Molecular Genetics & Genomic Medicine Open Access WILEY

ORIGINAL ARTICLE

Identification of de novo variants in nonsyndromic cleft lip with/without cleft palate patients with low polygenic risk scores

Nina Ishorst[1] | Leonie Henschel[1] | Frederic Thieme[1] | Dmitriy Drichel[2] |
Sugirthan Sivalingam[3,4,5] | Sarah L. Mehrem[1] | Ariane C. Fechtner[1] | Julia Fazaal[1] |
Julia Welzenbach[1] | André Heimbach[1] | Carlo Maj[4] | Oleg Borisov[4] |
Jonas Hausen[3,4,5] | Ruth Raff[1] | Alexander Hoischen[6,7,8] | Michael Dixon[9] |
Alvaro Rada-Iglesias[10] | Michaela Bartusel[11,12] | Augusto Rojas-Martinez[13,14] |
Khalid Aldhorae[15,16] | Bert Braumann[17] | Teresa Kruse[17] | Christian Kirschneck[18] |
Gerrit Spanier[19] | Heiko Reutter[1,20] | Stefanie Nowak[1] | Lina Gölz[21,22] |
Michael Knapp[5] | Andreas Buness[3,4,5] | Peter Krawitz[4] | Markus M. Nöthen[1] |
Michael Nothnagel[2,23] | Tim Becker[24] | Kerstin U. Ludwig[1] |
Elisabeth Mangold[1]

[1]Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany

[2]Cologne Center for Genomics, University of Cologne, Cologne, Germany

[3]Core Unit for Bioinformatic Analysis, Medical Faculty, University of Bonn, Bonn, Germany

[4]Institute for Genomic Statistics and Bioinformatics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany

[5]Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany

[6]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands

[7]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

[8]Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

[9]Faculty of Biology, Medicine & Health, University of Manchester, Manchester, UK

[10]Institute of Biomedicine and Biotechnology of Cantabria (IBBTEC), CSIC/University of Cantabria, Santander, Spain

[11]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany

[12]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[13]Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Monterrey, Mexico

[14]Centro de Investigacion y Desarrollo en Ciencias de la Salud, Universidad Autonoma de Nuevo Leon, Monterrey, Mexico

[15]Department of Orthodontics, College of Dentistry, Thamar University, Thamar, Yemen

[16]Department of Orthodontics, College of Dentistry, University of Ibn al-Nafis for Medical Sciences, Sanaa, Yemen

[17]Faculty of Medicine and University Hospital Cologne, Department of Orthodontics, University of Cologne, Cologne, Germany

[18]Department of Orthodontics, University of Regensburg, Regensburg, Germany

[19]Department of Cranio-Maxillofacial Surgery, University Hospital Regensburg, Regensburg, Germany

[20]Division of Neonatology and Pediatric Intensive Care, Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Erlangen, Germany

[21]Department of Orthodontics, University of Erlangen-Nürnberg, Erlangen, Germany

[22]Department of Orthodontics, University of Bonn, Bonn, Germany

[23]University Hospital Cologne, Cologne, Germany

[24]Institute of Community Medicine, University of Greifswald, Greifswald, Germany

Correspondence
Nina Ishorst and Elisabeth Mangold, Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany.
Email: nina.ishorst@uni-bonn.de, e.mangold@uni.bonn.de

Abstract

Background: Nonsyndromic cleft lip with/without cleft palate (nsCL/P) is a congenital malformation of multifactorial etiology. Research has identified >40 genome-wide significant risk loci, which explain less than 40% of nsCL/P heritability. Studies show that some of the hidden heritability is explained by rare penetrant variants.

Methods: To identify new candidate genes, we searched for highly penetrant de novo variants (DNVs) in 50 nsCL/P patient/parent-trios with a low polygenic risk for the phenotype (discovery).

We prioritized DNV-carrying candidate genes from the discovery for resequencing in independent cohorts of 1010 nsCL/P patients of diverse ethnicities and 1574 population-matched controls (replication). Segregation analyses and rare variant association in the replication cohort, in combination with additional data (genome-wide association data, expression, protein–protein-interactions), were used for final prioritization.

Conclusion: In the discovery step, 60 DNVs were identified in 60 genes, including a variant in the established nsCL/P risk gene *CDH1*. Re-sequencing of 32 prioritized genes led to the identification of 373 rare, likely pathogenic variants. Finally, *MDN1* and *PAXIP1* were prioritized as top candidates. Our findings demonstrate that DNV detection, including polygenic risk score analysis, is a powerful tool for identifying nsCL/P candidate genes, which can also be applied to other multifactorial congenital malformations.

KEYWORDS
candidate genes, de novo variants, exome sequencing, nonsyndromic cleft lip with/without cleft palate, polygenic risk, single-molecule molecular inversion probes

1 | INTRODUCTION

Nonsyndromic cleft lip with/without cleft palate (nsCL/P [MIM:119530]) is a common form of orofacial clefting (OFC), with a prevalence of around 1/1000 live births in Europeans (Mangold et al., 2011). The estimated heritability is around 90%, and epidemiological data indicate that high-penetrance variants in "major" genes act on a multifactorial background (Grosen et al., 2011; Marazita, 2012). To date, >40 common risk loci have been identified across diverse populations, mainly via genome-wide association studies (GWAS) and follow-up investigations (Leslie et al., 2017; Ludwig et al., 2017; Welzenbach et al., 2021; Yu et al., 2017). However, GWAS-based estimates indicate that common variation explains only a fraction of the estimated heritability, e.g.,<40% among Europeans (Ludwig et al., 2017).

Studies have shown that a fraction of the missing heritability of nsCL/P is attributable to rare, highly penetrant variants in genes that may be implicated in craniofacial development. Sequencing studies have detected functionally-relevant dominant mutations in genes that were selected as candidate genes on the basis of evidence from syndromic CL/P (Basha et al., 2018; Khandelwal et al., 2017), or their location at previously identified linkage/GWAS loci (Leslie et al., 2015; Marini et al., 2019; Savastano et al., 2017). In a more systematic approach, Cox et al. identified novel, highly penetrant susceptibility genes via exome sequencing (ES), and detected dominant mutations of relevance to nsCL/P in multiply affected families (Cox et al., 2018; Cox et al., 2019).

A complementary approach to the identification of highly penetrant susceptibility genes is a search for rare,

highly penetrant de novo variants (DNVs). Two lines of evidence support the hypothesis that DNVs are implicated in nsCL/P. First, the offspring of nsCL/P patients show higher recurrence risks than their parents (Grosen et al., 2010). Second, research indicates that in populations with high rates of mortality among affected infants, the incidence of OFC is stable. In rural India, for example, starvation secondary to breastfeeding difficulties appears to be a frequent cause of death among OFC infants, and most patients therefore fail to reach reproductive age, thus leading to selective pressure (Christensen et al., 2004). In the first study to address DNVs in OFC, Bishop et al. identified a significant enrichment of loss-of-function DNVs in genes that are highly expressed in craniofacial tissues (Bishop et al., 2020).

The aim of the present investigation was to identify highly penetrant candidate genes for nsCL/P using a three-step approach. First, an ES-based search for rare DNVs was performed in patient/unaffected parent trios of European ethnicity with a low polygenic risk for nsCL/P and no history of OFC in siblings or ancestors (discovery step; Figure 1 and Table 1). Second, replication of prioritized DNV-carrying genes from the discovery step was performed in an independent sample comprising nsCL/P patients from diverse ethnicities and population-matched controls (replication step; Figure 1 and Table 1). Third, to prioritize genes and provide additional evidence for a true contribution of the candidate genes to nsCL/P, a series of adjunct analyses were performed (additional evidence; Figure 1).

2 | MATERIALS AND METHODS

2.1 | Patient and control samples

Written informed consent was obtained from all participants, or the respective parents/legal guardians, prior to inclusion. Detailed information on all cohorts is provided in Table 1.

2.2 | Discovery step

2.2.1 | Selection of trios

nsCL/P patients and their unaffected parents were drawn from a large, phenotypically well characterized Central European cohort (Mangold et al., 2010). The discovery cohort (Table 1) was enriched for potential de novo events by: (i) selecting index patients with no history of OFC in siblings or ancestors; and (ii) prioritizing individuals with severe nsCL/P phenotypes, as defined by the number of

affected structures (Grosen et al., 2010). Preference was given to individuals with affected descendants and available genome-wide SNP data for the performance of polygenic risk score (PRS) analysis.

2.2.2 | PRS analysis

The PRS was established in the European subset of the published GWAS data of Beaty et al. (phs000094.v1.p1) (Beaty et al., 2010). This score was then applied to each of the following three subgroups of our in-house European nsCL/P GWAS dataset (Mangold et al., 2010): (i) nsCL/P patients who were selected as index patients for the present ES analyses ($n = 47$); (ii) healthy controls ($n = 1318$); and (iii) all nsCL/P patients, with the exception of individuals from multiplex families from a previous linkage study (Mangold et al., 2009). For three of the present discovery index cases, no array-based genotype data were available.

PRS computation was performed using PRSice-2 v. 2.2.11 (https://www.prsice.info). Variants for PRS computation were selected using a clumping procedure, with a radius of 250 kb and a variant correlation threshold r^2 of 0.1. For each of the three GWAS subgroups, the PRS was computed as the weighed sum of the clumped variants. The effect size estimates (i.e., the β-values from association values) were taken as weights, and the full model was considered (i.e., no filtering based on association p-values). Differences in PRS distributions across the three GWAS subgroups were assessed using t-tests.

2.2.3 | Exome sequencing

Enrichment of the coding sequence (50.4 Mb) was performed using 3 μg high molecular dsDNA, in accordance with the standard procedure of the Sure SelectXT Human All Exon v5 Kit (Agilent). Libraries were sequenced on a HiSeq2500 platform (Illumina), using paired-end sequencing with 2×125 bp read-length. Demultiplexed raw data were processed using the default Burrows-Wheeler Alignment Tool (BWA)/Genome Analysis Toolkit (GATK) v.3.4 pipeline (https://github.com/broadinstitute/gatk-docs/tree/master/gatk3-methods-and-algorithms), which includes variant quality score recalibration (VQSR) and the genotype refinement workflow (Van der Auwera et al., 2013). Annotation was performed using ANNOVAR (Wang et al., 2010), and trio status was confirmed using Identity by descent (IBD) computing. The detailed filtering and candidate gene selection strategy is shown in Figure 3a. Briefly, DNVs located in segmental duplications were excluded. Reads encompassing the remaining DNVs were visually inspected using Integrative Genomics

Discovery

PRS analysis for selection of Discovery cohort

↓

Exome Sequencing (50 trios)

↓

De novo mutation?

→ yes

Additional evidence

Interactions at protein level

GWAS associations in respective TADs

Expression in relevant mouse tissues/human cell lines

60 candidate genes

↓

Validated and protein altering? → no → 27 candidate genes excluded

↓ yes

33 candidate genes†

↓

Replication

Resequencing with smMIPs (1010 patients, 1574 controls)

→ Gene-based test

if possible ↓

Additional *de novo*/co-segregating variants?‡ → no → 28 candidate genes excluded

↓ yes

4 candidate genes

↓

Prioritization → 2 candidate genes excluded
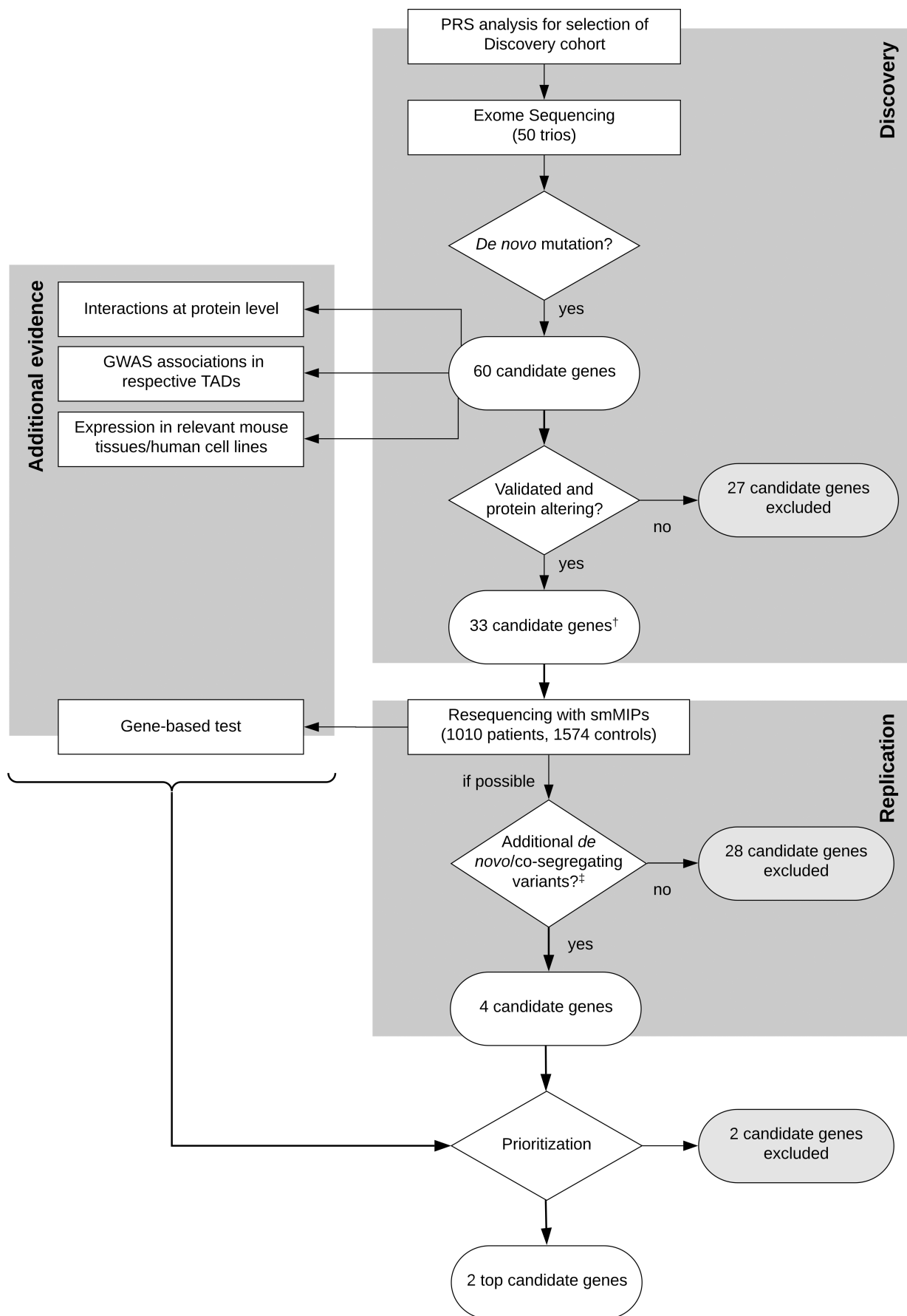
↓

2 top candidate genes

FIGURE 1 Present workflow. Discovery and replication were performed sequentially. The analyses indicated in the gray box labeled 'additional evidence' were performed to generate further evidence for a contribution of the candidate genes to nsCL/P. The results of these analyses were not used as filter criteria but for a final prioritization. Discovery: The PRS was established in the European subset of the published GWAS data of Beaty et al. (2010). This score was then used to compare the polygenic risk of the nsCL/P trios to nsCL/P patients from GWAS. Exome sequencing was performed in 50 selected trios from the in-house European nsCL/P cohort. This cohort was first described in Mangold et al. (2010). Replication: Resequencing with smMIPs was performed in 1010 patients and 1547 controls from the in-house nsCL/P cohorts, who were of European, Yemenite, or Mexican ancestry. GWAS, Genome-wide association study; nsCL/P, nonsyndromic cleft lip with/without cleft palate; PRS, Polygenic risk score; smMIPs, Single-molecule molecular inversion probes; TAD, topologically associating domain. [†]A total of 33 candidate genes fulfilled the criteria for inclusion in the replication step. For one gene, however, assay design was not possible. [‡]Segregation analysis was only possible for the European replication cohort. For the Mexican and the Yemenite cohorts, no DNA from additional family members was available.

TABLE 1 Overview of present samples

| | Sample numbers | | | | Sex | | | |
| | | | | | Patients | | Controls | |
	Trios	Patients	Controls	Unaffected parents	f	m	f	m
Discovery cohort								
ES trios from Central Europe	50	50[b]	—	100	22	28	—	—
Replication cohort[a]								
Resequencing Europe		643[c] (651)	855 (858)		232	411	403	452
Resequencing Yemen		217 (217)	418 (421)		82	135	11	407
Resequencing Mexico		150 (150)	301[d] (312)		51	99	198	98
Total		1010	1574		365	645	612	957

Note: The discovery cohort comprised 50 trios (index with nsCL/P and unaffected parents) of central European ancestry. The replication cohort comprised independent nsCL/P patients and controls of diverse ethnicities.

Abbreviations: ES, exome sequencing; f, female; m, male; nsCL/P, nonsyndromic cleft lip with/without cleft palate.

[a]Numbers in brackets indicate numbers pre quality control.

[b]Discovery cohort comprised the following phenotypes: bilateral cleft lip and palate (*n* = 26); unilateral cleft lip and palate (*n* = 15); bilateral cleft lip only (*n* = 3); and unilateral cleft lip only (*n* = 6).

[c]337 cases suitable for de novo detection via Sanger sequencing (resequencing was performed in independent individuals with nsCL/P only).

[d]Sex of five Mexican controls unknown.

Viewer (IGV) (https://software.broadinstitute.org/software/igv/), and those with poor read quality were excluded. DNVs that were validated using Sanger sequencing (true DNVs; Figure 3a), and those for which validation was precluded for technical reasons (possible DNVs; Figure 3a) were used in the subsequent analyses (Replication and Additional evidence).

2.3 | Replication step

2.3.1 | Description of replication cohort

The replication cohort comprised nsCL/P patients of European (Mangold et al., 2010), Mexican (Rojas-Martinez et al., 2010), and Yemeni (Aldhorae et al., 2014) ancestry, as well as population-matched controls (Table 1). The European samples were drawn from the same cohort as the ES samples (Mangold et al., 2010), and included all independent nsCL/P patients not included in the discovery cohort (*n* = 651). For the European replication cohort, DNA from additional family members was available for segregation analysis. For 337 of these 651 index patients, DNA was available from both unaffected parents. For a further 88 European index patients, DNA from two or more family members (either affected or unaffected) was available for segregation analysis. Due to the lack of DNA from additional family members, the Mexican and Yemeni cohorts were used for the gene-based analysis only (Figure 1).

2.3.2 | Prioritization of candidate genes for replication

A subset of DNV-carrying genes from the discovery step was prioritized based on combined evidence from: (i) functional annotation (Wang et al., 2010); (ii) quality

control; (iii) minor allele frequency (MAF) of the DNVs in the candidate genes in GnomAD (https://gnomad.broadinstitute.org/) and GoNL (http://www.nlgenome.nl/); and (iv) presence of the DNV in the candidate gene in denovo-db (http://denovo-db.gs.washington.edu/denovo-db/) (Figure 3a).

2.3.3 | Resequencing

Targeted sequencing of the prioritized gene subset was performed in the multiethnic replication cohort (Table 1) using single molecule molecular inversion probes (smMIPs), in accordance with previously reported procedures (Eijkelenboom et al., 2016; Thieme et al., 2021). SmMIP libraries were sequenced on a HiSeq2500 platform (Illumina), using paired-end sequencing with a 2×125 bp read-length.

Trimming and collapsing of smMIPs was performed using scripts provided by the Shendure lab (https://github.com/shendurelab/MIPGEN/tree/master/tools). Unified Genotyper was used for variant calling. Retrieved variants were annotated using ANNOVAR (Wang et al., 2010). A schematic representation of dataset quality control/filtering is provided in Figure 3b. Reads were manually inspected with IGV.

Using this dataset, two analytical approaches were applied: (i) segregation analysis in the European cohort only; and (ii) gene-based analysis in all three cohorts.

2.3.4 | Segregation analysis in European replication cohort

Filtering was performed for missense and nonsense variants that fulfilled all three of the following criteria: (i) a CADD score ≥15 (Rentzsch et al., 2019); (ii) an MAF ≤0.1% in the GnomAD non-Finnish European exomes; and (iii) a presence in nsCL/P patients only in the present cohorts (Figure 3b). To determine whether the detected variants encompassed further DNVs, Sanger sequencing was performed in unaffected parents for whom DNA was available. For families with more than one affected family member, co-segregation analysis using Sanger sequencing was performed when DNA was available. Primer sequences are available upon request. For index patients carrying a DNV, paternity testing was performed using the Powerplex 16 HS System (Promega).

2.4 | Additional evidence

2.4.1 | Interactions between proteins encoded by the identified candidate genes

Protein–protein interaction (PPI) analyses were performed using STRING v11.5, with changes to the default settings (https://string-db.org/). Edges with evidence from 'STRING active interaction source'-categories textmining, experiments, and curated databases are shown. Associations based on evidence from 'STRING active interaction source'-categories co-expression, neighborhood, gene fusion, and co-occurrence were excluded. Disconnected nodes representing proteins with no associations were hidden. Edges represent protein–protein associations. These do not necessarily imply actual physical binding.

2.4.2 | GWAS associations in the respective candidate gene TADs

P-values for common variants located in topologically associating domains (TADs) encompassing DNV-carrying genes were extracted from our recent European nsCL/P GWAS meta-analysis dataset ($n = 1860$) (Ludwig et al., 2017). Candidate gene regions ($n = 60$) were defined as the exons of all transcripts, including intronic regions and flanking regions. TAD structures were drawn from an Hi-C human embryonic stem cell dataset (Dixon et al., 2012). Bonferroni correction was used to adjust the single-test p-value threshold for all markers from all TADs ($p_{corrected} = 0.05/113{,}749 = 4.4 \times 10^{-7}$). The dataset comprised 56 TADs (3 TADs with 2 DNV-carrying genes). This included 113,749 common SNPs with an info score ≥0.8 after genotype imputation.

2.4.3 | Assessment of candidate gene expression

The expression of candidate genes was inspected in two RNA-Seq datasets from relevant embryonic time-points in tissues of relevance to lip/palate development. First, reads per kilobase per million (RPKM) values from a previously generated RNA-Seq dataset from murine frontonasal processes (E10.5 to E12.5) and secondary palatal shelves (E13.5 to E14.5) were assessed (accession number E-MTAB-3157). Human gene names were assigned to their mouse orthologs using HGNC Comparison of Orthology Predictions (https://www.genenames.org/). Second, fragments per kilobase per million (FPKM) values from a previously published RNA-Seq dataset from human

neural crest cells (hNCC) and human cranial neural crest cells (hCNCC) were assessed (Laugsch et al., 2019).

2.4.4 | Rare variant association (gene-based analysis)

The Optimized Sequence Kernel Association Test (SKAT-O) was performed on smMIP data from the replication cohort (1010 nsCL/P patients; 1574 population-matched controls) (Table 1) (Wu et al., 2011). The analysis was performed using the R package SKAT v1.3.2.1 in each of the three ethnic cohorts, and included all variants that passed the pipeline quality criteria and which had a CADD score >20 ($n = 941$ from all three cohorts). Based on the assumption of an inverse correlation between the effect sizes and the MAF, variants were weighed according to the MAF using the β-density function. The regions were defined by the coordinates of the re-sequenced genes.

Results are reported as p-values. The Holm-Bonferroni method was used to correct for multiple testing.

3 | RESULTS

3.1 | Discovery step

3.1.1 | PRS analysis

The PRS analysis was used to demonstrate that the discovery cohort had a low polygenic risk, as based on common variation. GWAS individuals with nsCL/P had a significantly increased polygenic risk compared to unaffected GWAS controls ($p = 1.6 \times 10^{-8}$), and the 47/50 nsCL/P index patients from the discovery cohort for whom genome-wide SNP data were available showed no significant difference in polygenic risk distribution compared to the unaffected GWAS controls ($p = 0.74$) (Figure 2).
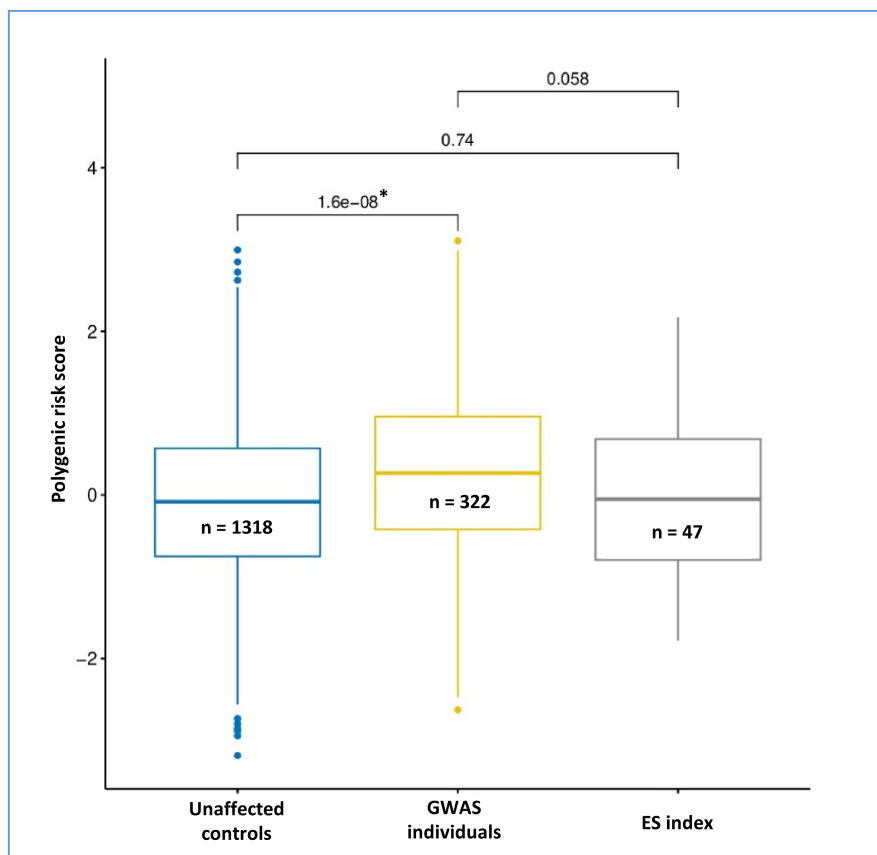

FIGURE 2 Comparison of polygenic risk for nsCL/P between unaffected controls from GWAS, GWAS individuals with nsCL/P (excluding multiplex families), and nsCL/P patients selected as index cases for the present ES trios. The polygenic risk score was established using the European component of the Beaty et al. (2010) dbGaP dataset (accession number phs000094.v1.p1), and was applied to different subgroups. The European component of our Ludwig et al. (2017) nsCL/P GWAS dataset was subdivided into three groups: (i) unaffected GWAS controls ($n = 1318$); (ii) individuals with nsCL/P who were not index patients in the present trios, and who had not been included in a previous in-house linkage study (Mangold et al., 2009) ($n = 322$); and (iii) individuals with nsCL/P who were index patients for the present trios ($n = 47$). The polygenic risk distributions of the three groups were compared, and the statistical significance of the differences was calculated using t-tests. ES, exome sequencing; GWAS, genome-wide association study; nsCL/P, nonsyndromic cleft lip with/without cleft palate.

3.1.2 | Exome sequencing

ES of the discovery cohort (50 trios) resulted in a mean target coverage of 116.3X (78.6X-173.8X; 94.7% of target regions covered at least 20X). IBD computation confirmed parental status in all 50 trios.

After quality control and filtering (Figure 3a), a total of 63 DNVs remained, including five nonsense DNVs (Figure S1, Table S1). In total, 3/63 DNVs were false positive findings. For 13/63 variants, validation was precluded for technical reasons, i.e., no primers could be designed or the PCR/sequencing yielded no conclusive results.
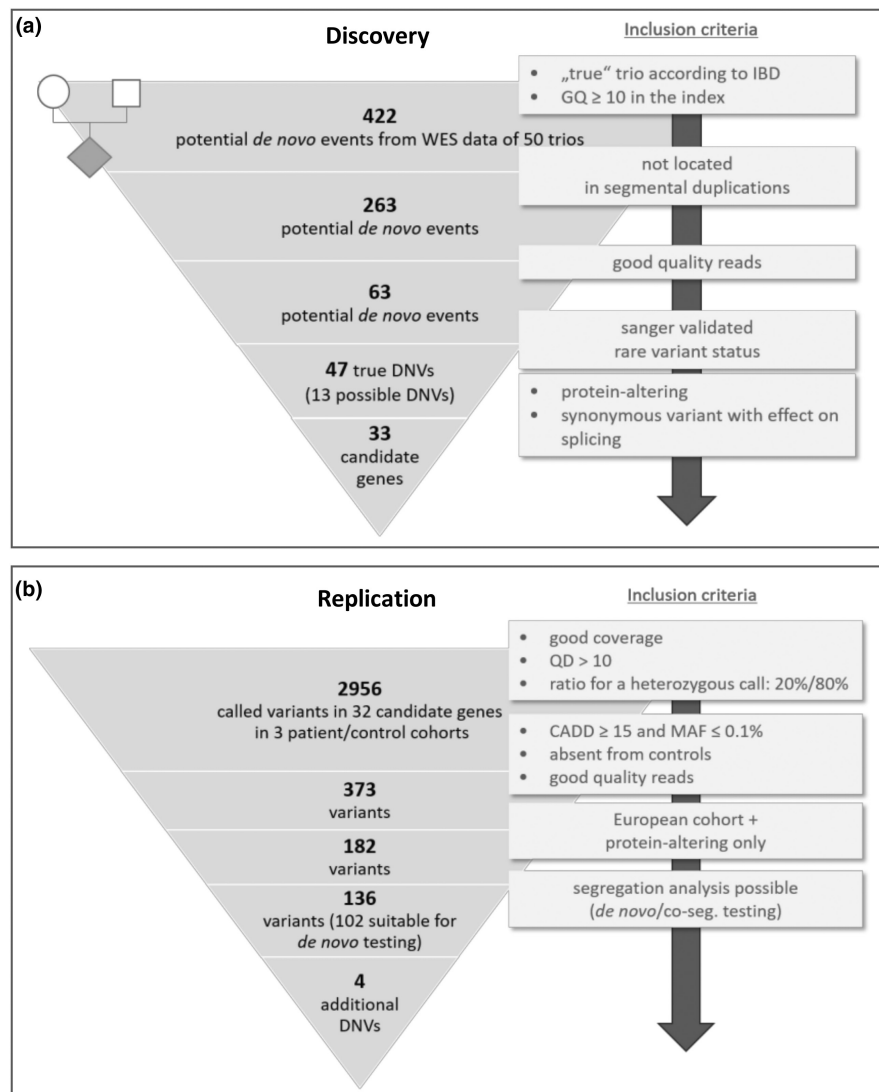

FIGURE 3 Filtering strategy and results for discovery (a) and replication (b) step. (a) Trio status was confirmed via IBD computation. De novo events with a GQ ≥10 in the index were called. Variants in segmental duplications were excluded, and read quality was assessed using integrative genomics viewer. In total, 63 potential de novo events with good quality reads underwent validation, resulting in 47 true DNVs and 13 variants that could not be validated for technical reasons (*n* = 60 in total). For this set of variants, rare variant status was confirmed using population frequencies from publicly available databases (GnomAD and GoNL). All synonymous variants among these 60 (*n* = 19) were analyzed for a potential effect on splicing using NNSplice. All synonymous variants with an effect on splicing and all protein-altering (nonsense, missense, and (non)-frameshift) variants were retained. (b) Samples for which ≥10% of the single molecule molecular inversion probes had poor coverage (<20x after collapsing) were excluded from the variant calling procedure. Called variants with a QD <10 were filtered out. Variants with 20% or more reads for the alternative allele and 80% or less reads for the reference allele were considered heterozygous. The final calls were filtered for a CADD score ≥15, a MAF ≤0.1%, and good quality reads according to manual inspection. Protein-altering variants (nonsense, missense, and (non)-frameshift) that were identified in the European cohort and that occurred in individuals for whom DNA of additional family members was available were subjected to a segregation/de novo analysis. CADD, combined annotation dependent depletion; co-seg., co-segregation; DNV, de novo variant; ES, exome sequencing; GQ, genotype quality; IBD, identity by state computation; MAF, minor allele frequency; QD, quality by depth.

However, these remain potentially true findings. Thus, 60 DNVs were found in 50 trios (range = 1–4 DNVs per individual), resulting in an average of 1.2 DNVs, which corresponds to the expected number of around 1–2 DNVs per exome (Samocha et al., 2014). This set of 60 DNVs, which included one DNV in Cadherin 1 (*CDH1* [MIM:192090]; NM_004360.3), was distributed over 60 candidate genes, which were then forwarded for further analysis (Table S1). A total of 33/60 candidate genes had DNVs with protein-altering effects (including effects on splicing) and were selected for the replication step (Materials and Methods, Figure 3a, Table 2).

3.2 | Replication

3.2.1 | Targeted-sequencing with single-molecule molecular inversion probes

Assay design for smMIPs was successful for 32/33 genes. Targeted-sequencing in the replication cohort (independent nsCL/P index patients and population-matched controls, Table 1) resulted in a mean target coverage of 1160x before, and 273x after, the collapsing of all reads originating from one PCR duplicate to one consensus read (Materials and Methods). A total of 2956 variants in the independent nsCL/P index patients from the replication cohort were called (Figure 3b).

3.2.2 | Segregation analysis in the European replication cohort

After filtering, a total of 373 variants in 402 individuals remained (Figure 3b). Since DNA from additional family members was only available for the European participants, the segregation analysis was restricted to this cohort. In total, 182 protein-altering variants were called in the 643 European nsCL/P patients in the replication step. A total of 12/182 variants were excluded due to technical problems during validation. A total of 170/182 variants were validated using standard Sanger sequencing. A final total of 162 variants were validated in 170 independent index patients (Table S2). In the European replication cohort, 144/643 nsCL/P patients fulfilled both criteria for inclusion in the segregation analysis. For 102/144 of these individuals, DNA was available from both unaffected parents, and analyses were performed to determine whether the variants had occurred de novo (Table S2). For the candidate genes Midasin AAA ATPase 1 (*MDN1* [MIM:618200]; NM_014611.1), CUB And Sushi Multiple Domains 1 (*CSMD1* [MIM:608397]; NM_033225.5), Ankyrin 1 (*ANK1* [MIM:612641]; NM_001142445.1, NM_000037.3), and Pax

Transcription Activation Domain-Interacting Protein 1 (*PAXIP1* [MIM:608254]; NM_007349.3), at least one additional DNV was found in an nsCL/P patient from the replication cohort (Table S3, Figure S2). All four individuals had a severe phenotype, i.e., unilateral cleft lip with a cleft palate ($n = 2$) or bilateral cleft lip with a cleft palate ($n = 2$). For all de novo findings, paternity was confirmed. For the index patient with the DNV in *CSMD1*, DNA from a successive generation was available. The DNV was not transmitted to the affected offspring of the index (Figure S2). The analyses also identified: (i) a variant co-segregating with the phenotype in *MDN1* and *CSMD1*; and (ii) a compound heterozygous index patient with two variants in *MDN1*, one of which had been inherited from the father and the other from the mother (Table S3, Figure S2).

3.3 | Analysis for additional evidence

No significant enrichment of PPIs was found among the candidate gene set of 60 DNV carrying genes ($p = 0.518$). However, the identified interactions included two networks (Figure 4a) that comprised more than two genes from our dataset.

Analyses to identify associations with common variants at the 60 gene loci identified one test-wide significant marker (rs3746101, $p_{\text{corrected}} = 2.30 \times 10^{-7}$) in TAD49, including two genes with DNVs (Signal Peptide Peptidase Like 2B (*SPPL2B* [MIM:608239]; no NM available); and Major Facilitator Superfamily Domain Containing 12 (*MFSD12* [MIM:617745]; NM_001287529.2)) (Figure 4b, Table S5). Notably, rs3746101 is the lead SNP at an established genome-wide significant nsCL/P risk locus (19p13.3) (Leslie et al., 2016). The second most significant *p*-value was found for rs4582663 ($p_{\text{corrected}} = 2.27 \times 10^{-5}$) in TAD25, which includes TRNA Methyltransferase O (*TRMO* [no MIM available]; NM_016481.4) (9q22 locus) (Figure 4c, Table S5) (Moreno et al., 2009). This marker is in moderate linkage disequilibrium with the lead SNP from the study by Moreno et al. (rs3758249, $r^2 = 0.67$).

The assessment of candidate gene expression in embryonic mouse tissues of relevance to lip/palate development was performed in 58/60 candidate genes with corresponding mouse orthologs (Table S6). Relevant human cell line data were available for all 60 candidate genes (Table S6). With respect to the four genes with more than one DNV (i.e., one from the discovery and one from the replication, respectively), *MDN1* and *PAXIP1* were highly expressed in all relevant mouse tissues at all embryonic time points (RPKM$_{Mdn1}$ = 922.02–1637.17; RPKM$_{Paxip1}$ = 429.83–1338.15; Table S6), and in hNCC and hCNCC (FPKM$_{MDN1(hNCC)}$ = 11.36; FPKM$_{PAXIP1(hNCC)}$ = 9.18; Table S6). *ANK1* showed a moderate expression in

TABLE 2 Overview of DNV–carrying candidate genes selected for the replication step

Candidate gene	Change[b]	Trio	MAF[c] gnomAD (NFE)	GoNL	Denovo-db
SDC3	NM_014654.3:c.166G>A (p.(Glu56Lys))	BN46	nf	nf	nf
NAV1	NM_001167738.1:c.3397G>A (p.(Val1133Ile))	BN104	4.87×10^{-5}	nf	nf
XDH	NM_000379.3:c.431A>C (p.(Gln144Pro))	BN192	3.96×10^{-5}	nf	nf
ATXN7	NM_001128149.3:c.1759G>A (p.(Glu587Lys))	BN82	nf	nf	nf
SENP2	NM_021627.2:c.967G>A (p.(Glu323Lys))	BN142	nf	nf	nf
GLRA3	NM_001042543.2:c.1018G>C (p.(Val340Leu))	BN192	nf	nf	nf
MAST4	NM_001290227.2:c.5394_5396del (p.(Ser1799del))	BN98d	1.80×10^{-5}	nf	nf
H1-3	NM_005320.3:c.115C>T (p.(Pro39Ser))	BN565	8.95×10^{-6}	nf	nf
MDN1	NM_014611.1:c.489C>G (p.(Phe163Leu))	BN82	nf	nf	nf
TAAR8	NM_053278.2:c.748_749del (p.(Lys250GlufsTer7))	BN591	2.06×10^{-4}	nf	nf
TMEM120A	NM_031925.2:c.870G>A (p.(Ala290=))	BN23	2.76×10^{-5}	nf	nf
CFTR	NM_000492.3:c.4333G>A (p.(Asp1445Asn))	BN462	2.93×10^{-4}	nf	nf
PAXIP1	NM_007349.3:c.2434G>A (p.(Asp812Asn))	BN46	0	nf	nf
CSMD1	NM_033225.5:c.2699A>G (p.(Asp900Gly))	BN172	nf	nf	nf
LONRF1	NM_152271.4:c.2227T>C (p.(Tyr743His))	BN192	nf	nf	nf
ANK1	NM_001142445.1:c.13G>A (p.(Val5Ile))	BN865	4.75×10^{-5}	nf	nf
VCP	NM_007126.4:c.475C>T (p.(Arg159Cys))	BN62d	8.95×10^{-6}	nf	nf
TRMO	NM_016481.4:c.1265C>T (p.(Pro422Leu))	BN98	1.79×10^{-5}	nf	N/A
CLPB	NM_001258393.2:c.913A>G (p.(Lys305Glu))	BN421	0	nf	nf
BSX	NM_001098169.2:c.556G>C (p.(Glu186Gln))	BN104	nf	nf	nf
KIRREL3	NM_001161707.1:c.1190T>C (p.(Val397Ala))	BN139	nf	nf	nf
CDH1	NM_004360.3:c.774_776del (p.(Asn258del))	BN86	nf	nf	nf
SMG6	NM_001282326.2:c.391C>T (p.(Gln131Ter))	BN82	nf	nf	nf
DHX40[a]	NM_001166301.2:c.1842G>A (p.(Trp614Ter))	BN137	nf	nf	nf
ABCA8	NM_007168.3:c.3103T>G (p.(Ser1035Ala))	BN119	nf	nf	nf
FDXR	NM_001258015.2:c.343C>T (p.(Arg115Trp))	BN54	7.91×10^{-6}	nf	nf
TMC6	NM_001127198.5:c.1870G>A (p.(Val624Ile))	BN249	9.67×10^{-6}	nf	nf
ZNF407	NM_001146189.1:c.3373G>A (p.(Ala1125Thr))	BN138	nf	nf	nf
EIF3G	NM_003755.4:c.3G>T (p.(Met1?))	BN497	nf	nf	nf
CASP14	NM_012114.3:c.674G>A (p.(Arg225Lys))	BN200	nf	nf	nf
GNAS	NM_016592.3:c.190_218del (p.(Phe64ProfsTer5))	BN565	nf	nf	nf
SEC14L3	NM_174975.4:c.268G>A (p.(Gly90Ser))	BN139	nf	nf	nf
NAP1L2	NM_021963.3:c.1090G>T (p.(Asp364Tyr))	BN240	nf	N/A	nf

Abbreviations: DNV, de novo variant; gnomAD, Genome aggregation database; MAF, minor allele frequency; N/A, not available (not covered); nf, not found; NFE, non-Finnish European; GoNL, Genomes of the Netherlands.

[a]Gene failed in the single molecule molecular inversion probe assay design.

[b]Gene accession number (NM); change on DNA level (c.); change on protein level (p.).

[c]Publicly available datasets accessed August 2018.

relevant mouse tissues (RPKM$_{Ank1}$ = 35.23–71.27), as well as in hNCC and hCNCC (FPKM$_{ANK1(hNCC)}$ = 0.34). Finally, the expression profile of *CSMD1* in the assessed mouse tissues ranged from not expressed at all to marginally expressed (RPKM$_{Csmd1}$ = 0–52.14), while only marginal expression was evident in the two human cell lines (FPKM$_{CSMD1(hNCC)}$ = 0.04).

In the gene-based analysis, four candidate genes (*CDH1*; Transmembrane Channel like 6 (*TMC6* [MIM:605828]; NM_001127198.5); Eukaryotic Translation Initiation Factor 3 Subunit G (*EIF3G* [MIM:603913]; NM_003755.4); and *TRMO*) showed a nominally significant enrichment for deleterious variants in the European patients (p_{CDH1} = 0.0019, p_{TMC6} = 0.0132, p_{EIF3G} = 0.0133,
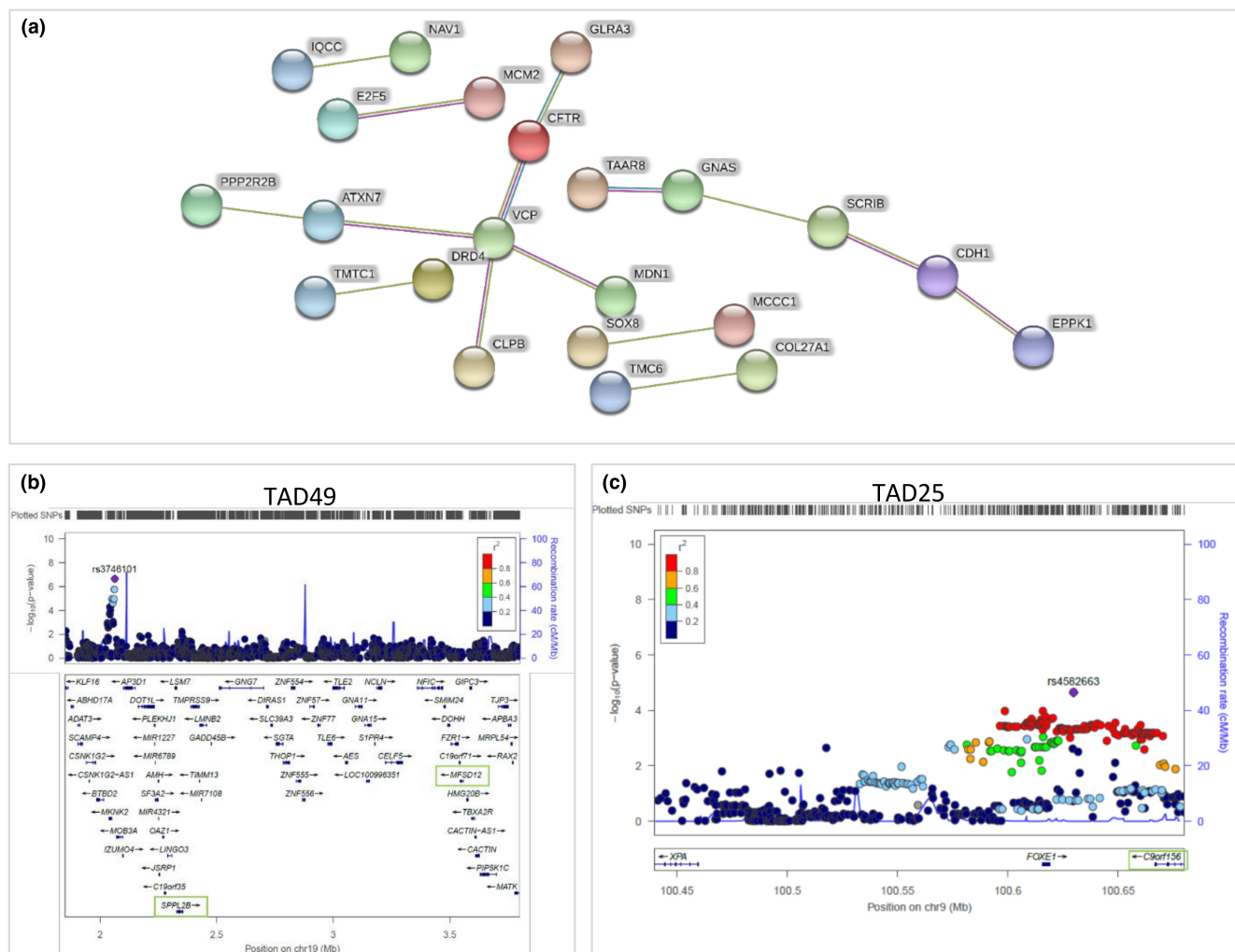
FIGURE 4   Results of adjunct analyses (additional evidence). (a) STRING based protein–protein interaction network. Edges represent protein–protein associations. Although specific, these do not necessarily imply actual physical binding. Edges with evidence from textmining (light green), experiments (pink), and curated databases (light blue) are shown. Disconnected nodes with no associations are hidden. The respective input genes are shown in Table S1 (excluding not true and mosaic findings) ($n = 60$). (b and c) marker association drawn from the European meta-analysis of nonsyndromic cleft lip with/without cleft palate (Ludwig et al., 2017). Plotted SNPs, plotted single nucleotide polymorphisms (termed single nucleotide variant in the present manuscript). (b) Regional association plot for TAD49, including *SPPL2B* and *MFSD12*. The purple diamond represents the top marker in the topologically associating domain (TAD; rs3746101). The candidate genes *SPPL2B* and *MFSD12* are indicated by green boxes. (c) Regional association plot of TAD25, including *TRMO* (*C9orf156*) at the 9q22 locus (including *FOXE1*). The candidate gene *TRMO* (*C9orf156*) is indicated by a green box. The purple diamond represents the top marker in the TAD (rs4582663). For respective gene accession numbers (NM) refer to Table S1.

$p_{TRMO} = 0.0399$) (Table S7). *CDH1* almost withstood correction for multiple testing ($p_{CDH1-corrected} = 0.0607$) (Table S7). In the Yemeni cohort, Brain Specific Homeobox (*BSX* [MIM:611074]; NM_001098169.2) showed a nominally significant overrepresentation of deleterious variants ($p_{BSX} = 0.0045$; $p_{BSX\_corrected} = 0.1446$) (Table S8). In the Mexican cohort, a nominally significant enrichment of deleterious variants was found for *EIF3G* ($p_{EIF3G} = 0.0115$; $p_{EIF3G\_corrected} = 0.3583$) (Table S9).

All results are summarized in Table S10.

4 | DISCUSSION

NsCL/P is a common congenital malformation, and recurrence risk is a major concern for many affected families. To date, >40 common nsCL/P risk loci have been identified. Although these explain a proportion of the etiology, they are not used for personalized risk estimations within clinical practice. The aim of the present study was to identify highly penetrant nsCL/P genes of potential clinical relevance in individuals with a low polygenic background via

the identification of DNVs. With this genetically-defined selection process, our study is the first nsCL/P patient/parent-trio investigation to employ PRS analysis. This approach led to the identification of 60 candidate genes, including the established nsCL/P susceptibility gene *CDH1* (Table S1) (Cox et al., 2018; Vogelaar et al., 2013). Furthermore, the PPI analysis identified one protein network that included CDH1. CDH1-interaction partners identified on the basis of textmining and experimental evidence were Scribble Planar Cell Polarity Protein (SCRIB [MIM:607733]; NM_015356.4) and Epiplakin 1 (EPPK1 [MIM:607553]; NM_031308.4). In addition, *CDH1* almost reached test-wide significance in the rare variant association in the European cohort. Reidentification of this established candidate gene corroborates our approach.

In contrast to Bishop and colleagues (Bishop et al., 2020), no enrichment of DNVs was evident in the present cohort. The lack of a statistical significant enrichment may have been attributable to sample size, since Bishop and colleagues included approximately fourteen times more nsCL/P trios than were available in our discovery cohort ($n_{Bishop} = 698$; $n_{Ishorst} = 50$).

In total, the discovery and the replication step identified four genes with at least two protein-altering DNVs and co-segregating/compound heterozygous variants (*ANK1*, *CSMD1*, *MDN1*, and *PAXIP1*). *MDN1* and *PAXIP1* were prioritized as top candidates on the basis of: (i) all results from the discovery/replication step and the additional analyses (Figure 1; Table S10), in particular those concerning the second DNV in the replication step; (ii) their intolerance to genetic variation (Constraint Score; Karczewski et al., 2020); and (iii) their strong expression in the precursor cells of craniofacial structures (hNCC) (Table S10).

Previous research shows that *MDN1* and *PAXP1* are promising candidate genes for nsCL/P. *MDN1* (Constraint Score = 2.01; Table S10) is a nuclear chaperone required for the maturation and nuclear export of pre-60S ribosome subunits (Raman et al., 2016). According to data from the International Mouse Phenotype Consortium, mice in which a homozygous allele has been introduced die prior to weaning, with complete penetrance (Dickinson et al., 2016). A plausible hypothesis is that this may be attributable to an orofacial cleft, which leads to feeding problems and death secondary to starvation. The protein encoded by *PAXIP1* forms a complex with human PAX-Interacting Protein 1 (PAXIP1)-associated glutamate rich protein 1 (PAGR1 [MIM:612033]; NM not relevant), lysine methyltransferase 2C (KMT2C [MIM:606833]; NM not relevant) and lysine methyltransferase 2D (KMT2D [MIM:602113]; NM not relevant). Loss of Pagr1 protein in mice, and thus the loss of the complex, leads to abnormalities in embryonic development, suggesting that the protein complex including Paxip1 plays a role in early embryogenesis (Kumar et al., 2014). Furthermore, *Paxip1* is expressed in the

branchial arches of mouse embryos (Fowles et al., 2003), and is intolerant to missense variation (Constraint Score = 3.28; Table S10). Interestingly, *PAXIP1* is associated with Kabuki syndrome, a disorder whose phenotypic spectrum includes OFC (McVeigh et al., 2015). To clarify whether the identified susceptibility genes make a true contribution to nsCL/P, functional studies are warranted using in vivo models suitable for the investigation of craniofacial development. Possible approaches include CRISPR/Cas9 knock-ins of the variants, or complete gene knock-outs in mouse or zebrafish (Van Otterloo et al., 2016).

Although the present results highlight two top candidate genes, all 60 DNV-carrying genes identified in the discovery cohort are plausible nsCL/P candidates. Genes with protein-altering DNVs were prioritized for the replication step due to limited capacity in the replication assay. However, genes with synonymous DNVs were nonetheless included in the 'additional analyses', since research has shown that this type of genetic variation might have a relevant influence on mRNA stability or mRNA/protein levels (Gaither et al., 2021; Sharma et al., 2019). Additional analyses identified: (i) genes encoding interaction partners of CDH1 (*SCRIB*, *EPPK1*); (ii) genes located at established nsCL/P GWAS risk loci (*MFSD12*, *SPPL2B*, *TRMO*); and (iii) genes that showed a nominally significant association with nsCL/P in at least one replication cohort (*CDH1*, *TMC6*, *EIF3G*, *BSX*) in a gene-based test (Table S10). In *CDH1*, for example, no further DNV was identified in the replication cohort, despite the fact that multiple layers of evidence, including data from in vivo models, implicate *CDH1* in OFC (Cox et al., 2018). In addition, our PPI analysis identified a network centered on CDH1, which involved the interaction partner SCRIB and EPPK1 (Figure 4a). A copy number variant at chromosome 8q24.3, which includes the gene encoding for SCRIB, causes syndromic phenotypes that encompass features such as cleft palate (Dauber et al., 2013). Although zebrafish with a *scrib* knock-down present with craniofacial defects, to date these have not been evaluated with regard to any defect of the anterior neurocranium, which is the homolog structure of the human palate (Dauber et al., 2013). In a study that identified candidate genes for the VATER/VACTERL association, in situ hybridization revealed that the sites of *Eppk1* expression in mouse embryos included two locations of interest: (i) the nasal passages at E12.5; and (ii) the tooth buds at E14.5, which is of relevance since OFC is occasionally associated with tooth agenesis (Hilger et al., 2013; Phan et al., 2016). Three other candidate genes from the discovery step (*TRMO*, *SPPL2B*, and *MFSD12*) are located at well-established nsCL/P risk loci (chromosome 9q22 and 19p13.3) (Leslie et al., 2016; Moreno et al., 2009). Furthermore, none of these were found to carry any further rare (de novo) variants in the

replication step. The gene-based test revealed a nominally significant enrichment of rare deleterious variants in *TMC6*, *CDH1*, *EIF3G*, and *TRMO* in individuals with ns-CL/P, with *CDH1* almost reaching test-wide significance in the European cohort (Table S7). The detection of *CDH1* in the gene-based analysis adds weight to the findings for *TMC6* and *EIF3G*. The paucity of test-wide significant findings – particularly in the Mexican and Yemenite cohorts – may have been attributable to the limited sample sizes. Resequencing of the 32 genes in further individuals from the respective ethnicities, and meta-analyses of cohorts from other ethnicities, are warranted.

Our study had three main limitations. First, the filtering criteria in the replication cohort were rather conservative, since they excluded all variants present in any control individual - irrespective of their ancestry - and did not consider the potential presence of individuals with OFC in the European control cohort. A less conservative filtering approach would: (i) only exclude variants present in the ethnically matched control cohort; and (ii) not exclude variants present in both one European control and one European patient, since this control individual could be a potential clefting patient. Second, the use of prediction tools (CADD-score) may have led to the loss of true positive findings. Finally, the occurrence of sporadic cases in families without a history of OFC and a low polygenic burden could also be explained by recessive inheritance. Thus, a promising future approach would be to re-analyze the discovery dataset for variants inherited from both parents, i.e., a situation that would lead to homozygosity or compound-heterozygosity in the index patient.

In summary, as defined in the aims, the present study could identify promising novel nsCL/P candidate genes. The results suggest that the employment of PRS analysis for cohort selection, in combination with DNV detection, is a powerful tool for candidate gene discovery in nsCL/P. This approach could also be applied to other multifactorial congenital malformations. The analyses identified two nsCL/P top candidate genes with recurrent DNVs (*MDN1* and *PAXIP1*), as well as further interesting candidate genes that represent potential candidates for functional studies. Following the future identification of susceptibility genes, molecular testing of rare variants in high-risk susceptibility genes may become a component of precision medicine and be of direct clinical relevance, e.g., in terms of genetic counseling.

Web Resources

https://github.com/broadinstitute/gatk-docs/tree/master/gatk3-methods-and-algorithms
https://software.broadinstitute.org/software/igv/
https://gnomad.broadinstitute.org/
http://www.nlgenome.nl/
http://denovo-db.gs.washington.edu/denovo-db/
https://github.com/shendurelab/MIPGEN/tree/master/tools
https://string-db.org/
https://www.genenames.org/

ETHICS STATEMENT

This study was approved by the Ethical Committee of the Medical Faculty of the University of Bonn, Germany (Ethics approval number 295/14, updated June, 2022).

ORCID

Nina Ishorst https://orcid.org/0000-0001-9836-554X
Frederic Thieme https://orcid.org/0000-0001-9839-1764
Alexander Hoischen https://orcid.org/0000-0002-8072-4476
Michaela Bartusel https://orcid.org/0000-0003-1008-1951
Augusto Rojas-Martinez https://orcid.org/0000-0003-3765-6778
Christian Kirschneck https://orcid.org/0000-0001-9473-8724
Gerrit Spanier https://orcid.org/0000-0002-0226-7580
Andreas Buness https://orcid.org/0000-0003-1889-8632
Michael Nothnagel https://orcid.org/0000-0001-8305-7114
Kerstin U. Ludwig https://orcid.org/0000-0002-8541-2519
Elisabeth Mangold https://orcid.org/0000-0002-3819-2576

REFERENCES

Aldhorae, K. A., Böhmer, A. C., Ludwig, K. U., Esmail, A. H. A., Al-Hebshi, N. N., Lippke, B., Gölz, L., Nöthen, M. M., Daratsianos, N., Knapp, M., Jäger, A., & Mangold, E. (2014). Nonsyndromic cleft lip with or without cleft palate in Arab populations: Genetic analysis of 15 risk loci in a novel case-control sample recruited in Yemen. *Birth Defects Research Part A - Clinical and Molecular Teratology*, *100*(4), 307–313. https://doi.org/10.1002/bdra.23221

Basha, M., Demeer, B., Revencu, N., Helaers, R., Theys, S., Bou Saba, S., Boute, O., Devauchelle, B., Francois, G., Bayet, B., & Vikkula, M. (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. *Journal of Medical Genetics*, *55*(7), 449–458. https://doi.org/10.1136/jmedgenet-2017-105110

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., Liang, K. Y., Wu, T., Murray, T., Fallin, M. D., Redett, R. A., Raymond, G., Schwender, H., Jin, S.-C., Cooper, M. E., Dunnwald, M., Mansilla, M. A., Leslie, E., Bullard, S., … 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*(6), 525–529. https://doi.org/10.1038/ng.580

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., Restrepo Muñeton, C. P., Wehby, G., Hecht, J. T., Deleyiannis, F., Weinberg, S. M., Wu-Chou, Y. H., Chen, P. K., Brand, H., Epstein, M. P., … Leslie, E. J. (2020). Genome-wide enrichment of De novo coding mutations in orofacial cleft trios. *The American Journal of Human Genetics.*, *107*, 124–136. https://doi.org/10.1016/j.ajhg.2020.05.018

Christensen, K., Juel, K., Herskind, A. M., & Murray, J. C. (2004). Long term follow up study of survival associated with cleft lip and palate at birth. *BMJ*, *328*(7453), 1405. https://doi.org/10.1136/bmj.38106.559120.7C

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., Yang, Y., Carstens, R. P., Anand, D., Lachke, S. A., Smith, J. D., Dorschner, M. O., Bedell, B., Kirk, E., Hing, A. V., … Roscioli, T. (2018). Mutations in the epithelial cadherin-p120-catenin complex cause mendelian non-syndromic cleft lip with or without cleft palate. *American Journal of Human Genetics*, *102*(6), 1143–1157. https://doi.org/10.1016/j.ajhg.2018.04.009

Cox, T. C., Lidral, A. C., McCoy, J. C., Liu, H., Cox, L. L., Zhu, Y., Anderson, R. D., Moreno Uribe, L. M., Anand, D., Deng, M., Richter, C. T., Nidey, N. L., Standley, J. M., Blue, E. E., Chong, J. X., Smith, J. D., Kirk, E. P., Venselaar, H., Krahn, K. N., … Roscioli, T. (2019). Mutations in GDF11 and the extracellular antagonist, Follistatin, as a likely cause of mendelian forms of orofacial clefting in humans. *Human Mutation*, *40*, 1813–1825. https://doi.org/10.1002/humu.23793

Dauber, A., Golzio, C., Guenot, C., Jodelka, F. M., Kibaek, M., Kjaergaard, S., Leheup, B., Martinet, D., Nowaczyk, M. J. M., Rosenfeld, J. A., Zeesman, S., Zunich, J., Beckmann, J. S., Hirschhorn, J. N., Hastings, M. L., Jacquemont, S., … Katsanis, N. (2013). SCRIB and PUF60 are primary drivers of the multisystemic phenotypes of the 8q24.3 copy-number variant. *The American Journal of Human Genetics*, *93*(5), 798–811. https://doi.org/10.1016/j.ajhg.2013.09.010

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H., Baker, C. N., Bower, L., Brown, J. M., Caddle, L. B., Chiani, F., Clary, D., Cleak, J., Daly, M. J., Denegre, J. M., … Murray, S. A. (2016). High-throughput discovery of novel developmental phenotypes. *Nature*, *537*(7621), 508–514. https://doi.org/10.1038/nature19356

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Eijkelenboom, A., Kamping, E. J., Kastner-van Raaij, A. W., Hendriks-Cornelissen, S. J., Neveling, K., Kuiper, R. P., Hoischen, A., Nelen, M. R., Ligtenberg, M. J., & Tops, B. B. J. (2016). Reliable next-generation sequencing of formalin-fixed, paraffin-embedded tissue using single molecule tags. *The Journal of Molecular Diagnostics*, *18*(6), 851–863. https://doi.org/10.1016/j.jmoldx.2016.06.010

Fowles, L. F., Bennetts, J. S., Berkman, J. L., Williams, E., Koopman, P., Teasdale, R. D., & Wicking, C. (2003). Genomic screen for genes involved in mammalian craniofacial development. *Genesis*, *35*(2), 73–87. https://doi.org/10.1002/gene.10165

Gaither, J. B. S., Lammi, G. E., Li, J. L., Gordon, D. M., Kuck, H. C., Kelly, B. J., Fitch, J. R., & White, P. (2021). Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. *GigaScience*, *10*(4). https://doi.org/10.1093/gigascience/giab023

Grosen, D., Bille, C., Petersen, I., Skytthe, A., von Bornemann Hjelmborg, J., Pedersen, J. K., Murray, J. C., & Christensen, K. (2011). Risk of oral clefts in twins. *Epidemiology (Cambridge, Mass.)*, *22*(3), 313–319. https://doi.org/10.1097/EDE.0b013e3182125f9c

Grosen, D., Chevrier, C., Skytthe, A., Bille, C., Mølsted, K., Sivertsen, A., Murray, J. C., & 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. *Journal of Medical Genetics*, *47*(3), 162–168. https://doi.org/10.1136/jmg.2009.069385

Hilger, A., Schramm, C., Pennimpede, T., Wittler, L., Dworschak, G. C., Bartels, E., Engels, H., Zink, A. M., Degenhardt, F., Müller,

A. M., Schmiedeke, E., Grasshoff-Derr, S., Märzheuser, S., Hosie, S., Holland-Cunz, S., Wijers, C. H. W., Marcelis, C. L. M., van Rooij, I. A. L. M., Hildebrandt, F., … Draaken, M. (2013). De novo microduplications at 1q41, 2q37.3, and 8q24.3 in patients with VATER/VACTERL association. *European Journal of Human Genetics*, *21*(12), 1377–1382. https://doi.org/10.1038/ejhg.2013.58

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., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., … Kosmicki, J. A. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Khandelwal, K. D., Ishorst, N., Zhou, H., Ludwig, K. U., Venselaar, H., Gilissen, C., Thonissen, M., van Rooij, I. A. L. M., Dreesen, K., Steehouwer, M., van de Vorst, M., Bloemen, M., van Beusekom, E., Roosenboom, J., Borstlap, W., Admiraal, R., Dormaar, T., Schoenaers, J., Poorten, V. V., … Carels, C. E. L. (2017). Novel IRF6 mutations detected in orofacial cleft patients by targeted massively parallel sequencing. *Journal of Dental Research*, *96*(2), 179–185. https://doi.org/10.1177/0022034516678829

Kumar, A., Lualdi, M., Loncarek, J., Cho, Y.-W., Lee, J.-E., Ge, K., & Kuehn, M. R. (2014). Loss of function of mouse Pax-interacting protein 1-associated glutamate rich protein 1a (Pagr1a) leads to reduced Bmp2 expression and defects in chorion and amnion development. *Developmental Dynamics*, *243*(7), 937–947. https://doi.org/10.1002/dvdy.24125

Laugsch, M., Bartusel, M., Rehimi, R., Alirzayeva, H., Karaolidou, A., Crispatzu, G., Zentis, P., Nikolic, M., Bleckwehl, T., Kolovos, P., van Ijcken, W. F. J., Šarić, T., Koehler, K., Frommolt, P., Lachlan, K., Baptista, J., & Rada-Iglesias, A. (2019). Modeling the pathological long-range regulatory effects of human structural variation with patient-specific hiPSCs. *Cell Stem Cell*, *24*(5), 736–752.e12. https://doi.org/10.1016/j.stem.2019.03.004

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Butali, A., Buxó, C. J., Castilla, E. E., Christensen, K., Deleyiannis, F. W.-B., Field, L. L., Hecht, J. T., Moreno, L., Orioli, I. M., Padilla, C., Vieira, A. R., Wehby, G. L., Feingold, E., Weinberg, S. M., Murray, J. C., Beaty, T. H., & Marazita, M. L. (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. *Human Genetics*, *136*(3), 275–286. https://doi.org/10.1007/s00439-016-1754-7

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Feingold, E., Wehby, G., Laurie, C. A., Jain, D., Laurie, C. C., Doheny, K. F., McHenry, T., Resick, J., Sanchez, C., Jacobs, J., Emanuele, B., Vieira, A. R., Neiswanger, K., Lidral, A. C., Valencia-Ramirez, L. C., Lopez, A. M., … 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*(13), 2862–2872. https://doi.org/10.1093/hmg/ddw104

Leslie, E. J., Taub, M. A., Liu, H., Steinberg, K. M., Koboldt, D. C., Zhang, Q., Carlson, J. C., Hetmanski, J. B., Wang, H., Larson, D. E., Fulton, R. S., Kousa, Y. A., Fakhouri, W. D., Naji, A., Ruczinski, I., Begum, F., Parker, M. M., Busch, T., Standley, J., … 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*(3), 397–411. https://doi.org/10.1016/j.ajhg.2015.01.004

Ludwig, K. U., Böhmer, A. C., Bowes, J., Nikolic, M., Ishorst, N., Wyatt, N., Hammond, N. L., Gölz, L., Thieme, F., Barth, S., Schuenke, H., Klamt, J., Spielmann, M., Aldhorae, K., Rojas-Martinez, A., Nöthen, M. M., Rada-Iglesias, A., Dixon, M. J., Knapp, M., & Mangold, E. (2017). Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. *Human Molecular Genetics*, *26*(4), 829–842. https://doi.org/10.1093/hmg/ddx012

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., de Assis, N. A., Al Chawa, T., Mattheisen, M., Steffens, M., Barth, S., Kluck, N., Paul, A., Becker, J., Lauster, C., Schmidt, G., Braumann, B., Scheer, M., … Nöthen, M. M. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, *42*(1), 24–26. https://doi.org/10.1038/ng.506

Mangold, E., Ludwig, K. U., & Nöthen, M. M. (2011). Breakthroughs in the genetics of orofacial clefting. *Trends in Molecular Medicine*, *17*(12), 725–733. https://doi.org/10.1016/j.molmed.2011.07.007

Mangold, E., Reutter, H., Birnbaum, S., Walier, M., Mattheisen, M., Henschke, H., Lauster, C., Schmidt, G., Schiefke, F., Reich, R. H., Scheer, M., Hemprich, A., Martini, M., Braumann, B., Krimmel, M., Opitz, C., Lenz, J.-H., Kramer, F.-J., Wienker, T. F., … Diaz Lacava, A. (2009). Genome-wide linkage scan of nonsyndromic orofacial clefting in 91 families of central European origin. *American Journal of Medical Genetics. Part A*, *149A*(12), 2680–2694. https://doi.org/10.1002/ajmg.a.33136

Marazita, M. L. (2012). The evolution of human genetic studies of cleft lip and cleft palate. *Annual Review of Genomics and Human Genetics*, *13*(1), 263–283. https://doi.org/10.1146/annurev-genom-090711-163729

Marini, N. J., Asrani, K., Yang, W., Rine, J., & Shaw, G. M. (2019). Accumulation of rare coding variants in genes implicated in risk of human cleft lip with or without cleft palate. *American Journal of Medical Genetics Part A*, *179*(7), ajmg.a.61183–ajmg.a.61269. https://doi.org/10.1002/ajmg.a.61183

McVeigh, T. P., Banka, S., & Reardon, W. (2015). Kabuki syndrome: Expanding the phenotype to include microphthalmia and anophthalmia. *Clinical Dysmorphology*, *24*(4), 135–139. https://doi.org/10.1097/MCD.0000000000000092

Moreno, L. M., Mansilla, M. A., Bullard, S. A., Cooper, M. E., Busch, T. D., Machida, J., Johnson, M. K., Brauer, D., Krahn, K., Daack-Hirsch, S., L'heureux, J., Valencia-Ramirez, C., Rivera, D., López, A. M., Moreno, M. A., Hing, A., Lammer, E. J., Jones, M., Christensen, K., … Lidral, A. C. (2009). FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Human Molecular Genetics*, *18*(24), 4879–4896. https://doi.org/10.1093/hmg/ddp444

Phan, M., Conte, F., Khandelwal, K. D., Ockeloen, C. W., Bartzela, T., Kleefstra, T., van Bokhoven, H., Rubini, M., Zhou, H., & Carels, C. E. L. (2016). Tooth agenesis and orofacial clefting: Genetic brothers in arms? *Human Genetics*, *135*(12), 1299–1327. https://doi.org/10.1007/s00439-016-1733-z

Raman, N., Weir, E., & Müller, S. (2016). The AAA ATPase MDN1 acts as a SUMO-targeted regulator in mammalian pre-ribosome remodeling. *Molecular Cell*, *64*(3), 607–615. https://doi.org/10.1016/j.molcel.2016.09.039

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. https://doi.org/10.1093/nar/gky1016

Rojas-Martinez, A., Reutter, H., Chacon-Camacho, O., Leon-Cachon, R. B. R., Munoz-Jimenez, S. G., Nowak, S., Becker, J., Herberz, R., Ludwig, K. U., Paredes-Zenteno, M., Arizpe-Cantú, A., Raeder, S., Herms, S., Ortiz-Lopez, R., Knapp, M., Hoffmann, P., Nöthen, M. M., & Mangold, E. (2010). Genetic risk factors for nonsyndromic cleft lip with or without cleft palate in a Mesoamerican population: Evidence for IRF6 and variants at 8q24 and 10q25. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, *88*(7), 535–537. https://doi.org/10.1002/bdra.20689

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Jr., … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*(9), 944–950. https://doi.org/10.1038/ng.3050

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, Beales, P., Lees, M., Moore, G. E., Stanier, P., & Passos-Bueno, M. R. (2017). Impact of rare variants in ARHGAP29 to the etiology of oral clefts: Role of loss-of-function vs missense variants. *Clinical Genetics*, *91*(5), 683–689. https://doi.org/10.1111/cge.12823

Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., Backofen, R., & Diederichs, S. (2019). A pan-cancer analysis of synonymous mutations. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10489-2

Thieme, F., Henschel, L., Hammond, N. L., Ishorst, N., Hausen, J., Adamson, A. D., Biedermann, A., Bowes, J., Zieger, H. K., Maj, C., Kruse, T., Buness, A., Hoischen, A., Gilissen, C., Kreusch, T., Jäger, A., Gölz, L., Braumann, B., Aldhorae, K., … Ludwig, K. U. (2021). Extending the allelic spectrum at noncoding risk loci of orofacial clefting. *Human Mutation*, *42*(8), 1066–1078. https://doi.org/10.1002/HUMU.24219

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Van Otterloo, E., Williams, T., & Artinger, K. B. (2016). The old and new face of craniofacial research: How animal models inform human craniofacial genetic and clinical data. *Developmental Biology*, *415*(2), 171–187. https://doi.org/10.1016/j.ydbio.2016.01.017

Vogelaar, I. P., Figueiredo, J., van Rooij, I. A. L. M., Simões-Correia, J., van der Post, R. S., Melo, S., Seruca, R., Carels, C. E. L., Ligtenberg, M. J. L., & Hoogerbrugge, N. (2013). Identification of germline mutations in the cancer predisposing gene CDH1 in patients with orofacial clefts. *Human Molecular Genetics*, *22*(5), 919–926. https://doi.org/10.1093/hmg/dds497

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/NAR/GKQ603

Welzenbach, J., Hammond, N. L., Nikolić, M., Thieme, F., Ishorst, N., Leslie, E. J., Weinberg, S. M., Beaty, T. H., Marazita, M. L., Mangold, E., Knapp, M., Cotney, J., Rada-Iglesias, A., Dixon, M. J., & Ludwig, K. U. (2021). Integrative approaches generate insights into the architecture of non-syndromic cleft lip with or without cleft palate. *Human Genetics and Genomics Advances*, *2*(3), 100038. https://doi.org/10.1016/J.XHGG.2021.100038

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Yu, Y., Zuo, X., He, M., Gao, J., Fu, Y., Qin, C., Meng, L., Wang, W., Song, Y., Cheng, Y., Zhou, F., Chen, G., Zheng, X., Wang, X., Liang, B., Zhu, Z., Fu, X., Sheng, Y., Hao, J., … Bian, Z. (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nature Communications*, *8*(2), 14364. https://doi.org/10.1038/ncomms14364

SUPPORTING INFORMATION

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