ORIGINAL ARTICLE

Accumulation of rare coding variants in genes implicated in risk of human cleft lip with or without cleft palate

Nicholas J. Marini[1] [ID] | Kripa Asrani[1] | Wei Yang[2] | Jasper Rine[1] | Gary M. Shaw[2] [ID]

[1]California Institute for Quantitative Biosciences, Department of Molecular and Cellular Biology, University of California, Berkeley, California

[2]Department of Pediatrics, Stanford University School of Medicine, Stanford, California

Correspondence
Nicholas J. Marini, California Institute for Quantitative Biosciences, Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720.
Email: nmarini@berkeley.edu

Abstract

Cleft lip with/without cleft palate (CLP) is a common craniofacial malformation with complex etiologies, reflecting both genetic and environmental factors. Most of the suspected genetic risk for CLP has yet to be identified. To further classify risk loci and estimate the contribution of rare variants, we sequenced the exons in 49 candidate genes in 323 CLP cases and 211 nonmalformed controls. Our findings indicated that rare, protein-altering variants displayed markedly higher burdens in CLP cases at relevant loci. First, putative loss-of-function mutations (nonsense, frameshift) were significantly enriched among cases: 13 of 323 cases (~4%) harbored such alleles within these 49 genes, versus one such change in controls ($p = 0.01$). Second, in gene-level analyses, the burden of rare alleles showed greater case-association for several genes previously implicated in cleft risk. For example, *BHMT* displayed a 10-fold increase in protein-altering variants in CLP cases ($p = .03$), including multiple case occurrences of a rare frameshift mutation (K400 fs). Other loci with greater rare, coding allele burdens in cases were in signaling pathways relevant to craniofacial development (*WNT9B*, *BMP4*, *BMPR1B*) as well as the methionine cycle (*MTRR*). We conclude that rare coding variants may confer risk for isolated CLP.

KEYWORDS
cleft lip, folic acid, genetic association, rare variants

1 | INTRODUCTION

Although their causes are largely unknown, orofacial clefts are suspected of being etiologically heterogeneous with both genetic and nongenetic risk factors (Dixon, Marazita, Beaty, & Murray, 2011). Clinical observations have long indicated that orofacial clefts should be classified into at least two distinct phenotypic groups: cleft lip with or without cleft palate (CLP) and cleft palate alone (CPO) (Fogh-Anderson, 1967). Collectively, these are among the most common birth defects with a worldwide prevalence of ~1 in 700 live births (WHO, 2001).

Genetic associations for CLP have been observed for several genes, including *IRF6*, *FGFR2*, *FOXE1*, *MSX1*, *NECTIN1*, *BMP4*, *TBX22*, and *TGFa* (Dixon et al., 2011) from either linkage analysis or candidate gene studies. In addition, genome-wide association studies (GWAS) have identified approximately 15 additional risk loci (Beaty et al., 2010; Birnbaum et al., 2009; Grant et al., 2009; Leslie et al., 2016; Ludwig et al., 2012; Mangold et al., 2010) with some of these replicating in subsequent studies (e.g., – Jia et al., 2015). For some of these loci, such as *IRF6* and *NOG*, follow-up sequencing efforts have identified compelling functional variants as putative risk alleles (Leslie et al., 2015; Rahimov et al., 2008). Interestingly, there is apparently little overlap between the most significantly-associated GWAS regions and genes previously implicated in CLP risk.

This is probably indicative of the complex genetic etiologies underlying the CLP phenotype and the difficulties detecting associated loci at genome-wide significance. Thus, there are likely to be numerous additional genetic risk factors yet to be identified. Indeed, the combined efforts of all these genetic approaches have been extremely fruitful, yet still explain only a fraction of the population burden of these human birth defects.

Several nongenetic factors also appear to contribute to cleft phenotypes (Mossey, Little, Munger, Dixon, & Shaw, 2009). Of particular relevance to the current study is that maternal use of multivitamin supplements containing folic acid in early pregnancy has been associated with decreased risk of CLP (Shaw, Lammer, Wasserman, O'Malley, & Tolarova, 1995; Wilcox et al., 2007). Although some reports do not observe a reduction in risk associated with maternal folate supplementation (see Little et al., 2008), a recent meta-analysis indicated a significant reduction in risk of CLP with maternal folic acid use (Jahanbin, Shadkam, Miri, Shirazi, & Abtahi, 2018). Because of this folic acid link, folate/one-carbon pathway genes have been a logical place to look for risk variants (Blanton et al., 2011; Boyles et al., 2009; Marini et al., 2016). However, many of the known common pathway SNPs have been evaluated, yet the results are still somewhat unconvincing. The exception to this seems to be the fairly consistent association of the *BHMT/BHMT2/DMGDH* locus on Chromosome 5 with CLP observed in a number of studies (Marini et al., 2016). Interestingly, while the locus often shows association, the local SNPs driving the association can vary between studies. Thus, the nature of the risk at this locus is still unclear.

In this study we hypothesized that some genetic risk for CLP lies in rare variants and this has contributed to the lack of consistent findings and attendant difficulty in unraveling risk alleles. Thus, we sequenced the exons in 30 genes involved in folate metabolism in a 534-member case–control population to determine if rare variant allele burden (in individual genes, subsets of genes, or the entire pathway) correlated with phenotype. We also sequenced the coding region of an additional 19 candidate genes (not related to folate metabolism) that are involved in lip/palate development or previously associated with CLP. Our results showed that, while there were no folate pathway-wide trends in rare variant allele burden, two genes involved in methionine synthesis (*BHMT* and *MTRR*) showed convincing case-associated trends for protein-altering variants. Likewise, genes involved in BMP/WNT signaling (*BMP4*, *BMPR1B*, *WNT9B*), which have been previously implicated in orofacial cleft development (Juriloff, Harris, Mager, & Gagnier, 2014; Suzuki et al., 2009; Zhang et al., 2002), displayed higher allele burdens in cases. CLP cases also presented a greater number of loss-of-function alleles (frameshifts, truncations) in this candidate gene set.

2 | METHODS

2.1 | Study population

This case–control study included data on deliveries that had estimated due dates from 1995 to 2003. The study included live-born infants with isolated CLP (N = 681), isolated CPO (N = 157), or without any structural malformation (controls; N = 706). Because the goal of the study was to determine whether rare coding variants in candidate genes underlie risk for CLP, variant discovery by gene sequencing was performed on an ethnically-mixed subset of this population comprising 323 CLP cases and 211 controls. A second group (358 CLP cases, 157 CPO cases, 485 controls) was used to further validate the CLP case-association of a frameshift mutation in *BHMT*. See Table 1 for race-ethnic breakdown of the study population.

Case information was abstracted from hospital reports and medical records following established procedures by the California Birth Defects Monitoring Program (Croen, Shaw, Jensvold, & Harris, 1991). Each medical record was further reviewed by a medical geneticist (Edward J. Lammer, MD). Infants with trisomies were ineligible. Nonmalformed control infants were selected randomly to represent the birth population from which the cases were derived in selected counties and birth periods. This study, including the collection and use of archived newborn bloodspots, was approved by the California State Committee for the Protection of Human Subjects as well as Institutional Review Boards at Stanford University and the University of California, Berkeley.

2.2 | Genomic DNA isolation

Genomic DNA (gDNA) was available from newborn screening dried blood spots obtained from linkage efforts made by the California Birth Defects Monitoring Program. gDNA was extracted from dried blood spots and whole-genome amplified using QIAmp DNA Mini kits and Replica-G Midi amplification kits (Qiagen), respectively, as described previously (Marini et al., 2011).

2.3 | Illumina paired-end sequencing of enriched target exons

3–5 μg of whole-genome amplified DNA was used as starting material for Illumina library construction and target exon capture with Sure Select (Agilent) probes according to the Agilent SureSelect[XT] protocol (https://www.agilent.com/cs/library/usermanuals/public/G7530-90000.pdf).

Briefly, DNA was fragmented with a Covaris sonicator, followed by end-repair and A-tailing, then ligation of sequencing adaptors and PCR amplification. Amplified libraries were then used to perform target exon capture by hybridization against a custom-made capture library of 120 bp biotinylated oligonucleotide probes covering the 555 coding exons in 49 candidate genes of interest (Table S1). Probes (Agilent SureSelect) were designed via the SureDesign web tool using the genomic coordinates of target exons (genome build GRCh37/hg19). To ensure capture and coverage of the entire coding region and splice sites, probes were designed with an additional 50 bp "pad" that covered intronic regions adjacent to coding exons or, in the case of the first or last coding exon, covered the 5′- or 3′- untranslated regions. Following hybridization and streptavidin capture of enriched exons, captured libraries were barcoded with 6 bp tags with an additional round of PCR, quantified, and 12–16 sample libraries were pooled for 100 bp paired-end sequencing on the Illumina HiSeq (v3 chemistry).

TABLE 1 Populations used in study

A. Discovery population		
Race/ethnicity	CLP cases (*collective percentage*)	Controls (*collective percentage*)
Asian	35 (*10.8*)	25 (*11.8*)
Black	13 (*4.0*)	8 (*3.8*)
White, Hispanic	190 (*58.8*)	126 (*59.7*)
White, non-Hispanic	79 (*24.5*)	48 (*22.7*)
Other	6 (*1.9*)	4 (*1.9*)
TOTAL	323	211

B. Validation population			
Race/ethnicity	CLP cases (*collective percentage*)	CPO cases (*collective percentage*)	Controls (*collective percentage*)
Asian	41 (*11.5*)	18 (*11.5*)	41 (*8.5*)
Black	9 (*2.5*)	9 (*5.7*)	29 (*6.0*)
White, Hispanic	188 (*52.5*)	58 (*36.9*)	231 (*47.6*)
White, non-Hispanic	54 (*15.1*)	32 (*20.4*)	83 (*17.1*)
Other	6 (*1.7*)	1 (*0.6*)	5 (*1.0*)
No race reported	59 (*16.5*)	39 (*24.8*)	96 (*19.8*)
Native American	1 (*0.3*)		
TOTAL	358	157	485

2.4 | Read mapping and variant filtering

Raw sequence files (FASTQ) were imported into CLC Genomics Workbench (v.6; Qiagen) and mapped onto the human genome (GRCh37/hg19). Only full-length, paired, uniquely-mapping reads with <= 3 mismatches were used for variant calling. Variants with depth of coverage <30, or with genotype quality scores <25, were excluded. In addition, alleles not observed on both strands (with a minimum frequency ≥ 0.1) or with allelic imbalances (heterozygote genotype calls with one allele observed with frequency < 0.25 or homozygote alleles observed with frequency < 0.85) were also excluded. After compilation of the dataset, positions not covered in at least 95% of the cases or controls were then removed.

2.5 | Variant calling accuracy

The accuracy of the variant calls in the filtered dataset was evaluated in two ways: (a) the concordance of genotype calls by Illumina sequencing and by TaqMan allelic discrimination assays (described in Marini et al., 2016) was determined for 8,537 variant calls (mostly common alleles) in 476 samples and (b) confirmation by Sanger sequencing for the subset of rare variants in genes with striking case-related trends (~15% of all protein-altering variants identified in the study). Verification sequencing was performed on original bloodspot gDNA preparations, prior to whole-genome amplification, using amplicon-specific M13-capped PCR primers designed by VariantSEQr methodology (ThermoFisher) and deposited into the NCBI probe database (https://www.ncbi.nlm.nih.gov/probe/docs/projvariantseqr/). Each sample for verification had a corresponding negative control for comparison.

2.6 | Analyses

To estimate the contribution of rare (MAF < 1%) coding variants (for which there may only be a single occurrence) to CLP risk, a simple allele burden test was applied on a gene-by-gene basis to highlight genes with prominent case–control skews. Alleles were summed by gene (or groups of genes) and case–control associations were tested by Fisher's exact test based on genotype distributions (i.e., carrier frequency). Although, this cannot be used as a strict test for association, statistical differences in case or control representation are readily observed with this metric. For common polymorphisms, variants were treated as categorical variables, that is, homozygous wild-type as referent versus heterozygous or homozygous variant. Odds ratios and 95% confidence intervals were used to estimate risks.

3 | RESULTS

3.1 | Dataset definition, variant calling, and quality control

We took a candidate gene approach, with a focus on folate pathway genes, to investigate the potential contribution of rare coding variants to CLP risk by determining whether putative loss-of-function alleles or rare allele burdens in certain genes (or pathways) correlated with phenotype. Thus we sequenced the coding exons and exon/intron boundaries in 30 genes related to folate metabolism and 19 additional genes involved in craniofacial development in a case–control study (Table 1). The full list of genes queried, including exon annotations and coverage statistics are in Table S1. 532 coding exons out of the

555 targeted for enrichment in these 49 genes (96%) passed quality control for adequate coverage and sequence quality. Mean coverage for each of these 532 exons was similar between the case group and the control group, indicating each group was sequenced with equal efficiency (Figure S1).

Variants were called in exons and in intron regions within 10 bp from exon/intron junctions as described in Methods. To evaluate the overall quality and accuracy of the variant calls, we performed two types of independent assessments on different subsets of variants. In the first, we established the concordance between the genotypes called from Illumina sequencing with those obtained from TaqMan allelic discrimination assays for many multiply-occurring loci (i.e., not singletons) using TaqMan assays previously described (Marini et al., 2016). 91 variant positions in 24 genes were compared in 476 samples. Out of 8,537 variant calls, we observed 99.7% concordance between the two methods (Table S2).

In a second quality-control check, we evaluated all putative loss-of-function alleles (frameshifts and truncations), and all rare and singleton missense variant calls from genes whose alleles showed higher case representation, by exon-specific Sanger resequencing (see Section 2). Out of 95 tested variants, all were confirmed except one frameshift in a low complexity region, which was subsequently removed (Table S3). Based on these analyses, we are confident in the sequencing accuracy of the data, necessary for rare variant genotyping.

The final analytical dataset contained data for 773 variant positions: 285 synonymous, 380 protein-altering, and 108 noncoding variants. One hundred fifty-five (54%) of the synonymous alleles were singletons, whereas 243 (64%) of the protein-altering alleles were singletons. The complete list of annotated variants, allele frequencies, and case–control distributions are in Table S4.

3.2 | Common allele associations

Although our focus was on the distribution of rare coding alleles, sequence data also provided information on associations of more common variants. We observed a small number of nominally significant associations that are in agreement with previous studies. For example, association of nonsynonymous SNPs in *DMGDH* (S279P, homozygous OR = 0.5 [0.3–0.9], p = .022) and *ALDH1L1* (I812V, heterozygous OR = 2.7 [1.3–5.5], p = .008) have been reported previously (Boyles et al., 2009; Marini et al., 2016).

By far, the strongest statistically significant signal of any variant with high enough frequency to calculate was the nonsynonymous SNP V274I in *IRF6* (heterozygous OR = 0.5 (0.3–0.7), p = 4E−04; homozygous OR = 0.5 (0.2–0.9), p = .046), a well-studied polymorphism in orofacial cleft studies with numerous reported associations (Beaty, Marazita, & Leslie, 2016). A second *IRF6* synonymous SNP (S153S) was also significantly associated in homozygotes (OR = 0.6 [0.4–0.9], p = .02). Finally, there were two linked synonymous SNPs in the *VANGL2* gene (G445G and P467P; heterozygous OR = 1.8 [1.2–2.7], p = .004), which has not previously been associated with CLP.

Odds ratios were calculated only from genotype frequencies in the entire, ethnically diverse population, and not in ethnically stratified sub-populations, because this study was not sufficiently powered for such analyses. Our focus was on rare variants which are often impervious to meaningful ethnic stratification analyses owing to the infrequency of occurrence. However, we have ensured that the race-ethnic breakdown was approximately the same in case and control groups (Table 1).

3.3 | Association of putative loss-of-function alleles

The candidate gene set interrogated in this study is a mix of both folate pathway genes and genes implicated in cleft occurrence because of their roles in craniofacial development. Thus, this set was enriched for potential cleft risk genes over what might be considered in a broader sampling of the exome. Therefore, it is noteworthy that we observed a significant enrichment of putative loss-of-function alleles (frameshifts, truncations) in CLP cases: 13 cases harbored one such allele from this gene set versus only 1 control (Fisher's exact p = .01; Table 2). All alleles in Table 2 were seen only once in this population except for the deletion/frameshift at Lysine 400 in *BHMT* (K400fs), which was seen in four cases of differing ethnicity. It is further noteworthy that several of the genes represented in Table 2 (*BHMT*, *DMGDH*, *MTRR*, *IRF6*) have displayed significant CLP associations in multiple studies (see Section 4). It should be emphasized that functional impact of these alleles is only inferred, though they represent good candidates for mutations of consequence for cleft risk.

3.4 | Folate pathway variation

Despite numerous epidemiological studies suggesting that maternal folic acid supplementation may modify risk for CLP, common SNP association studies with folate/one-carbon pathway variants have not been definitive in linking pathway loci to phenotype. Suggestive evidence exists for several loci from this group (e.g., Blanton et al., 2011; Boyles et al., 2009; Marini et al., 2016), though reproducibility across populations and small effect sizes remain problematic. The role of rare variants in folate pathway genes has not been adequately queried and may account for genetic risk yet to be identified. Indeed, the folate/one-carbon pathway may be particularly susceptible to the aggregate mutation burden throughout the pathway, since there are many co-dependent and interacting components. Thus, we hypothesized that the aggregate burden of rare variants in the folate pathway (or related genes in pathway sub-compartments) may confer risk for CLP and help explain the folate link. We have found such pathway analyses useful in understanding the role of folate pathway variation, common and rare, with respect to neural tube defect risk (Marini et al., 2011).

By allele summing of rare variants (defined here as MAF < 1%), the group of folate/one-carbon pathway genes (N = 30) did not show an increased aggregate burden of either synonymous mutations (case aggregate allele frequency = 0.27, control aggregate allele frequency = 0.28; p = .57, Mann–Whitney test) or intron mutations within 10 bp of the exon boundary (case aggregate allele frequency = 0.12, control aggregate

TABLE 2 Putative loss-of-function alleles

Gene	Variant	Type[a]	Case appearance[b]	Control appearance[b]
BHMT	K400fs	Deletion/frameshift	Hispanic	
			Hispanic	
			Non-Hispanic, White	
			Non-Hispanic, White	
DMGDH	W495term[b]	SNV/stop-gain	Hispanic	
GGH	L190fs	Insertion/frameshift	Non-Hispanic, White	
IRF6	Q112term[b]	SNV/stop-gain	Non-Hispanic, White	
MTRR	R386term[b]	SNV/stop-gain	Non-Hispanic, White	
SARDH	R457term[b]	SNV/stop-gain	Hispanic	
SARDH	R514term[b]	SNV/stop-gain	Hispanic	
SP8	S261term[b]	SNV/stop-gain	Hispanic	
SHMT1	R99term[b]	SNV/stop-gain	Hispanic	
SHMT1	R207term[b]	SNV/stop-gain	Asian	
SHMT1	Y457term[b]	SNV/stop-gain		Asian
TOTAL			**13**	**1**

[a]Abbreviation: SNV, single nucleotide variation.
[b]Ethnicity of newborn.

allele frequency = 0.10; p = .34, Mann–Whitney test) that might distinguish cases from controls. There was as slight increase in aggregated protein-altering allele frequency for cases (case aggregate allele frequency = 0.38, control aggregate allele frequency = 0.31); though the distributions were not significantly different (p = .08).

Upon sub-dividing pathway genes into compartments with metabolically related genes (e.g., – purine synthesis, see (Marini et al., 2011)), it became obvious that the strongest case-associated signals for rare, protein-altering variants were restricted to two genes involved in methionine synthesis: BHMT (betaine-homocysteine methyl transferase) and MTRR (methionine synthase reductase), both of which are necessary for the remethylation of homocysteine to methionine (Figure 1, Table 3). In addition to the four case occurrences of the K400fs allele in BHMT, nonsynonymous variants were found in an additional 7 cases and only a single control (Fisher's exact p = .03). Similarly, protein-altering variants for MTRR were found in 13 cases versus 2 controls (p = .06).

Using a minor allele frequency cutoff of 5%, rather than 1%, yielded similar folate pathway-wide trends: there were no significant differences in the aggregate burden of noncoding variants (p = .38) or synonymous variants (p = .81), but the aggregate burden of protein-altering variants was significantly higher in cases (case aggregate allele frequency = 0.63, control aggregate allele frequency = 0.50; p = .02). No additional folate pathway genes, besides BHMT and MTRR, displayed significant case–control differences.

3.5 | Individual gene analysis

For each of the 49 genes in this study, rare synonymous or protein-altering variants were collapsed by summing and genotype frequencies for this allele burden metric were subject to Fisher's exact testing (Figure 1). Consistent with the absence of a si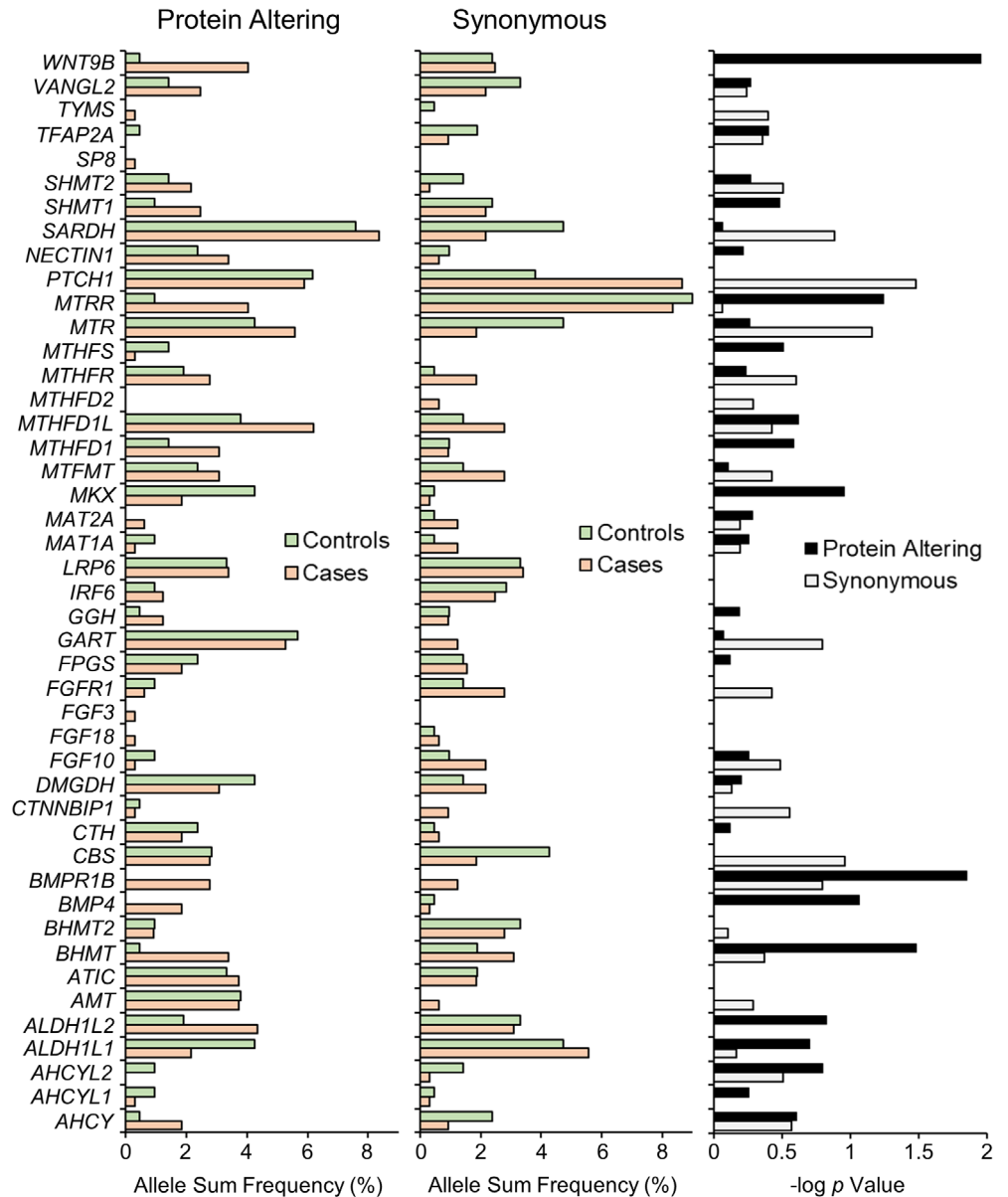gnal seen from synonymous variation in the folate pathway, PTCH1 was the only gene with a statistically significant altered case–control distribution of rare synonymous variants: case carrier frequency (0.20) was approximately twofold higher than that in controls (0.12; p = .03).

In contrast, there were several genes with higher carrier frequency of protein-altering variants in cases with more striking case–control differentials (Figure 1 and Table 3). In addition to the BHMT and MTRR allele distributions discussed above, there were 13 case appearances (all heterozygotes) of missense alleles in the WNT9B gene, versus only one missense occurrence in controls (p = .01). One missense change, A126T, was found exclusively in three cases. Likewise, in the BMPR1B gene, which encodes a receptor for bone morphogenetic protein (BMP) secreted ligands, there were nine occurrences of missense alleles in cases without a single observation in controls (p = .01). Also noteworthy is the BMP4 gene, which encodes a specific, secreted bone morphogenetic ligand involved in embryonic head development. We observed six different missense alleles of BMP4 in case heterozygotes versus zero protein-altering variants in controls (p = .09). These last three genes were included based on their roles in craniofacial development (and, in some cases, isolated cleft risk (Juriloff et al., 2014; Suzuki et al., 2009; Yang et al., 2017)). Thus, it was interesting to see striking accumulation of rare variation in CLP cases.

3.6 | Replication of the case association of BHMT K400fs in a second population

As described above, we observed multiple case occurrences of a frameshift (delA) at lysine 400 in BHMT. This lesion occurs only 7 amino acids from the C-terminus and is inferred (by nucleotide sequence) to replace these last 7 residues (KQKFKSQ) with 13 different amino acids (NKNSNHSSLDRSYF). Thus, the functional impact of this mutation is

FIGURE 1 The occurrences
of all rare coding region variants
(MAF <1.0%) were summed on a
gene-by-gene basis to yield the
total allele burden for
synonymous (left panel) and
protein-altering (middle panel)
variants. Protein-altering variants
encompass frameshifts,
truncations, and missense alleles.
Allele sum frequency is the carrier
frequency (% heterozygote
carriers) for each gene. p values
were calculated (right panel)
based on 2 × 2 Fisher's exact test
of the summed carrier frequency
for synonymous and protein-
altering changes and expressed as
−log p [Color figure can be
viewed at wileyonlinelibrary.com]

4 | DISCUSSION

In this study we took a candidate gene approach to investigate the
potential contribution of rare coding variants to CLP risk, with the
reasoning that at least some of the genetic risk not yet identified in
common allele studies may reside in rare variation. The contribution
of rare, coding mutations in CLP risk has been previously suggested in

difficult to confidently discern. We took the same set of Exon 8-specific
resequencing primers used for variant verification (see Section 2) to
genotype-by-sequence this variant in a second, case–control population
(validation population in Table 1). We observed this variant as a hetero-
zygote three additional times in the validation population: two additional
CLP cases (one Hispanic, one non-Hispanic, White) and one control
(non-Hispanic, White). We did not observe this variant in the cleft palate
only (CPO) subset of this population (N = 157).

multiple sequencing studies (Al Chawa et al., 2014; Leslie & Murray,
2013; Suzuki et al., 2009), though this has only recently begun to be
systematically addressed (Aylward et al., 2016; Leslie et al., 2017). We
focused on 49 candidate genes with an emphasis on nearly complete
coverage of folate metabolism (N = 30 genes), since rare folate pathway
variants have not been explored for CLP risk, particularly when con-
sidered in aggregate across the pathway or within related genesets.
We queried 19 additional genes unrelated to folate metabolism, that
have been implicated in disease risk by vertebrate models of clefting
(e.g., WNT9B) or craniofacial development, or by human genetic associa-
tion studies (e.g., BMP4, IRF6). From a catalog of the full spectrum of alle-
lic variation within the coding regions of this target-enriched gene set, we
found two lines of evidence for the contribution of rare variants to CLP
risk: (a) cases displayed an overabundance of putative loss-of-function
alleles at these loci and (b) several mechanistically relevant genes showed
similar case over-representation for protein-altering variants, in general.

TABLE 3 Genes with strongest case representations of rare, protein-altering variants

Gene	Case appearances[a] Variant	#	Control appearances[a] Variant	#
BHMT	D105N	1	V122M	1
	E136K	1		
	P197S	3		
	G199S	2		
	K400fs	4		
BMP4	D56E	1		
	R139H	1		
	R162Q	1		
	R223H	1		
	R226Q	1		
	R269Q	1		
BMPR1B	P23S	1		
	V67D	1		
	V140I	2		
	R149W	1		
	Q153R	1		
	M301V	1		
	R371Q	1		
	R413S	1		
MTRR	H36Y	1		
	N218K	1		
	V265G	1		
	T286M	3		
	I290T	1		
	K352N	1		
	R386[a]	1		
	R418Q	1		
	T490A	2	T490A	1
	V674A	1	V607I	1
WNT9B	A42V	1		
	Q47R	2		
	R94Q	2	R94Q	1
	N99S	1		
	A126T	3		
	K184N	1		
	R222H	1		
	S244L	2		

[a]All variants were seen as heterozygotes; only *MTRR* T490A and *WNT9B* R94Q were seen in both cases and controls.

4.1 | Folate pathway variation: Methionine cycle genes

We sequenced the majority of folate/one-carbon pathway genes, which enabled us to interpret rare variation in the context of pathway function. This approach was useful in identifying folate-related risk signatures for neural tube defects (Marini et al., 2011). While we did not see any pathway-wide trends in allele burdens or distributions that significantly differentiated CLP cases from controls, we did observe two genes directly involved in methionine biosynthesis with a considerable over-abundance of protein altering variants in cases: *BHMT* and *MTRR* (Table 3). Betaine-homocysteine methyltransferase (BHMT) and Methionine synthase reductase (MTRR) are both involved in the methylation of homocysteine to yield methionine. They operate in different pathways; however, and while BHMT utilizes methyl groups generated from choline catabolism, the reaction involving MTRR utilizes 5-methyl-tetrahydrofolate.

The CLP-association of rare variants at *BHMT* is intriguing for two reasons: First, this locus, which is part of the metabolically-related *BHMT/BHMT2/DMGDH* gene cluster on chromosome 5, is one of the most consistently associated loci in one-carbon pathway gene-CLP association studies (reviewed in Marini et al., 2016). However, while the locus is often associated, the regional SNPs that drive the association can be different, as can the directionality of risk (i.e., whether odds ratios are < or >1). We postulated that these phenomena may reflect the contribution of underlying rare variants at this locus (which may occur on the haplotype of the major or minor allele of a nearby common variant) as has been previously suggested for clefts (Leslie & Murray, 2013) and as has been seen for regions highlighted by GWAS studies (Leslie et al., 2015). The distribution of rare variants at *BHMT* is also intriguing because of the appearance of a frameshift mutation (K400fs) that was observed in multiple CLP cases (discussed below).

The increased mutation burden in these two genes brings attention to homocysteine clearance as relevant for modifying CLP risk, which has been previously suggested (Kumari, Ali, Sukla, Singh, & Raman, 2013; Wong et al., 1999). Furthermore, interference of methionine synthesis suggests perturbations in cellular methylation reactions (via S-adenosylmethionine levels), which may manifest in DNA methylation changes associated with CLP, as has been observed (Alvizi et al., 2017; Sharp et al., 2016).

4.2 | Putative loss-of-function alleles

Within this target gene set, CLP cases displayed a significantly greater frequency of indel frameshift and truncation alleles (p = .01; Table 2). One of the most intriguing alleles is the K400 fs deletion allele of *BHMT*, which was observed in four cases (both Hispanic and non-Hispanic, White) but not in controls in the discovery population. A second, validation population revealed two additional CLP case occurrences and one control occurrence. Although the functional impact of this allele is yet to be determined, its observance in multiple cases is noteworthy particularly because global population frequencies of this variant are estimated at 2.2E−04 in the Genome Aggregation Database (www.gnomad.broadinstitute.org), substantially lower than our overall case frequency (MAF = 0.004 in all cases queried). The low number of occurrences for this allele does not allow an effective statistical test in populations of this size, nevertheless, multiple case observations in two separate populations is higher than expected

based on the reported global MAF. It should be noted that this variant is described as low-confidence in the gnomAD database. However, the signal was strong in our data and all instances clearly replicated in verification Sanger sequencing (see Figure S2, for example). Thus, we are confident of these variant calls.

In addition to *BHMT*, we also observed case-associated truncations in several other genes relevant to methyl donor production (*MTRR* R386*, *DMGDH* W495*, and *SARDH* R457*, R514*). Based on protein lengths (MTRR 698aa, DMGDH 866aa, SARDH 918aa), these changes are likely to be consequential. BHMT, DMGDH (dimethylglycine dehydrogenase), and SARDH (sarcosine dehydrogenase) are sequential enzymatic steps in the catabolism of choline to glycine, which generates methyl donors for cellular methylation reactions as noted above. Interestingly, increased intake of choline in the maternal diet has been associated with decreased CLP risk (Shaw, Carmichael, Laurent, & Rasmussen, 2006).

Outside of folate/one-carbon pathway genes, two case truncations were observed in genes previously implicated in orofacial cleft risk: *IRF6* Q112* and *SP8* S261*. Significant associations of *IRF6* with CLP have been reported in multiple populations (Beaty et al., 2016) and may account for up to 12% of the genetic component of CLP in some populations (Zucchero et al., 2004). SP8 is a transcription factor in which mutations display severe craniofacial malformations including cleft palate (Kasberg, Brunskill, & Steven Potter, 2013).

4.3 | Genes unrelated to folate metabolism

Three additional genes involved in craniofacial development also displayed striking case-overrepresentation of missense variants: *BMP4*, *BMPR1B*, and *WNT9B* (Table 3). WNT9B and BMP4 are secreted signaling proteins regulating critical interacting signaling pathways in craniofacial development (Alexander, Piloto, Le Pabic, & Schilling, 2014). *BMPR1B* encodes a receptor for BMP4 and other bone morphogens. Significantly, all three genes have been implicated in orofacial cleft risk.

BMP4 is one of the most promising candidate genes for CLP, with evidence shown in animal experiments (e.g., Liu et al., 2005) and association studies (e.g., Chen et al., 2014). Furthermore, a gene sequencing study (Suzuki et al., 2009) demonstrated that rare *BMP4* coding mutations were enriched in CLP cases at similar frequencies to those shown here. Only the R162Q case-specific nonsynonymous mutation was identified in both studies.

Although rare coding variants in *WNT9B* or *BMPR1B* have yet to be reported as contributing to CLP risk, these genes have been implicated by other means. *WNT9B* is at a locus identified in a genome-wide association study (Yu et al., 2017) and a murine model demonstrated that transcriptional repression of Wnt9b via epigenetic mechanisms leads to the development of cleft lip in mouse embryos (Juriloff et al., 2014). *BMPR1B* is at a locus associated with CLP by linkage analysis (Schultz et al., 2004) and has been identified as a genetic cause of Pierre Robin syndrome, in which cleft palate is a prominent component (Yang et al., 2017).

4.4 | Rare versus private mutations

Of all the variants described in Tables 2 and 3, most have been seen before in the population, as measured by their annotation in the Genome Aggregation Database (www.gnomad.broadinstitute.org). Novel variants in this study are restricted to two missense variants in *BMPR1B* (P23S and V67D), and three putative loss-of-function truncations (*DMGDH* W495*, *IRF6* Q112*, *SP8* S261*). The frameshift allele in *GGH* (insT at L190) was previously observed once in over 250,000 alleles. All others have multiple occurrences, though at significantly lower frequencies in the general population than seen here in CLP cases. However, most of our observations are singletons (with a notable exception being the K400fs allele of *BHMT*), precluding true case frequency estimates.

If rare alleles such as these contribute to CLP risk, it is likely that they will be uncovered only by sequencing/discovery approaches rather than array-based genotyping, even for large array sizes. Furthermore, the degree to which risk loci identified by traditional association studies will reveal loci with underlying rare variant burdens (as we have suggested for the *BHMT/BHMT2/DMGDH* gene cluster) is still unknown. For example, Leslie et al. (2015) failed to find convincing evidence of CLP-related rare allele burdens at GWAS loci in 1500 trios. It may be that the underlying genetic heterogeneity for CLP risk is sufficiently complex that many such burdened loci do not display genome-wide significance in traditional association studies. Indeed, nearly all variants discussed here were heterozygous and none co-occurred with any others. Thus, if they are involved in CLP risk, they must be part of a more complex etiology.

4.5 | Strengths/limitations

Strengths of this study lie in the case–control design, which allows for a broader sampling of rare alleles than family-based studies, and the assurance of accuracy in the sequencing data, which is essential for rare variant discovery. Although CLP risk loci identified in this study are consistent with previous studies and with biological mechanisms, the study is limited by both sample size and the ability to draw reliable associations from rare variants. Thus, statistical measures used here are not a true measure of association but of highlighting genes with differential case burdens. Therefore, replication of these findings in additional, larger populations is needed to afford significance of the potential associations observed here.

Another limitation is that the functional impact of variants is unknown. We have assumed that the striking mutation skews are indicative of a contribution to CLP etiology. In support of this assumption, we used two computational algorithms to predict the functional impact of all missense mutations identified from the five genes in Table 3: CADD (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019) and PolyPhen2 (Adzhubei et al., 2010). Both tools estimated that mutations occurring only in cases were significantly enriched for deleterious changes when compared against variants in the same genes seen in controls (Figure S3). Nevertheless, given the potential fallibility of predictive algorithms, we await empirical determination of functional impact to enable more robust mechanistic conclusions with

respect to phenotype. Despite these limitations, these results are consistent with several reports implicating rare coding variants in CLP risk. The clustering of risk genes reinforces focus on one-carbon metabolism and the BMP/WNT signaling pathways.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Nicholas J. Marini 🆔 https://orcid.org/0000-0001-8782-2573
Gary M. Shaw 🆔 https://orcid.org/0000-0001-7438-4914

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7, 248–249.

Al Chawa, T., Ludwig, K. U., Fier, H., Pötzsch, B., Reich, R. H., Schmidt, G., ... Mangold, E. (2014). Nonsyndromic cleft lip with or without cleft palate: Increased burden of rare variants within Gremlin-1, a component of the bone morphogenetic protein 4 pathway. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 100, 493–498.

Alexander, C., Piloto, S., Le Pabic, P., & Schilling, T. F. (2014). Wnt signaling interacts with bmp and edn1 to regulate dorsal-ventral patterning and growth of the craniofacial skeleton. *PLoS Genetics*, 10, e1004479.

Alvizi, L., Ke, X., Brito, L. A., Seselgyte, R., Moore, G. E., Stanier, P., & Passos-Bueno, M. R. (2017). Differential methylation is associated with non-syndromic cleft lip and palate and contributes to penetrance effects. *Scientific Reports*, 7, 2441.

Aylward, A., Cai, Y., Lee, A., Blue, E., Rabinowitz, D., Haddad, J., Jr., & University of Washington Center for Mendelian Genomics. (2016). Using whole exome sequencing to identify candidate genes with rare variants in Non-syndromic cleft lip and palate. *Genetic Epidemiology*, 40, 432–441.

Beaty, T. H., Marazita, M. L., & Leslie, E. J. (2016). Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. *F1000Res*, 5, 2800.

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., ... Scott, A. F. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genetics*, 42, 525–529.

Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Steffens, M., Rubini, M., ... Mangold, E. (2009). Key susceptibility locus for non-syndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature Genetics*, 41, 473–477.

Blanton, S. H., Henry, R. R., Yuan, Q., Mulliken, J. B., Stal, S., Finnell, R. H., & Hecht, J. T. (2011). Folate pathway and nonsyndromic cleft lip and palate. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 91, 50–60.

Boyles, A. L., Wilcox, A. J., Taylor, J. A., Shi, M., Weinberg, C. R., Meyer, K., ... Lie, R. T. (2009). Oral facial clefts and gene polymorphisms in metabolism of folate/one-carbon and vitamin a: A pathway-wide association study. *Genetic Epidemiology*, 33, 247–255.

Chen, Q., Wang, H., Schwender, H., Zhang, T., Hetmanski, J. B., Chou, Y. H., ... Beaty, T. H. (2014). Joint testing of genotypic and gene-environment interaction identified novel association for BMP4 with non-syndromic CL/P in an Asian population using data from an international cleft consortium. *PLoS One*, 9, e109038.

Croen, L. A., Shaw, G. M., Jensvold, N. G., & Harris, J. A. (1991). Birth defects monitoring in California: A resource for epidemiological research. *Paediatric and Perinatal Epidemiology*, 5, 423–427.

Dixon, M. J., Marazita, M. L., Beaty, T. H., & Murray, J. C. (2011). Cleft lip and palate: Understanding genetic and environmental influences. *Nature Reviews. Genetics*, 12, 167–177.

Fogh-Anderson, P. (1967). Genetic and non-genetic factors in the etiology of facial clefts. *Scandinavian Journal of Plastic and Reconstructive Surgery*, 1, 22–29.

Grant, S. F., Wang, K., Zhang, H., Glaberson, W., Annaiah, K., Kim, C. E., ... Hakonarson, H. (2009). A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *The Journal of Pediatrics*, 155, 909–913.

Jahanbin, A., Shadkam, E., Miri, H. H., Shirazi, A. S., & Abtahi, M. (2018). Maternal folic acid supplementation and the risk of Oral clefts in offspring. *The Journal of Craniofacial Surgery*, 29, e534–e541.

Jia, Z., Leslie, E. J., Cooper, M. E., Butali, A., Standley, J., Rigdon, J., ... Murray, J. C. (2015). Replication of 13q31.1 association in non-syndromic cleft lip with cleft palate in Europeans. *American Journal of Medical Genetics*, 167, 1054–1060.

Juriloff, D. M., Harris, M. J., Mager, D. L., & Gagnier, L. (2014). Epigenetic mechanism causes Wnt9b deficiency and nonsyndromic cleft lip and palate in the a/WySn mouse strain. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 100, 772–788.

Kasberg, A. D., Brunskill, E. W., & Steven Potter, S. (2013). SP8 regulates signaling centers during craniofacial development. *Developmental Biology*, 381, 312–323.

Kumari, P., Ali, A., Sukla, K. K., Singh, S. K., & Raman, R. (2013). Lower incidence of nonsyndromic cleft lip with or without cleft palate in females: Is homocysteine a factor? *Journal of Biosciences*, 38, 21–26.

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Buxó, C. J., Castilla, E. E., Christensen, K., ... Marazita, M. L. (2017). Association studies of low-frequency coding variants in nonsyndromic cleft lip with or without cleft palate. *American Journal of Medical Genetics. Part A*, 173, 1531–1538.

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Feingold, E., Wehby, G., Laurie, C. A., ... Marazita, M. L. (2016). A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24.2, 17q23 and 19q13. *Human Molecular Genetics*, 25, 2862–2872.

Leslie, E. J., & Murray, J. C. (2013). Evaluating rare coding variants as contributing causes to non-syndromic cleft lip and palate. *Clinical Genetics*, 84, 496–500.

Leslie, E. J., Taub, M. A., Liu, H., Steinberg, K. M., Koboldt, D. C., Zhang, Q., ... Murray, J. C. (2015). Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *American Journal of Human Genetics*, 96, 397–411.

Little, J., Gilmour, M., Mossey, P. A., Fitzpatrick, D., Cardy, A., Clayton-Smith, J., ... ITS MAGIC collaboration. (2008). Folate and clefts of the lip and palate-a U.K.-based case-control study: Part I: Dietary and supplemental folate. *The Cleft Palate-Craniofacial Journal*, 45, 420–427.

Liu, W., Sun, X., Braut, A., Mishina, Y., Behringer, R. R., Mina, M., & Martin, J. F. (2005). Distinct functions for bmp signaling in lip and palate fusion in mice. *Development*, 132, 1453–1461.

Ludwig, K. U., Mangold, E., Herms, S., Nowak, S., Reutter, H., Paul, A., … Nöthen, M. M. (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nature Genetics*, *44*, 968–971.

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., … Nöthen, M. M. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, *42*, 24–26.

Marini, N. J., Hoffmann, T. J., Lammer, E. J., Hardin, J., Lazaruk, K., Stein, J. B., … Rine, J. (2011). A genetic signature of spina bifida risk from pathway-informed comprehensive gene-variant analysis. *PLoS One*, *6*, e28408.

Marini, N. J., Yang, W., Asrani, K., Witte, J. S., Rine, J., Lammer, E. J., & Shaw, G. M. (2016). Sequence variation in folate pathway genes and risks of human cleft lip with or without cleft palate. *American Journal of Medical Genetics*, *170*, 2777–2787.

Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J., & Shaw, W. C. (2009). Cleft lip and palate. *Lancet*, *374*, 1773–1785.

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., … Murray, J. C. (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nature Genetics*, *40*, 1341–1347.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894.

Schultz, R. E., Cooper, M. E., Daack-Hirsch, S., Shi, M., Nepomucena, B., Graf, K. A., … Murray, J. C. (2004). Targeted scan of fifteen regions for nonsyndromic cleft lip and palate in Filipino families. *American Journal of Medical Genetics. Part A*, *125A*, 17–22.

Sharp, G. C., Ho, K., Davies, A., Stergiakouli, E., Humphries, K., McArdle, W., … Relton, C. L. (2016). Distinct DNA methylation profiles in subtypes of orofacial cleft. *Clinical Epigenetics*, *9*, 63.

Shaw, G. M., Carmichael, S. L., Laurent, C., & Rasmussen, S. A. (2006). Maternal nutrient intakes and risk of orofacial clefts. *Epidemiology*, *17*, 285–291.

Shaw, G. M., Lammer, E. J., Wasserman, C. R., O'Malley, C. D., & Tolarova, M. M. (1995). Risks of orofacial clefts in children born to women using multivitamins containing folic acid periconceptionally. *Lancet*, *346*, 393–396.

Suzuki, S., Marazita, M. L., Cooper, M. E., Miwa, N., Hing, A., Jugessur, A., … Murray, J. C. (2009). Mutations in BMP4 are associated with sub-epithelial, microform, and overt cleft lip. *American Journal of Human Genetics*, *84*, 406–411.

WHO. 2001. *Global registry and database on craniofacial anomalies.* Report of a WHO Registry Meeting on Craniofacial Anomalies (Vol. 12., pp. 1–3). Baura Brazil.

Wilcox, A. J., Lie, R. T., Solvoll, K., Taylor, J., McConnaughey, D. R., Abyholm, F., … Drevon, C. A. (2007). Folic acid supplements and risk of facial clefts: National population based case-control study. *BMJ*, *334*, 464.

Wong, W. Y., Eskes, T. K., Kuijpers-Jagtman, A. M., Spauwen, P. H., Steegers, E. A., Thomas, C. M., … Steegers-Theunissen, R. P. (1999). Nonsyndromic orofacial clefts: Association with maternal hyper-homocysteinemia. *Teratology*, *60*, 253–257.

Yang, Y., Yuan, J., Yao, X., Zhang, R., Yang, H., Zhao, R., … Zhu, Y. (2017). BMPR1B mutation cases Pierre Robin sequence. *Oncotarget*, *8*, 25864–25871.

Yu, Y., Zuo, X., He, M., Gao, J., Fu, Y., Qin, C., … Bian, Z. (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nature Communications*, *8*, 14364.

Zhang, Z., Song, Y., Zhao, X., Zhang, X., Fermin, C., & Chen, Y. (2002). Rescue of cleft palate in Msx1-deficient mice by transgenic Bmp4 reveals a network of BMP and Shh signaling in the regulation of mammalian palatogenesis. *Development*, *129*, 4135–4146.

Zucchero, T. M., Cooper, M. E., Maher, B. S., Daack-Hirsch, S., Nepomuceno, B., Ribeiro, L., … Murray, J. C. (2004). Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *The New England Journal of Medicine*, *351*, 769–780.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.