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

Nina Ishorst1 | Leonie Henschel1 | Frederic Thieme1 | Dmitriy Drichel2 | Sugirthan Sivalingam3,4,5 | Sarah L. Mehrem1 | Ariane C. Fechtner1 | Julia Fazaal1 | Julia Welzenbach1 | André Heimbach1 | Carlo Maj4 | Oleg Borisov4 | Jonas Hausen3,4,5 | Ruth Raff1 | Alexander Hoischen6,7,8 | Michael Dixon9 | Alvaro Rada-Iglesias10 | Michaela Bartusel11,12 | Augusto Rojas-Martinez13,14 | Khalid Aldhorae15,16 | Bert Braumann17 | Teresa Kruse17 | Christian Kirschneck18 | Gerrit Spanier19 | Heiko Reutter1,20 | Stefanie Nowak1 | Lina Gözl21,22 | Michael Knapp5 | Andreas Buness3,4,5 | Peter Krawitz4 | Markus M. Nöthen1 | Michael Nothnagel2,23 | Tim Becker24 | Kerstin U. Ludwig1 | Elisabeth Mangold1

1Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany
2Cologne Center for Genomics, University of Cologne, Cologne, Germany
3Core Unit for Bioinformatic Analysis, Medical Faculty, University of Bonn, Bonn, Germany
4Institute for Genomic Statistics and Bioinformatics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany
5Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany
6Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands
7Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands
8Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands
9Faculty of Biology, Medicine & Health, University of Manchester, Manchester, UK
10Institute of Biomedicine and Biotechnology of Cantabria (IBBTEC), CSIC/University of Cantabria, Santander, Spain
11Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany
12Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
13Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Monterrey, Mexico
14Centro de Investigacion y Desarrollo en Ciencias de la Salud, Universidad Autonoma de Nuevo Leon, Monterrey, Mexico
15Department of Orthodontics, College of Dentistry, Thamar University, Thamar, Yemen
16Department of Orthodontics, College of Dentistry, University of Ibn al-Nafis for Medical Sciences, Sanaa, Yemen
17Faculty of Medicine and University Hospital Cologne, Department of Orthodontics, University of Cologne, Cologne, Germany
18Department of Orthodontics, University of Regensburg, Regensburg, Germany
19Department of Cranio-Maxillofacial Surgery, University Hospital Regensburg, Regensburg, Germany
20Division of Neonatology and Pediatric Intensive Care, Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Erlangen, Germany
21Department of Orthodontics, University of Erlangen-Nürnberg, Erlangen, Germany

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. Molecular Genetics & Genomic Medicine published by Wiley Periodicals LLC.
INTRODUCTION

Nonsyndromic cleft lip with/without cleft palate (nsCL/P) 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 $2 \times 125$ 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
PRS analysis for selection of Discovery cohort

Exome Sequencing (50 trios)

De novo mutation?

yes

60 candidate genes

Validated and protein altering?

no

27 candidate genes excluded

yes

33 candidate genes

Resequencing with smMIPs (1010 patients, 1574 controls)

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

Additional evidence

Interactions at protein level

GWAS associations in respective TADs

Expression in relevant mouse tissues/human cell lines

Gene-based test

 Discovery

 Replication
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

<table>
<thead>
<tr>
<th>Sample numbers</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients</td>
</tr>
<tr>
<td></td>
<td>f</td>
</tr>
<tr>
<td>Discovery cohort</td>
<td></td>
</tr>
<tr>
<td>ES trios from Central Europe</td>
<td>50</td>
</tr>
<tr>
<td>Replication cohort</td>
<td></td>
</tr>
<tr>
<td>Resequencing Europe</td>
<td>643 (651)</td>
</tr>
<tr>
<td>Resequencing Yemen</td>
<td>217 (217)</td>
</tr>
<tr>
<td>Resequencing Mexico</td>
<td>150 (150)</td>
</tr>
<tr>
<td>Total</td>
<td>1010</td>
</tr>
</tbody>
</table>

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.

Numbers in brackets indicate numbers pre quality control.

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).

337 cases suitable for de novo detection via Sanger sequencing (resequencing was performed in independent individuals with nsCL/P only). 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 (Alldorae 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 × 125bp 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 associated 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 × 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 \( \beta \)-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](https://onlinelibrary.wiley.com/doi/10.1002/mgg3.2109)

**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 $\geq 10$ 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 $\geq 10\%$ of the single molecule molecular inversion probes had poor coverage ($<20\times$ 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 $\geq 15$, a MAF $\leq 0.1\%$, 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 across 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 total final 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 Midsain AAA ATPase 1 (MDNI [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 (PAIPI [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 MDNI and CSMD1; and (ii) a compound heterozygous index patient with two variants in MDNI, 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 NM 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), MDNI and PAIPI were highly expressed in all relevant mouse tissues at all embryonic time points (RPKMMDNI = 292.02–1637.17; RPKMPAIPI = 429.83–1338.15; Table S6), and in hNCC and hCNCC (FPKMMDSNi(hNCC) = 11.36; FPKMPAIPI(hNCC) = 9.18; Table S6). ANKI showed a moderate expression in...
<table>
<thead>
<tr>
<th>Candidate gene</th>
<th>Change</th>
<th>Trio</th>
<th>MAF&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Denovo-db</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDC3</td>
<td>NM_014654.3:c.166G&gt;A (p.Glu56Lys))</td>
<td>BN46</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>NAV1</td>
<td>NM_00167738.1:c.3397G&gt;A (p.Val1133Ile))</td>
<td>BN104</td>
<td>4.87×10⁻⁵</td>
<td>nf</td>
</tr>
<tr>
<td>ATXN7</td>
<td>NM_001128149.3:c.1759G&gt;A (p.Glu587Lys))</td>
<td>BN82</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>SEPN2</td>
<td>NM_021267.2:c.967G&gt;A (p.Glu323Lys))</td>
<td>BN142</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>GLRA3</td>
<td>NM_001042543.2:c.1018G&gt;A (p.Val340Leu))</td>
<td>BN192</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>MAST4</td>
<td>NM_001290227.2:c.5394_5396del (p.Ser1799del))</td>
<td>BN98d</td>
<td>1.80×10⁻⁵</td>
<td>nf</td>
</tr>
<tr>
<td>H1-3</td>
<td>NM_005320.3:c.115C&gt;T (p.(Pro39Ser))</td>
<td>BN565</td>
<td>8.95×10⁻⁶</td>
<td>nf</td>
</tr>
<tr>
<td>MDN1</td>
<td>NM_014611.1:c.489C&gt;G (p.Phe163Leu))</td>
<td>BN82</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>TAA8R</td>
<td>NM_005327.2:c.748_749del (p.Lys250GlufsTer7))</td>
<td>BN591</td>
<td>2.06×10⁻⁴</td>
<td>nf</td>
</tr>
<tr>
<td>TME1M20A</td>
<td>NM_031925.2:c.870G&gt;A (p.(Ala290=))</td>
<td>BN23</td>
<td>2.76×10⁻⁶</td>
<td>nf</td>
</tr>
<tr>
<td>CFTR</td>
<td>NM_000492.3:c.4333G&gt;A (p.(Asp1445Asn))</td>
<td>BN62</td>
<td>2.93×10⁻⁴</td>
<td>nf</td>
</tr>
<tr>
<td>PAXIP1</td>
<td>NM_007349.3:c.2434G&gt;A (p.(Asp812Asn))</td>
<td>BN46</td>
<td>0</td>
<td>nf</td>
</tr>
<tr>
<td>CSMD1</td>
<td>NM_033225.5:c.2699A&gt;G (p.(Asp900Gly))</td>
<td>BN172</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>LONRF1</td>
<td>NM_152271.4:c.2227T&gt;C (p.(Tyr743His))</td>
<td>BN192</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>ANK1</td>
<td>NM_001142445.1:c.13G&gt;A (p.(Val51Leu))</td>
<td>BN865</td>
<td>4.75×10⁻⁵</td>
<td>nf</td>
</tr>
<tr>
<td>VCP</td>
<td>NM_007126.4:c.475C&gt;T (p.(Arg159Cys))</td>
<td>BN62d</td>
<td>8.95×10⁻⁶</td>
<td>nf</td>
</tr>
<tr>
<td>TRMO</td>
<td>NM_016481.4:c.1265C&gt;T (p.(Pro422Leu))</td>
<td>BN98</td>
<td>1.79×10⁻⁵</td>
<td>nf</td>
</tr>
<tr>
<td>CLPB</td>
<td>NM_001258393.2:c.913A&gt;G (p.(Lys305Glu))</td>
<td>BN421</td>
<td>0</td>
<td>nf</td>
</tr>
<tr>
<td>BSX</td>
<td>NM_001098169.2:c.556G&gt;C (p.(Glu186Gln))</td>
<td>BN104</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>KIRREL3</td>
<td>NM_001161707.1:c.1190T&gt;C (p.(Val397Ala))</td>
<td>BN139</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>CDH1</td>
<td>NM_004360.3:c.774_776del (p.(Asn258del))</td>
<td>BN82</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>SMG6</td>
<td>NM_001282326.2:c.391C&gt;T (p.(Gln131Ter))</td>
<td>BN82</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>DHX40&lt;sup&gt;α&lt;/sup&gt;</td>
<td>NM_001166301.2:c.1842G&gt;A</td>
<td>BN137</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>ABCA8</td>
<td>NM_007168.3:c.3103T&gt;G (p.(Ser1035Ala))</td>
<td>BN119</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>FDXR</td>
<td>NM_001258015.2:c.343C&gt;T (p.(Arg113Trp))</td>
<td>BN54</td>
<td>7.91×10⁻⁶</td>
<td>nf</td>
</tr>
<tr>
<td>TMC6</td>
<td>NM_001127198.5:c.1870G&gt;A (p.(Val624Ile))</td>
<td>BN249</td>
<td>9.67×10⁻⁴</td>
<td>nf</td>
</tr>
<tr>
<td>ZNF407</td>
<td>NM_001146119.1:c.3373G&gt;A (p.(Ala1125Thr))</td>
<td>BN138</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>EIF3G</td>
<td>NM_003755.4:c.3G&gt;T (p.(Met1?))</td>
<td>BN497</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>CASP14</td>
<td>NM_0012114.3:c.674G&gt;A (p.(Arg225Lys))</td>
<td>BN200</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>GNAS</td>
<td>NM_016592.3:c.190_218del (p.(Phe64ProfsTer5))</td>
<td>BN655</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>SEC14L3</td>
<td>NM_174975.4:c.268G&gt;A (p.(Gly90Ser))</td>
<td>BN139</td>
<td>nf</td>
<td>nf</td>
</tr>
<tr>
<td>NAP1L2</td>
<td>NM_001963.3:c.1090G&gt;T (p.(Asp364Tyr))</td>
<td>BN240</td>
<td>nf</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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.

<sup>a</sup>Gene failed in the single molecule molecular inversion probe assay design.

<sup>b</sup>Gene accession number (NM); change on DNA level (c.); change on protein level (p.).

<sup>c</sup>Publicly available datasets accessed August 2018.

Relevant mouse tissues (RPKM<sub>Ank1</sub> = 35.23–71.27), as well as in hNCC and hCNCC (FPKM<sub>ANKI(hNCC)</sub> = 0.34). Finally, the expression profile of CSMD1 in the assessed mouse tissues ranged from not expressed at all to marginally expressed (RPKM<sub>CSMD1</sub> = 0–52.14), while only marginal expression was evident in the two human cell lines (FPKM<sub>CSMD1(hNCC)</sub> = 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<sub>CDH1</sub> = 0.0019, p<sub>TMC6</sub> = 0.0132, p<sub>EIF3G</sub> = 0.0133, p<sub>TRMO</sub> = 0.0133).
$P_{\text{TRMO}} = 0.0399$ (Table S7). CDH1 almost withstood correction for multiple testing ($P_{\text{CDH1-corrected}} = 0.0607$) (Table S7). In the Yemeni cohort, Brain Specific Homeobox (BSX [MIM:611074]; NM_001098169.2) showed a nominally significant overrepresentation of deleterious variants ($P_{\text{BSX}} = 0.0045; P_{\text{BSX-corrected}} = 0.1446$) (Table S8). In the Mexican cohort, a nominally significant enrichment of deleterious variants was found for EIF3G ($P_{\text{EIF3G}} = 0.0115; P_{\text{EIF3G-corrected}} = 0.3583$) (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 employPRS 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 Epiphan1 (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 ($n_{\text{Bishop}} = 698; n_{\text{ Ishorst}} = 50$).

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, MDM1, and PAXP1). MDM1 and PAXP1 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 MDM1 and PAXP1 are promising candidate genes for nsCL/P. MDM1 (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 PAXP1 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, BXS) 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 (MDNI 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/

AUTHOR CONTRIBUTIONS


ACKNOWLEDGMENTS

The authors thank all patients, relatives, and control individuals for their participation. We thank the German support group for individuals with cleft lip and/or palate (Wolfgang Rosenthal Gesellschaft) for assistance with recruitment. We acknowledge the invaluable assistance of all clinical, laboratory, and bioinformatic personnel. The authors thank the Next Generation Sequencing Core Facility of the Medical Faculty of the University of Bonn for sequencing the samples that were used in this study. DbGaP datasets were accessed through dbGaP accession number phs000094.v1.p1 (Supplemental Acknowledgments). Finally, the authors thank the Genome Aggregation Database (gnomAD), and all groups that provided exome and genome variant data to this resource. A full list of gnomAD contributors is provided in the gnomAD flagship paper (Karczewski et al., 2020). Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The present study was supported by the German Research Foundation (DFG)-Grants BE 3828/8-1, LU 1944/2-1, MA 2546/5-1, and LU1944/3-1.

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).
REFERENCES


**SUPPORTING INFORMATION**

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