. et al. Genome-wide enrichment of De novo coding mutations in orofacial cleft trios. *Am. J. Hum. Genet.* **107**, 124–136 (2020).
14. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
15. Jurgens, J. A. et al. Gene identification for ocular congenital cranial motor neuron disorders using human sequencing, zebrafish screening, and protein binding microarrays. *Investig. Ophthalmol. Vis. Sci.* **66**, 62 (2025).
16. Kock, K. H. et al. DNA binding analysis of rare variants in homeodomains reveals homeodomain specificity-determining residues. *Nat. Commun.* **15**, 3110 (2024).
17. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
18. Richter, F. et al. Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.* **52**, 769–777 (2020).
19. Wilson, K., Newbury, D. F. & Kini, U. 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 (2023).
20. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
21. He, X. et al. Integrated model of De novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet*. **9**, e1003671 (2013).
22. Fu, J. M. et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).
23. Awotoye, W. et al. Whole-genome sequencing reveals de-novo mutations associated with nonsyndromic cleft lip/palate. *Sci. Rep.* **12**, 11743 (2022).
24. Pedersen, B. S. et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genom. Med.* **6**, 60 (2021).
25. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
26. Qi, H. et al. MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun*. **12**, 510 (2021).
27. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688 (2012).
28. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
29. Dong, C. et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
30. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
31. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**, (2013).
32. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353. https://doi.org/10.1101/148353 (2017).
33. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
34. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020).
35. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
36. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
37. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
38. Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
39. Egbe, A., Lee, S., Ho, D., Uppu, S. & Srivastava, S. Prevalence of congenital anomalies in newborns with congenital heart disease diagnosis. *Ann. Pediatr. Cardiol.* **7**, 86–91 (2014).
40. Stoll, C., Alembik, Y. & Roth, M.-P. Co-occurring anomalies in congenital oral clefts. *Am. J. Med. Genet. A* **188**, 1700–1715 (2022).
41. Hartill, V. et al. Molecular diagnoses and candidate gene identification in the congenital heart disease cohorts of the 100,000 genomes project. *Eur. J. Hum. Genet*. https://doi.org/10.1038/s41431-024-01744-2 (2024).
42. Marino, B. S. et al. Neurodevelopmental outcomes in children with congenital heart disease: evaluation and management: a scientific statement from the American Heart Association. *Circulation* **126**, 1143–1172 (2012).
43. Tillman, K. K. et al. Increased risk for neurodevelopmental disorders in children with orofacial clefts. *J. Am. Acad. Child Adolesc. Psychiatry* **57**, 876–883 (2018).
44. Diaz Perez, K. K. et al. Rare variants found in clinical gene panels illuminate the genetic and allelic architecture of orofacial clefting. *Genet. Med.* **25**, 100918 (2023).
45. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
46. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
47. Clark, K. L., Yutzey, K. E. & Benson, D. W. Transcription factors and congenital heart defects. *Annu. Rev. Physiol.* **68**, 97–121 (2006).
48. Moretti, F. et al. A regulatory feedback loop involving p63 and IRF6 links the pathogenesis of 2 genetically different human ectodermal dysplasias. *J. Clin. Investig.* **120**, 1570–1577 (2010).
49. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
50. Ji, W. et al. De novo damaging variants associated with congenital heart diseases contribute to the connectome. *Sci. Rep.* **10**, 7046 (2020).
51. Douglas, G. et al. 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 (2018).
52. Konrad, E. D. H. et al. CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum. *Genet. Med.* **21**, 2723–2733 (2019).
53. Faundes, V. et al. Histone lysine methylases and demethylases in the landscape of human developmental disorders. *Am. J. Hum. Genet.* **102**, 175–187 (2018).
54. Chang, S.-W. et al. Genetic abnormalities in FOXP1 are associated with congenital heart defects. *Hum. Mutat.* **34**, 1226–1230 (2013).
55. Goddard, L. M. et al. Hemodynamic forces sculpt developing heart valves through a KLF2-WNT9B paracrine signaling axis. *Dev. Cell* **43**, 274–289.e5 (2017).
56. Crane-Smith, Z. et al. 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 (2023).
57. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

58. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol*. https://doi.org/10.1371/journal.pcbi.1002195 (2011).
59. Southard, A. E., Edelmann, L. J. & Gelb, B. D. Role of copy number variants in structural birth defects. *Pediatrics* **129**, 755–763 (2012).
60. Diaz Perez, K. K. et al. Rare variants found in multiplex families with orofacial clefts: Does expanding the phenotype make a difference? *Am. J. Med. Genet. A* **191**, 2558–2570 (2023).
61. Pinz, H. et al. 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 (2018).
62. Vivante, A. et al. Mutations in TBX18 cause dominant urinary tract malformations via transcriptional dysregulation of ureter development. *Am. J. Hum. Genet.* **97**, 291–301 (2015).
63. Berger, M. F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
64. Gabriella Miller Kids First Data Resource Center. https://kidsfirstdrc.org/.
65. Yun, T. et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2020).
66. TOPMed Consortium, T. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
67. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
68. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
69. Kim, J.-M. et al. Uncovering potential causal genes for undiagnosed congenital anomalies using an in-house pipeline for trio-based whole-genome sequencing. *Hum. Genom.* **19**, 1 (2025).
70. Frank, D. et al. Cardiac α-actin (ACTC1) gene mutation causes atrial-septal defects associated with late-onset dilated cardiomyopathy. *Circ. Genom. Precis. Med.* **12**, e002491 (2019).
71. Pan, X. et al. De novo variants in FRYL are associated with developmental delay, intellectual disability, and dysmorphic features. *Am. J. Hum. Genet.* **111**, 742–760 (2024).
72. Borroto, M. C. et al. A genotype/phenotype study of KDM5B-associated disorders suggests a pathogenic effect of dominantly inherited missense variants. *Genes* **15**, 1033 (2024).

Acknowledgements

We thank members of the Bulyk lab for helpful discussion. This work was funded by NIH grant 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 and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-025-00525-0.

Correspondence and requests for materials should be addressed to Martha L. Bulyk.