Rare variants found in clinical gene panels illuminate the genetic and allelic architecture of orofacial clefting

Kimberly K. Diaz Perez,[1] Sarah W. Curtis,[1] Alba Sanchis-Juan,[2] Xuefang Zhao,[2] S. Taylor Head,[3] Samantha Ho,[1] Bridget Carter,[1,4] Toby McHenry,[5] Madison R. Bishop,[1] Luz Consuelo Valencia-Ramirez,[6] Claudia Restrepo,[6] Jacqueline T. Hecht[7], Lina Moreno Uribe[8], George Wehby[9], Seth M. Weinberg[5], Terri H. Beaty[10], Jeffrey C. Murray[11], Eleanor Feingold[12], Mary L. Marazita[5,12], David J. Cutler[1], Michael P. Epstein[1], Harrison Brand[2], Elizabeth J. Leslie[1*]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA
[2]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA; Department of Neurology and Surgery, Massachusetts General Hospital and Harvard Medical School, Boston, MA
[3]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA
[4]Agnes Scott College, Decatur, GA
[5]Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, University of Pittsburgh School of Dental Medicine, Pittsburgh, PA
[6]Fundación Clinica Noel, Medellin, Colombia
[7]Department of Pediatrics, McGovern Medical, School and School of Dentistry, UT Health at Houston, Houston, TX 77030, USA
[8]Department of Orthodontics, University of Iowa, Iowa City, IA
[9]Department of Health Management and Policy, University of Iowa, Iowa City, IA
[10]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
[11]Department of Pediatrics, University of Iowa, Iowa City, IA
[12]Department of Human Genetics, University of Pittsburgh School of Public Health, Pittsburgh, PA

*Corresponding author: ejlesli@emory.edu

ABSTRACT

Orofacial clefts (OFCs) are common craniofacial birth defects including cleft lip (CL), cleft lip with cleft palate (CLP), and cleft palate (CP). The etiological heterogeneity of OFCs complicates clinical diagnostics as it is not always readily apparent if the cause is Mendelian, environmental, or multifactorial. Although sequencing could aid diagnosis, it is not commonly used for the 60-70% of OFC cases that appear isolated or lack a strong family history. We aimed to estimate the diagnostic yield evaluating 503 genes using whole-genome sequencing data from 841 cases and 294 controls. After curating variants, we evaluated them according to the American College of Medical Genetics pathogenicity criteria blinded to case-control status. We found 'likely pathogenic' variants in 9.04% of cases and 1.36% of controls ($p<0.0001$), which was almost exclusively driven by dominant variants in autosomal genes. The yield was highest in CP (17.6%) and CLP (9.09%) cases, while CL (2.80%) cases were not significantly different from controls. We identified 'likely pathogenic' variants in cases in 39 genes, underscoring the genetic heterogeneity of OFCs. Notably, nine genes, including *CTNND1*, *IRF6*, and *COL2A1*, were recurrently mutated, accounting for more than half of the yield (occurring in 4.64% of OFC cases). Most reviewed variants (60.6%) were 'variants of uncertain significance' (VUS) and were more frequent in cases ($p=7.31 \times 10^{-4}$), but no individual gene showed a significant excess of VUS. Cumulatively, these results underscore the etiological heterogeneity of OFCs and suggest clinical sequencing could help reduce the diagnostic gap in OFC etiology.

INTRODUCTION

Orofacial clefts (OFCs, [MIM: 119530]) arise when the primordial tissues of the face fail to fuse, creating an opening in the lip and/or the palate. The three main types of OFCs are cleft lip (CL), cleft palate (CP), and cleft lip with cleft palate (CLP). OFCs affect 1/1000 individuals worldwide, making OFCs the most common craniofacial birth defects [1]. Although OFCs can be surgically repaired in infancy, treatment and orthodontic/maxillofacial care can extend into adulthood, exceeding a $200,000 lifetime cost [2]. Furthermore, individuals with OFCs experience higher rates of mortality from all causes, and higher rates of certain comorbidities, such as cancer and congenital heart disease, and psychosocial effects [2; 3].

OFCs are etiologically heterogeneous [4; 5]. Genetic factors, such as point mutations, copy number variants, and chromosomal abnormalities, contribute significantly to the etiology of OFCs, especially in Mendelian OFC syndromes where an OFC is typically accompanied by other structural anomalies, cognitive anomalies, or intellectual disability. Hundreds of Mendelian syndromes involving OFCs have been described in the Online Mendelian Inheritance in Man [6], but each is individually rare. In contrast, most OFC cases occur as isolated birth defects (often termed non-syndromic) and have been considered etiologically complex disorders where both genetic and environmental factors influence risk. It is also possible, however, that both syndromic and isolated cases have an etiology caused by genetics (e.g., single genes), environment, or the combined effect of multiple genetic and environmental risk factors. This etiological heterogeneity of OFCs is further compounded by phenotypic heterogeneity, incomplete penetrance, and variable expressivity, making clinical diagnostics challenging.

To illustrate this, consider Van der Woude syndrome (VWS [MIM: 119300]), a dominant Mendelian disorder thought to be the most common OFC syndrome (affecting 1/35,000 individuals). Although the diagnostic criteria for VWS is an OFC with lower lip pits, ~15% of individuals diagnosed with Van der Woude syndrome and confirmed to have a causal mutation have only an OFC [7; 8]. Thus, they are indistinguishable from approximately 70% of children with an OFC that do not present any additional structural, cognitive, or developmental anomalies. Further, recurrence risk estimates based on epidemiological studies vary by cleft type and gender with an approximate sibling recurrence of ~4% [9]. This risk is much lower than the absolute sibling risk for an autosomal dominant disorder or the empirical risk of being affected in incompletely penetrant autosomal dominant disorder but is also significantly higher than expected if risk were driven by a *de novo* mutations alone. It is therefore important to be able to determine the cause of OFCs as such knowledge informs recurrence risk estimates, plus the approaches for genetic counseling and clinical management.

In the absence of a differential diagnosis based on the phenotype alone, molecular diagnoses based in clinical genetic testing could offer a solution. However, genetic testing is not common for individuals with apparently isolated OFCs without any family history, which make up the majority of individuals with OFCs. Furthermore, existing commercial clinical testing panels are highly variable in their content, leading to the possibility of missed diagnoses. Multiple studies have explored the use of sequencing to identify pathogenic variation to improve diagnostics for OFCs. We previously investigated the proportion of isolated OFC cases that could be attributed to mutations in *IRF6* alone; however, our estimate of 0.2-0.4% was too low to recommend broad screening of this one gene [10]; at that time, whole exome sequencing (WES) and whole genome sequencing (WGS) were prohibitively expensive and had not yet been embraced by either the craniofacial research or clinical genetics communities. The uptake of these technologies has increased in recent years. Basha et al. tested the

diagnostic rate from WES in 46 multiplex OFC families, finding 10% of index cases carry 'likely pathogenic' variants, primarily in genes associated with recognized autosomal dominant OFC syndromes[11]. Recently, we investigated *de novo* mutations in 756 OFC trios through WGS and found that 6% of sequenced trios had a *de novo* mutation in genes recognized as being associated with OFCs[12].

In the present study, we aimed to estimate the diagnostic yield of a set of 503 genes associated with OFCs using whole-genome sequencing in 841 OFC cases and 294 controls. Using WGS, we are able to interrogate all types of genetic variation, from single-nucleotide variants (SNVs) to structural variation. Ultimately, these results can provide insights into the genetic architecture and heterogeneity in OFCs, aid in estimating more accurate recurrence risks, and inform the clinical utility of sequencing for OFCs.

SUBJECTS AND METHODS

Study Population

The case sample consisted of 841 total OFC cases (765 case-parent trios, 60 parent-child dyads, and 16 singleton cases) sequenced as part of the Gabriella Miller Kids First (GMKF) Pediatric Research Program. The case sample was sequenced in three cohorts based on recruitment site/ancestry: 1) "Europeans" from the United States, Argentina, Turkey, Hungary, and Spain; 2) "Latinos" from Colombia; and 3) "Asians" from Taiwan (Supplemental Table 1). The collection of cases was recruited over many years using different research protocols, but each protocol generally included a physical exam to screen out individuals with multiple major anomalies or known intellectual disability that might be characteristic of some OFC syndrome. Individuals with known diagnoses were excluded from sequencing. This case sample is therefore enriched for isolated OFCs and depleted for multiple congenital anomalies and severe manifestations of syndromes. The case sample includes probands with cleft lip only (CL; 107 cases), cleft lip and cleft palate (CLP; 660 cases), and cleft palate only (CP; 74 cases). The cases were primarily male (56% CL, 65% CLP, 53% CP) reflecting the large proportion of the study cases having CLP where males are known to be overrepresented.

A total of 621 probands reported no family history of OFCs, defined here as any affected relative within the 3rd degree, and are we considered these simplex cases. An additional 220 probands reported having at least one other affected individual (1st, 2nd or 3rd degree) and were classified multiplex families (Supplemental Table 2). This included 63 probands with at least one affected parent. All probands were confirmed with kinship calculations using KING to exclude close relatives (1st, 2nd, or 3rd degree).

The control sample was comprised of 294 child-parent trios from the 1000 Genomes Project (1KGP) [13]. Because a number of 1KGP samples are derived from cell lines, some samples have excessive numbers of *de novo* mutations acquired as a result of multiple passages and are not comparable to the pattern of variation in the GMKF samples[14]. Therefore, we selected 1KGP trios through a quality control process (described below) to have approximately the same amount of total variation including *de novo* mutation rates (all selected 1KGP trios had fewer than 138 *de novo* events per trio) as the case cohorts. The control cohort included control-parent trios from multiple ancestries: 84 African, 116 US and European, 52 East Asian, and 42 South Asian trios. Although phenotype information is unavailable for 1KGP and we cannot know with certainty that these trios have no individuals with OFCs, we would

expect at most 1 OFC in the 294 trios (882 total individuals) based on an overall prevalence rate of OFCs at 1 in 1000 live births worldwide. Therefore, 1KGP can serve as a suitable comparison group as it is unlikely to be enriched for pathogenic variants influencing risk to OFCs.

Sequencing and Quality Control

The OFC cohort was sequenced, aligned, and SNVs and small insertions/deletions (indels) were called using the procedures detailed by Bishop et al [12]. The control cohort was sequenced as described in Byrska-Bishop et al. [15]. The same quality control procedures were performed on the separate case and control VCF files. We retained genotype calls with a genotype quality ≥ 20 and read depth ≥ 10 and biallelic variants passing VSQR with a Quality Normalized by Depth (QD) score > 4 using VCFTools (v0.1.13) and BCFtools (v1.9). Variants with > 2 Mendelian errors, > 5% missingness, or severe deviations from Hardy-Weinberg equilibrium ($p < 10^{-7}$) in unaffected samples were dropped. For *de novo* mutations, we required an allele balance of at least 0.3 – 0.7 in each proband and less than 0.05 in both parents.

Selection of Gene List

We created a comprehensive gene set of 503 genes (Supplemental Table 3, Supplemental Figure 1A) to extract and prioritize genetic variants possibly associated with OFCs from five sources (all downloaded September 4th, 2020): 1) the Prevention Genetics CL/P clinical genetic testing panel; 2) the Developmental Disorders Genotype-to-Phenotype (DDG2P) database, which defines genes for disorders in which the most prominent pathogenic mechanism occurs during embryogenesis; we selected genes in which the organ specificity contained "Face"; 3) the National Health Service (NHS) Genomic Medicine Service cleft panel (v2.2, current signed off version, March 2020), an expert-curated list of genes for familial cleft lip and/or cleft palate, familial isolated clefting, and syndromic clefting; 4) Clinical synopses/genes from the Online Mendelian Inheritance in Man (OMIM) where the clinical synopses included orofacial clefts with a known inheritance and molecular basis. OMIM clinical synopses search terms included: "cleft lip," "cleft palate," "oral cleft," "orofacial cleft," and "cleft lip and/or palate;" 5) a manually curated list from recent literature on genetic studies of OFCs. The NHS panel included an indicator for level of evidence corresponding to expert consensus for a gene to be on the panel: "green" for genes of known clinical utility and scientific validity, "amber" for moderate evidence levels of evidence, and "red" indicating little evidence [16]. We classified genes as autosomal dominant (AD), autosomal recessive (AR), and X-linked (XL) based on the mechanism by which variants in these genes lead to OFC phenotypes. Genes associated with multiple or unspecified modes of inheritance were considered in both dominant and recessive analyses. The average depth of coverage for each gene was comparable between cases and controls and between each of the three case cohorts (Supplemental Figure 1B).

SNV and Indel Annotation and Variant Filtering

Variants were annotated using ANNOVAR (version 201707) and Variant Effect Predictor (VEP, release 102, 103, 106). Protein-altering variants were extracted for each of the 503 genes. Variants were filtered using a maximum allele frequency threshold of 0.1% for variants in autosomal dominant genes and 0.5% for variants in autosomal recessive genes using gnomAD (v2 and v3) and ExAC (v0.3), and an allele count ≤ 10 in the dataset. Variant-level annotations used in the prioritization and interpretation included nine *in silico* pathogenicity predictions (SIFT [17], PolyPhen [18], LRT [19], MutationTaster [20],

FATHMM [21], PROVEAN [22], MetaSVM [23], MetaLR [23], and M.CAP [24]), CADD [25] scores, variant pathogenicity classifications from ClinVar [26], and allele frequencies (AF) in gnomAD (v2 and v3) [27] and ExAC [28]. Variants were also annotated for constrained regions within genes [29].

Structural Variants (SVs) Identification and Filtering

We detected SVs in the OFC cohort with the GATK-SV discovery pipeline as previously described [30; 31]. Briefly, GATK-SV is a publicly available, comprehensive cloud-based SV detection pipeline (https://github.com/broadinstitute/gatk-sv). It relies on an ensemble approach that harmonizes SV detection from multiple tools followed by machine learning to remove likely false positive events and then performs joint genotyping and refined variant resolution for complex events. For this study, we ran Manta [32], Wham [33], MELT [34], Gatk-gCNV [35], and cn.MOPS [36]. Upon completion of the discovery pipeline, the derived VCF file was annotated by svtk [30] to predict functional impact of a given SV and compare allele frequencies against gnomAD SV (v2.1) [31]. We developed an interpretation framework to identify disease-associated SVs involved in OFC and applied this to our annotated SV VCF. First, we obtained SVs that overlapped with genes in our gene list, and these were filtered by allele frequency (OFC cohort AF ≤ 0.03 and gnomAD SV AF ≤ 0.01). Structural variants overlapping recurrent genomic disorder regions[37] were investigated independently on the allele frequency and here we report the gene(s) overlapping our gene list from those regions in this study. Further inheritance-specific genotype and frequency filters were applied to identify possible relevant *de novo* (gnomAD SV AF ≤ 1e-3, cohort AC ≤ 10 and cohort sample count ≤ 5), homozygous (cohort homozygous allele count ≤ 10 and absent in healthy individuals from the cohort), compound heterozygous and X-linked recessive SVs. We also considered SVs inherited from unaffected parents if the AC in the cohort was ≤ 10, of which ≤ 5 were unaffected individuals. Candidate SVs (Supplemental Table 4) were manually reviewed and visually inspected in normalized read depth plots using Integrative Genomics Viewer [38].

Classification into the Tier System

Rare SNVs and indels located within the 503 genes in our list were classified into a ranked tier system designed to minimize the total number of variants undergoing manual American College of Medical Genetics and Genomics (ACMGG) review while retaining as many potential 'likely pathogenic' variants as possible. Each tier was based on gene or variant annotation criteria, including variant type, *in silico* pathogenicity predictions, and gene constraint (Supplemental Figure 2A). Tiers were ranked based on qualitative assessments of their likelihood to contain 'likely pathogenic' variants. Assessments were made by KDP, MRB, and EJL and the final tiers were formed from a consensus of these assessments.

After sorting variants into tiers, we identified a cutoff point above which all variants would be manually reviewed according to ACMGG criteria. To determine the cutoff point, we extracted all variants in ClinVar from 503 genes classified as either pathogenic ('likely pathogenic' or 'pathogenic') or benign ('likely benign' or 'benign'). We sorted the resulting 527 pathogenic and 303 benign variants into tiers (Supplemental Figure 2B). We identified Tier 1B as a point at which 95% of pathogenic variants but only 48% of benign variants would be retained for review.

ACMGG Variant Classification

All variants meeting the Tier 1B threshold on this tier system were assessed using ACMGG criteria blinded to case or control status [39]. We considered variants with "damaging" pathogenicity predictions

from ≥ 5 out of 9 algorithms to meet the criteria for PP3 (multiple lines of computational evidence support a deleterious effect on the gene/gene product), while variants with ≥ 5 out 9 "tolerant" predictions met the BP4 criteria (multiple lines of computational evidence suggest no impact on a gene/gene product) (Supplemental Table 5). For criteria based on allele frequencies alone (PM2, BS1), we used the maximum AF observed in any population across gnomAD v2 genomes, gnomAD v2 exomes, gnomAD v3 genomes, and ExAC exomes. Variants with an AF < 0.001% met criteria PM2, and variants with AF ≥ 0.005% (heterozygous) and AF ≥ 0.2% (homozygous) were considered to be higher than expected for the disorder, meeting BS1. We estimated the maximum credible AF for a variant by considering an OFC prevalence of 1 in 1,000, 5% of allelic heterogeneity, 100% of genetic heterogeneity, and 50% penetrance [40]. All genetic variants classified as 'pathogenic' or 'likely pathogenic' were counted towards the diagnostic yield calculation and are referred to as 'likely pathogenic' throughout the manuscript.

Statistical Analysis

Statistical tests were carried out to calculate differences between groups using two-sided chi-square and Fisher exact tests, which were conducted using R (version 3.6.3). We performed 10,000 permutations for the chi-square tests comparing cases and controls (overall, by cleft subtype, population, and sex) to adjust for multiple hypothesis testing under the null hypothesis of no association between the number of individuals with 'likely pathogenic' variants and case-control status. The significance level was set at $P < 0.05$ for these tests. Odds ratios and 95% confidence intervals were estimated through chi-square tests in R (version 3.6.3).

We tested gene-based associations in genes with 'variants of uncertain significance' (VUS) using the Optimal Sequence Kernel Association test (SKAT-O), which unites the Sequence Kernel Association test (SKAT) and the burden test objectives to maximize statistical power while allowing for variants of opposite effects [41]. Data were converted to binary PLINK files and imported into the SKAT package (version 2.0.1) [42] in R (version 3.6.3). First, we performed a SKAT-O test for 161 genes with more than one VUS or 'likely pathogenic' variant. We then excluded all "solved" cases and controls with 'likely pathogenic' variants and carried out SKAT-O tests on 151 genes with more than one VUS in the remaining cases and controls. We used a Bonferroni correction to adjust for multiple testing in all analyses.

RESULTS

We identified a total 3,004 SNVs, small indels, and SVs from 841 OFC cases and 294 controls. After sorting variants into tiers designed to prioritize variants, we narrowed our list to 1,725 genetic variants for manual review under ACMGG criteria (Supplemental Figure 2C). On average, we reviewed 1.53 variants per case and 1.48 variants per control ($p=0.22$).

After ACMGG review, 80 of the variants (4.64%) were classified as 'likely pathogenic' or 'pathogenic' (hereafter referred to as 'likely pathogenic') (Supplemental Table 6). These 80 'likely pathogenic' variants were dominated by variants presumed to be loss-of-function (LoF): 46.3% were stop-gain, frameshifting indels, and canonical splice site variants, and 15% were SVs. Overall, 9.04% of cases and 1.36% of controls had 'likely pathogenic' variants ($p<0.0001$, Figure 1). Stratifying our gene list by the mode of inheritance, we found 'likely pathogenic' variants were almost exclusively in autosomal dominant genes (8.80% of cases vs. 1.36% of controls; $p<0.0001$). Consistent with our previous analysis demonstrating an excess of *de novo* mutations in clinically relevant genes among OFC

cases [12], 3.69% of cases (versus 0.340% of controls; p=0.003) had a *de novo* 'likely pathogenic' variant. Notably, we did not identify any 'likely pathogenic' homozygous or compound heterozygous variants in autosomal recessive genes among cases or controls. This absolute lack of signal was unexpected because a subset of the trios came from consanguineous families from Turkey and Colombia. Similarly, there was a limited contribution from X-linked genes. Only two individuals (0.24% of cases) had 'likely pathogenic' variants on the X chromosome: a hemizygous male with a loss-of-function variant in *PHF8* (MIM: 300560) inherited from his unaffected mother and a heterozygous female with a *de novo* in-frame deletion in *FLNA* (MIM: 300017).

Epidemiology, developmental embryology, and recent association studies suggest some differences in the genetic architecture of specific OFC subtypes [43-45]. We, therefore, stratified the case cohort into CL, CLP, and CP to test for any differences in diagnostic yield across subtypes (Figure 1B). Among CLP cases, which represent a majority (79%) of the OFC cohort, 9.09% had a 'likely pathogenic' variant (60 out of 660 total CLP cases) (p=0.0003 vs. controls). The diagnostic yield was much higher among CP cases (13 out of 74 total CP cases), where 17.6% had a 'likely pathogenic' variant (p<0.0001 vs. controls and p=0.035 vs. CLP). Equally striking was the difference between CL and CLP, which have historically been viewed as a variation in severity of the same disorder and are commonly analyzed together. Only 2.80% of CL cases (3 out of 107 total CL cases) had a 'likely pathogenic' variant, which was not significantly different than controls (p=0.395) and only nominally different from CLP (p=0.045). These data suggest the differences in genetic architecture between CL and CLP seen in several studies [46-48] may extend to rare variants.

There are characteristic sex biases among OFC subtypes where CP occurs twice as frequently in females than males, and CL/P occurs twice as frequently in males than females [49; 50]. We considered whether the documented sex bias was also reflected in the yields for each sex and OFC subtype. Although the less frequently affected sex had consistently higher diagnostic yields within each subtype, none of these differences were statistically significant (Figure 1C). These results could be consistent with a "protective effect" model; when there are disease prevalence differences between sexes, affected individuals among the less commonly affected sex have, on average, greater enrichment for disease causing alleles or alleles of much larger effect than members of the more commonly affected sex. This would also be consistent with the observation that sex biases are not as commonly observed in Mendelian OFC syndromes. We also observed small (but non-significant) differences in yield when broadly stratifying by population (Supplemental Figure 3). Although this is loosely correlated with OFC prevalence rates [51], it is more likely that these differences are due to the differences in the representation of these populations in reference databases that impact the filtering of variants based on allele frequency.

The 76 'likely pathogenic' variants in individuals with OFCs were found across 39 genes, constituting 7.75% of the gene list (Figure 2). Overall, we found sixteen recurrently mutated genes, and nine of these had at least three 'likely pathogenic' variants in cases. These nine genes: *CTNND1* (6 cases [MIM: 601045]), *ARHGAP29* (5 cases [MIM: 610496]), *COL2A1* (5 cases [MIM: 120140]), *IRF6* (5 cases [MIM: 607199]), *TFAP2A* (5 cases [MIM: 107580]), *CDH1* (4 cases [MIM: 192090]), *CHD7* (3 cases [MIM: 608892]), *PDGFC* (3 cases [MIM: 608452]), and *TBX1* (3 cases, all 22q11.2 deletions [MIM: 602054]) accounted for 4.64% of OFC cases alone. Eight of these genes (and 35 out of 39 genes with 'likely pathogenic' variants) were described as autosomal dominant genes in the literature. Of the 205 genes we studied that were associated with autosomal dominant disease, 17.1% had at least one 'likely pathogenic' variant, demonstrating the substantial genetic heterogeneity of OFCs. Although previous

studies have found 'likely pathogenic' variants in two or more disease loci in the same individual[52], we did not identify any cases with more than one 'likely pathogenic' variant in this study.

Previous exome studies in OFCs report incomplete penetrance for several genes, including *CTNND1*, *CDH1, and TP63* (MIM: 603273) [11; 53], but few have studied large datasets drawn from both simplex and multiplex families, giving us an opportunity to weigh the contribution of *de novo* and transmitted variants, as well as estimate penetrance for various autosomal dominant variants. A total of 220 probands were drawn from multiplex families, defined as having at least one other affected relative (up to the 3rd-degree). There was no difference in diagnostic yield between individuals from multiplex and simplex families (11.8% multiplex vs. 7.73% simplex, p=0.089). However, there were notable differences in the types of variants identified (Supplemental Figure 4). Twenty of the 26 'likely pathogenic' variants in individuals from multiplex families were transmitted (Supplemental Figure 5); the rest were *de novo* (Supplemental Figure 6). In contrast, 52.1% (25 out of 48) of 'likely pathogenic' variants in simplex families were *de novo*. The aligned reads for all six *de novo* variants identified in multiplex families were visually inspected, confirming the *de novo* call (Supplemental Figure 6). In three of the six families, one parent was also affected, and we cannot exclude the possibility of mosaicism in other tissues. But it is also possible these *de novos* are not the only genetic variants conferring risk in these pedigrees.

Among the transmitted variants in multiplex families, we asked how often the variant was transmitted by the parent with a personal or family history of OFC. 82.4% (14/17) of variants were transmitted by the parent with a family history of OFCs with no differences by maternal vs. paternal history; 64% (9/14) of these variants were transmitted by an affected parent. We can therefore estimate the global penetrance to be approximately 60% among multiplex families. Including all transmitting parents, including those from simplex families (who are all unaffected), the penetrance of transmitted variants falls to 25% (9/36). Interestingly, almost all of these transmitted variants are predicted to be loss-of-function, and impacted genes included *ARHGAP29*, *CDH1*, *CTNND1*, *TFAP2A*, and *TP63*, each of which is considered haploinsufficient with reduced penetrance [54-56].

The vast majority (60.6%) of all classified variants were variants of uncertain significance (VUS). Overall, we found a significant enrichment of VUSs among OFC cases compared to controls (61.8% cases vs. 50.3% controls, $p=7.31 \times 10^{-4}$). This result was consistent across populations but not OFC subtypes (Supplemental Table 7). VUSs were not clustered either in cases with 'likely pathogenic' variants or in cases without such variants as removing "solved" cases/controls resulted in an almost identical enrichment: specifically, 61.4% of 765 "unsolved" cases have at least one VUS vs. 50% of 290 controls ($p=9.88 \times 10^{-4}$).

One possible hypothesis to explain the excess of VUSs in cases is that there are cryptic 'likely pathogenic' variants among this set of VUSs. For both cases and controls, VUSs were overwhelmingly missense variants, which is not surprising given the challenges of interpreting this class of variation. We expected that VUSs among cases would have greater "damaging" prediction scores, but there was no difference in the distribution pathogenicity predictions aggregated under nine different algorithms (see variant annotation methods, Supplemental Figure 7).

We next hypothesized that VUSs would be more likely occur in genes with 'likely pathogenic' variants. Collectively, VUS were similarly enriched among genes with 'likely pathogenic' variants (OR 1.61, p=0.008) as they were among genes without 'likely pathogenic' variants (OR 1.48, p=0.005) (Figure 3A). To try to further parse which sets of genes were contributing the VUS signal, we categorized genes based of their evidence level on the NHS cleft panel, which corresponds to three

levels of support ("green" for high evidence, "amber" for moderate evidence, and "red" for low evidence) as reviewed by an expert panel. The enrichment of VUSs was strongest among 69 "green" genes (OR 2, $p=1.36 \times 10^{-4}$) but there was also an enrichment of VUSs among the 129 genes that were not on the UK Panel App list (OR 1.45, p=0.011) (Figure 3A). We then performed SKAT-O tests for all autosomal genes with 'likely pathogenic' variants and/or VUS to pinpoint individual genes where VUSs are contributing to the signal. Although no genes individually reached formal significance (due to an unbalanced sample size favoring cases), *PRICKLE1* (MIM: 608500) was nominally significant with an odds ratio indicating an increased risk for OFC (Supplemental Figure 8). Furthermore, for many genes with multiple 'likely pathogenic' variants, there was an increase in odds ratio from the VUSs (Figure 3B).

Lastly, given that our gene list was created through curating five different sources, we asked whether there was a specific gene source harboring genes with 'likely pathogenic' variants that could be used to prioritize sequencing analyses. We found genes from our manually curated gene panel, the Prevention Genetics CL/P panel, and the National Health Service PanelApp lists had the highest proportion of autosomal dominant and X-linked genes with 'likely pathogenic' variants (Supplemental Figure 9). The majority of autosomal dominant genes with 'likely pathogenic' variants (27 out of 35) were shared across at least three sources, including the manually curated genes, the Prevention Genetics, and the NHS panels. In contrast, the list derived from the Developmental Disorders Genotype-to-Phenotype database (DDG2P) had the lowest percentage (6.77%) of 'likely pathogenic' variants, indicating genes associated more generally with facial dysmorphism are not likely to be a fruitful source of variants causing OFCs.

DISCUSSION

Current genetic diagnostic approaches for OFCs are usually performed on individuals with a family history of OFCs consistent with Mendelian inheritance patterns and individuals with syndromes. Consequently, diagnostic testing may be conducted in only a small fraction of OFC cases, creating a potential clinical diagnostic gap. Previous studies in OFCs estimated the genetics diagnostic yield to be ~10% using whole-exome sequencing in multiplex families [11]. Because multiplex families are more likely to be enriched for rare 'likely pathogenic' variants, it is possible this number could be an overestimate when applied to a larger, more diverse cohort. In our sample of 841 cases from multiple populations and different family structures, we also estimated the yield to be ~10%. Therefore, our results replicate previous studies of OFCs reporting 'likely pathogenic' variants in autosomal dominant genes with incomplete penetrance [11; 53; 57]. We found estimates for diagnostic yield varied greatly by OFC subtype. We observed a yield of 17.6% in CP and 9.09% in CLP cases, but only a 2.80% yield in CL cases, the latter of which was not significantly different from controls. Our results reinforce the etiologic heterogeneity in OFC subtypes observed recently from multiple genome-wide association studies [45; 58-61] and extend it to rare variants. They also suggest that when choosing to pursue clinical sequencing, proband cleft type might be a consideration. However, future investigation is still needed to replicate these findings as we focused on a selected gene set and our findings of the etiological heterogeneity in OFCs suggests there might be additional candidate genes beyond these well-established genes. Our study was also not population based and so the true diagnostic yield in unselected populations is yet to be determined.

Our estimated yield is similar to those from exome or genome-based studies of other pediatric conditions including congenital heart disease (12.7%) [62] and autism spectrum disorder (7.5%) [37],

although some estimates of diagnostic yield are as high as 40-50%. A major determining factor in these studies is the approach used and the ascertainment of the cohort. For example, Lowther and colleagues estimated the yield from WES or WGS in autism spectrum disorder at 7.5% using a panel of 907 neurodevelopmental genes [37]. They found a similar yield (12%) for a heterogeneous group of fetal structural anomalies using a panel of 2,535 genes but noted these cases had been pre-screened by karyotype and microarray, lowering their diagnostic yield. Targeted investigations such as ours may favor specificity (but lose sensitivity) since the overall yield will be lower than studies evaluating the entire exome or including other first-tier techniques such as karyotyping and microarray. Further, diagnostic yields are typically higher in syndromic cases or those with multiple congenital anomalies [63-65].

We identified 'likely pathogenic' variants in a total of 39 of the 503 genes in our selected gene list. Although several genes showed multiple suspected variants, no individual variant recurred in these data, underscoring the extensive allelic and genetic heterogeneity of OFCs. Among the recurrently mutated genes were several expected genes based on previous literature citing frequent mutations in OFCs, including *IRF6*, *CTNND1*, and *COL2A1*. We observed patterns of mutation across OFC subtypes consistent with the literature. For example, we found 'likely pathogenic' *COL2A1* mutations exclusively among CP cases [66], *TFAP2A* mutations were found exclusively among CLP cases [67], and *IRF6* mutations were found in individuals with either CP or CLP[68]. Interestingly, despite a strong genotype-phenotype correlation between *GRHL3* (MIM: 608317) and CP [69-71], the only 'likely pathogenic' variant identified in that gene was found in an individual with CL. Due to differences in sample sizes for each OFC subtype, we were not powered to quantitatively analyze genotype-phenotype correlations for each gene, so these remain anecdotal observations requiring follow-up in much larger datasets.

Interpretation of VUSs is a considerable challenge. Nearly 60.6% of variants were classified as VUSs, the majority of which were missense mutations. The effect of single amino-acid substitutions is often difficult to interpret without independent functional evidence, especially for genes with high allelic heterogeneity. *In silico* pathogenicity prediction tools have limited utility, as VUS in cases and controls were equally likely to be predicted to be damaging under multiple algorithms. There were multiple genes for which VUSs were identified in cases while no such variants were identified in controls; however, none were significant after correcting for multiple comparisons. One example is *PRICKLE1,* which was previously evaluated through family-based association studies and showed evidence of association with OFCs[72], and showed a nominally significant deleterious effect on OFC risk in the present study. Therefore, it is likely these datasets are underpowered to detect genes with a moderate burden of rare variants and the top-ranked genes should be considered candidates for further analysis pending future functional testing to sort out the effect of identified variants[73].

It is important to note this study was conducted on a cross-sectional cohort collected under multiple recruitment protocols intended for research, not diagnostics and is not representative of all OFC cases that may be referred to genetics from a craniofacial clinic. Clinical diagnostics and differential diagnoses are aided by detailed phenotyping and collection of family histories, but the availability is limited for specific populations in this cohort. Although those with multiple congenital anomalies and significant developmental delays were likely excluded from most contributing studies and such individuals should represent a minority of the dataset, the timing of recruitment, varying skills of the clinical and research teams, and different recruitment goals make this a highly heterogeneous cohort with incomplete phenotypic data needed for the clinical setting. We note, however, that many syndromes that include OFCs show incomplete penetrance and variable

expressivity, which can complicate a diagnosis based on phenotype alone even when detailed phenotyping is available. In this study, we were able to estimate penetrance of OFC for transmitted alleles but were not able to estimate the extent of variable expressivity of other phenotypic features. Moreover, these penetrance estimates represent global penetrance, rather than gene-level, which would require additional investigation in larger cohorts. Nonetheless, the low penetrance of 'likely pathogenic' variants was quite striking, as many of these variants were predicted to be 'loss of function'. The ideal cohort to fully evaluate penetrance and expressivity would be a prospectively recruited, deeply phenotyped and sequenced cohort of sequential cases, but such cohorts are difficult and costly to assemble. Although imperfect, sequencing studies such as this one as well as those investigating the functional consequences of variants in model systems will be necessary to advance translation of research to clinical practice.

In conclusion, we evaluated the diagnostic yield of OFC cases and controls. We utilized a framework to prioritize SNVs/indels and SVs in 503 selected candidate OFC genes using whole-genome sequencing. This work provides in-depth insight into the heterogeneous etiology of OFCs and aids in gene and variant interpretation as the clinical application of sequencing in OFCs becomes more widespread. In addition to the potential clinical applications of this work, we highlight the critical need for high-throughput validation of genetic variants to quantitatively distinguish the effects of individual rare variants. Future work in this area should allow for improved variant interpretation in a clinical setting, greater understanding of the function of genes influencing craniofacial birth defects, and may help explain the variable penetrance observed in this study.

Declaration of Interests:
The authors declare no competing interests.

Acknowledgements
These studies are part of the Gabriella Miller Kids First Pediatric Research Program, supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH). Sequencing of the European trios was completed at Washington University's McDonnell Genome Institute (3U54HG003079-12S1 and X01-HL132363 [M.L.M., E.F.]) and the Colombian and Taiwanese trios were sequenced at the Broad Institute Sequencing Center (U24-HD090743, X01-HL136465 [M.L.M., E.F.], X01-HL140516 [T.H.B.]). The sequencing centers plus the Kids First Data Resource Center, supported by the NIH Common Fund through grant U2CHL138346, provided technical and analytical support of this project. The assembling of the sample of child-parent trios, collection of the phenotypic data and samples, and data analysis were supported by NIH grants: R01-DE016148 [M.L.M. and S.M.W.], R03-DE026469 [E.F. and M.L.M.], R03-DE027193 [E.J.L.], R03-DE027103 [E.J.L], R00-DE025060 [E.J.L.], R01-DE027983 [E.J.L.], R01-DE028342 [E.J.L], R01-DE030342 [E.J.L], R01-DE011931 [J.T.H.], U01-DD000295 [G.W.], R03-DE027121 [T.H.B], and T32-GM008490 [K.D.P], R01-DE031261 [H.B.], and R00-DE026824 [H.B.]. This work was supported in part by a grant to Emory University from the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program [K.D.P]. This study would not be possible without the dedication of many families, study teams, and colleagues worldwide.

Data Availability
The case data analyzed and reported in this manuscript were accessed from the database of Genotypes and Phenotypes (dbGaP; European trios, dbGaP: phs001168.v2.p2; Colombian trios, dbGaP: phs001420.v1.p1; Taiwanese trios, dbGaP: phs000094.v1.p1) and from the Kids First Data Resource Center. The control data is available from public data repositories as described in https://www.internationalgenome.org/data-portal/data-collection/30x-grch38.

FIGURES

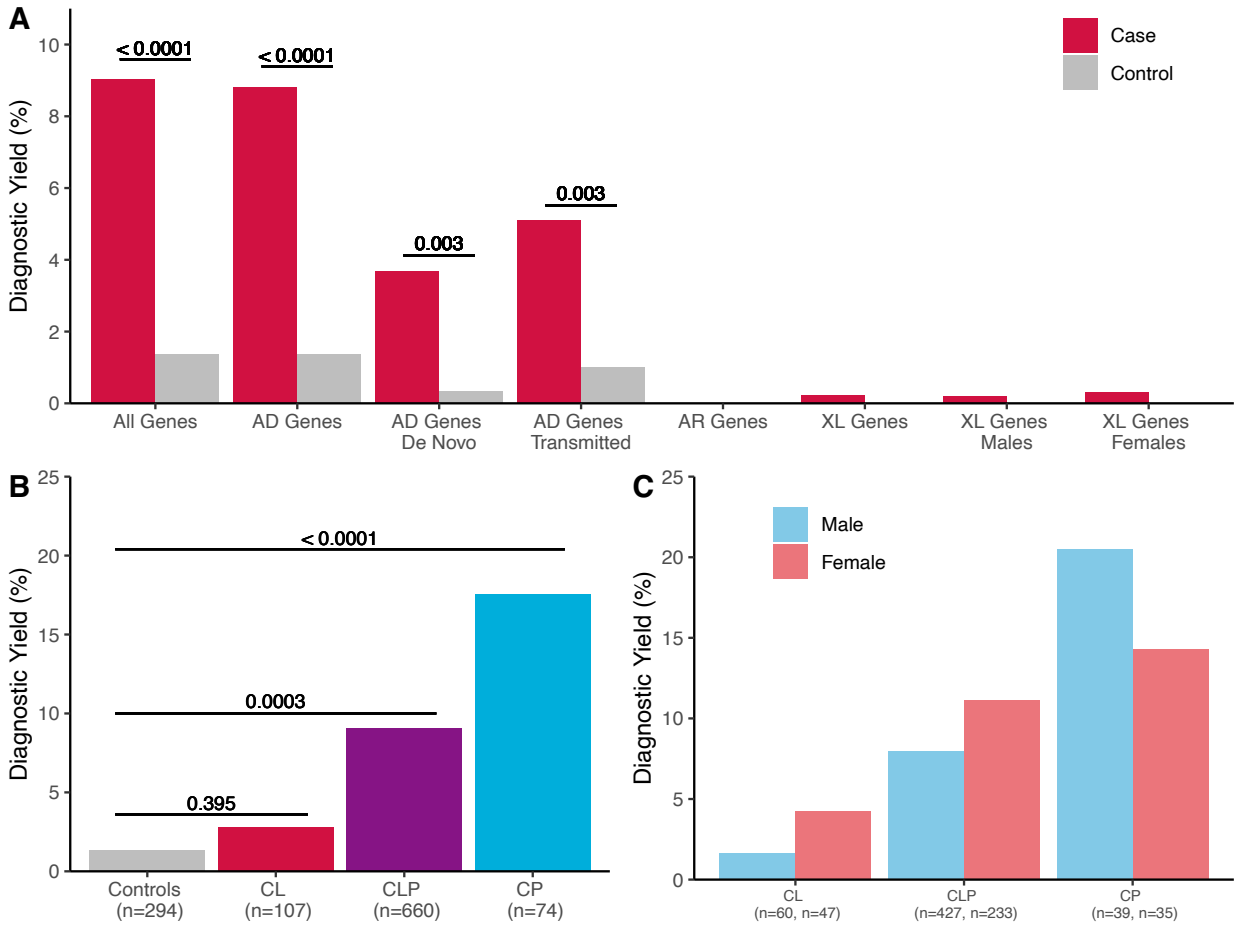


Figure 1. Diagnostic Yield of 503 OFC Genes. (A) The percentage of cases (red) and controls (gray) with at least one 'likely pathogenic' variant by mode of inheritance. (B) The percentage of individuals with 'likely pathogenic' variants in controls (gray) and in cases by OFC subtype: cleft lip (CL, red), cleft lip and palate (CLP, purple), and cleft palate (CP, blue). (C) Percentage of cases with a 'likely pathogenic' variant stratified by cleft type and proband sex (female (pink) and male (blue)). Yields were not significantly different between males and females in any cleft subtype. P-values were calculated using chi-square tests with 10,000 permutations in R.

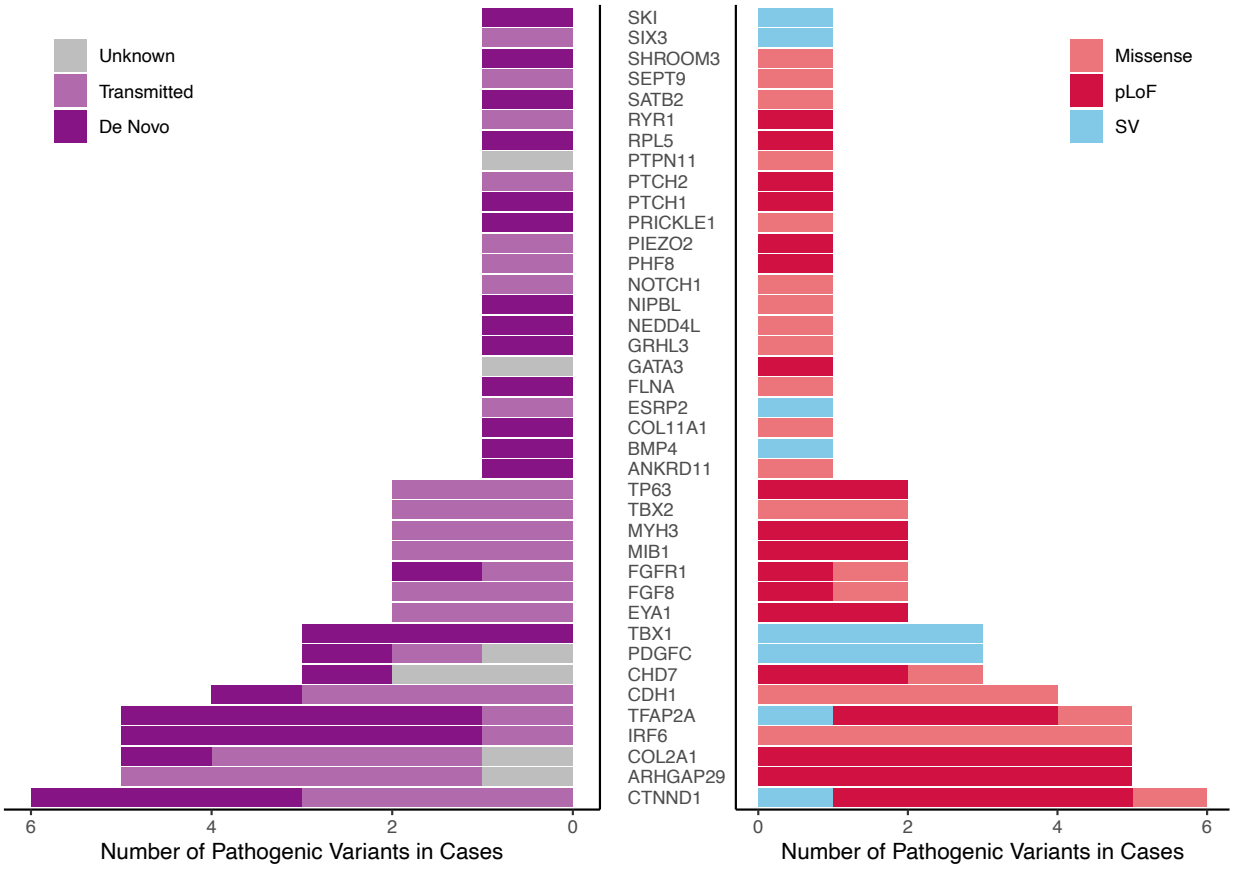


Figure 2. Genes with Likely Pathogenic Variants in Cases. The count of 'likely pathogenic' variants in cases in each of the 39 genes with 'likely pathogenic' variants. Variants are colored in the left panel based on the mode of inheritance: *de novo* (dark purple), transmitted (light purple), or were unknown (gray), and in the right panel based on variant consequence: missense (red) variants, predicted loss-of-function (pLoF, dark red), and structural variants (blue).

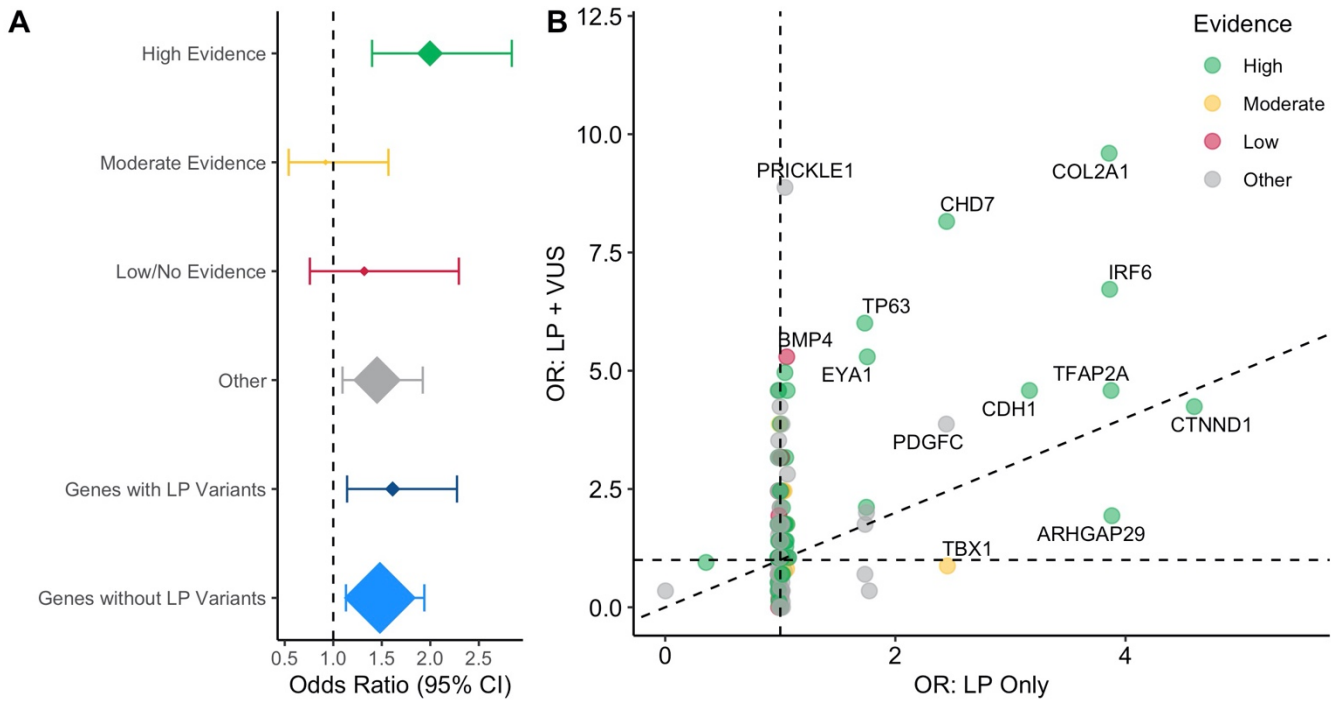


Figure 3. Variants of Uncertain Significance (VUS) are enriched in cases vs. controls. (A) Enrichment of VUSs in cases for sets of genes. Odds ratios and 95% confidence intervals are calculated from a chi-square test. Point estimates are scaled by the number of genes in the gene set. (B) Odds ratios for individual gene for 'likely pathogenic' (LP) variants only vs. 'likely pathogenic' and variants of uncertain significance (LP + VUS). The dotted lines show OR = 1 for the x and y axes and y=x. In both panels, genes/gene sets are colored by level of evidence in the UK NHS Cleft Lip and Palate PanelApp: high evidence (green), moderate evidence (amber), low evidence (red), or not on the PanelApp gene list (gray).

SUPPLEMENTAL FIGURES

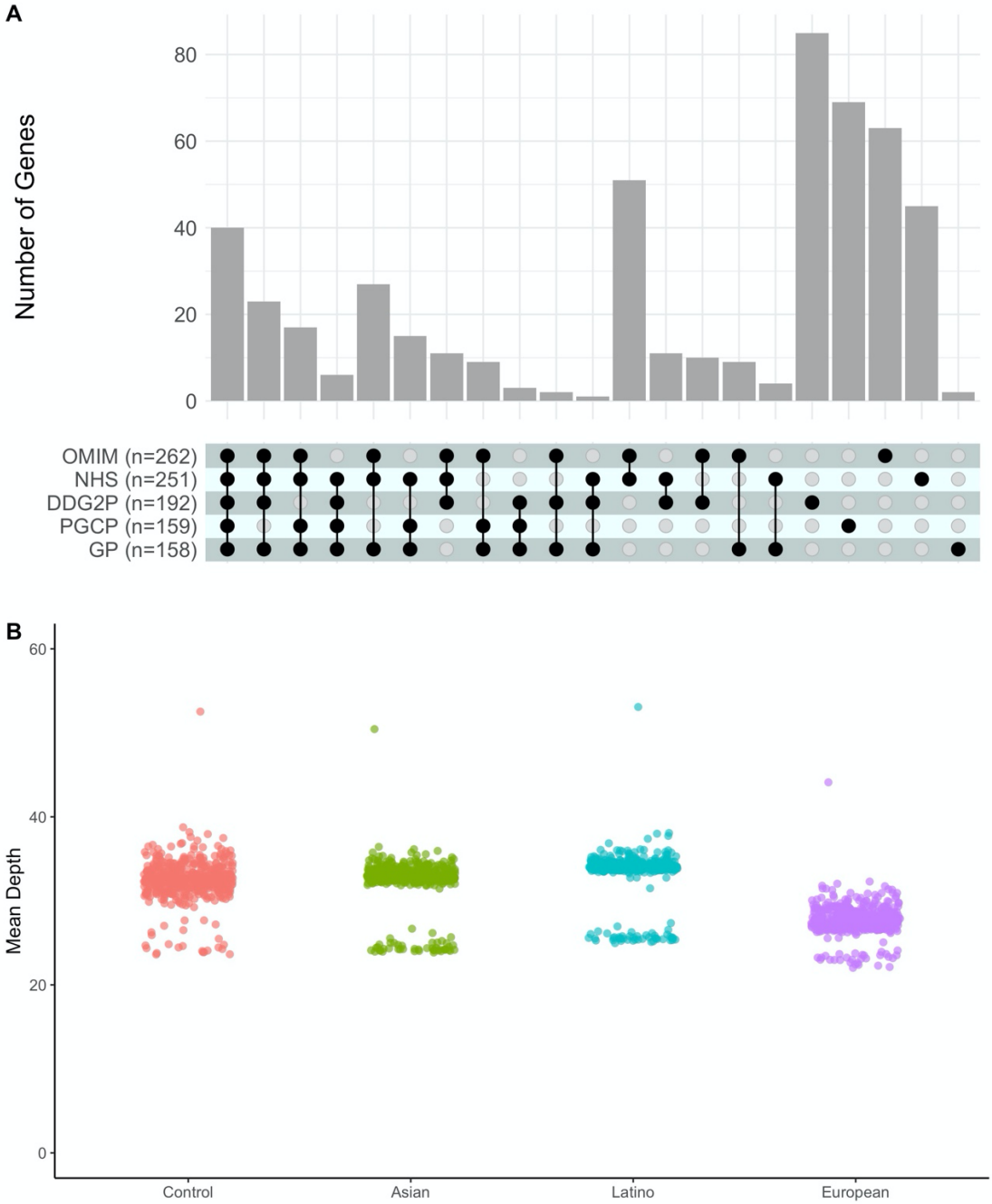


Figure S1. OFC Gene List. (A) We created a gene list containing 503 genes associated with OFCs and craniofacial development obtained from five sources: the Online Mendelian Inheritance in Man (OMIM), the National Health Service (NHS), the Developmental Disorders Genotype to Phenotype (DDG2P), Prevention Genetics Cleft Lip & Palate Panel (PGCP), and a manually curated gene panel (GP). The overlap between sources is shown by the connecting lines on the lower panel while the number of genes in each overlap and individual source is shown by the gray bars. **(B)** The sequencing coverage of the OFC genes (circles) by study population: Controls (pink), Asian (green), Latino (blue), and European (purple).

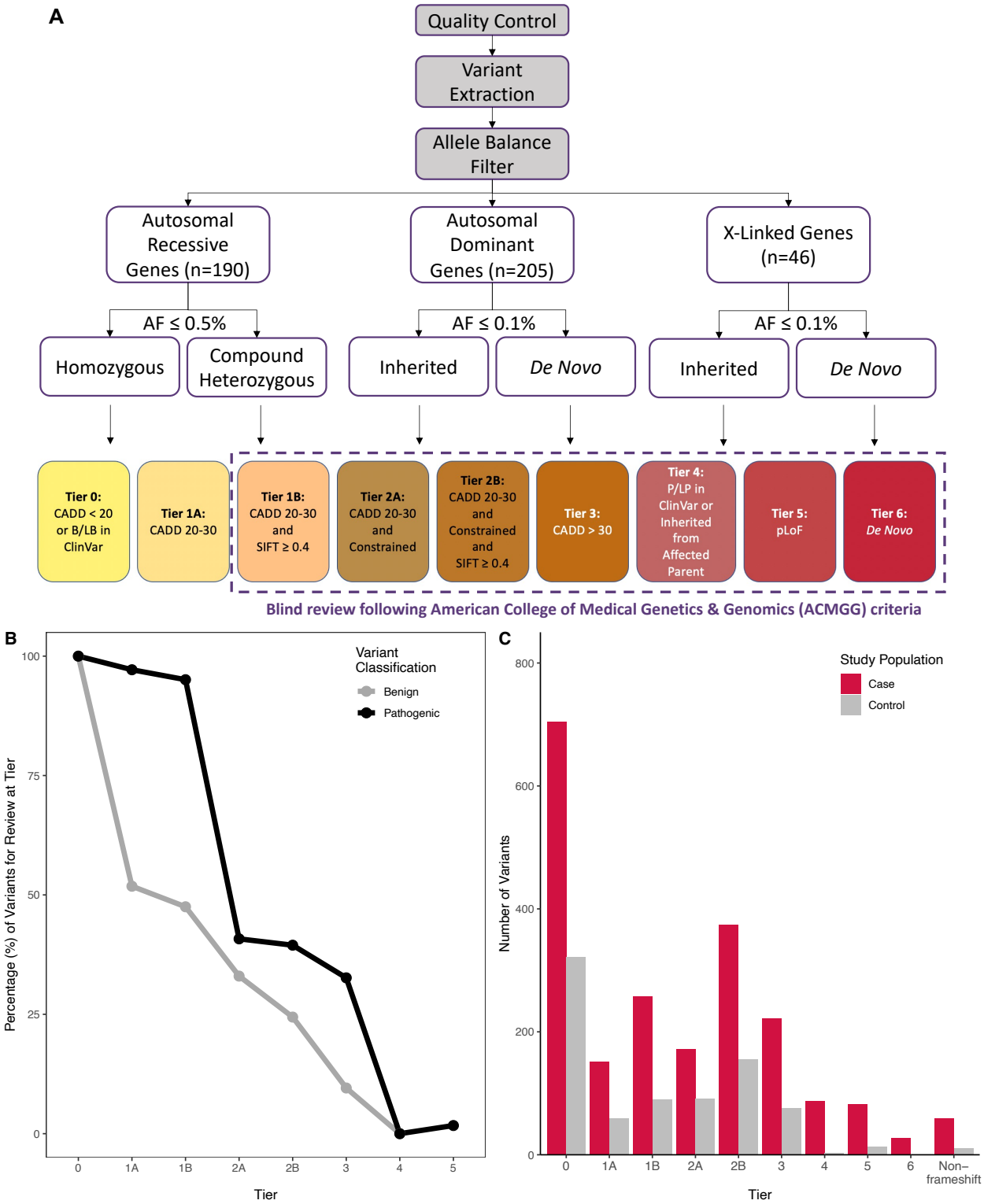


Figure S2. Whole-Genome Sequencing and Variant Filtering Pipeline. (A) Variants in 503 genes were sorted into categories and filtered on allele frequency and mode of inheritance. Variants were then sorted into bins in increasing likelihood of pathogenicity. Variants in Tier 1B or greater were manually reviewed using the American College of Medical Genetics & Genetics criteria. (B) The percentage of

pathogenic (red) and benign (gray) ClinVar variants in genes from the OFC gene list at each tier threshold. (C) The number of variants from OFC cases and controls sorted into tiers.

CADD: Combined Annotation Dependent Depletion; B/LB: 'Benign' or 'Likely Benign' Variants; SIFT: Sorting Intolerant from Tolerant; P/LP: 'Pathogenic' or 'Likely Pathogenic' Variants; pLoF: Predicted Loss-of-Function Variants

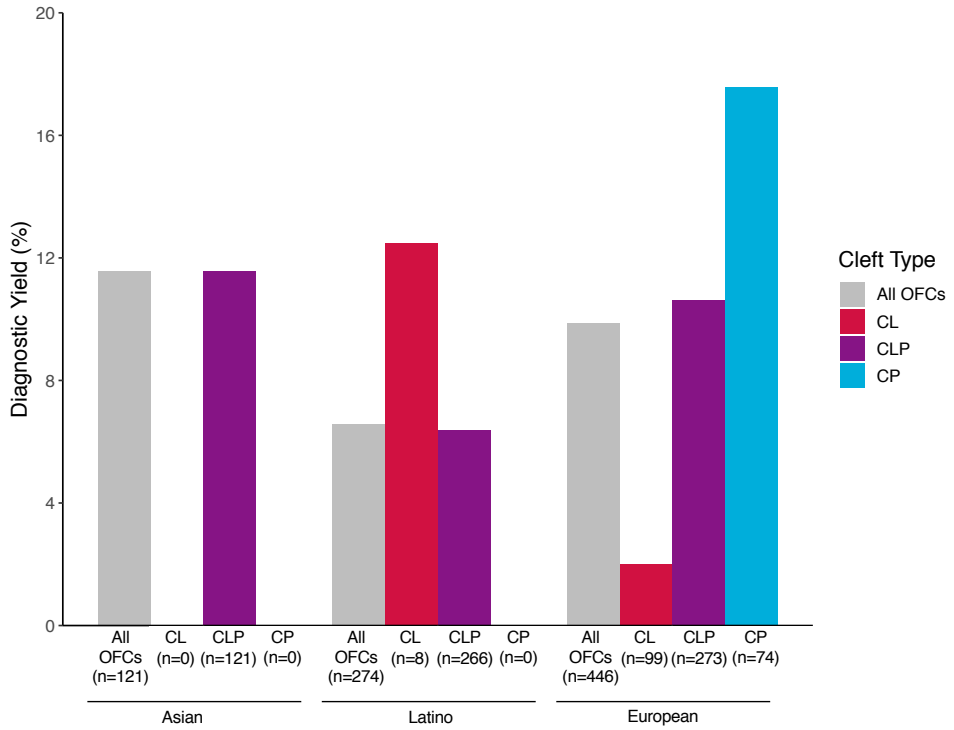


Figure S3. Diagnostic Yield by Cleft Type and Population. The sample size for each OFC subtype and population group is denoted below each bar.

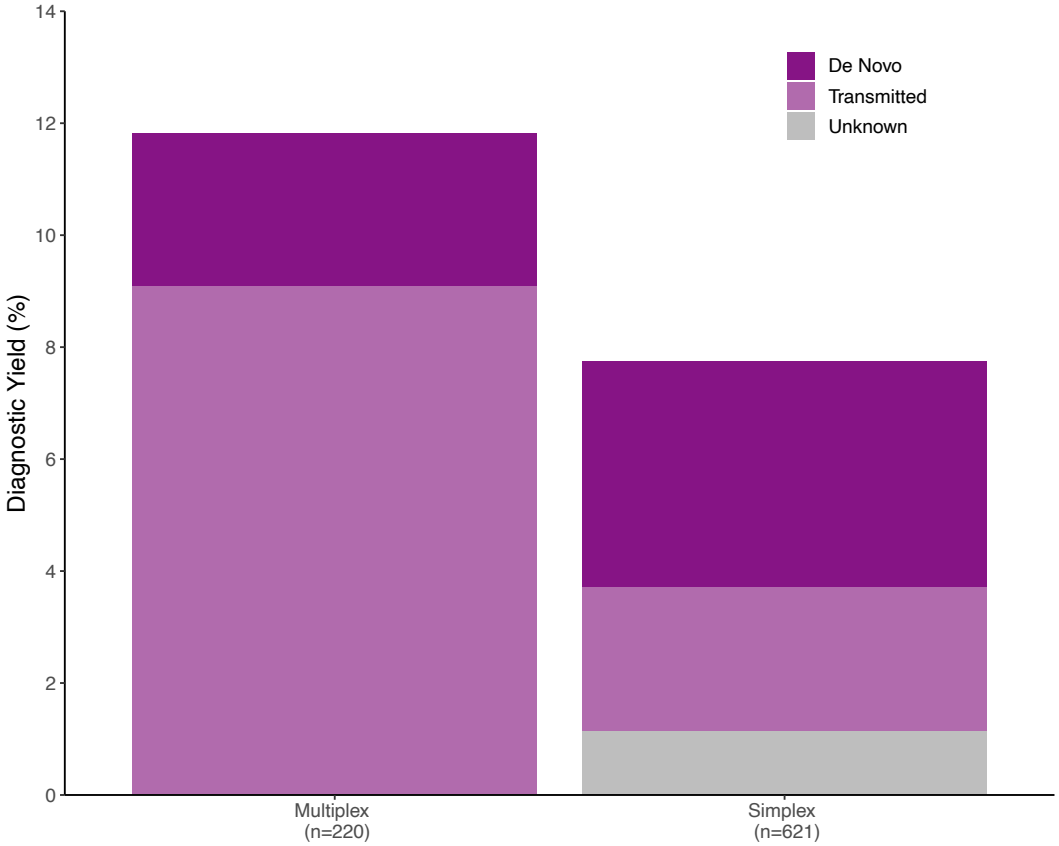


Figure S4. Diagnostic Yield by Family Type. The diagnostic yield from multiplex families (N=220) versus simplex families (N=621). 'Likely pathogenic' variants are classified by mode of inheritance: *de novo* (dark purple), transmitted from a parent (light purple), or unknown (gray).

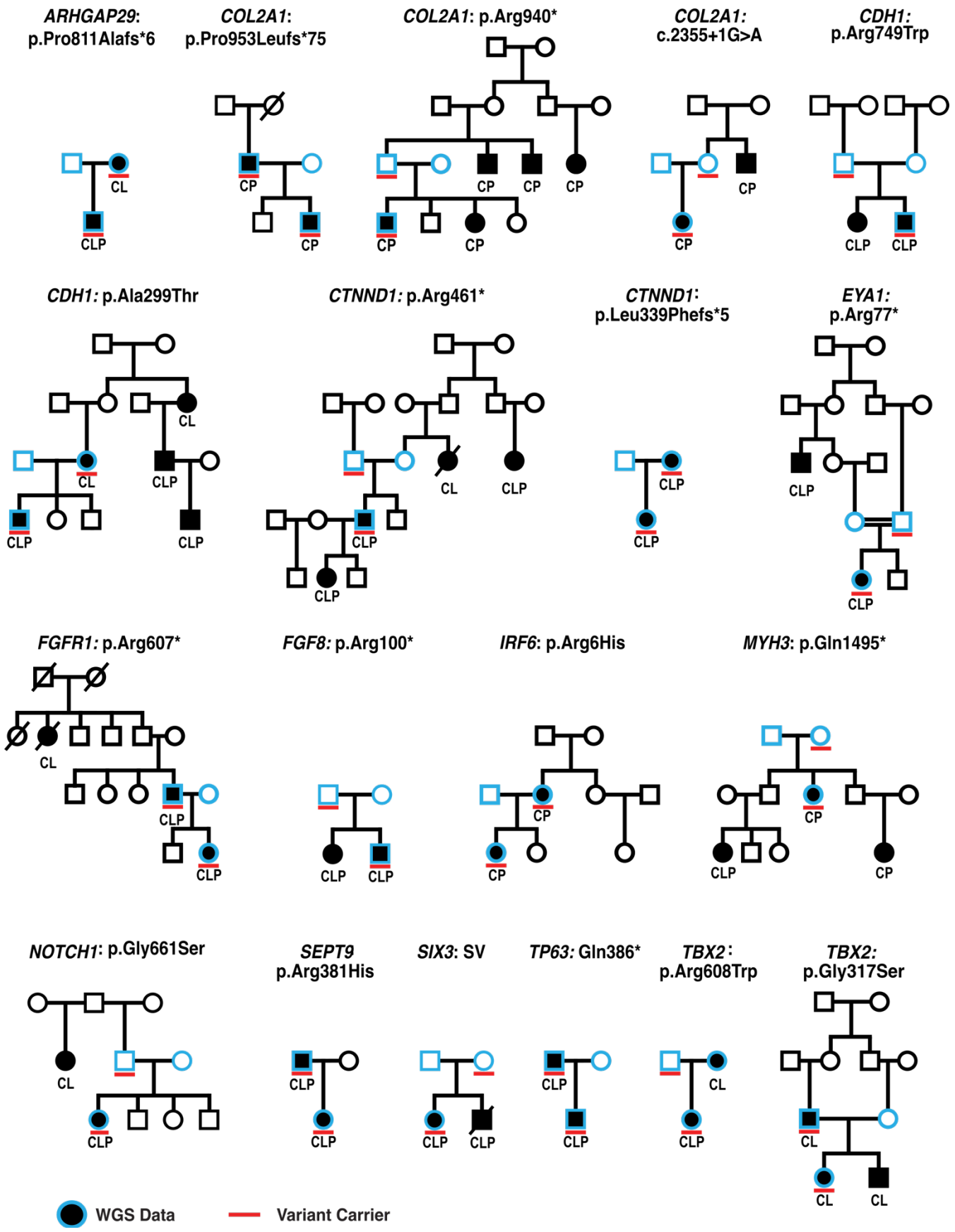

ARHGAP29: p.Pro811Alafs*6
COL2A1: p.Pro953Leufs*75
COL2A1: p.Arg940*
COL2A1: c.2355+1G>A
CDH1: p.Arg749Trp
CDH1: p.Ala299Thr
CTNND1: p.Arg461*
CTNND1: p.Leu339Phefs*5
EYA1: p.Arg77*
FGFR1: p.Arg607*
FGF8: p.Arg100*
IRF6: p.Arg6His
MYH3: p.Gln1495*
NOTCH1: p.Gly661Ser
SEPT9 p.Arg381His
SIX3: SV
TP63: Gln386*
TBX2: p.Arg608Trp
TBX2: p.Gly317Ser
CL
CP
CLP
WGS Data
Variant Carrier

Figure S5. Inherited Likely Pathogenic Variants in Multiplex Families. The mutated gene and consequence are noted above each pedigree (refer to Table S6 for variant details). Sex symbols with solid black indicate the phenotype of the individual: CL (cleft lip), CP (cleft palate), and CLP (cleft lip and palate). The red solid lines below individuals indicate variant carriers while the blue outline of the sex symbols indicates individuals with WGS data.

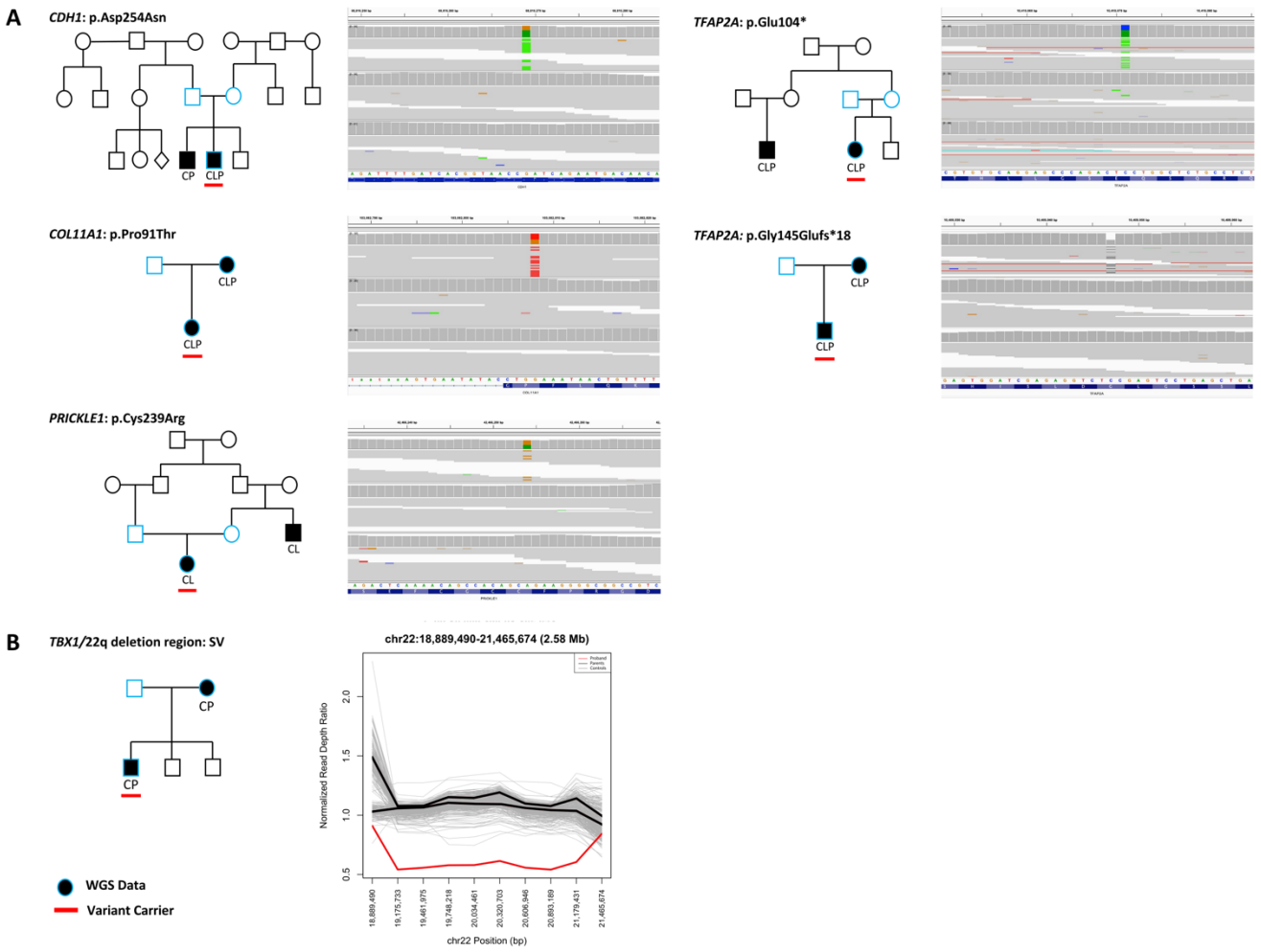


Figure S6. *De Novo* Mutations in Multiplex Families. (A) We identified 'likely pathogenic' *de novo* variants in *CDH1*, *COL11A1*, *PRICKLE1*, and *TFAP2A* (refer to Table S6 for variant details). We confirmed the *de novo* mutation by visual inspection of the proband (top segment), father (middle segment), and mother (bottom segment) reads using the Integrative Genomics Viewer (IGV) [38]. **(B)** We identified a *de novo* deletion encompassing the *TBX1*/22q deletion region. On the right, we the read depth ratio of the proband (red), parents (black), and controls (gray) of the SV region on the right. In each pedigree, the red solid lines below symbols indicate variant carriers while the blue outline correspond to sequenced individuals.

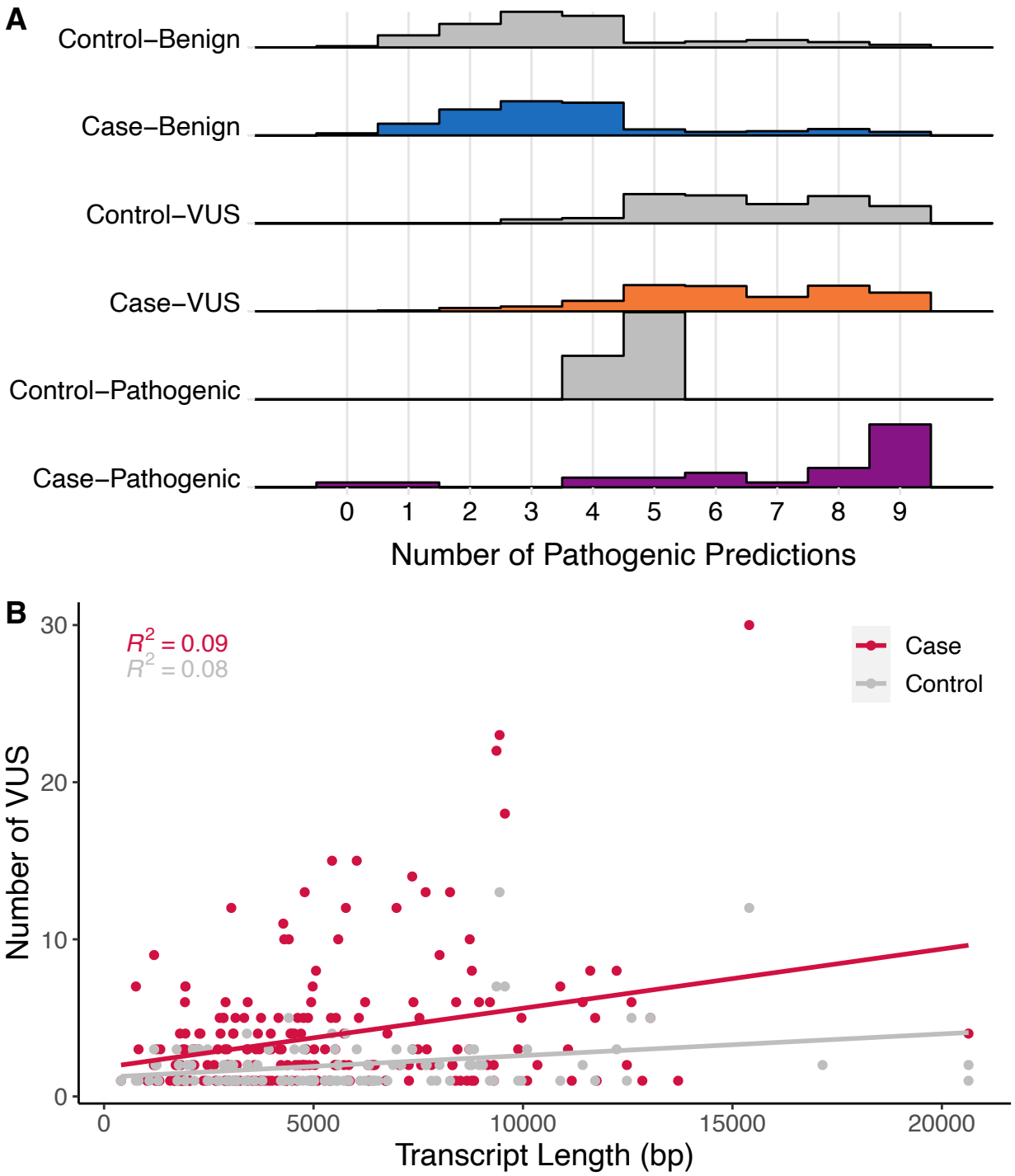


Figure S7. **Variants of Uncertain Significance in Cases and Controls.** (A) Distribution of the number of *in silico* prediction tools from nine different algorithms predicting a missense variant to be likely pathogenic/damaging for 'likely benign', VUS, and 'likely pathogenic' variants. (B) The number of VUS in genes is correlated with transcript length in cases (red, $p=1.23 \times 10^{-05}$) and controls (gray, $p=2.11 \times 10^{-03}$).

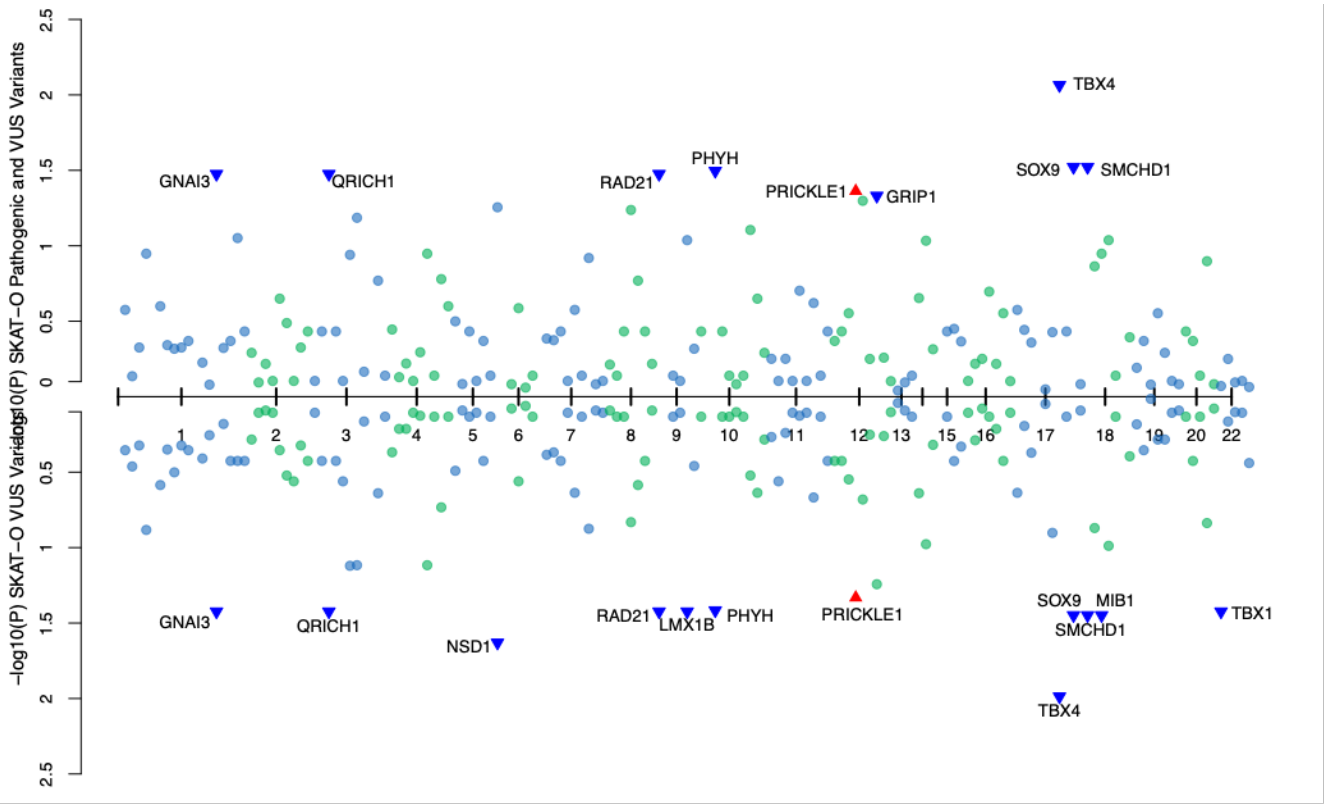
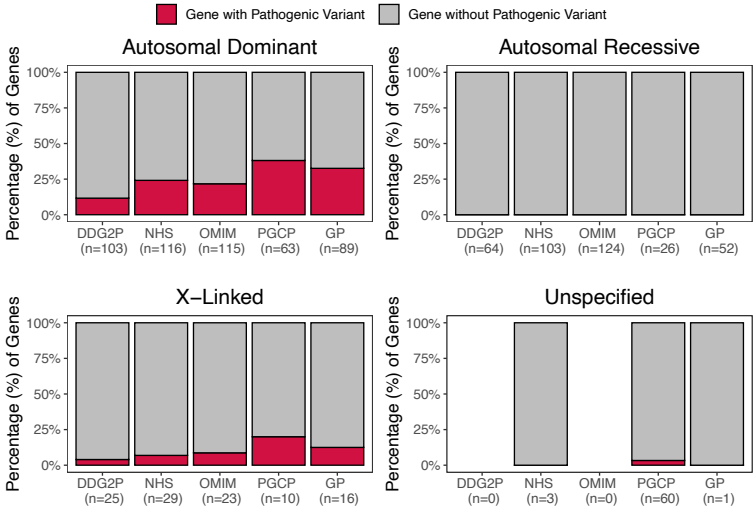


Figure S8. Gene-Based Association Tests of VUS. SKAT-O gene-based association tests for 161 genes with VUS and/or 'likely pathogenic' variants (top) and 151 genes with VUS variants (excluding individuals with 'likely pathogenic' variants) (bottom). Each gene with $p < 0.05$ is labeled according to the direction of effect with a triangle: decreased risk for cases (dark blue) and increased risk for cases (red). No gene reached a formal Bonferroni significance threshold ($p < 3.11 \times 10^{-4}$ (top) and $p < 3.31 \times 10^{-4}$ (bottom)).

A



B

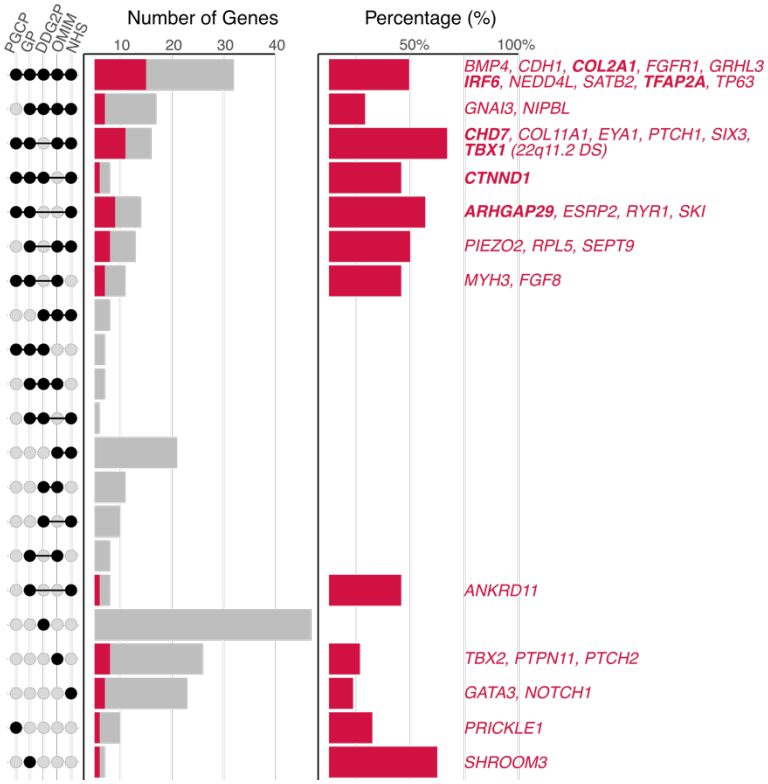


Figure S9. Genes with Likely Pathogenic Variants by Gene List Source. (A) The proportion of autosomal dominant, autosomal recessive, X-linked, and unspecified genes with 'likely pathogenic' variants (red). (B) Genes with 'likely pathogenic' variants by source.

SUPPLEMENTAL TABLES

Table S1. Demographics of OFC Cohort.
Table S2. Family History Information for Study Populations.
Table S3. OFC Gene List
Table S4. Analyzed Structural Variants
Table S5. American College of Medical Genetics & Genomics Classification Modifications
Table S6. Likely Pathogenic Variants in Present Study
Table S7. VUS Comparisons by Population and Cleft Type

REFERENCES

1. Beaty, T., Marazita, M., and Leslie, E. (2016). Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities [version 1; peer review: 2 approved]. F1000Research 5.
2. Wehby, G.L., and Cassell, C.H. (2010). The impact of orofacial clefts on quality of life and healthcare use and costs. Oral diseases 16, 3-10.
3. Setó-Salvia, N., and Stanier, P. (2014). Genetics of cleft lip and/or cleft palate: Association with other common anomalies. Eur J Med Genet 57, 381-393.
4. Alade, A., Awotoye, W., and Butali, A. (2022). Genetic and epigenetic studies in non-syndromic oral clefts. Oral Dis.
5. Leslie, E.J. (2022). Genetic models and approaches to study orofacial clefts. Oral Dis 28, 1327-1338.
6. Medicine, M.-N.I.o.G. Online Mendelian Inheritance in Man, OMIM. In. (Johns Hopkins University (Baltimore, MD).
7. Schutte, B.C., Saal, H.M., Goudy, S., and Leslie, E.J. (1993). IRF6-Related Disorders. In GeneReviews(®), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K.W. Gripp, G.M. Mirzaa, andA. Amemiya, eds. (Seattle (WA), University of Washington, Seattle
Copyright © 1993-2022, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.
8. Burdick, A.B., Bixler, D., and Puckett, C.L. (1985). Genetic analysis in families with van der Woude syndrome. J Craniofac Genet Dev Biol 5, 181-208.
9. Grosen, D., Chevrier, C., Skytthe, A., Bille, C., Mølsted, K., Sivertsen, A., Murray, J.C., and Christensen, K. (2010). A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. J Med Genet 47, 162-168.
10. Leslie, E.J., Koboldt, D.C., Kang, C.J., Ma, L., Hecht, J.T., Wehby, G.L., Christensen, K., Czeizel, A.E., Deleyiannis, F.W., Fulton, R.S., et al. (2016). IRF6 mutation screening in non-syndromic orofacial clefting: analysis of 1521 families. Clin Genet 90, 28-34.
11. Basha, M., Demeer, B., Revencu, N., Helaers, R., Theys, S., Bou Saba, S., Boute, O., Devauchelle, B., Francois, G., Bayet, B., et al. (2018). Whole exome sequencing identifies mutations in 10% of patients with familial non-syndromic cleft lip and/or palate in genes mutated in well-known syndromes. J Med Genet 55, 449-458.
12. 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. The American Journal of Human Genetics.
13. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature 526, 68-74.
14. Ng, J.K., Vats, P., Fritz-Waters, E., Sarkar, S., Sams, E.I., Padhi, E.M., Payne, Z.L., Leonard, S., West, M.A., Prince, C., et al. (2021). de novo variant calling identifies cancer mutation profiles in the 1000 Genomes Project. bioRxiv, 2021.2005.2027.445979.
15. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv, 2021.2002.2006.430068.

16. 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.
17. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31, 3812-3814.
18. 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, 248-249.
19. Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res 19, 1553-1561.
20. Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11, 361-362.
21. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 34, 57-65.
22. 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.
23. 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, 2125-2137.
24. Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper, D.N., Bernstein, J.A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 48, 1581-1586.
25. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2018). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47, D886-D894.
26. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46, D1062-d1067.
27. 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. bioRxiv, 531210.
28. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45, D840-d845.
29. Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2019). A map of constrained coding regions in the human genome. Nature Genetics 51, 88-95.
30. Werling, D.M., Brand, H., An, J.Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet 50, 727-736.
31. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444-451.

32. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220-1222.
33. Kronenberg, Z.N., Osborne, E.J., Cone, K.R., Kennedy, B.J., Domyan, E.T., Shapiro, M.D., Elde, N.C., and Yandell, M. (2015). Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput Biol 11, e1004572.
34. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res 27, 1916-1929.
35. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
36. Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40, e69.
37. Lowther, C., Valkanas, E., Giordano, J.L., Wang, H.Z., Currall, B.B., O'Keefe, K., Collins, R.L., Zhao, X., Austin-Tse, C.A., Evangelista, E., et al. (2020). Systematic evaluation of genome sequencing as a first-tier diagnostic test for prenatal and pediatric disorders. bioRxiv, 2020.2008.2012.248526.
38. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol 29, 24-26.
39. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405-424.
40. Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., MacArthur, D., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med 19, 1151-1158.
41. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Team, N.G.E.S.P.E.L.P., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American journal of human genetics 91, 224-237.
42. Seunggeun Lee, Z.Z., with contributions from Larisa Miropolsky and Michael Wu. (2020). SKAT: SNP-Set (Sequence) Kernel Association Test. In. (
43. Rahimov, F., Jugessur, A., and Murray, J.C. (2012). Genetics of nonsyndromic orofacial clefts. Cleft Palate Craniofac J 49, 73-91.
44. Sivertsen, A., Wilcox, A.J., Skjaerven, R., Vindenes, H.A., Abyholm, F., Harville, E., and Lie, R.T. (2008). Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. BMJ (Clinical research ed) 336, 432-434.
45. Carlson, J.C., Anand, D., Butali, A., Buxo, C.J., Christensen, K., Deleyiannis, F., Hecht, J.T., Moreno, L.M., Orioli, I.M., Padilla, C., et al. (2019). A systematic genetic analysis and visualization of phenotypic heterogeneity among orofacial cleft GWAS signals. Genet Epidemiol 43, 704-716.
46. Leslie, E.J., Carlson, J.C., Shaffer, J.R., Butali, A., Buxo, C.J., Castilla, E.E., Christensen, K., Deleyiannis, F.W., Leigh Field, L., Hecht, J.T., et al. (2017). Genome-wide meta-analyses of nonsyndromic

orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. Hum Genet 136, 275-286.
47. Harville, E.W., Wilcox, A.J., Lie, R.T., Vindenes, H., and Abyholm, F. (2005). Cleft lip and palate versus cleft lip only: are they distinct defects? Am J Epidemiol 162, 448-453.
48. Marazita, M.L., Lidral, A.C., Murray, J.C., Field, L.L., Maher, B.S., Goldstein McHenry, T., Cooper, M.E., Govil, M., Daack-Hirsch, S., Riley, B., et al. (2009). Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. Hum Hered 68, 151-170.
49. Marazita, M.L. (2012). The evolution of human genetic studies of cleft lip and cleft palate. Annu Rev Genomics Hum Genet 13, 263-283.
50. Burdi, A.R., and Silvey, R.G. (1969). Sexual differences in closure of the human palatal shelves. Cleft Palate J 6, 1-7.
51. Mossey, P. (2007). Epidemiology underpinning research in the aetiology of orofacial clefts. Orthod Craniofac Res 10, 114-120.
52. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. The New England journal of medicine 376, 21-31.
53. Cox, L.L., Cox, T.C., Moreno Uribe, L.M., Zhu, Y., Richter, C.T., Nidey, N., Standley, J.M., Deng, M., Blue, E., Chong, J.X., et al. (2018). Mutations in the Epithelial Cadherin-p120-Catenin Complex Cause Mendelian Non-Syndromic Cleft Lip with or without Cleft Palate. Am J Hum Genet 102, 1143-1157.
54. Savastano, C.P., Brito, L.A., Faria Á, C., Setó-Salvia, N., Peskett, E., Musso, C.M., Alvizi, L., Ezquina, S.A., James, C., Gosgene, et al. (2017). Impact of rare variants in ARHGAP29 to the etiology of oral clefts: role of loss-of-function vs missense variants. Clin Genet 91, 683-689.
55. Selvanathan, A., Nixon, C.Y., Zhu, Y., Scietti, L., Forneris, F., Moreno Uribe, L.M., Lidral, A.C., Jezewski, P.A., Mulliken, J.B., Murray, J.C., et al. (2020). CDH1 Mutation Distribution and Type Suggests Genetic Differences between the Etiology of Orofacial Clefting and Gastric Cancer. Genes 11.
56. Leslie, E.J., Mansilla, M.A., Biggs, L.C., Schuette, K., Bullard, S., Cooper, M., Dunnwald, M., Lidral, A.C., Marazita, M.L., Beaty, T.H., et al. (2012). Expression and mutation analyses implicate ARHGAP29 as the etiologic gene for the cleft lip with or without cleft palate locus identified by genome-wide association on chromosome 1p22. Birth defects research Part A, Clinical and molecular teratology 94, 934-942.
57. Bureau, A., Parker, M.M., Ruczinski, I., Taub, M.A., Marazita, M.L., Murray, J.C., Mangold, E., Noethen, M.M., Ludwig, K.U., Hetmanski, J.B., et al. (2014). Whole exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. Genetics 197, 1039-1044.
58. Curtis, S.W., Chang, D., Sun, M.R., Epstein, M.P., Murray, J.C., Feingold, E., Beaty, T.H., Weinberg, S.M., Marazita, M.L., Lipinski, R.J., et al. (2021). FAT4 identified as a potential modifier of orofacial cleft laterality. Genet Epidemiol 45, 721-735.
59. Curtis, S.W., Chang, D., Lee, M.K., Shaffer, J.R., Indencleef, K., Epstein, M.P., Cutler, D.J., Murray, J.C., Feingold, E., Beaty, T.H., et al. (2021). The PAX1 locus at 20p11 is a potential genetic modifier for bilateral cleft lip. HGG Adv 2.

60. Carlson, J.C., Standley, J., Petrin, A., Shaffer, J.R., Butali, A., Buxó, C.J., Castilla, E., Christensen, K., Deleyiannis, F.W., Hecht, J.T., et al. (2017). Identification of 16q21 as a modifier of nonsyndromic orofacial cleft phenotypes. Genet Epidemiol 41, 887-897.
61. Mukhopadhyay, N., Feingold, E., Moreno-Uribe, L., Wehby, G., Valencia-Ramirez, L.C., Restrepo Muñeton, C.P., Padilla, C., Deleyiannis, F., Christensen, K., Poletta, F.A., et al. (2022). Genome-wide association study of multiethnic nonsyndromic orofacial cleft families identifies novel loci specific to family and phenotypic subtypes. Genet Epidemiol.
62. Mone, F., Eberhardt, R.Y., Morris, R.K., Hurles, M.E., McMullan, D.J., Maher, E.R., Lord, J., Chitty, L.S., Giordano, J.L., Wapner, R.J., et al. (2021). COngenital heart disease and the Diagnostic yield with Exome sequencing (CODE) study: prospective cohort study and systematic review. Ultrasound Obstet Gynecol 57, 43-51.
63. Scott, T.M., Campbell, I.M., Hernandez-Garcia, A., Lalani, S.R., Liu, P., Shaw, C.A., Rosenfeld, J.A., and Scott, D.A. (2022). Clinical exome sequencing data reveal high diagnostic yields for congenital diaphragmatic hernia plus (CDH+) and new phenotypic expansions involving CDH. J Med Genet 59, 270-278.
64. Tammimies, K., Marshall, C.R., Walker, S., Kaur, G., Thiruvahindrapuram, B., Lionel, A.C., Yuen, R.K., Uddin, M., Roberts, W., Weksberg, R., et al. (2015). Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder. JAMA 314, 895-903.
65. Hathaway, J., Heliö, K., Saarinen, I., Tallila, J., Seppälä, E.H., Tuupanen, S., Turpeinen, H., Kangas-Kontio, T., Schleit, J., Tommiska, J., et al. (2021). Diagnostic yield of genetic testing in a heterogeneous cohort of 1376 HCM patients. BMC Cardiovasc Disord 21, 126.
66. Hoornaert, K.P., Vereecke, I., Dewinter, C., Rosenberg, T., Beemer, F.A., Leroy, J.G., Bendix, L., Björck, E., Bonduelle, M., Boute, O., et al. (2010). Stickler syndrome caused by COL2A1 mutations: genotype-phenotype correlation in a series of 100 patients. European journal of human genetics : EJHG 18, 872-880.
67. Reiber, J., Sznajer, Y., Posteguillo, E.G., Müller, D., Lyonnet, S., Baumann, C., and Just, W. (2010). Additional clinical and molecular analyses of TFAP2A in patients with the branchio-oculo-facial syndrome. Am J Med Genet A 152a, 994-999.
68. Kondo, S., Schutte, B.C., Richardson, R.J., Bjork, B.C., Knight, A.S., Watanabe, Y., Howard, E., de Lima, R.L., Daack-Hirsch, S., Sander, A., et al. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nat Genet 32, 285-289.
69. Peyrard-Janvid, M., Leslie, E.J., Kousa, Y.A., Smith, T.L., Dunnwald, M., Magnusson, M., Lentz, B.A., Unneberg, P., Fransson, I., Koillinen, H.K., et al. (2014). Dominant mutations in GRHL3 cause Van der Woude Syndrome and disrupt oral periderm development. Am J Hum Genet 94, 23-32.
70. Leslie, E.J., Liu, H., Carlson, J.C., Shaffer, J.R., Feingold, E., Wehby, G., Laurie, C.A., Jain, D., Laurie, C.C., Doheny, K.F., et al. (2016). A Genome-wide Association Study of Nonsyndromic Cleft Palate Identifies an Etiologic Missense Variant in GRHL3. American journal of human genetics 98, 744-754.
71. Mangold, E., Böhmer, A.C., Ishorst, N., Hoebel, A.K., Gültepe, P., Schuenke, H., Klamt, J., Hofmann, A., Gölz, L., Raff, R., et al. (2016). Sequencing the GRHL3 Coding Region Reveals Rare Truncating Mutations and a Common Susceptibility Variant for Nonsyndromic Cleft Palate. Am J Hum Genet 98, 755-762.

72. Yang, T., Jia, Z., Bryant-Pike, W., Chandrasekhar, A., Murray, J.C., Fritzsch, B., and Bassuk, A.G. (2014). Analysis of PRICKLE1 in human cleft palate and mouse development demonstrates rare and common variants involved in human malformations. Mol Genet Genomic Med 2, 138-151.
73. Li, E.B., Truong, D., Hallett, S.A., Mukherjee, K., Schutte, B.C., and Liao, E.C. (2017). Rapid functional analysis of computationally complex rare human IRF6 gene variants using a novel zebrafish model. PLoS Genet 13, e1007009-e1007009.